BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

NGUYỄN QUỐC LONG

NHẬN DẠNG TIẾNG NÓI TIẾNG VIỆT SỬ DỤNG MẠNG NƠ-RON NHÂN TẠO VÀ MÔ HÌNH MARKOV ẨN

Chuyên ngành: Khoa học máy tính
Mã số: 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2011

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học : **PGS.TS Phan Huy Khánh**

Phản biện 1: **PGS.TS. Võ Trung Hùng**

Phản biện 2: **PGS.TS. Đoàn Văn Ban**

Luận văn được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 11 tháng 9 năm 2011

** Có thể tìm hiểu luận văn tại:*

- Trung tâm Thông tin Học liệu, Đại học Đà Nẵng
- Trung tâm Học liệu, Đại học Đà Nẵng

MỞ ĐẦU

1. Lý do chọn đề tài

Tiếng nói là phương tiện giao tiếp cơ bản và rộng rãi nhất của loài người, nó hình thành và phát triển song song với quá trình tiến hóa của loài người. Đối với con người, sử dụng lời nói là một cách diễn đạt đơn giản và hiệu quả nhất. Ưu điểm của việc giao tiếp bằng tiếng nói trước tiên là ở tốc độ giao tiếp, tiếng nói từ người nói được người nghe hiểu ngay lập tức sau khi được phát ra. Từ khi ngành công nghiệp máy tính phát triển, nhiều công trình nghiên cứu trên tiếng nói nhằm khai thác các thông tin từ tiếng nói để ứng dụng trong nhiều lĩnh vực như hệ thống trả lời điện thoại tự động, dịch vụ tra cứu thông tin du lịch bằng tiếng nói, và ứng dụng nhận dạng tiếng nói trong các hệ thống bảo mật… đã đem lại nhiều lợi ích và cách thức giao tiếp thuận tiện hơn cho con người.

Lĩnh vực nghiên cứu nhận dạng tiếng nói đã được bắt đầu từ cuối thập kỷ 40, các nghiên cứu và ứng dụng về xử lý ngôn ngữ nói chung trên thế giới và nhiều nước khác đã trải qua nhiều giai đoạn, và điều quan trọng hơn cả là nhiều cách tiếp cận và cách thức xử lí ngôn ngữ đã được trải nghiệm và thừa nhận. Ở Việt Nam, lĩnh vực nhận dạng và xử lý tiếng nói tiếng Việt vẫn còn khá mới, theo người viết luận văn được biết, các tập thể làm nghiên cứu đã có những kết quả gần đây là Viện Công nghệ Thông tin, Trường Đại học KHTN TPHCM và Trung tâm nghiên cứu quốc tế Thông tin đa phương tiện, truyền thông và ứng dụng (MICA) – ĐHBK Hà nội, cộng với một số đề tài nghiên cứu thạc sĩ, tiến sĩ trên cả nước; nhìn chung các đề tài tập trung xử lý tiếng nói tiếng Việt trên tập dữ liệu nhỏ và vừa, phụ thuộc và độc lập người nói, khả năng xử lý nhiễu của tín hiệu còn thấp,

thường áp dụng hướng tiếp cận nhận dạng đối sánh mẫu như nắn chỉnh thời gian động (DTW), các mô hình Markov ẩn rời rạc…dẫn đến một số kết quả chỉ mang tính chất tìm hiểu, chưa hệ thống và định hướng rõ ràng, có hiệu suất nhận dạng từ 88% - 96% [1][2][3].

Vì ý nghĩa đó và được sự đồng ý hướng dẫn của Thầy PGS.TS Phan Huy Khánh, tôi đã chọn đề tài "*Nhận dạng tiếng nói tiếng Việt sử dụng mạng nơ-ron nhân tạo và mô hình Markov ẩn*" thực hiện với mong muốn đóng góp một giải pháp trong lĩnh vực nhận dạng tiếng nói tiếng Việt.

2. Mục đích nghiên cứu

Mục tiêu của đề tài là nghiên cứu chung các vấn đề về nhận dạng tiếng nói và ứng dụng mô hình Markov ẩn kết hợp mạng nơ-ron trong nhận dạng tiếng nói tiếng Việt. Đồng thời, xây dựng chương trình nhận dạng nhằm mục đích kiểm tra giải pháp và đánh giá hiệu suất nhận dạng của hệ thống.

Về lý thuyết, thực hiện nghiên cứu tổng quan về nhận dạng tiếng nói bao gồm các hướng tiếp cận nhận dạng tiếng nói, các mô hình và kỹ thuật phân lớp, tiếp đến trình bày các bước tiền xử lý tín hiệu tiếng nói, phương pháp phân tích trích đặc trưng tiếng nói. Đối với bài toán nhận dạng, nghiên cứu chi tiết, triển khai và ứng dụng mô hình Markov ẩn trong nhận dạng tiếng nói.

