Xây Dựng Mô-đun Điều Khiển Bằng Giọng Nói Trong Ứng Dụng Đọc Báo Điện Tử Cho Người Khiếm Thị

Lê Vũ Công Hòa[1], Hoàng Thị Minh Khanh[1], Lê Quang Tam[1], Ninh Khánh Duy[1]

[1] Trường Đại học Bách Khoa, Đại học Đà Nẵng
conghoacntt13t1@gmail.com, minhkhanhhoang2105@gmail.com,
lesan1995@gmail.com, nkduy@dut.udn.vn

Tóm tắt. Nhận dạng tiếng nói dùng mô hình Markov ẩn (HMM) đã được ứng dụng rộng rãi trong các hệ thống giao tiếp người-máy bằng giọng nói. Bài báo này mô tả các bước đầu tiên trong việc xây dựng một mô-đun điều khiển máy tính bằng giọng nói nhằm trợ giúp người khiếm thị điều khiển ứng dụng đọc báo điện tử. Để tạo ra hệ thống dễ sử dụng cho người khiếm thị, chúng tôi đã thiết kế tập lệnh điều khiển gồm 4 nhóm với 46 lệnh là các từ đơn. Để huấn luyện và kiểm thử hệ thống nhận dạng tiếng nói, chúng tôi đã thu âm dữ liệu tiếng nói của 42 người với các chất giọng khác nhau trong điều kiện môi trường thực tế và tiến hành các thử nghiệm nhận dạng. Thực nghiệm cho thấy việc thiết lập các tham số của HMM và kích thước dữ liệu huấn luyện ảnh hưởng không nhỏ đến kết quả nhận dạng. Ở chế độ offline, hệ thống nhận dạng tiếng nói rời rạc của chúng tôi đạt độ chính xác cao nhất lần lượt là 99,42% và 91,14% trong các thử nghiệm nhận dạng phụ thuộc người nói và độc lập người nói. Ở chế độ online, hệ thống đạt độ chính xác trên 80% khi nhận dạng độc lập người nói trong điều kiện phòng tương đối yên tĩnh và phần cứng máy tính có tài nguyên hạn chế.

Từ khóa: Điều khiển bằng giọng nói, Nhận dạng tiếng nói rời rạc, Mô hình Markov ẩn, Đọc báo cho người khiếm thị.

1 Giới thiệu

Nhận dạng tiếng nói ra đời đã góp phần thay đổi cách người dùng điều khiển máy tính cũng như các thiết bị điện tử khác. Không cần phải thao tác trên màn hình hay bàn phím như thông thường, hệ thống nhận dạng tiếng nói giúp chuyển đổi tín hiệu tiếng nói từ người dùng thành câu lệnh tương ứng. Dựa vào khả năng này, việc áp dụng nhận dạng tiếng nói cho người khiếm thị điều khiển máy tính là hoàn toàn phù hợp.

Hiện nay, khi thế giới đang ngày càng phẳng dần, mọi người ai cũng có nhu cầu tiếp cận nguồn thông tin vô tận trên Internet, kể cả người khiếm thị. Ý tưởng tạo ra ứng dụng đọc báo điện tử cũng được hình thành từ đó. Việc tương tác với ứng dụng bằng giọng nói là cần thiết vì người khiếm thị không có khả năng dùng màn hình. Do đó, cần tạo ra một hệ thống điều khiển bằng giọng nói mà có thể thay thế các thao tác trên giao diện.

Trong các hướng tiếp cận cho việc huấn luyện và nhận dạng tiếng nói, hướng tiếp cận học máy dùng mô hình Markov ẩn là vượt trội hơn cả. Được nghiên cứu và phát triển từ những năm 50 và 60, mô hình Markov ẩn đã trở nên phổ biến trong những năm gần đây vì sự dồi dào trong cấu trúc toán học và áp dụng tốt trong các ứng dụng thực tiễn [1][2]. Vì thế chúng tôi chọn hướng tiếp cận này để thực hiện công việc nhận dạng tiếng nói phục vụ cho mục tiêu của mình. Tuy nhiên, trong quá trình áp dụng, công đoạn chuẩn bị dữ liệu huấn luyện và cấu hình các tham số cho mô hình cần được nghiên cứu và thực hiện kỹ lưỡng. Dữ liệu cần đủ nhiều và tham số cần được lựa chọn cho thích hợp để đem lại kết quả khả quan nhất.

