BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

Ngô Minh Dũng

NGHIÊN CỨU KỸ THUẬT NHẬN DẠNG NGƯỜI NÓI DỰA TRÊN TỪ KHÓA TIẾNG VIỆT

Chuyên ngành : Công nghệ phần mềm
Mã số : 62.48.10.01

Tóm tắt
LUẬN ÁN TIẾN SỸ KỸ THUẬT

HÀ NỘI - 2010

Công trình được hoàn thành tại trường **Đại học Bách khoa Hà Nội**

Người hướng dẫn khoa học:

1. **PGS. TS. Đặng Văn Chuyết**
2. **PGS. TS. Vũ Kim Bảng**

Phản biện 1: **PGS. TS. Nguyễn Quang Hoan**

Phản biện 2 : **GS. TS. Nguyễn Văn Khang**

Phản biện 3: **PGS. TS. Ngô Quốc Tạo**

Luận án được bảo vệ trước Hội đồng chấm luận án cấp trường tại Trường Đại học Bách khoa Hà Nội
Vào hồi 14 giờ , ngày 15 tháng 9 năm 2010

Có thể tìm hiểu luận án tại:

- Thư viện Quốc gia
- Thư viện trường Đại học Bách khoa Hà Nội

DANH MỤC CÔNG TRÌNH KHOA HỌC LIÊN QUAN ĐÃ CÔNG BỐ CỦA TÁC GIẢ

1. Ngô Minh Dũng, Đặng Văn Chuyết (2004) , *Khảo sát tính ổn định của một số đặc trưng ngữ âm trong nhận dạng người nói* - Bưu chính viễn thông, Chuyên san Các công trình nghiên cứu, triển khai viễn thông và công nghệ thông tin, số12, 2004, Tr: 70-74
2. Ngô Minh Dũng, Đặng Văn Chuyết (2006) , *Khả năng phân biệt người nói của các âm tiết tiếng Việt ,* Tuyển tập các báo cáo khoa học, Phân ban Công nghệ thông tin, Hội nghị khoa học lần thứ 20 ĐHBKHN, Nhà xuất bản Bách khoa Hà nội, 10/2006. Tr: 135-141
3. Ngô Minh Dũng, Đặng Văn Chuyết (2007) , *Xây dựng và khảo sát độ dài từ khóa trong nhận dạng người nói phụ thuộc từ khóa tiếng Việt theo mô hình Markov ẩn ,* Tạp chí bưu chính viễn thông và công nghệ thông tin, Chuyên san: Các công trình nghiên cứu khoa học, nghiên cứu triển khai Công nghệ thông tin và truyền thông, số 18. 10/2007. Tr: 93-99
4. Ngo Minh Dung, Dang Van Chuyet (2007) , *Mean spectrum of many speakers for robust speaker recognition ,* Proceeding of the 2nd Asia Pacific International conference on information science and technology, Hanoi, 12/2007, pp 139 – 145.

A. THÔNG TIN CHUNG CỦA LUẬN ÁN

1. Tính cấp thiết của đề tài

Tiếng nói là phương tiện trao đổi thông tin phổ biến nhất của con người. Nhận dạng người từ giọng nói hay nhận dạng người nói *(speaker recognition)* cùng với nhận dạng tiếng nói *(speech recognition)* là những lĩnh vực nhận dạng liên quan đến xử lý tiếng nói đang được quan tâm nghiên cứu hiện nay. Tiếng nói, ngoài thông tin ngữ nghĩa mà người nói muốn truyền đạt cho người nghe (những thông tin có thể ghi lại dưới dạng chữ viết), còn chứa những thông tin khác như phương ngữ, trạng thái tình cảm khi nói cũng như những thông tin riêng của giọng nói. Trong khi nhận dạng tiếng nói dựa trên thông tin ngữ nghĩa thì nhận dạng người nói lại dựa vào các thông tin riêng của giọng nói.

Các lĩnh vực ứng dụng nhận dạng người nói hiện nay như xác thực quyền truy cập vào các hệ thống an ninh bằng mật khẩu nói, giám sát người qua giọng nói hay tách tiếng nói của từng người từ môi trường nhiều người nói. Ứng dụng xác thực người nói trong giao dịch sử dụng thẻ tín dụng hay trong giao tiếp điện tử bằng hộp thư thoại có sử dụng kỹ thuật nhận dạng người nói để giúp nhận dạng tiếng nói có được các tham số nhận dạng thích hợp. Ngoài ra, nhận dạng người nói còn có một lĩnh vực ứng dụng khá quan trọng đó là giám định pháp lý nhận dạng người nói *(forensic speaker recognition)*.

Ở nước ta hiện nay, nhận dạng người nói mới bước đầu được ứng dụng trong lĩnh vực giám định pháp lý nhận dạng người nói phục vụ cho công tác điều tra và xét xử tội phạm. Lĩnh vực giám định này chủ yếu liên quan tới quá trình xác thực người nói giữa mẫu tiếng nói được ghi âm xong chưa biết ai nói (*unknown speaker*) và mẫu tiếng nói của những người bị nghi vấn (*suspect speakers*). Đây là một lĩnh vực giám định pháp lý mới với nhiều vấn đề liên quan tới kỹ thuật nhận dạng người nói cần giải quyết khi xây dựng cũng như nâng cao độ tin cậy của kết luận giám định. Cho đến trước năm 2004 chưa có công trình nghiên cứu nào về vấn đề này cho người nói tiếng Việt được công bố. Trước tình hình đó, luận án đã chọn vấn đề nhận dạng người nói tiếng Việt ứng dụng trong giám định pháp lý để nghiên cứu.

2. Mục tiêu nghiên cứu của luận án

Mục tiêu nghiên cứu của luận án là nghiên cứu các kỹ thuật nhận dạng người nói nhằm giải *quyết* các vấn đề liên quan tới nhận dạng người nói tiếng Việt ứng dụng trong giám định pháp lý tại Việt Nam. Các kỹ thuật nhận dạng người nói liên quan tới tiếng Việt như nghiên cứu phạm vi ổn định của một số các tham số tiếng nói đối với mỗi người nói, lựa chọn đơn vị ngữ âm thích hợp để tiến hành so sánh nhận dạng người nói, hay đánh giá khả năng nhận dạng người nói của các đơn vị ngữ âm tiếng Việt… Tất cả nhằm tới mục đích cuối cùng là xây dựng và hoàn thiện một quy trình giám định pháp lý nhận dạng người nói tiếng Việt phục vụ công tác điều tra và xét xử tội phạm tại Việt Nam.

3. Đối tượng và phạm vi nghiên cứu

Để tiến hành nghiên cứu nhận dạng người nói tiếng Việt, đối tượng được luận án chọn để nghiên cứu là tiếng Việt của những người nói giọng Bắc Bộ có tham khảo thêm một số người nói giọng Bắc Trung Bộ (Nghệ Tĩnh) để so sánh. Xong các kết quả nghiên cứu được áp dụng thử nghiệm cho cả những người nói giọng Nam bộ và Trung bộ để đánh giá..

Tất cả các nội dung nghiên cứu trong luận án chỉ giới hạn trong phạm vi điều kiện người nói trong trạng thái bình thường, các vấn đề người nói cố tình cải trang hay giả giọng nói đều nằm ngoài phạm vi nghiên cứu của luận án.

4. Ý nghĩa khoa học và thực tiễn của luận án

Việc nghiên cứu các kỹ thuật nhận dạng người nói tiếng Việt ứng dụng trong giám định pháp lý như phạm vi ổn định một số các tham số tiếng nói đối với mỗi người nói hay lựa chọn đơn vị ngữ âm thích hợp cho tiếng Việt cũng như nghiên cứu về khả năng nhận dạng người nói của các đơn vị ngữ âm này… sẽ đóng góp vào bức tranh tổng thể về nghiên cứu nhận dạng người nói nói chung, phục vụ cho các ứng dụng khác nhau của nhận dạng người nói tiếng Việt.

Kết quả nghiên cứu của luận án góp phần trực tiếp xây dựng, phát triển lĩnh vực giám định pháp lý nhận dạng người nói tại Việt Nam. Điều này có ý nghĩa thực tiễn to lớn trong công tác điều tra và xét xử tội phạm liên quan tới người nói tiếng Việt, nhất là trong bối cảnh số vụ án có liên quan tới tiếng nói tại nước ta tăng nhanh trong những năm qua và sẽ còn tiếp tục tăng trong những năm tới theo sự phát triển mạnh mẽ của các thiết bị thông tin viễn thông.

5. Kết cấu luận án

Nội dung luận án được chia thành 4 chương, 110 trang, 5 bảng số liệu, 31 hình vẽ và đồ thị, 49 tài liệu tham khảo và 40 trang phụ lục.

B. NỘI DUNG CHÍNH

Chương 1: Tổng quan về nhận dạng người nói

1.1 Cơ sở khoa học của nhận dạng người nói

Tiếng nói tự nhiên do cơ quan cấu âm của con người tạo ra. Đặc tính riêng trong giọng nói của mỗi người hay đặc tính riêng của người nói là một hiện tượng phức tạp được hình thành từ 2 yếu tố: cấu tạo giải phẫu sinh lý cơ quan cấu âm của con người và những đặc điểm phát âm mà con người học được trong cuộc sống. Một yếu tố đặc trưng cho cấu trúc vật lý của cơ quan cấu âm còn yếu tố kia đặc trưng cho hành vi hoạt động của nó.

1.2. Thông tin đặc trưng giọng nói mỗi người

Các thông tin đặc trưng cho giọng nói của mỗi người được thể hiện ở nhiều mức khác nhau, từ các đặc trưng mức cao như phong cách nói, cách sử dụng cú pháp hay từ vừng khi nói, đến các đặc trưng mức thấp hơn như ngôn điệu, ngữ âm, cho tới mức thấp nhất là các đặc trưng âm thanh. Các thông tin đặc trưng mức cao có ưu điểm là ít bị ảnh hưởng bởi nhiễu và kênh truyền xong rất khó trích chọn tự động, mô hình hóa phức tạp và thường phải yêu cầu thời gian phát âm đủ lớn, trong khi đó thông tin đặc trưng mức thấp thì ngược lại rất dễ bị tác động bởi nhiễu và kênh truyền xong trích chọn tự động dễ dàng hơn, mô hình hóa cũng đơn giản hơn và thường không yêu cầu nhiều về thời gian phát âm.