Về thực tiễn, nghiên cứu và phát triển các giải thuật cho hệ thống nhận dạng tiếng nói trên môi trường Matlab sử dụng các công cụ sẵn có như Auditory ToolBox, HMM Toolbox, CLSU.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài là nhận dạng tiếng nói tiếng Việt. Phạm vi nghiên cứu của đề tài là các phương pháp phát hiện

tiếng nói, rút trích đặc trưng tiếng nói, mô hình Markov ẩn rời rạc và liên tục, kết hợp mạng nơ-ron trong nhận dạng tiếng nói và tiếp đến là xây dựng ứng dụng mô hình Markov ẩn nhằm kiểm tra và đánh giá hiệu suất nhận dạng. Cơ sở dữ liệu dùng cho nhận dạng và kiểm thử chỉ dừng ở tập dữ liệu gồm 10 chữ số tiếng Việt được thu từ 15 người.

4. Phương pháp nghiên cứu

Các phương tiện và công cụ dùng để có thể triển khai đề tài là các tài liệu liên quan đến xử lý tín hiệu tiếng nói, và cách thức lập trình trong môi trường Matlab liên quan đến đề tài.

5. Ý nghĩa khoa học và thực tiễn của đề tài

Sau khi thực hiện nghiên cứu và xây dựng hệ thống nhận dạng tiếng nói tiếng Việt, góp phần cung cấp một giải pháp nhận dạng tiếng nói tiếng Việt, cung cấp cơ sở lý thuyết cho việc phát triển các ứng dụng nhận dạng tiếng nói về sau.

6. Cấu trúc của luận văn

Bố cục của luận văn được tổ chức thành 3 chương, có nội dung như sau:

- Chương 1: Thống kê tình hình nghiên cứu xử lý ngôn ngữ, tìm hiểu tổng quan về lý thuyết nhận dạng, các hướng tiếp cận nhận dạng tiếng nói, phân tích và thống kê đặc điểm cơ bản của tiếng Việt.
- Chương 2: Trình bày chi tiết một hệ thống nhận dang tiếng nói từ giai đoạn phân tích rút đặc trưng tín hiệu tiếng nói, cho đến ứng dụng mô hình Markov ẩn trong nhận dạng tiếng nói bao gồm đặc tả mô hình, các bài toán cơ bản cho đến các giải thuật để giải quyết bài toán nhận dạng.

- Chương 3: Giới thiệu các phương pháp nhận dạng đã được triển khai, phân tích đánh giá ưu và nhược điểm của mỗi phương pháp, từ đó đề xuất giải pháp cho đề tài. Tiếp đến trình bày các bước xây dựng hệ thống nhận dạng ứng dụng mô hình Markov ẩn kết hợp mạng nơ-ron. Cuối chương, tiến hành đánh giá thử nghiệm các kết quả nhận dạng tiếng nói tiếng Việt phụ thuộc người nói và độc lập người nói.

CHƯƠNG 1 - NGHIÊN CỨU TỔNG QUAN

1.1. LỊCH SỬ NHẬN DẠNG

1.1.1. Xu hướng phát triển

Giao tiếp người-máy là một lĩnh vực nghiên cứu lớn và khó nhưng lại có nhiều ứng dụng thực tiễn. Tiếng nói là một phương tiện giao tiếp tự nhiên nhất của con người và vì vậy, nghiên cứu để máy tính có thể hiểu tiếng nói của con người, hay còn gọi là nhận dạng tiếng nói tự động (Automatic Speech Recognition – ASR), đã trải qua quá trình 50 năm phát triển.

Những nỗ lực nghiên cứu đầu tiên về ASR đã được tiến hành trong thập niên 50 với ý tưởng chính là dựa trên ngữ âm. Trong giai đoạn này, có các hệ thống đáng chú ý như: hệ thống nhận dạng ký số rời rạc của Bell-lab (1952), bộ nhận dạng 13 âm vị của trường đại học College–Anh (1958)…

Trong thập kỉ 1960, điểm đáng ghi nhận nhất là ý tưởng của tác giả người Nga, Vintsyuk khi ông đề xuất phương pháp nhận dạng tiếng nói dựa trên qui hoạch động theo thời gian - Dynamic Time Warping.

Nghiên cứu về ASR trong thập kỉ 80 đánh dấu phép dịch chuyển trong phương pháp luận: từ cách tiếp cận đối sánh mẫu sang cách tiếp cận sử dụng mô hình thống kê. Ngày nay, hầu hết các hệ thống ASR đều dựa trên mô hình thống kê được phát triển ở thập kỉ này, cùng với những cải tiến ở thập kỉ 90. Một trong những phát minh quan trọng nhất ở thập kỉ 80 là mô hình Markov ẩn (Hidden Markov Model – HMM).

Các hệ thống ASR ra đời trong thời gian này có thể kể đến: hệ thống Sphinx của trường đại học CMU, Byblos của công ty BBN,

Decipher của viện SRI, và các hệ thống khác của Lincoln Labs, MIT và AT&T Bell Labs.