Đề tài đọc báo điện tử cũng như nhận dạng tiếng nói tiếng Việt là không hề mới. Gần đây đã có ứng dụng đọc báo điện tử tiếng Việt đáp ứng việc đọc nội dung trang báo thành tiếng tên là VNR4B [3], nhưng ứng dụng này còn hạn chế ở chỗ chưa có công cụ nhận lệnh bằng giọng nói. Điều này gây khó khăn cho người khiếm thị khi sử dụng. Vì thế, cần thiết phải kết hợp tính năng đọc văn bản thành tiếng với tính năng điều khiển bằng tiếng nói, đặc biệt là tiếng Việt, để tạo ra ứng dụng đọc báo phục vụ cho người khiếm thị Việt Nam, và cả những người Việt khác muốn dùng ứng dụng mà chỉ thông qua việc nghe và nói. Trong bài báo này, chúng tôi sẽ chú trọng đến việc nghiên cứu điều khiển bằng giọng nói cho ứng dụng đọc báo điện tử.

Từ những vấn đề trên, chúng tôi tiến hành tìm hiểu về mô hình Markov ẩn, cụ thể là trong ứng dụng nhận dạng tiếng nói rời rạc, từ đó áp dụng vào đề tài của nhóm. Đóng góp của chúng tôi trong đề tài này là: thiết kế tập lệnh hướng tới sự dễ sử dụng cho người khiếm thị; thu âm tập lệnh đã thiết kế để chuẩn bị dữ liệu cho việc huấn luyện và nhận dạng; ứng dụng hệ thống nhận dạng tiếng nói rời rạc dùng mô hình Markov ẩn để thực nghiệm trên dữ liệu đã thu âm và đánh giá kết quả.

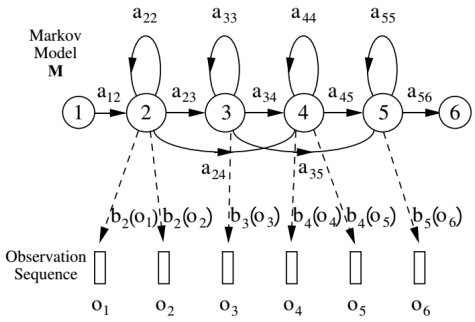
Bài báo được tổ chức thành các phần như sau. Phần 2 là phần giới thiệu ngắn gọn về mô hình Markov ẩn và ứng dụng trong nhận dạng tiếng nói rời rạc. Chúng tôi mô tả thiết kế tập lệnh trong phần 3. Phần 4 trình bày thực nghiệm và kết quả. Phần 5 đưa ra kết luận và hướng phát triển.

2 Mô hình Markov ẩn

2.1 Giới thiệu

Mô hình Markov ẩn (Hidden Markov Model – HMM) là phương pháp thống kê phổ biến dùng để mô hình hóa chuỗi vector đặc trưng của tiếng nói. Một mô hình Markov ẩn có thể biểu diễn cho một đơn vị âm thanh (như là từ hay âm vị). Trong nhận dạng tiếng nói, HMM giải quyết việc phân lớp tín hiệu tiếng nói một cách hiệu quả.

Mô hình Markov ẩn gồm chuỗi các trạng thái (state), được nối với nhau bởi các dây cung hay còn gọi là xác suất chuyển đổi trạng thái. Mỗi trạng thái có thể sinh ra các quan sát (observation) theo các xác suất nhất định (Hình 1). Ta gọi đây là mô hình Markov ẩn vì các trạng thái đã bị ẩn đi, chuỗi quan sát không cho biết cụ thể mỗi quan sát được sinh từ trạng thái nào. Các tham số của mô hình HMM được mô tả đầy đủ trong [1]. Trong phần thực nghiệm chúng tôi sẽ khảo sát chủ yếu 2 tham số sau:



Hình 1. Mô hình Markov ẩn sinh ra chuỗi quan sát [4].

- nState: số trạng thái của một mô hình HMM.
- nMix: số hỗn hợp (mixture) của phân bố Gauss, là phân bố xác suất sinh ra quan sát tại mỗi trạng thái.