1.3. Các phương pháp nhận dạng người nói hiện nay trên thế giới

Có 3 phương pháp nhận dạng người nói hiện nay:

- Nhận dạng người nói bằng bằng cơ quan thính giác của con người.
- Phương pháp thủ công : so sánh ảnh phổ của hai mẫu tiếng nói để quyết định xem liệu chúng có phải do cùng một người nói ra không.
- Phương pháp tự động: nhận dạng người nói được thực hiện tự động dựa trên việc mô hình hoá tín hiệu tiếng nói bằng cách trích chọn thông tin đặc trưng người nói và sử dụng các thuật toán máy tính phân lớp nhận dạng các mô hình người nói này.

1.4 Nguyên lý làm việc hệ nhận dạng người nói

Như mọi hệ nhận dạng thông thường, cấu trúc của một hệ nhận dạng người nói cũng bao gồm hai modul cơ bản là trích chọn đặc trưng và phân lớp nhận dạng, trong đó modul phân lớp nhận dạng gồm hai thành phần là đối sánh mẫu và quyết định nhận dạng.

Cơ sở dữ liệu bao gồm các mô hình người nói được tạo ra trong pha huấn luyện. Trong pha nhận dạng, mẫu tiếng nói của người chưa biết sẽ được đối sánh với các mô hình người nói có trong cơ sở dữ liệu để ra quyết định nhận dạng.

Hiện có nhiều phương pháp phân lớp nhận dạng người nói xong chủ yếu là sử dụng các mô hình thống kê như mô hình Markov ẩn (HMM) hay mô hình hỗn hợp Gauss (GMM).

1.5 Các nguyên nhân gây lỗi trong nhận dạng người nói

- Tính không ổn định của tiếng nói của mỗi người theo sức khỏe thể chất và tâm lý
- Cải trang hay giả giọng là cố tình làm thay đổi giọng nói.
- Các tác nhân kỹ thuật (được gọi chung là nhiễu) làm mất tính trung thực của tiếng nói. Ngoài ra điều kiện ghi âm khác nhau cũng là nguyên nhân gây lỗi trong nhận dạng người nói.

1.6 Sơ lược tình hình nghiên cứu nhận dạng người nói

1.6.1 Nghiên cứu nhận dạng người nói bằng phương pháp thủ công

Đầu những năm 60 của thế kỷ trước, Lawrence Kersta đã lần đầu tiên thực hiện nhận dạng người từ tiếng nói bằng cách so sánh ảnh phổ ba chiều của tiếng nói tại phòng thí nghiệm tiếng nói của hãng Bell Telephone. Về cơ bản, nguyên tắc nhận dạng người nói bằng phương pháp thủ công này vẫn được giữ nguyên cho đến nay.

1.6.2 Nghiên cứu nhận dạng người nói bằng phương pháp tự động

Hiện các vấn đề về nhận dạng người nói chủ yếu tập trung vào việc nghiên cứu nâng cao khả năng nhận dạng của các hệ nhận dạng người nói đặc biệt trong điều kiện tiếng nói bị suy giảm (méo) do các tác nhân kỹ thuật gây ra. Hướng nghiên cứu chính là khai thác các thông tin mức cao của tiếng nói, hay áp dụng cải tiến các kỹ thuật sẵn có…

1.7 Giám định pháp lý nhận dạng người nói và vấn đề tiếng Việt

Giám định pháp lý nhận dạng người nói là một ứng dụng quan trọng các phương pháp nhận dạng người nói trong điều tra và xét xử tội phạm. Hiện trên thế giới tồn tại hai phương pháp giám định nhận dạng người nói: Phương pháp nghe-phân tích phổ âm thanh (phương pháp kinh điển) và phương pháp tự động.

1.7.1 Phương pháp kinh điển giám định pháp lý nhận dạng người nói

Đây là một phương pháp giám định nhận dạng người nói tổng hợp, kết hợp phương pháp nhận dạng người nói bằng cảm thụ của cơ quan thính giác con người với phương pháp nhận dạng người nói thủ công và đo lường tự động một số các tham số tiếng nói để đối sánh. Ưu điểm của phương pháp này thường cho kết luận giám định với độ chính xác và độ tin cậy cao. Nhược điểm là chậm và tốn nhiều công sức.

1.7.2 Phương pháp tự động giám định pháp lý nhận dạng người nói

Đây là phương pháp giám định nhận dạng người nói hoàn toàn dựa vào sự phân tích và so sánh các mẫu tiếng nói bằng máy tính theo nguyên tắc làm việc của các phương pháp nhận dạng người nói tự động. Ưu điểm của phương pháp giám định tự động là thời gian thực hiện nhanh, ít tốn sức người. Nhược điểm của của phương pháp này là rất nhạy cảm với các loại nhiễu do các mô hình người nói được xây dựng chủ yếu dựa trên các thông tin mức thấp của tiếng nói, những thông tin rất nhảy cảm với nhiễu.

1.7.3 Các vấn đề đặt ra cho giám định nhận dạng người nói tiếng Việt

Phương pháp kinh điển chủ yếu áp dụng khi giám định so sánh hai mẫu tiếng nói có phải do cùng một người nói ra hay không, nên về hoạt động nhận dạng người nói phương pháp này giống một hệ xác thực người nói (đối sánh 1:1). Vì vậy để áp dụng phương pháp giám định kinh điển cho người nói tiếng Việt, cần xác định các ngưỡng nhận dạng cho các tham số tiếng nói tiếng Việt mang thông tin về người nói được sử dụng theo phương pháp này.

Phương pháp tự động giám định nhận dạng người nói được áp dụng chủ yếu khi giám định nhận dạng người nói trên tập dữ liệu nhiều người nói. Về bản chất đây chính là hoạt động của một hệ định danh người nói (đối sánh 1:N). Việc áp dụng các hệ tự động nhận dạng người nói trong thực tế còn gặp nhiều trở ngại, đặc biệt là do các tác nhân kỹ thuật như nhiễu hay điều kiện đối sánh khác nhau gây ra. Ngoài ra, với nhận dạng người nói phụ thuộc từ khóa tiếng Việt, các vấn đề đặt ra như nên chọn những câu, từ tiếng Việt một cách ngẫu nhiên hay có chủ định từ trước, hay chọn đơn vị ngữ âm như thế nào để xây dựng tập từ điển từ khóa tiếng Việt...

Chương 2:Giám định nhận dạng người nói tiếng Việt bằng phương pháp nghe-phân tích phổ âm thanh

2.1 Ngữ âm tiếng Việt với nhận dạng người nói

2.1.1 Một số đặc trưng ngữ âm tiếng Việt

Tiếng Việt là ngôn ngữ đơn âm tiết và có thanh điệu. Trong tiếng Việt đơn vị phát âm nhỏ nhất đồng thời cũng là đơn vị ngôn ngữ có ý nghĩa nhỏ nhất. Đặc điểm của ngữ âm tiếng Việt là tính cố định về vị trí của âm vị trong âm tiết tạo nên tính thống nhất trong cấu trúc âm tiết. Khi nghiên cứu về cấu âm, trong tiếng Anh vai trò âm tiết khá mờ nhạt so với âm vị, còn trong tiếng Việt âm tiết đóng vai trò quan trọng không kém so với âm vị.

2.1.2 Đặc trưng ngữ âm tiếng Việt với nhận dạng người nói

Đơn vị ngôn ngữ có ý nghĩa nhỏ nhất (hình vị) có vai trò như những viên gạch để xây nên các từ, các câu trong ngôn ngữ nói. Do vậy, trong nhận dạng người nói phụ thuộc từ khóa, nghiên cứu khả năng phân biệt người nói của hình vị đóng một vai trò quan trong trong việc chọn lựa từ khóa. Việc nghiên cứu này cũng có ý nghĩa quan trọng tương tự như trong việc lựa chọn từ để so sánh trong giám định pháp lý nhận dạng người nói bằng phương pháp kinh điển.

Trong tiếng Việt, đơn vị ngữ âm đóng vai trò hình vị không phải là âm vị mà là âm tiết [49], nên bên cạnh việc nghiên cứu khả năng phân biệt người nói của các âm vị với tư cách là đơn vị ngữ âm

nhỏ nhất, cần tập trung nghiên cứu khả năng phân biệt người nói của các âm tiết với vai trò là đơn vị phát âm nhỏ nhất đồng thời cũng là đơn vị ngôn ngữ có ý nghĩa nhỏ nhất.

Do thường có nhiều âm vị trong từ (đa âm tiết) và các âm vị của từ không có tính thống nhất trong cấu trúc từ nên giá trị formant xác định trong toàn bộ từ tiếng Anh ít được quan tâm chú ý. Ngược lại, âm tiết tiếng Việt có tính thống nhất trong cấu trúc: âm đầu, (âm đệm), âm chính, âm cuối. Do cách cấu âm của âm tiết tiếng Việt luôn bắt đầu bằng động tác kép dần lại tại một bộ phận nào đó của cơ quan cấu âm dẫn đến chỗ cản trở luồng khí từ phổi đi lên, sau đó mở ra, nên năng lượng âm phát ra của phần đầu âm tiết (âm đầu) luôn nhỏ sau đó mới mạnh lên ở phần trung tâm (âm chính) và giảm dần ở phần cuối âm tiết (âm cuối). Chính cách phân bố năng lượng có quy luật như vậy làm cho ranh giới giữa các âm tiết trong tiếng Việt tương đối rõ ràng. Bên cạnh đó, mỗi âm tiết tiếng Việt lại có một thanh điệu riêng nên âm tiết càng được phân tách rõ ràng hơn, dẫn đến không có hiện tượng nối âm, luyến âm hay nuốt âm khi phát âm hai âm tiết tiếng Việt đứng cạnh nhau như tiếng Anh. Điều này gợi ý có thể sử dụng âm tiết làm đơn vị so sánh hai mẫu tiếng Việt trong giám định nhận dạng người nói theo phương pháp kinh điển thay vì ở mức từ, hoặc cụm từ như tiếng Anh.