Thập niên 90 ghi nhận một số kết quả nghiên cứu mới trong lĩnh vực phân lớp mẫu. Cụ thể, bài toán phân lớp theo mô hình thống kê (dựa trên luật quyết định Bayes), đòi hỏi phép ước lượng các phân bố cho dữ liệu, được chuyển thành bài toán tối ưu, bao gồm phép cực tiểu lỗi phân lớp bằng thực nghiệm.

Đến những năm đầu của thế kỷ 21, các nghiên cứu tập trung vào việc nâng cao kết quả nhận dạng tiếng nói, thông qua chương trình có tên gọi EARS (Effective Affordable Reusable Speech-to-Text).

Đích hướng tới của chương trình này là khả năng nhận dạng, tóm tắt và chuyển ngữ các đoạn audio, giúp cho người đọc hiểu nhanh nội dung của chúng thay vì phải nghe toàn bộ. Chủ yếu, các nghiên cứu tập trung vào 3 nhóm chính:

- Nhận dạng tiếng nói tự nhiên
- Nhận dạng tiếng nói dựa trên nhiều kênh thông tin.

Về mặt kinh tế và thương mại, công nghệ nhận dạng tiếng nói đã thay đổi cách con người tương tác với hệ thống và thiết bị, không còn bó buộc trong cách thức tương tác truyền thống (như thông qua bàn phím của máy tính hay điện thoại) mà chuyển sang tương tác trực tiếp bằng giọng nói.

Về mặt nghiên cứu khoa học, các hệ thống nhận dạng tiếng nói hiện tại đều dựa trên phương pháp thống kê và so khớp mẫu. Phương pháp này đòi hỏi các tri thức về ngữ âm và một lượng lớn dữ liệu huấn luyện, bao gồm cả dạng âm thanh và dạng văn bản, để huấn luyện bộ nhận dạng. Lượng dữ liệu huấn luyện càng lớn, bộ nhận dạng càng có nhiều khả năng đưa ra kết quả chính xác hơn.

1.1.2. Tình hình nghiên cứu ở Việt Nam

Tại Việt Nam, có 2 nhóm nghiên cứu chính về bài toán nhận dạng tiếng nói [3]. Nhóm đầu tiên thuộc Viện Công nghệ Thông tin do GS.TSKH Bạch Hưng Khang đứng đầu. Nhóm tập trung nghiên cứu các vấn đề sau:

- Nghiên cứu, phân tích các đặc trưng ngữ âm, thông số của tiếng Việt, văn phạm tiếng Việt phục vụ cho nhận dạng tiếng nói
- Nghiên cứu để tạo lập CSDL các mẫu câu để tạo tham số huấn luyện cho mô hình 3 mức: âm tiết – âm vị - âm học.
- Nghiên cứu bài toán nhận dạng tiếng nói liên tục trên CSDL từ vựng cỡ nhỏ, trung bình, tiến tới CSDL lớn

Nhóm thứ hai thuộc trường Đại học Khoa học Tự nhiên thành phố Hồ Chí Minh do Tiến sĩ Vũ Hải Quân đứng đầu. Các nghiên cứu của nhóm tập trung vào bài toán truy vấn thông tin cho bản tin thời sự tiếng Việt.

Ngoài ra, gần đây có nghiên cứu của LIG (Laboratoire Informatique de Grenoble) hợp tác với phòng thí nghiệm MICA ở Hà Nội về sự khả chuyển của các mô hình ngữ âm (acoustic model portability)

Một số hệ thống nhận dạng tiếng Việt hiện nay có thể liệt kê như sau:

- VnCommand: Chương trình nhận dạng lệnh, trình diễn khả năng điều khiển chương trình ứng dụng trên Windows.
- Chương trình nhận dạng lệnh 10 chữ số tiếng Việt liên tục qua điện thoại.
- VnDictator: chương trình đọc chính tả.

1.2. NHẬN DẠNG TIẾNG NÓI

1.2.1. Tổng quan

Nhận dạng đối với con người là quá trình mô phỏng lại sự nhận biết các sự vật hiện tượng xung quanh não người. Một hệ nhận dạng với các thành phần cơ bản sau:

1) Module thu nhận tín hiệu và trích đặc trưng.
2) Module học mẫu.
3) Module tra cứu – so khớp

Việc nhận dạng tiếng nói thực chất chính là quá trình nghiên cứu tiếng nói để đưa ra tập các đặc tính và quá trình nhận dạng sau đó sẽ so sánh tiếng nói cần được nhận dạng với tập các đặc tính trên để phán đoán.

Phân loại một số hê thống nhận dạng tiếng nói khác nhau như:

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

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…

1.2.2. Các hướng tiếp cận

1.2.2.1. Tiếp cận dựa vào âm học và 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. Một công đoạn quan trọng của phương pháp này là sự phân đoạn và gán nhãn bởi nó liên quan đến sự phân đoạn tiếng nói ra những vùng rời rạc (về thời gian) trên đó những thuộc tính ngữ âm của tín hiệu tương trưng cho một (hoặc nhiều) đơn vị ngữ âm (hoặc lớp ngữ âm).