Với số lượng âm vị trong mỗi âm tiết tương đối ít nên các formant, được xác định trong phạm vi toàn âm tiết tiếng Việt, ngoài phản ảnh chủ yếu âm sắc của âm chính (nơi tập trung nhiều năng lượng nhất của âm tiết), còn có thể chỉ ra được sự ảnh hưởng của âm đầu, âm cuối và cả âm đệm (nếu có) lên âm sắc của âm chính. Nếu thực sự giá trị các formant này (tạm gọi là formant của âm tiết hay formant trong âm tiết) có khả năng phân biệt được người nói, sẽ làm cho việc xác định và so sánh các formant trong giám định nhận dạng người nói tiếng Việt trở nên đơn giản hơn so với tiếng Anh.

2.2 Các tham số tiếng nói trong nhận dạng người nói

Các tham số tiếng nói thường được sử dụng trong giám định pháp lý nhận dạng người nói thực hiện theo phương pháp giám định kinh điển là formant, tần số cơ bản và phổ trung bình thời gian dài. Với các ngôn ngữ đa âm tiết như tiếng Anh, các khúc đoạn để xác định và so sánh các formant thường thuộc phạm vị âm vị. Phân tích ngữ âm tiếng Việt cho thấy có thể sử dụng giá trị formant trong phạm vi âm tiết để so sánh.

2.3 Các formant trong âm tiết tiếng Việt

Các formant được định nghĩa là các tần số cộng hưởng của tuyến phát âm, do vậy liên quan trực tiếp tới hình dạng, kích thước của cơ quan cấu âm và vì thế chúng cung cấp nhiều thông tin đặc trưng về người nói.

2.3.1 Một số đặc điểm cấu trúc formant trong âm tiết tiếng Việt

Với các âm tiết có âm chính là nguyên âm dòng trước, formant thứ nhất nằm ở vùng tần số khoảng 300 - 600 Hz, formant thứ 2 nằm ở vùng tần số khoảng 1600 - 2200 Hz., formant thứ ba và thứ tư nằm ở vùng tần số khoảng từ 2000 - 3600 Hz. Với các âm tiết có âm chính là nguyên âm dòng giữa, formant thứ nhất nằm ở vùng tần số khoảng 600 - 1200 Hz, formant thứ 2 nằm ở vùng tần số khoảng 1200 - 1800 Hz., formant thứ ba và thứ tư nằm ở vùng tần số khoảng từ 2000 - 3600 Hz. Với các âm tiết có âm chính là nguyên âm dòng sau, formant thứ nhất nằm ở vùng tần số khoảng 300 - 800 Hz, formant thứ 2 nằm ở vùng tần số khoảng 700 - 1200 Hz., formant thứ ba và thứ tư nằm ở vùng tần số khoảng từ 1800 - 3600 Hz.

Trong mỗi âm tiết tiếng Việt, cấu trúc formant của nguyên âm bị thay đổi khi đi với âm đầu hoặc/và âm cuối. Sự ảnh hưởng của âm đầu lên cấu trúc formant của nguyên âm ít hơn so với âm cuối.

2.3.2 Đánh giá các phương pháp xác định formant

Vì tuyến âm được coi là không đổi trong khoảng thời gian 10-30ms, nên thông thường các formant được xác định trong mỗi 10-30ms của tiếng nói. Tuy nhiên, việc so sánh định lượng giữa các formant trên từng khúc đoạn nhỏ 10-30ms rất khó thực hiện, do tính không ổn định của tiếng nói nên việc căn lề xác định các khúc đoạn tương ứng giữa các mẫu tiếng nói gặp rất nhiều khó khăn.

Để khắc phục vấn đề này, giá trị các formant có thể được xác định và so sánh trên các khúc đoạn lớn hơn và thường ở mức phạm vi âm vị như trong nhận dạng người nói tiếng Anh vẫn sử dụng. Tuy vậy, việc so sánh này vẫn chưa thực sự dễ dàng vì có sự ảnh hưởng lẫn nhau giữa các âm vị đứng cạnh nhau, nên không có ranh giới rõ ràng giữa các âm vị này.Với tiếng Việt, việc so sánh các formant được xác định trong các khúc đoạn tương ứng thuộc phạm vi âm tiết sẽ dễ dàng

hơn so với phạm vi âm vị hay nhỏ hơn. Vấn đề là đánh giá khả năng phân biệt người nói khi sử dụng giá trị các formant trong phạm vi âm tiết tiếng Việt.

2.3.3 Xây dựng cơ sở dữ liệu người nói tiếng Việt

Để tiến hành nghiên cứu nhận dạng người nói trên các âm tiết tiếng Việt, luận án đã tiến hành xây dựng một cơ sở dữ liệu người nói với 17 âm tiết sau để khảo sát so sánh, đó là 10 âm tiết số "Một", "Hai", "Ba", "Bốn", "Năm", "Sáu", "Bẩy", "Tám", "Chín", "Không" và 7 âm tiết khác là các âm tiết : "Có", "Tôi", "Đã", "Luôn", "Sợ", "Hết", "Tiền".

Cơ sở dữ liệu người nói được xây dựng với 150 người và được chia thành 2 tập dữ liệu người nói (100 người và 50 người). Tất cả những người này tham gia thực nghiệm nói trong 6 phiên. Trong mỗi phiên, mỗi người được yêu cầu đếm từ 1 đến 9, rồi nói cụm từ "*Không có*" và câu "*Tôi đã luôn sợ hết tiền*" trong trạng thái bình thường và nói với tốc độ vừa phải. Trong 5 phiên đầu, mỗi người được ghi âm hai lần. Riêng trong phiên thứ 6, mỗi người được ghi âm 5 lần. Việc ghi âm được thực hiện trực tiếp điều kiện phòng thí nghiệm nhiễu nền thấp, sau đó các âm tiết này được cắt thủ công ra khỏi chuỗi lời nói và lưu vào từng file. Như vậy mỗi người phát âm các âm tiết trên 15 lần trong dòng ngữ lưu rồi được cắt thành các âm tiết đơn lẻ lưu trong các file âm thanh riêng.

2.3.4 Phạm vi thay đổi của các formant trong âm tiết tiếng Việt

Để xác định phạm vi thay đổi của các formant trong âm tiết tiếng Việt đối với mỗi người nói, luận án đã tiến hành khảo sát trên tập dữ liệu người nói thứ nhất được xây dựng ở trên với 100 nói và sử dụng 10 lần phát âm đầu để đánh giá. Với mỗi người, phạm vi biến đổi của từng formant trong 10 lần phát âm cùng một âm tiết được xác định theo công thức sau:

$$T(i) = STD(i) / Mean(i) \ (\%)$$

Với: Mean(i) : Giá trị trung bình của formant thứ i trong âm tiết.

STD(i) : Độ lệch chuẩn của formant thứ i trong âm tiết.

T(i) : phạm vi biến đổi tương đối của formant thứ i trong âm tiết.

Để so sánh với phạm vi biến đổi của từng formant giữa những người nói khác nhau, luận án đã chia 100 người nói với 10 lần phát âm đầu trong tập dữ liệu người nói thứ nhất thành 10 nhóm, mỗi nhóm 10 người. Trong mỗi nhóm này, trên mỗi âm tiết, lần phát âm thứ nhất của từng người trong mỗi nhóm được cho thành một nhóm nhỏ. Tiến hành tương tự như vậy với 9 lần phát âm còn lại, như vậy trong mỗi nhóm sẽ có 10 nhóm nhỏ trên từng âm tiết. Tổng cộng có 100 nhóm nhỏ cho mỗi âm tiết. Với mỗi nhóm nhỏ này, phạm vi biến đổi của từng formant trong 10 lần phát âm cùng một âm tiết của 10 người được xác định tương tự như khi khảo sát trên mỗi người ở trên. Kết quả khảo sát cho trong bảng 1

Bảng 1: Phạm vi biến đổi trung bình formant trong âm tiết

Formant và bề rộng dải thông tương ứng	Phạm vi biến đổi trung bình trong mỗi người nói (%)	Độ lệch chuẩn trung bình phạm vi biến đổi trong mỗi người nói (%)	Phạm vi biến đổi trung bình giữa nhiều người nói khác nhau (%)	Độ lệch chuẩn trung bình phạm vi biến đổi giữa nhiều người nói khác nhau(%)
F1	15.4	10.1	25.3	8.4
F2	10.0	5.7	15.9	5.1
F3	6.3	4.1	10.7	3.2
F4	5.2	2.6	8.6	1.9
B1	25.9	11.8	40.1	10.9
B2	23.7	8.9	34.8	8.4
B3	23.5	8.5	36.2	8.6
B4	22.9	8.3	32.5	8.3

Khảo sát phạm vi thay đổi của các formant được xác định trong các khúc đoạn tương ứng thuộc phạm vi âm tiết cho thấy: Các formant bậc cao có xu hướng ổn định hơn so với các formant bậc thấp. Với mỗi người, phạm vi biến đổi trung bình của các formant từ thứ nhất đến thứ tư vào khoảng 15,4%; 10%; 6,3%,; 5,2%; trong khi đó phạm vi biến đổi trung bình giữa những người nói khác nhau có các giá trị tương ứng là 25,3%; 15,9%; 10,7%; 8,6%. Phạm vi biến đổi trung bình

của bề rộng formant lớn hơn giá trị formant tương ứng. Phạm vi biến đổi trung bình của bề rộng formant của mỗi người cũng lớn hơn phạm vi biến đổi giữa những người nói khác nhau.

Tóm lại, với tiếng Việt, việc so sánh các formant được xác định trong các khúc đoạn tương ứng thuộc phạm vi âm tiết không chỉ dễ dàng hơn trong việc phân tách giới hạn giữa các khúc đoạn, mà còn có thể sử dụng để giám định nhận dạng người nói như các phương pháp đang được áp dụng rộng rãi hiện nay trên các khúc đoạn âm vị.

2.4 Phạm vi thay đổi trung bình của tần số cơ bản

Tiếng Viêt, với đặc thù là ngôn ngữ có thanh điệu, tần số cơ bản luôn thay đổi trong mỗi âm tiết, nên ngoài việc khảo sát phạm vi thay đổi của tần số trung bình đối với mỗi người nói, cần khảo sát thêm yếu tố độ dài thời gian phát âm cần thiết để có thể xác định chính xác giá trị tần số cơ bản trung bình của mỗi người.

Để xác định phạm vi thay đổi của tần số cơ bản đối với mỗi người nói, luận án sử dụng đại lượng độ lệch chuẩn của phân bố thống kê tần số cơ bản trung bình trong khoảng thời gian phát âm. Đại lượng này sẽ biểu thị phạm vi thay đổi hay độ ổn định của tần số cơ bản trung bình của mỗi người nói.

Tiến hành khảo sát trên 35 người độ tuổi từ 25-55 cho thấy với mỗi người nói, mặc dù tần số cơ bản thay đổi liên tục trong mỗi âm tiết do thanh điệu, xong giá trị trung bình của tần số này trong khoảng thời gian phát âm lại có xu hướng ổn định. Thời gian tính tần số cơ bản trung bình càng dài, phạm vi thay đổi trung bình càng có xu hướng giảm dần. Phạm vi thay đổi trung bình của F0 trong các khoảng thời gian khác nhau thể hiện trong bảng 2 của hai giọng nam, nữ (F_{0tb} trong bảng được tính theo khoảng thời gian 6 giây).

Bảng 2. Khảo sát phạm vi thay đổi trung bình của F0 (Hz)

	F_{0tb}	2s	3s	4s	5s	6s	8s	10s	15s
Nam	*132,2*	43,4	37,6	26,5	12,3	10,7	12,1	10,9	9,6
Nữ	*215,3*	47,5	40,2	31,4	23,6	16,3	14,3	15,6	16,1

Kết quả khảo sát cho thấy,

Giọng nam, thời gian tính trung bình từ 5 giây trở lên, tần số cơ bản trung bình thay đổi trong phạm vi khoảng 12 Hz .

Giọng nữ, thời gian tính trung bình từ 6 giây trở lên, tần số cơ bản trung bình thay đổi trong phạm vi khoảng 16 Hz.

2.5. Phổ trung bình trong thời gian dài

Các nghiên cứu về phổ trung bình trong thời gian dài cho thấy đây là một đặc trưng khá ổn định đối với giọng nói của mỗi người ngay cả khi người đó đã cố tình giả giọng nói khác đi so với khi nói bình thường. Khảo sát trên máy phân tích âm thanh Sonagraph DSP với những người nói tự do cho thấy, khi thời gian phát âm tăng phổ trung bình dần tiến tới khá ổn định ở khoảng thời gian 15-30 giây tùy mỗi người. So sánh định tính cho thấy, hình dáng phổ LTA của những người khác nhau thì khác nhau. Để đánh giá sự sai khác này luận án đã sử dụng khoảng cách O'clid để đo khoảng cách giữa 2 phổ LTA trên 50 người phát âm 5 lần thời lượng 20 giây bằng thiết bị phân tích phổ CSL4500.

Bảng 3. Kết quả khảo sát độ ổn đinh của phổ LTA

	Sai khác trên mỗi người (dB/Hz)	Sai khác trung bình giữa 2 người với nhau (dB/Hz)
Giá tri trung bình	6,46	23,26
Độ lệch chuẩn	4,12	10,89

Kết quả khảo sát cho thấy, phổ LTA khá ổn định đối với mỗi người, sự thay đổi của phổ này đối với mỗi người nhỏ hơn sự sai khác giữa 2 người nói với nhau. So sánh định lượng giữa hai phổ LTA, nếu độ sai khác giữa hai phổ này nhỏ hơn ngưỡng được chọn bằng ((6,46 + 4,12) + (23,26-10,89))/2 = 11,475 thì kết luận hai phổ LTA đó thuộc về cùng một người nói, ngược lại chúng có thể thuộc hai người khác nhau.

Kết quả khảo sát các formants, tần số cơ bản, phổ trung bình thời gian dài đối với người nói tiếng Việt cho thấy phạm vi thay đổi của các tham số tiếng nói này đối với mỗi người nói nhỏ hơn so với phạm vi thay đổi giữa những người nói khác nhau. Điều này cho phép sử dụng các tham số tiếng nói trên để bổ xung định lượng cho việc so sánh nhận dạng người nói định tính bằng phương pháp thủ công.

2.6 Quy trình giám định nhận dạng người nói tiếng Việt

Một quy trình giám định pháp lý nhận dạng người nói tổng quát có thể chia thành hai pha. Pha thứ nhất: lọc từ tập dữ liệu những người nói nghi vấn ra một hoặc một vài người nói giống với tiếng nói mẫu cần giám định nhất. Pha thứ hai: so sánh nhận dạng người nói bằng phương pháp kinh điển giữa tiếng nói cần giám định với các mẫu tiếng nói của những người bị nghi vấn đã được pha thứ nhất lọc ra.

Pha thứ nhất, các cơ sở dữ liệu người nói nghi vấn có thể được chia làm 2 loại dựa trên thông tin về tiếng nói. Loại thứ nhất là những người trong cơ sở dữ liệu nói một số câu, từ chọn trước (từ khóa), loại thứ hai là người nói tự do trong khoảng thời gian đủ lớn.

Pha thứ hai, quy trình giám định nhận dạng người nói tiếng Việt theo phương pháp kinh điển giữa hai mẫu tiếng nói cần giám định và nghi vấn, thực hiện theo các bước sau.

Bước 1: So sánh nhận dạng người nói theo phương pháp cảm thụ bằng cơ quan thính giác của con người. Nếu ít nhất một mẫu tiếng nói được đánh giá là phát âm không bình thường, có biểu hiện giả giọng thì dừng và không đưa ra kết luận giám định. Ngược lại, tập trung so sánh các thông tin mức cao giữa hai mẫu tiếng nói như *Phương ngữ; Cao độ giọng nói; Các đặc trưng từ vựng; Đặc trưng ngữ điệu; Đặc điểm ngữ âm; Tật phát âm*. Nếu nhận thấy có nhiều điểm giống nhau giữa các mẫu thì chuyển sang bước 2, ngược lại thì kết luận phủ định (không đồng nhất) và dừng.

Bước 2: So sánh tần số cơ bản trung bình (F0) trong khoảng thời gian tối thiểu 6 giây của *hai* mẫu tiếng nói. Nếu độ sai khác tần số cơ bản trung bình nhỏ hơn 12 Hz (với giọng nam) hay 16 Hz (với giọng nữ) thì chuyển sang bước 3, ngược lại thì kết luận phủ định (không đồng nhất) và dừng.

Bước 3: Trường hợp cả hai mẫu tiếng nói được ghi âm trong cùng điều kiện thì so sánh định lượng phổ LTA trong khoảng thời gian ít nhất là 20 giây giữa hai mẫu tiếng nói. Nếu khoảng cách O'clid giữa hai phổ LTA nhỏ hơn 11,475 thì kết luận khẳng định (hai mẫu tiếng nói cùng do một người nói), ngược lại kết luận phủ định (không đồng nhất) và dừng. Trường hợp hai mẫu tiếng nói được ghi âm trong các điều kiện khác nhau hoặc không xác định được điều kiện ghi âm thì chuyển sang bước 4

Bước 4: Tìm các âm tiết (từ đơn) hay cụm từ đồng âm giữa hai mẫu tiếng nói để so sánh bằng phương pháp thủ công. Đánh giá độ giống nhau của các âm tiết đồng âm khi so sánh các vệt formant trên phổ ba chiều của các âm tiết này dựa trên diễn tiến của các formant, bề rộng và tỷ lệ tương đối giữa chúng. So sánh định lượng formant của các âm tiết này với nhau, nếu sai khác giữa các formant 1, 2, 3, 4 lần lượt nhỏ hơn 15,4%; 10%; 6,3%; 5,2% và bề rộng formant nhỏ hơn khoảng 23% thì có thể kết luận hai âm tiết đồng âm này là đồng nhất. Nếu số lượng âm tiết đồng nhất vượt quá một ngưỡng nhất định thì có thể kết luận khẳng định (hai mẫu tiếng nói này do cùng một người nói), ngược lại kết luận khả năng hoặc phủ định nếu số âm tiết đồng nhất quá ít...

Vấn đề đặt ra ở đây là, với số lượng âm tiết đồng nhất bằng bao nhiêu đối với giám định nhận dạng người nói tiếng Việt thì có thể kết luận hai mẫu tiếng nói là đồng nhất.

Chương 3: Xác suất nhận dạng người nói của âm tiết tiếng Việt

3.1 Cơ sở đánh giá khả năng phân biệt người nói đối với âm tiết

Việc khảo sát phạm vi biến đổi của các formant trong âm tiết đối với mỗi người nói và giữa những người nói khác nhau ở chương 2 dựa trên sự đánh giá phạm vi biến đổi của tỷ số giữa độ lệch chuẩn và trị trung bình của từng formant khi phát âm cùng một âm tiết đối với mỗi người và giữa nhiều người nói. Vì việc đánh giá dựa trên sự thay đổi của một biến (tỷ số giữa độ lệch chuẩn và trị trung bình), tức xác suất xuất hiện giá trị của biến đó, nên để xác định khả năng phân biệt người nói của mỗi âm tiết cần xác định luật xác suất xuất hiện của tập hợp các giá trị của biến này.