1.2.2.2. Tiếp cận dựa theo mẫu

Phương pháp tiếp cận dựa vào nhận dạng mẫu trong nhận dạng tiếng nói về cơ bản là sử dụng trực tiếp những mẫu tiếng nói mà không xác định rõ ràng các đặc tính âm – ngữ học và sự phân đoạn. Phương pháp này có hai bước: huấn luyện mẫu tiếng nói và nhận dạng các mẫu chưa biết thông qua việc so sánh với các mẫu đã huấn luyện. Vấn đề là nếu cung cấp đầy đủ các diễn tả của mẫu dùng để nhận dạng gọi là tập huấn luyện thì sau khi huấn luyện, mẫu tham khảo sẽ có thể mô tả đủ những đặc tính âm học của mẫu. Tiện lợi của phương pháp này là giai đoạn so sánh mẫu: so sánh trực tiếp tiếng nói chưa biết với mỗi mẫu đã huấn luyện và tìm ra tiếng nói chưa biết tùy theo tính chất của mẫu phù hợp.

1.2.2.3. Tiếp cận dựa theo hướng trí tuệ nhân tạo

Phương pháp tiếp cận dựa vào trí tuệ nhân tạo thực chất là sự kết hợp giữa hai phương pháp trên, nó khai thác cả ý tưởng và các khái niệm của hai phương pháp này. Phương pháp này cố gắng máy móc hóa thủ tục nhận dạng theo cách của con người áp dụng trí thông

minh của mình để hình dung, phân tích và cuối cùng tạo một quyết định trên những đặc tính âm học đo được.

Ý tưởng cơ bản của phương pháp này là biên soạn và kết hợp những tri thức từ nhiều nguồn tri trúc:

- Tri thức học (acoustic knowledge).
- Tri thức từ vựng học (lexical knowledge).
- Tri thức cú pháp học (syntactic knowledge).
- Tri thức ngữ nghĩa (semantic knowledge_.
- Tri thức thực tế (pragmatic knowledge).

1.3. ĐỘ ĐO HIỆU SUẤT NHẬN DẠNG

1.3.1. Độ chính xác

Độ chính xác nhận dạng là thước đo đơn giản và quan trọng nhất để đánh giá hiệu suất nhận dạng tiếng nói. Vì vậy, mục tiêu xây dựng hệ thống làm sao giảm thiểu tỉ lệ lỗi nhận dạng trên cả tập huấn luyện và hiệu suất khác nhau trên cả tập huấn luyện và tập kiểm tra.

1.3.2. Độ phức tạp

Độ phức tạp cũng là một vấn đề cần xem xét trong hầu hết các hệ thống nhận dạng thương mại, đặc biệt khi chi phí phần cứng là một tiêu chí cho sự thành công của hệ thống. Thông thường, độ phức tạp của hệ thống nhận dạng đề cập đến độ phức tạp tính toán và độ phức tạp mô hình. Việc giảm độ phức tạp mô hình có thể tiết kiệm bộ nhớ và tính toán một cách hiệu quả trong khi độ chính xác nhận dạng sẽ giảm xuống.

1.3.3. Độ đo khả năng

Các khía cạnh quan trọng của các điều kiện hoạt động bao gồm mức độ nhiễu, kênh nhiễu và độ méo tín hiệu, các người nói khác nhau, cú pháp và ngữ nghĩa khác nhau…Trong thực tế, sự chênh lệch của những ràng buộc này từ những giả định trong giai đoạn thiết

kế có thể dẫn đến sự giảm sút đáng kể đến hiệu năng hoạt động của hệ thống.

1.4. ĐẶC TRƯNG ÂM HỌC

1.4.1. Bản chất của âm

Tất cả các âm đều bắt nguồn từ dao động thuộc kiểu này hay khác, những người chơi nhạc biểu diễn các hành động kiểu như cử động tay hay thổi bằng miệng, và hoạt động của họ tạo ra nhiều kiểu loại dao động khác nhau mà chúng ta nghe thành các âm.

Để tạo ra âm nghe được, ba tiêu chí đi kèm sau đây phải được thoả mãn đồng thời.

- Phương tiện lan truyền.
- Một âm phải nằm ở trong vùng tần số nghe được.
- Biên độ của âm đủ lớn để có thể thu nhận được.

Về chất lượng các âm không được tiếp nhận hoàn toàn giống nhau. Chúng ta có thể phân biệt hai bình diện cơ bản.

- Phân biệt giữa các âm liên tục và các âm rời rạc.
- Phân biệt các âm nhạc tính (musical sounds) từ các âm ồn (noise - like sound).

Một phương cách quan trọng nữa mà nhờ đó các âm phân biệt nhau là ở chất lượng hay âm sắc của âm.

1.4.2. Ngữ âm tiếng Việt

Tiếng việt được xem là một ngôn ngữ đơn lập tiêu biểu mà đặc điểm cơ bản của nó là: âm tiết giữ một vai trò cơ bản trong hệ thống các đơn vị ngôn ngữ; vốn từ vựng cơ bản của tiếng Việt đều là từ đơn tiết và mỗi âm tiết đều có khả năng tiềm tàng trở thành từ; các từ không biến hình.