Quan sát sự phân bố các giá trị biến đổi tương đối của các formant xung quanh trị trung bình với từng âm tiết khảo sát cho phép đưa ra giả thiết: luật xác suất xuất hiện của tập các giá trị này đối với từng formant tuân theo luật phân bố chuẩn (phân bố Guass) với hàm phân bố xác suất có trị trung bình và phương sai (bình phương độ lệch chuẩn) được xác định như trong bảng 21. Nếu giả thiết về mặt lý thuyết này đúng thì sự sai khác giữa 2 hàm phân bố chuẩn, biểu diễn xác suất xuất hiện giá trị biến đổi của từng formant trong âm tiết đối với mỗi người và giữa nhiều người nói, sẽ là cơ sở để đánh giá khả năng phân biệt người nói của từng âm tiết được khảo sát.

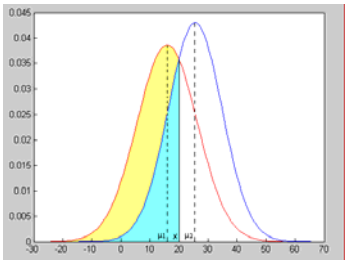
3.2 Kiểm định giả thiết thống kê đối với phạm vi biến đổi tương đối của các formant trong âm tiết

Để kiểm định giả thiết *phạm vi biến đổi tương đối của các formant trong âm tiết tuân theo luật phân bố chuẩn*, luận án đã sử dụng tiêu chuẩn χ^2 để đánh giá sự phù hợp giữa số liệu thực nghiệm phạm vi biến đổi tương đối của các formant với giả thiết lý thuyết này

Tiêu chuẩn phù hợp χ^2 được tính cho từng formant của từng âm tiết khảo sát. Đánh giá tiêu chuẩn phù hợp χ^2 với độ tin cậy $\alpha = 0,99$ thì có tới 250/272 = 92% tập hợp các giá trị thỏa mãn tiêu chuẩn χ^2. Nếu sử dụng độ tin cậy $\alpha = 0,95$ thì có tới 269/272 = 99% tập hợp các giá trị thỏa mãn tiêu chuẩn χ^2.

Kết quả đánh giá theo tiêu chuẩn phù hợp χ^2 có thể khẳng định giả thiết phạm vi biến đổi tương đối của các formant trong âm tiết tuân theo luật phân bố chuẩn là đúng.

Trên hình 1 biểu diễn quan hệ giữa 2 hàm phân bố chuẩn. Trên hình này, hàm phân bố xác suất phạm vi biến đổi của từng formant trong âm tiết đối với mỗi người nói được minh họa bằng đường cong màu đỏ, còn hàm phân bố xác suất phạm vi biến đổi của từng formant trong âm tiết đối với nhiều người nói khác nhau được minh họa bằng đường cong màu xanh (luôn nằm phía bên phải đường đỏ).



Hình 1: Minh họa quan hệ 2 hàm phân bố chuẩn.

3.3 Phân tích lý thuyết về khả năng phân biệt người nói của các âm tiết tiếng Việt

Một điều dễ chấp nhận là khả năng phân biệt người nói của từng formant trong âm tiết sẽ phụ thuộc vào quan hệ giữa hai hàm phân bố xác suất trên. Nếu hàm phân bố xác suất phạm vi biến đổi của formant đối với mỗi người càng cách xa hàm phân bố xác suất phạm vi biến đổi của formant đối với nhiều người, tức giá trị trung bình μ_1 của đường màu đỏ trên hình 1 càng khác xa so với μ_2 của đường màu xanh thì khả năng phân biệt người nói của formant đó càng lớn, vì điều đó chứng tỏ càng có sự khác biệt giữa một người nói với những người nói khác.

Từ đó, có thể nhận định: Khả năng phân biệt người nói của một formant trong âm tiết có thể được xác định thông qua vùng diện tích nằm dưới hàm phân bố xác suất phạm vi biến đổi tương đối của formant này trong âm tiết đối với cùng một người nói và nằm trên hàm phân bố xác suất phạm vi biến đổi tương đối của formant này giữa những người nói khác nhau. Trên hình 1, diện tích vùng này (vùng màu vàng) có thể được tính bằng hiệu của 2 hàm phân phối tích lũy:

$$S = F(x; \mu_1, \sigma_1) - F(x; \mu_2, \sigma_2)$$

Với:

$$F(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du$$

x : điểm giao nhau giữa 2 hàm phân bố xác suất

Vì diện tích nằm dưới đường cong phân bố xác suất biểu thị xác suất sự kiện nên có thể đưa ra một định nghĩa định lượng về khả năng phân biệt người nói của các âm tiết như sau : *Khả năng phân biệt người nói của âm tiết có thể định lượng bằng xác suất nhận dạng người nói của âm tiết đó, xác suất này được xác định bằng hiệu của các hàm phân phối tích lũy của phân bố xác suất phạm vi biến đổi tương đối của các formant trong âm tiết đối với mỗi người và nhiều người khác nhau.*. Phân bố xác suất suất phạm vi biến đổi tương đối của các formant trong âm tiết ở đây được xác định là phân bố chuẩn.

Áp dụng công thức trên cho các hàm phân bố xác suất của từng formant trong các âm tiết được khảo sát để xác định xác suất nhận dạng người nói của từng âm tiết này.

3.4 Một số nhận xét từ phân tích xác suất nhận dạng người nói

3.4.1 Số lượng âm tiết đồng nhất

Kết quả tính toán trên cho thấy xác suất nhận dạng người nói trung bình của một âm tiết tiếng Việt là 0,3795. Điều đó có thể hiểu là, nếu 2 mẫu tiếng nói có 2 âm tiết giống nhau (cả trên phương diện âm thanh nghe được và phổ của chúng) thì xác suất trung bình 2 mẫu tiếng nói đó do cùng một người nói (đồng nhất) là 37,95%. Hai âm tiết giống nhau về phương diện âm thanh nghe được chỉ có thể là các âm tiết đồng âm. Hai âm tiết có phổ âm thanh giống nhau khi 2 âm tiết đó có cấu trúc formant thể hiện trên phổ 3 chiều giống nhau và sự sai khác giữa các giá trị các formant tương ứng trong âm tiết đó thỏa mãn phạm vi biến đổi trung bình trong mỗi người như trên bảng 1. Hai âm tiết giống nhau như vậy được cho là đồng nhất.

Nếu gọi xác suất đồng nhất hai mẫu tiếng nói có 1 âm tiết đồng nhất là P(1) thì xác suất đồng nhất hai mẫu tiếng nói có n âm tiết đồng nhất P(n) có thể được tính theo công thức đệ quy với giả thiết n âm tiết đó khác nhau và độc lập với nhau :

$$P(1) = 0,3795$$

$$P(n) = P(n-1) + 0,3795*(1 - P(n-1))$$

Kết quả tính được : P(10) = 0,9915; … P(20) = 0,9999

Như vậy 2 mẫu tiếng nói tiếng Việt sẽ được coi là do cùng một người nói ra với xác suất trên 99% khi 2 mẫu tiếng nói đó có ít nhất là 10 âm tiết đồng nhất và với xác suất trên 99,99% khi 2 mẫu tiếng nói đó có ít nhất là 20 âm tiết đồng nhất.

3.4.2 Xác suất trung bình nhận dạng người nói của formant

Biểu diễn trị trung bình xác suất nhận dạng người nói của từng formant trong tất cả các âm tiết được khảo sát dưới dạng biểu đồ cho thấy: các formant bậc cao nhận dạng người nói tốt hơn các formant bậc thấp, đặc biệt là formant 3 có xác suất nhận dạng người nói cao hơn hẳn so với các formant khác, chứng tỏ thông tin về người nói được tập trung nhiều nhất ở formant 3.

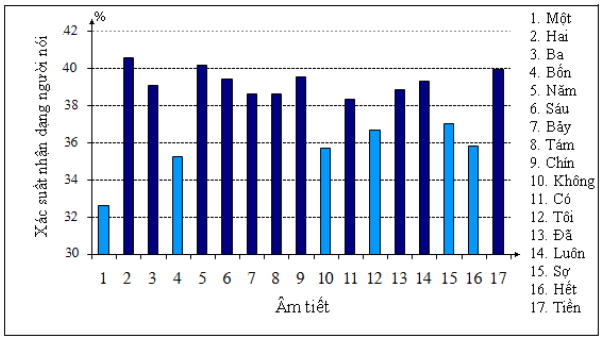
3.4.3 Khả năng phân biệt người nói của các âm tiết tiếng Việt

Hình 2 biểu diễn xác suất nhận dạng người nói của tất cả các âm tiết được khảo sát dưới dạng biểu đồ. Có thể rút ra một nhận xét là, các âm tiết khác nhau có khả năng phân biệt người nói khác nhau, một số nhận dạng người nói tốt, một số kém hơn. Nếu dựa trên xác suất nhận dạng trung bình của một âm tiết (0,3795) có thể chia các âm tiết được khảo sát ra làm hai nhóm:

Nhóm 1: các âm tiết có khả năng phân biệt người nói tốt gồm các âm tiết “Hai”, “Ba”, “Năm”, “Sáu”, “Bẩy”, “Tám”, “Chín”, “Có”, “Đã”, ”Luôn”, ”Tiền” . Trong đó các âm tiết “Hai”, “Năm”, “Sáu”, “Chín”, “Luôn”,”Tiền” phân biệt người nói tốt hơn các âm tiết còn lại.

Nhóm 2: các âm tiết có khả năng phân biệt người nói kém gồm các âm tiết “Một”, “Bốn”, “Không”, “Tôi”, “Sợ”, “Hết”. Trong đó kém nhất là âm tiết “Một”.