Trên phương diện ngữ âm, âm tiết tiếng Việt được xem là một đơn vị cơ bản. Âm tiết tiếng Việt có cấu trúc đơn giản, luôn gắn liền với thanh điệu, được tách biệt trong chuỗi lời nói.

Tóm lại, trong chương này tác giả luận văn đã tập trung tìm hiểu xu hướng phát triển lĩnh vực xử lý ngôn ngữ, đặc điểm của một hệ thống nhận dạng và các phương pháp tiếp cận nhận dạng tiếng nói. Tiếp đến trình bày các tiêu chí cụ thể để đánh giá hiệu suất của một hệ thống nhận dạng. Phần cuối chương, tập trung tìm hiểu về các đặc trưng cơ bản của âm học, và ngữ âm tiếng Việt.

CHƯƠNG 2 - HỆ THỐNG NHẬN DẠNG TIẾNG NÓI

Trong chương này, tác giả luận văn tập trung trình bày các kỹ thuật tiền xử lý tín hiệu tiếng nói nhằm trích chọn các đặc trưng của tín hiệu tiếng nói phù hợp cho giai đoạn nhận dạng, cụ thể cách thức xác định dữ liệu tiếng nói, phát hiện điểm đầu và điểm cuối của tín hiệu, phương pháp rút trích đặc trưng MFCC phổ biến trong các hệ thống nhận dạng hiện nay. Tiếp đến trình bày chi tiết ứng dụng mô hình Markov ẩn trong nhận dạng tiếng nói, và các phương pháp ứng dụng khác, thực hiện so sánh một số kết quả nhận dạng tiếng nói trước đây.

2.1. TIỀN XỬ LÝ TÍN HIỆU

Đây là một giai đoạn quan trọng ảnh hướng rất nhiều đến kết quả nhận dạng, nhất là khi hệ thống được đem ra sử dụng ngoài thực tế. Bởi vì nếu xử lý không tốt sẽ không nhận được dữ liệu tốt, mà dữ liệu đầu vào không đúng thì hệ thống cho ra kết quả sai là điều khó tránh khỏi.

2.1.1. Xác định dữ liệu tiếng nói

Dữ liệu thu được không phải lúc nào cũng là tiếng nói, nhất là khi thu động dữ liệu sẽ thường xuyên là khoảng lặng và nhiễu. Vì hệ thống nhận dạng được thiết kế theo dạng mô hình hóa nhằm so khớp tìm mẫu có xác suất tín hiệu quan sát là lớn nhất nên dù dữ liệu thu được không phải là tiếng nói mà được đưa vào thì hệ thống vẫn gán đó là một trong các tiếng đã học mẫu, điều này là sai hoàn toàn.

2.1.2. Phát hiện điểm đầu và cuối của một từ

Một trong những vấn đề cơ bản của xử lý tiếng nói là xác định điểm bắt đầu và kết thúc của một từ. Điều này khó thực hiện chính xác nếu tín hiệu được nói trong môi trường nhiễu. Việc phát hiện điểm đầu và cuối của một từ tốt, cho hiệu quả nhận dạng tối ưu.

2.2. RÚT TRÍCH ĐẶC TRƯNG

Giải pháp trích đặc trưng tín hiệu tiếng nói được hiểu như là một quá trình biến đổi từ vector có kích thước lớn sang vector có kích thước nhỏ hơn. Như vậy, về mặt hình thức, rút trích đặc trưng có thể được định nghĩa như một ánh xạ f:

$$f : R^N \rightarrow R^d, \text{ trong đó } d << N.$$

Một đặc trưng được cho là tốt cần phải có các tính chất sau:

- Sai biệt giữa các vector đặc trưng của những người nói khác nhau phải lớn.
- Sai biệt giữa các vectors đặc trưng của cùng một người nói phải nhỏ.
- Độc lập với các đặc trưng khác

2.2.1. Pre-emphasis

Mục tiêu của bước pre-emphasis là để củng cố các tần số cao bị mất trong quá trình thu nhận tín hiệu.

2.2.2. Phân khung

Dữ liệu tiếng nói thường không ổn định, nên thông thường phép biến đổi Fourier được thực hiện trên từng đoạn tín hiệu ngắn. Mục tiêu của bước chia khung là chia dữ liệu tiếng nói thành từng khung nhỏ có kích thước khoảng từ 20ms đến 30ms.

Việc nhân mỗi khung với hàm cửa sổ sẽ giúp củng cố tính liên tục ở 2 biên của khung và tạo tính chu kỳ cho toàn bộ tín hiệu trong khung.

2.2.3. Biến đổi Fourier rời rạc (Discrete Fourier Transform – DFT)

Sau khi tín hiệu được đưa qua hàm cửa sổ, biến đổi Fourier rời rạc (DFT) được sử dụng để chuyển đổi mẫu tín hiệu từ miền thời gian sang miền tần số.

2.2.4. Bộ lọc Mel

Bộ lọc Mel là một dãy các bộ lọc dạng tam giác chồng lên nhau với tần số cắt của mỗi bộ lọc được xác định bởi tần số trung tâm của hai bộ lọc kề với nó. Mục tiêu của bước áp dụng các bộ lọc Mel là để lọc lấy các tần số mà tai người có thể nghe được hoặc để nhấn mạng tần số thấp trên tần số cao, đồng thời rút ngắn kích thước của vector đặc trưng.

2.2.5. Biến đổi Cosine rời rạc (Discrete Cosine Transform – DCT)

2.3. MÔ HÌNH MARKOV ẨN

2.3.1. Quá trình Markov

Xét một hệ thống mà ở đó tại bất kì thời điểm nào ta cũng có thể mô tả nó bởi một trong N trạng thái phân biệt $S_1, S_2,\ldots,S_N$ *(N=3)*. Tại thời điểm t bất kỳ, hệ thống có thể đo được xác suất chuyển từ trạng thái S_i hiện hành sang một trong N-1 trạng thái còn lại hoặc chuyển trở lại chính trạng thái S_i.

Kết xuất của hệ thống là một chuỗi các trạng thái tại các thời điểm t tương ứng.

2.3.2. Mô hình markov ẩn

HMM gồm các thành phần sau đây:

1) N – số lượng trạng thái của mô hình.
2) M – số lượng tín hiệu có thể quan sát được trong mỗi trạng thái.
3) Các xác suất chuyển trạng thái A = $\{a_{ij}\}$
4) Các hàm mật độ xác suất trong mỗi trạng thái B = { $b_j(k)$ }
5) Xác xuất khởi đầu của mỗi trạng thái $\pi = \{\pi_i\}$.

Để thuận tiện, ta quy ước mỗi mô hình HMM sẽ được đại diện bởi bộ tham số λ = (A, B, π).

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

2.3.3.1. Bài toán 1 – Đánh giá xác suất

Một tiêu của bài toán thứ nhất là tính p(O| λ) – xác suất phát sinh O từ mô hình λ.

2.3.3.2. Bài toán 2 – Tìm chuỗi trạng thái tối ưu

Mục tiêu của bài toán 2 là tìm ra chuỗi trạng thái "tối ưu" nhất Q = $q_1\ q_2\ \ldots\ q_T$ đã phát sinh ra O.

2.3.3.3. Bài toán 3 – Vấn đề huấn luyện

Mục tiêu của bài toán thứ 3, cũng là bài toán phức tạp nhất trong ba bài toán, là tìm cách cập nhật lại các tham số của mô hình λ = (A, B, π) sao cho cực đại hóa xác suất p(O| λ) – xác suất quan sát được chuỗi tín hiệu O từ mô hình.

2.4. MỘT SỐ HỆ THỐNG NHẬN DẠNG TIẾNG NÓI

2.4.1. Hệ thống VQ

Hệ thống Vector Quantization sẽ ước lượng codebook cho từng mẫu tiếng nói từ tập dữ liệu huấn luyện. Trong bước nhận dạng, sai số quantization error (khoảng cách euclid) giữa mẫu test với codeword gần nó nhất trong codebook của từng mẫu tiếng nói sẽ được tính; và mẫu test sẽ được phân vào lớp có sai số lỗi lượng tử thấp nhất.

2.4.2. Hệ thống GMM

Đối với hệ thống GMM, đây cũng là một phương pháp gom cụm giống như VQ, mỗi dữ liệu tiếng nói sẽ được mô hình hóa bằng một GMM. Một mô hình GMM có kích thước M sẽ gồm M hàm mật độ Gauss với các tham số là vector trung bình μ và ma trận hiệp phương sai Σ.

2.4.3. Một số hệ thống nhận dạng khác

Ngoài hai phương pháp truyền thống là GMM và VQ, các công trình nghiên cứu gần đây đã tiếp cận bài toán theo một số hướng khác như Support Vector Machine (SVM), mạng neural (NN).

CHƯƠNG 3 - ĐỀ XUẤT GIẢI PHÁP VÀ CÀI ĐẶT THỬ NGHIỆM

3.1. ĐỀ XUẤT GIẢI PHÁP

3.1.1. So sánh các loại mô hình Markov ẩn

Có nhiều cách phân loại các mô hình Markov ẩn, trong đó người ta thường phân biệt dựa vào đặc trưng của ma trận chuyển trạng thái A_{ij}, có thể phân loại thành mô hình Markov ẩn có liên kết đầy đủ và mô hình Markov ẩn trái phải (Bakis). Hoặc là dựa vào tính chất của hàm mật độ xác xuất quan sát $B_j(k)$, người ta phân loại thành mô hình Markov ẩn rời rạc (DHMM), mô hình Markov ẩn liên tục (CDHMM), mô hình Markov ẩn bán liên tục (SCHMM):

- **DHMM:** Đối với mô hình Markov ẩn rời rạc, không gian vector đặc trưng của tín hiệu tiếng nói được chia vào hữu hạn các vùng (cluster) bằng một thủ tục phân nhóm chẳng hạn như lượng hóa vector (VQ).
- **CDHMM:** Lỗi lượng tử hóa vector đã được loại trừ bằng cách sử dụng hàm mật độ liên tục thay vì lượng hóa vector. Trong CDHMM, phân bố xác suất trên không gian vector âm học được mô hình hóa trực tiếp sử dụng hàm mật độ xác suất liên tục (PDF) chẳng hạn như hàm trộn của các hàm Gaussian.
- **SCHMM:** Mô hình này cung cấp chi tiết dữ liệu mô hình hóa thông qua việc chia sẽ các tham số. Mô hình này là một sự kết hợp giữa DHMM và CDHMM.