So sánh đặc trưng ngữ âm của các âm tiết trong từng nhóm và giữa hai nhóm với nhau có thể đưa ra nhận xét: Các âm tiết thuộc nhóm 1 hầu hết là các âm tiết có âm chính là nguyên âm hàng trước hoặc nguyên âm đôi (trừ âm tiết “Có”), còn nhóm 2 chủ yếu là các nguyên âm hàng sau và âm tiết khép.



Hình 2: Xác suất nhận dạng người nói của các âm tiết được khảo sát

Từ đây, có thể xác định khả năng phân biệt người nói của các âm tiết tiếng Việt như sau: Các âm tiết có âm chính là các nguyên âm hàng trước hoặc các nguyên âm đôi, âm tiết nửa mở, âm đầu hoặc/và cuối là các âm mũi có khả năng phân biệt người nói tốt nhất, các âm tiết khác khả năng phân biệt người nói kém hơn, kém nhất là các âm tiết khép.

3.5 Kiểm nghiệm khả năng phân biệt người nói của âm tiết tiếng Việt

Xuất phát từ quan điểm cho rằng, có thể đánh giá khả năng phân biệt người nói của một âm tiết thông qua việc đánh giá độ chính xác nhận dạng của một hệ nhận dạng người nói phụ thuộc từ khóa là chính âm tiết đó. Việc tiến hành đánh giá được thực hiện trên cơ sở dữ liệu người nói với 17 âm tiết đã được lựa chọn trong mục 2.3.2.

3.5.1 Hệ nhận dạng người nói phụ thuộc từ khóa cơ sở

Để khảo sát khả năng phân biệt người nói của các âm tiết tiếng Việt, luận án đã tiến hành xây dựng một hệ nhận dạng người nói phụ thuộc từ khóa cơ sở được phân lớp nhận dạng bằng mô hình HMM, vector đặc trưng trích chọn là các hệ số MFCC và được thực hiện cài đặt bằng ngôn ngữ máy tính MATLAB.

Để huấn luyện hệ nhận dạng người nói này, luận án đã xử dụng các phần mềm mã nguồn mở trong bộ công cụ H2M của Olivier Cappo, bộ công cụ này có thể download miễn phí từ địa chỉ http://www.tsi.enst.fr/~cappe/h2m/h2m.html. H2M là một tập hợp các hàm viết trên MATLAB thực hiện thuật toán EM để xây dựng các mô hình GMM hoặc HMM. Các hệ số MFCC được xác định bằng hàm mfcc lấy từ bộ công cụ xử lý âm thanh của Malcolm Slaney, bộ công cụ này có thể download từ địa chỉ : http://www.slaney.org/malcolm/pubs.html.

Việc đánh giá khả năng phân biệt người nói của từng âm tiết được thực hiện thông qua việc đánh giá độ chính xác nhận dạng người nói của từng hệ nhận dạng này cho từng âm tiết. Với từng âm tiết, sử dụng thuật toán Viterbi để xác định likelihood tương ứng của mỗi người trong cơ sở dữ liệu. Người có likelihood lớn nhất sẽ được nhận dạng. Thuật toán Viterbi là một hàm có trong bộ công cụ H2M.

3.5.2 Khảo sát độ chính xác nhận dạng của hệ nhận dạng người nói cơ sở với các âm tiết khác nhau

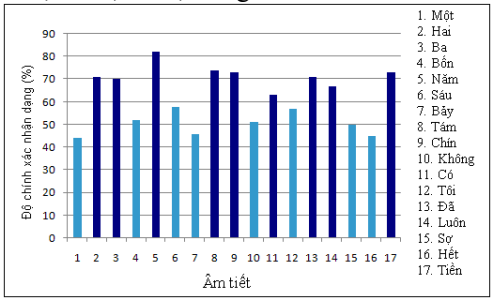
Kết quả khảo sát cho thấy, độ chính xác nhận dạng của hệ nhận dạng người nói cơ sở không chỉ phụ thuộc từ khóa là các âm tiết khác nhau mà còn phụ thuộc vào số trạng thái HMM và số hệ số MFCC. Nhìn chung, hệ nhận dạng sử dụng mô hình HMM có nhiều trạng thái và có số hệ số MFCC nhiều hơn thì nhận dạng người nói tốt hơn.

Hình 3 biểu diễn độ chính xác nhận dạng người nói của hệ nhận dạng người nói phụ thuộc từ khóa cơ sở phân lớp nhận dạng bằng mô hình HMM 7 trạng thái với 19 hệ số MFCC làm đặc trưng trích chọn đối với từng âm tiết tiếng Việt được khảo sát.

Nếu lấy độ chính xác nhận dạng người nói trung bình (61,6%) làm cơ sở, có thể chia các âm tiết được khảo sát ra làm hai nhóm:

Nhóm 1: các âm tiết có khả năng phân biệt người nói tốt gồm các âm tiết: “Hai”, “Ba”, “Năm”, “Tám”, “Chín”, “Có”,“Đã”, ”Luôn”, ”Tiền” .

Nhóm 2: các âm tiết có khả năng phân biệt người nói kém hơn gồm các âm tiết “Một”, “Bốn”, “Sáu”, “Bảy”, “Không”, “Tôi”, “Sợ”, Hết”, trong đó kém nhất là các âm tiết “Một”, “Hết” .



Hình 3: Độ chính xác nhận dạng của hệ nhận dạng người nói phụ thuộc từ khóa là các âm tiết được khảo sát

So sánh với xác suất nhận dạng người nói của từng âm tiết tính được theo lý thuyết xác suất thống kê trong mục 3.4.3 (hình 2), về cơ bản hai nhóm nhận dạng người nói tốt và kém được phân

chia khá giống nhau, điểm khác biệt chỉ là hai âm tiết nửa mở "Sáu", "Bẩy" về lý thuyết thì thuộc nhóm nhận dạng người nói tốt xong thực tế khi làm từ khóa trong các hệ nhận dạng người nói tự động lại thuộc về nhóm nhận dạng người nói kém.

Từ đây có thể đưa ra một quy tắc xác định khả năng phân biệt người nói của các âm tiết tiếng Việt áp dụng cho mọi trường hợp đó là:

Các âm tiết có âm chính là các nguyên âm hàng trước hoặc các nguyên âm đôi, âm đầu *hoặc/và cuối là các âm mũi có khả năng phân biệt người nói tốt nhất, các âm tiết khác khả năng phân biệt người nói kém hơn, kém nhất là các âm tiết khép.*

3.6 Ý nghĩa thực tiễn việc xác định khả năng phân biệt người nói của các âm tiết tiếng Việt

Xác định khả năng phân biệt người nói của các âm tiết tiếng Việt cho phép hoàn thiện quy trình giám định pháp lý nhận dạng người nói tiếng Việt xây dựng trong chương 2. Ngoài ra, để nâng cao độ tin cậy của kết luận giám định, các giám định viên cần thực hiện theo quy tắc tìm và so sánh các âm tiết đồng âm có khả năng phân biệt người nói tốt từ các mẫu tiếng nói. Việc đối sánh các âm tiết giữa các mẫu tiếng nói thực hiện chủ yếu là so sánh cấu trúc formant đặc biệt là các formant 3 trong trường hợp không thể xác định được đầy đủ các formant của âm tiết. Ngoài ra, quy tắc xác định khả năng phân biệt người nói của âm tiết cũng rất có ý nghĩa khi lựa chọn các từ hay âm tiết thích hợp để xây dựng cơ sở dữ liệu người nói tiếng Việt.

Quy trình giám định pháp lý nhận dạng người nói này đã được áp dụng trong thức tế, số vụ giám định nhận dạng người nói sử dụng quy trình này là 186 với tổng cộng 198 mẫu tiếng nói cần giám định trong đó nói giọng Bắc bộ là 61 mẫu, Trung bộ 52 mẫu và Nam bộ 85 mẫu. Kết quả giám định cho kết luận đồng nhất (khẳng định) là 168 mẫu, kết luận không đồng nhất (phủ định) là 30 mẫu. Tất cả các trường hợp này đều có kết luận giám định nhận dạng người nói đúng, chưa ghi nhận trường hợp nào phản hồi lại là kết luận khẳng định sai.

Tuy nhiên, quy trình này mới chỉ thực hiện tốt ở pha thứ 2, đối sánh hai mẫu tiếng nói bằng phương pháp kinh điển, còn trong pha thứ nhất, tự động lọc ra từ cơ sở dữ liệu người nói nghi vấn một hoặc một vài mẫu tiếng nói để đối sánh với mẫu tiếng nói cần giám định vẫn còn nhiều vấn đề cần giải quyết khi áp dụng các hệ tự động nhận dạng người nói trong giám định pháp lý.

Chương 4: Giám định tự động nhận dạng người nói tiếng Việt

4.1 Các vấn đề tồn tại của giám định tự động nhận dạng người nói tiếng Việt

Giám định tự động nhận dạng người nói là phương pháp giám định hoàn toàn dựa vào sự phân tích và so sánh các mẫu tiếng nói bằng máy tính trên nguyên lý làm việc của phương pháp nhận dạng người nói tự động. Ưu điểm chính của phương pháp này là thời gian thực hiện nhanh, do vậy thường được áp dụng khi giám định nhận dạng người nói trên tập dữ liệu nhiều người nói nghi vấn.

Tùy từng vụ việc cụ thể mà tập dữ liệu người nói nghi vấn được xây dựng như một hệ nhận dạng người nói phụ thuộc từ khóa hay không phụ thuộc từ khóa. Với các hệ nhận dạng người nói phụ thuộc từ khóa thì ngoài việc lựa chọn từ khóa nào cũng cần lựa chọn mô hình đơn vị ngữ âm thích hợp để từ đó xây dựng nên tập từ điển từ khóa. Với tiếng Việt, là ngôn ngữ đơn âm tiết, nên tập từ điển từ khóa chủ yếu là một số các âm tiết đã được chọn lọc từ trước, vấn đề sẽ chỉ còn là lựa chọn đơn vị ngữ âm là âm vị tạo nên các âm tiết đã được chọn hay sử dụng ngay âm tiết làm đơn vị ngữ âm khi xây dựng mô hình người nói trong các hệ nhận dạng người nói phụ thuộc từ khóa.