3.1.2. So sánh các phương pháp nhận dạng đã được triển khai

3.1.2.1. Phương pháp DTW

Hướng tiếp cận DTW là phương thức đối sánh mẫu, trong đó thuật toán thực hiện so sánh mẫu kiểm thử với mẫu tham chiếu để có số điểm tối thiểu.

3.1.2.2. Phương pháp ANN

Mạng nơ ron nhân tạo (NN) là một kiến trúc mạnh mẽ và linh hoạt để giải quyết vấn đề phân lớp. NN có thể học một cách hiệu quả và theo một cách riêng biệt.

3.1.3. Hướng tiếp cận và phát triển của đề tài

Hướng tiếp cận nghiên cứu của luận văn tập trung vào giải quyết một số phần sau đây:

- Tiền xử lý tín hiệu tiếng nói nhằm khử nhiễu và phát hiện tín hiệu dữ liệu tiếng nói. Sau đó tiến hành rút trích đặc trưng dữ liệu tiếng nói theo MFCC bao gồm các hệ số ceptral, năng lượng chuẩn hóa cùng với các hệ số đạo hàm bậc một, bậc hai của chúng (delta và Delta-delta)
- Nghiên cứu mạng nơ ron và mô hình Markov ẩn trong nhận dạng tiếng nói tiếng Việt.
- Đối với nhận dạng các chữ số rời rạc, sử dụng mạng nơ ron huấn luyện dữ liệu thực hiện sự phân lớp các phổ tín hiệu tiếng nói (gán nhãn cưỡng bức), sau đó thực hiện thuật toán Viterbi để nhận dạng dữ liệu.
- Thực hiện đánh giá tỉ lệ lỗi nhận dạng.

3.2. CÀI ĐẶT HỆ THỐNG

Hoạt động của hệ thống được thực hiện như sau:

- Đầu tiên phân chia tín hiệu tiếng nói thu được thành các khung tín hiệu.
- Tính toán các đặc trưng của mỗi khung tín hiệu. Những đặc trưng này có thể được dùng để biểu diễn vùng bao phủ đặc trưng phổ của tiếng nói tại khung tín hiệu đó và một số nhỏ các khung tín hiệu xung quanh gọi là “cửa sổ phạm vi”.

- Phân lớp các đặc trưng trong mỗi khung vào trong mỗi loại dựa trên âm học sử dụng mạng nơ ron. Đầu ra của mạng nơ ron là các ước lượng xác suất của mỗi loại ngữ âm, ứng với các đặc trưng tiếng nói tại khung tín hiệu này. Khi mạng nơ ron được sử dụng để phân lớp tất cả các khung, tạo ra một ma trận xác suất, với F cột và C hàng, trong đó F là số lượng các frame và C là số lượng phân loại.
- Sử dụng ma trận xác suất, tập các mô hình ngữ âm để xác định các từ cần nhận dạng thích hợp nhất sử dụng thuật toán tìm kiếm Viterbi trong mô hình HMM.

3.2.1. Mô hình âm vị

Trong từ điển phát âm, mỗi từ được phiên âm thành các âm vị và một từ có thể bao gồm một vài định nghĩa khác nhau. Để xây dựng các đơn vị nhận dạng phụ thuộc ngữ cảnh, các âm vị được chia thành một, hai hoặc ba phần, mỗi phần như vậy được gọi là category và là đơn vị nhận dạng cơ bản của hệ thống nhận dạng. Mỗi category phụ thuộc vào ngữ cảnh ở bên trái hoặc bên phải của nó.

3.2.2. Huấn luyện

Quá trình huấn luyện mạng nơron được thực hiện với từng phát âm dùng thủ tục truyền ngược sai số. Với mỗi phát âm, thông tin nhãn thời gian trong cơ sở dữ liệu tiếng nói cho ta các khoảng thời gian thuộc về âm vị cần huấn luyện. Như vậy với mỗi category các khoảng thời gian dành cho chúng được xác định trong mỗi phát âm. Các thông tin này được dùng để huấn luyện cho mạng ANN.

3.2.3. Nhận dạng

3.2.3.1. Mạng từ

Mạng từ (word network) được dùng để định nghĩa một ngữ pháp, mối liên hệ thứ tự giữa các từ được nhận dạng bởi hệ thống. Một tệp định nghĩa mạng từ chứa một danh sách các nút biểu diễn các từ và một danh sách các cung biểu diễn chuyển dịch giữa các từ.