Bên cạnh đó, một trong những nhược điểm của giám định tự động nhận dạng người nói là độ tin cậy của kết luận giám định chưa cao. Nguyên nhân do hiện tại nhận dạng tự động vẫn chủ yếu dựa trên các thông tin mức thấp của tiếng nói, mà các thông tin này rất nhạy cảm với nhiễu cũng như khi thay đổi điều kiện ghi âm.

4.2 Mô hình âm tiết và mô hình âm vị trong nhận dạng người nói tiếng Việt

Lựa chọn đơn vị ngữ âm nào thích hợp và hiệu quả đối với hoạt động nhận dạng người nói tiếng Việt phụ thuộc từ khóa. Tiếng Việt, như đã phân tích, là ngôn ngữ đơn âm tiết nên tập từ điển từ để xây dựng từ khóa thích hợp nhất là các âm tiết, vấn đề là là lựa chọn đơn vị ngữ âm là âm vị tạo nên các âm tiết đã được chọn làm tập từ điển hay sử dụng ngay âm tiết làm đơn vị ngữ âm khi xây dựng mô hình người nói trong các hệ nhận dạng người nói phụ thuộc từ khóa.

Để đánh giá so sánh các hệ nhận dạng người nói phụ thuộc từ khóa dựa trên các mô hình đơn vị ngữ âm là âm tiết và âm vị, luận án đã chọn các âm tiết số tiếng Việt làm tập từ điển từ để tiến hành khảo sát. Câu nói được dùng làm từ khóa sẽ là chuỗi số ngẫu nhiên. Độ dài từ khóa được xác định bằng số chữ số có trong chuỗi số đó.

Sử dụng các âm tiết số "Không", "Một", "Hai", "Ba", "Bốn", "Năm", "Sáu", "Bẩy", "Tám", "Chín" có trong các tập dữ liệu người nói đã được xây dựng trong chương 3 mục 3.2 để đánh giá các hệ nhận dạng người nói này.

4.2.1 Hệ nhận dạng người nói dựa trên các mô hình âm tiết

Với đơn vị ngữ âm là các âm tiết, để xây dựng hệ nhận dạng người nói phụ thuộc từ khóa là chuỗi số tiếng Việt với một tập từ điển từ là 10 âm tiết số, mỗi người nói cần huấn luyện đủ 10 mô hình HMM cho 10 âm tiết được dùng làm tập từ điển này. Chọn số trạng thái của mô hình HMM bằng 5 để biểu diễn các âm tiết số tiếng Việt.

Trong pha nhận dạng, sử dụng một hàm ngẫu nhiên tạo chuỗi số dùng làm từ khóa để kiểm tra nhận dạng hệ nhận dạng người nói trên sau khi đã được huấn luyện. Likelihood của chuỗi số làm từ khóa được tính bằng tổng các likelihood của từng âm tiết số thành phần.

4.2.2 Hệ nhận dạng người nói dựa trên các mô hình âm vị

Để xây dựng hệ nhận dạng người nói với bộ từ điển từ là 10 âm tiết số tiếng Việt dựa trên đơn vị ngữ âm là các âm vị, luận án đã xây dựng 28 mô hình HMM tương ứng với 28 âm vị ba (gồm các âm vị tạo thành các âm tiết số tiếng Việt và một âm vị đặc biệt để mô hình hóa khoảng lặng trong quá trình phát âm) cho mỗi người nói. Mỗi âm vị được biểu diễn bằng một mô hình HMM 3 trạng thái

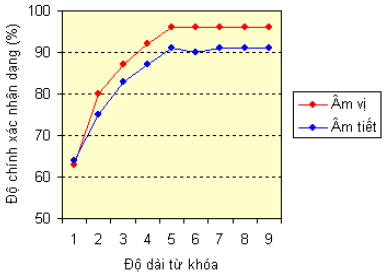
Trong pha huấn luyện, do ranh giới giữa các âm vị trong mỗi âm tiết rất khó xác định tự động, luận án đã sử dụng phương pháp gãn nhãn cưỡng bức để gãn nhãn cho từng âm vị ba trên.

Trong pha nhận dạng, câu nói được dùng làm từ khóa là chuỗi số được tạo từ một hàm ngẫu nhiên và sử dụng các tập dữ liệu người nói để kiểm tra nhận dạng người nói trên tập đóng và trên tập mở tương tự như với mô hình âm tiết. Chỉ có điều thay vì likelihood của các âm tiết được tính toán trực tiếp từ mô hình các âm tiết của mỗi người, likelihood lại được xác định theo các mô hình âm vị có trong từng âm tiết thành phần của chuỗi số. Likelihood của chuỗi số làm từ khóa được tính bằng tổng các likelihood của tất cả các âm vị của các âm tiết số thành phần.

Để cài đặt các hệ thống nhận dạng người nói này, luận án đã sử dụng các phần mềm mã nguồn mở viết bằng ngôn ngữ máy tính MATLAB có trong bộ công cụ H2M.

4.2.3 So sánh các hệ nhận dạng người nói dựa trên mô hình âm tiết và âm vị

Khảo sát trên tập đóng: Kết quả khảo sát *độ* chính xác nhận dạng cho trên hình 4.

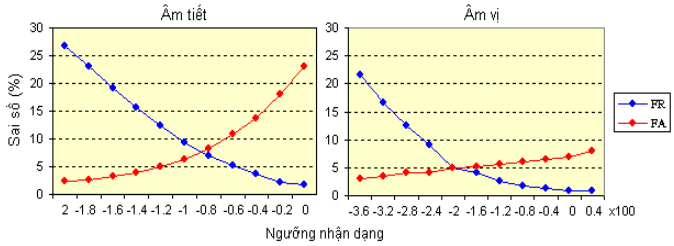


Hình 4: Kết quả khảo sát độ chính xác nhận dạng theo độ dài từ khóa của các hệ nhận dạng người nói dựa trên mô hình đơn vị âm tiết và âm vị

Có thể thấy, hệ nhận dạng người nói dựa trên các mô hình đơn vị âm vị có độ chính xác cao hơn hệ dựa trên các mô hình đơn vị âm tiết. Một nhận xét nữa là cả hai hệ nhận dạng người nói này đều có độ chính xác nhận dạng tăng theo độ dài từ khóa, tuy nhiên khi độ dài từ khóa bằng 5 trở lên độ chính xác nhận dạng của cả hai hệ thống đều không tăng và gần như không đổi, với hệ thống sử dụng mô hình đơn vị âm tiết độ chính xác nhận dạng khi đó đạt khoảng 91% , còn với mô hình đơn vị âm vị độ chính xác cao hơn, đạt mức 96%.

Khảo sát trên tập mở : Sai số cân bằng EER của các hệ nhận dạng người nói dựa trên các mô hình đơn vị ngữ âm là âm tiết và âm vị tiếng Việt được xây dựng ở trên được xác định trên tập mở với từ khóa có độ dài bằng 5.Kết quả khảo sát trên hình 5 cho thấy, sai số cân bằng EER của hệ

nhận dạng người nói dựa trên các mô hình đơn vị âm tiết là 7,6%, còn đối với mô hình đơn vị âm vị EER thấp hơn, khoảng 5%.



Hình 5: Sai số từ chối (FR) và sai số chấp nhận (FA) của các hệ nhận dạng người nói dựa trên các mô hình đơn vị âm tiết và âm vị

Nhận xét chung, với cùng một bộ từ điển, mô hình đơn vị âm tiết cần ít số mô hình HMM để mô hình hóa người nói hơn mô hình đơn vị âm vị xong khả năng nhận dạng người nói của hệ nhận dạng người nói dựa trên mô hình đơn vị âm vị tốt hơn dựa trên mô hình đơn vị âm tiết, tuy nhiên xây dựng hệ nhận dạng dựa trên mô hình đơn vị âm vị lại phức tạp hơn nhất là khi tăng số lượng từ trong từ điển.

Khi chuỗi số từ khóa được nói vào để nhận dạng người nói là thực (tức là không phải lấy từ cơ sở dữ liệu được xây dựng) thì, mô hình âm tiết cần bổ xung thêm một thuật toán tự động nhận và cắt các âm tiết từ chuỗi số từ khóa đưa vào trước khi trích chọn đặc trưng. Trong khi đó, do có sử dụng thêm âm vị đặc biệt /sil/ để mô hình hóa khoảng lặng nên mô hình âm vị không cần bổ xung thêm thuật toán cắt rời kiểu như vậy.

4.3 Chuẩn hóa điều kiện ghi âm trong giám định tự động nhận dạng người nói

4.3.1 Giám định tự động nhận dạng người nói trong các điều kiện ghi âm khác nhau

Một trong những nhược điểm của giám định tự động nhận dạng người nói so với giám định bằng phương pháp kinh điển là độ tin cậy của kết luận giám định không cao. Nguyên nhân do hiện tại nhận dạng tự động vẫn chủ yếu dựa trên các thông tin mức thấp của tiếng nói, mà các thông tin này rất nhạy cảm với nhiễu cũng như khi thay đổi điều kiện ghi âm. Các phương pháp lọc nhiễu hiện nay có thể khắc phục tương đối ảnh hưởng của nhiễu. Tuy nhiên, với điều kiện ghi âm thay đổi thì khác, dễ hình dung trong khi tiếng nói cần giám định thường được ghi bí mật trong bất cứ môi trường nào thì mẫu tiếng nói của đối tượng nghi vấn thường được ghi âm một cách công khai trong môi trường văn phòng. Đây là một trong những nguyên nhân chính đưa đến kết luận sai trong giám định tự động nhận dạng người nói hiện nay.

Để khắc phục vấn đề này, các phương pháp chuẩn hóa hay bù suy giảm do điều kiện đối sánh khác nhau trên kênh thông tin đã được nghiên cứu cho nhận dạng người nói.

4.3.2 Cơ sở của phương pháp chuẩn hóa theo phổ trung bình

Phạm vi nghiên cứu ở đây chủ yếu tập trung vào tìm hiểu ảnh hưởng của kênh thông tin lên quá trình nhận dạng người nói. Một cách lý tưởng là giả sử hoàn toàn loại bỏ được nhiễu cộng bằng các bộ lọc nhiễu trước khi đưa vào bộ tiền xử lý. Khi đó nếu biết được trước đặc tuyến tần số của kênh thông tin, về lý thuyết hoàn toàn có thể xác định lại tín hiệu tiếng nói sạch từ tín hiệu tiếng nói đã bị suy giảm bởi kênh truyền.

Các khảo sát thực nghiệm trên các thiết bị phân tích phổ tiếng nói đều chỉ ra rằng khi lấy trung bình phổ của tín hiệu tiếng nói của một người trong thời gian đủ dài, phổ trung bình sẽ không còn phụ thuộc vào nội dung nói nữa, khi đó nó chỉ còn mang thông tin đặc trưng về người nói. Đứng trên góc độ cấu âm có thể lý giải phổ trung bình tiếng nói của một người tương ứng với vị trí hoạt động trung bình của tuyến âm trong suốt quá trình cấu âm và do vậy sẽ mang thông tin về người đó.

Mở rộng ra, nếu lấy trung bình phổ tiếng nói không phải là của một người mà là của nhiều người nói trên cùng kênh thông tin thì phổ trung bình tiếng nói sẽ không còn mang thông tin đặc trưng về một người cụ thể mà chỉ còn là thông tin về đặc trưng của kênh thông tin. Do vậy, đặc

tuyến tần số của một kênh thông tin có thể được xác định gần đúng bằng phổ trung bình của nhiều người nói trên kênh đó.

Phương pháp chuẩn hóa theo phổ trung bình (Mean Spectrum - MS) dựa trên chuẩn hóa phổ tín hiệu tiếng nói bằng cách chia cho đặc tuyến tần số này của kênh thông tin trước khi tính các hệ số ceptrum. Phương pháp chuẩn hóa MS có thể áp dụng cho cả nhận dạng người nói không phụ thuộc từ khóa và phụ thuộc từ khóa.

4.3.3 Xây dựng tập dữ liệu khảo sát người nói trong điều kiện ghi khác nhau

Cơ sở dữ liệu người nói để khảo sát đánh giá các phương pháp chuẩn hóa gồm 140 người được ghi âm trong môi trường văn phòng, chủ yếu nói giọng Bắc bộ, mỗi người nói một cách tự nhiên 3 lần, mỗi lần 20 giây trong những khoảng thời gian khác nhau. Lần thứ nhất và lần thứ 2 được ghi trong cùng một điều kiện (HT1), lần thứ 3 được ghi trong điều kiện khác (HT2). Cơ sở dữ liệu này được chia làm 2 tập dữ liệu Data100 và Data40. Tập Data100 được dùng làm dữ liệu khảo sát trong khi tập Data40 chủ yếu được sử dụng như những người người nói mạo danh.

4.3.4 Hệ nhận dạng người nói cơ sở để khảo sát

Hệ nhận dạng người nói không phụ thuộc từ khóa sử dụng mô hình GMM có số thành phần bằng 32 với đặc trưng là các hệ số MFCC.

Trong pha huấn luyện, sử dụng các phần mềm mã nguồn mở trong bộ công cụ H2M để thực hiện thuật toán EM xác định bộ các tham số của mô hình GMM cho mỗi người nói trong tập dữ liệu

Trong pha nhận dạng, lần phát âm thứ hai được dùng khi khảo sát trường hợp ghi trong cùng kênh thông tin và lần phát âm thứ ba được dùng khi khảo sát trường hợp ghi trong điều kiện khác kênh thông tin.

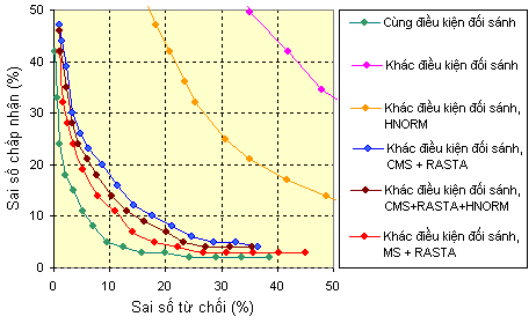
4.3.5 Đánh giá phương pháp chuẩn hóa theo phổ trung bình

Kết quả khảo sát cho thấy, khi cùng kết hợp với phương pháp RASTA, phương pháp MS cho kết quả tốt hơn phương pháp CMS (cải thiện được khoảng (13,15-11,45)/13,15 ≈ 12,9%).

Bảng 4: Kết quả khảo sát một số phương pháp chuẩn hóa

			Độ chính xác trên tập đóng	Sai số cân bằng trên tập mở
Không chuẩn hóa	*Cùng điều kiện đối sánh*		98 %	7,55 %
	Khác điều kiện đối sánh		3 %	41,97 %
Chuẩn hóa khi khác điều kiện đối sánh	*Các phương pháp thông dụng*	HNORM	25%	27,81 %
		RASTA	54 %	
		CMS	56 %	
		CMS, RASTA	61 %	13,15 %
		CMS, RASTA, HNORM	63 %	12,06 %
	Phương pháp đề xuất	MS	55%	
		MS, RASTA	63	11,45

Nhận xét : Nguyên tắc thực hiện phương pháp chuẩn hóa đề xuất MS giống với phương pháp CMS (cùng dựa trên nguyên tắc trừ trung bình phổ) nhưng được thực hiện trên miền phổ thay vì trên miền ceptre như CMS. Xong điểm khác biệt giữa hai phương pháp này chủ yếu là ở chỗ phương pháp MS có thể thực hiện trong thời gian thực còn phương pháp CMS thì không. Phương pháp CMS đơn giản hơn khi thực hiện xong bù lại phương pháp MS lại hiệu quả hơn trong những trường hợp thu được mẫu tiếng nói của nhiều người trên cùng kênh thông tin.



Hình 5: Các đường quan hệ sai số

4.4 Sơ đồ khối quy trình giám định nhận dạng người nói tổng quát

Nhìn một cách tổng quát, toàn bộ quy trình này hoạt động như một phương pháp giám định nhận dạng người nói tổng hợp, đó là sự kết hợp cả ba phương pháp nhận dạng người nói: tự động, cảm thụ thính giác và thủ công trong một hoạt động giám định nhận dạng người nói. Sơ đồ khối toàn bộ quy trình giám định này được thể hiện trên hình 6.

Hình 6: Sơ đồ khối quy trình giám định nhận dạng người nói tổng quát

Kết luận và kiến nghị

Với mục tiêu nghiên cứu nhận dạng người nói và ứng dụng trong giám định pháp lý nhận dạng người nói tiếng Việt, luận án đã đạt được một số kết quả chính như sau:

1. Đề xuất một quy trình giám định pháp lý nhận dạng người nói giữa hai mẫu tiếng nói tiếng Việt với âm tiết là đơn vị ngữ âm chính so sánh các mẫu tiếng nói. Quy trình giám định nhận dạng người nói này được xây dựng dựa trên phương pháp cảm thụ bằng cơ quan thính giác của con người kết hợp phân tích phổ âm thanh và ngưỡng nhận dạng được xác định giữa phạm vi thay đổi của một số tham số tiếng nói đối với mỗi người và giữa nhiều người nói.

2. Xác định được số lượng âm tiết đồng nhất tối thiểu giữa hai mẫu tiếng nói tùy thuộc độ chính xác nhận dạng yêu cầu để có thể kết luận hai mẫu tiếng nói tiếng Việt do cùng một người nói ra.

3. Đưa ra một quy tắc xác định khả năng phân biệt người nói của các âm tiết tiếng Việt như sau: *Các âm tiết có âm chính là các nguyên âm hàng trước hoặc các nguyên âm đôi, âm đầu hoặc/và cuối là các âm mũi có khả năng phân biệt người nói tốt nhất, các âm tiết khác khả năng phân biệt người nói kém hơn, kém nhất là các âm tiết khép.*

4. Các hệ giám định tự động nhận dạng người nói phụ thuộc từ khóa tiếng Việt dựa trên các mô hình đơn vị ngữ âm là âm vị cho kết quả nhận dạng chính xác hơn và phù hợp với cách nói tự nhiên hơn các mô hình âm tiết.

5. Đề xuất một phương pháp chuẩn hóa phổ tiếng nói trong nhận dạng người nói khi điều kiện đối sánh khác nhau dựa trên nguyên tắc trừ trung bình phổ của nhiều người trên kênh thông tin.

Các kết quả nghiên cứu trên đã được đưa vào ứng dụng trong thực tế công tác giám định nhận dạng người nói tiếng Việt từ năm 2007 và đã phát huy tác dụng tích cực trong điều tra và xét xử tội phạm tại Việt Nam. Để nâng cao hơn nữa khả năng giám định pháp lý nhận dạng người nói tiếng Việt cần tiếp tục nghiên cứu theo các hướng sau:

1. Mở rộng phạm vi nghiên cứu của luận án trong các trường hợp người nói cố tình cải trang, giả giọng nói hay trong các trạng thái không bình thường (như hồi hộp, lo lắng hay sợ hãi...) cho người nói tiếng Việt.

2. Tiếp tục nghiên cứu chuẩn hóa trong nhận dạng người nói theo hướng dựa trên các thông tin mức cao, những thông tin ít bị tác động của kênh truyền.

3. Nghiên cứu nhận dạng tự động phương ngữ thuộc các vùng miền khác nhau của đất nước.