3.2.3.2. Sử dụng mạng từ trong hệ thống nhận dạng

Khi mạng từ được nạp vào trong hệ thống nhận dạng, một từ điển phiên âm của hệ thống sẽ được dùng để tạo ra một mạng tương đương bao gồm các đơn vị nhận dạng cơ bản của hệ thống, các âm đơn hoặc các âm ba.

3.2.3.3. Giải mã

Nhiệm vụ của quá trình giải mã là tìm ra một đường đi trong mạng HMM có xác suất lớn nhất. Để thực hiện công việc này, thực hiện thuật toán Viterbi đã được trình bày.

3.3. KẾT QUẢ THỬ NGHIỆM

3.3.1. Dữ liệu tiếng nói

Hệ thống nhận dạng tiếng nói tiếng Việt được xây dựng và đánh giá hiệu suất nhận dạng dựa trên tập dữ liệu các chữ số rời rạc tiếng Việt phụ thuộc người nói (speaker-dependent). Tập dữ liệu huấn luyện bao gồm 1000 phát biểu rời rạc cho các chữ số từ 0 đến 9, được thu âm từ 10 người, tốc độ đọc 0.8 giây/1 từ, tần số lấy mẫu 8000Hz, độ phân giải 16 bits. Đối với nhận dạng phụ thuộc người nói, tập dữ liệu kiểm tra được lấy từ tập dữ liệu huấn luyện.

3.3.2. Kết quả nhận dạng phụ thuộc người nói

Thử nghiệm đã được thực hiện đối với nhận dạng chữ số rời rạc tiếng Việt phụ thuộc người nói để đánh giá độ chính xác khác nhau giữa CDHMM và HMM/ANN trong nhận dạng. Kết quả thử nghiệm như trong bảng 3.1 cho thấy độ chính xác nhận dạng của HMM/ANN tốt hơn so với CDHMM.

Bảng 3.1 So sánh kết quả nhận dạng phụ thuộc người nói

Mô hình nhận dạng	Độ chính xác (%)
CDHMM/BW	96,62
HMM/ANN	99,25

Trong chương này, tác giả luận văn đã tập trung phân tích và so sánh các phương pháp triển khai ứng dụng nhận dạng tiếng nói, từ đó đề xuất hướng giải quyết bài toán nhận dạng sử dụng HMM/ANN. Phần cài đặt hệ thống, tác giả đã giới thiệu chi tiết về mô hình hệ thống, các giai đoạn từ thu thập đến huấn luyện và nhận dạng sử dụng HMM/ANN. Cuối cùng, thực hiện thực nghiệm nhận dạng trên tập dữ liệu tiếng nói.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Kết quả thực hiện luận văn "Nhận dạng tiếng nói tiếng Việt sử dụng mạng nơ-ron và mô hình Markov ẩn" đã tập trung giải quyết một số nội dung về nhận dạng tiếng nói tiếng Việt. Ở chương 1 trình bày xu hướng phát triển lĩnh vực xử lý ngôn ngữ, nghiên cứu các hướng tiếp cận nhận dạng, các tiêu chí đánh giá ảnh hưởng đến hiệu suất nhận dạng, cuối chương tập trung tìm hiểu đặc trưng cơ bản của tiếng Việt như cấu trúc âm tiết, loại hình âm tiết. Chương 2, tác giả tập trung trình bày các bước xử lý tín hiệu tiếng nói từ giai đoạn thu thập, khử nhiễu, phát hiện tiếng nói cho đến rút trích các tham số đặc trưng. Tiếp đến, nghiên cứu đầy đủ và chi tiết ứng dụng mô hình Markv ẩn trong nhận dạng tiếng nói. Trong chương 3, tác giả luận văn thực hiện so sánh các phương pháp nhận dạng sử dụng mô hình Markov ẩn kết hợp mạng nơ-ron, với các phương pháp khác đã được triển khai, từ đó đề xuất hướng tiếp cận phát triển của đề tài. Phần cuối chương trình bày hệ thống nhận dạng tiếng nói được triển khai, từ việc khởi tạo mô hình, huấn luyện và nhận dạng tiếng nói. Thực hiện so sánh và đánh giá kết quả thử nghiệm trên tập dữ liệu rời rạc 10 chữ số.

Với nền tảng kiến thức đã được nghiên cứu và kết quả của luận văn, một số định hướng phát triển của luận văn có thể thực hiện trong thời gian đến như:

- Nghiên cứu quá trình xử lý tiếng nói làm sao để có thể tách được tiếng nói trong môi trường nhiễu (tiếng ồn) lớn.
- Trên cơ sở xác định mẫu tiếng nói, tiến tới mở rộng phát triển hệ thống xác định danh tính người nói phục vụ cho ứng dụng bảo mật.

- Mở rộng tập dữ liệu huấn luyện với số lượng người nói và số từ nói nhiều hơn nữa tận dụng tối đa ưu điểm của mô hình CDHMM.
- Phát triển hệ thống nhận dạng từ liên tục.