2.2 Ba bài toán cơ bản

Từ mô hình được biểu diễn như trên, có ba bài toán được đặt ra để ứng dụng vào các hệ thống sử dụng mô hình Markov ẩn. Ba bài toán và cách giải quyết được trình bày cụ thể trong [1].

Bài toán đánh giá. Cho chuỗi quan sát $O = o_1 o_2 \dots o_T$ và mô hình HMM λ. Tính xác suất mô hình sinh ra chuỗi quan sát $P(O|\lambda)$.

Bài toán này dùng trong giai đoạn nhận dạng bằng cách chọn ra mô hình tiếng nói sinh ra chuỗi quan sát tốt nhất. Bài toán đã được nghiên cứu giải quyết bằng thuật toán tiến-lui (Forward-Backward Procedure).

Bài toán giải mã. Cho mô hình HMM λ và chuỗi quan sát $O = o_1 o_2 \dots o_T$. Tìm chuỗi trạng thái $Q = q_1 q_2 \dots q_T$ tối ưu nhất.

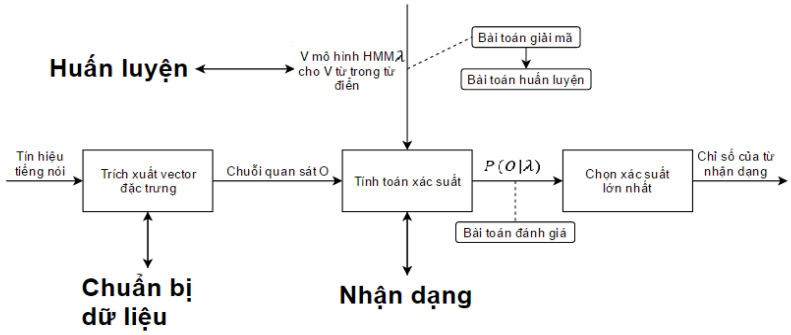
Đây còn được gọi là bài toán tìm ra phần ẩn, được dùng để tìm hiểu về cấu trúc của mô hình. Thuật toán Viterbi được áp dụng để giải bài toán.

Bài toán huấn luyện. Điều chỉnh các tham số của mô hình HMM λ để mô tả tốt nhất cách mà chuỗi quan sát được tạo ra bằng cách tối đa hóa xác suất $P(O|\lambda)$.

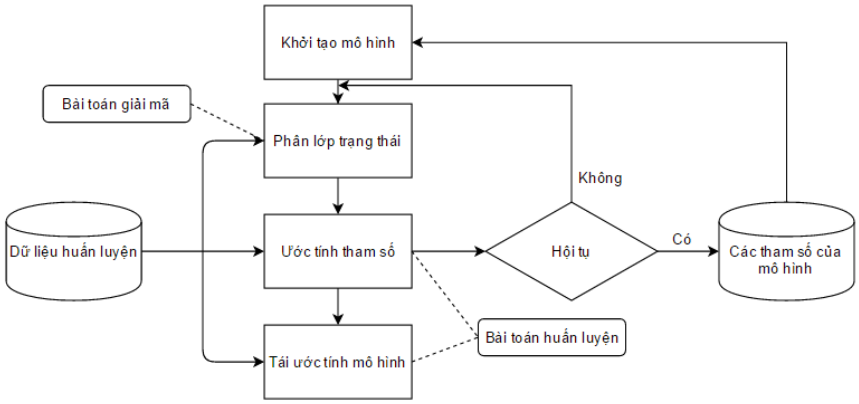
Áp dụng thuật toán Baum-Welch để giải quyết bài toán vào việc huấn luyện mô hình từ dữ liệu là các chuỗi quan sát.

2.3 Ứng dụng trong nhận dạng tiếng nói rời rạc

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.



Hình 2. Sơ đồ khối của một hệ nhận dạng tiếng nói rời rạc [1]



Hình 3. Giai đoạn huấn luyện mô hình [1]

1. Với mỗi từ trong từ điển, ta cần xây dựng một mô hình HMM λ bằng cách tính toán các tham số của mô hình sao cho nó biểu diễn tốt nhất K chuỗi quan sát trong bộ dữ liệu huấn luyện của nó. Bài toán giải mã được áp dụng để tìm ra chuỗi trạng thái tối ưu, sau đó thực hiện bài toán huấn luyện để điều chỉnh tham số (Hình 3).
2. Với mỗi từ cần nhận dạng, cần giải quyết bài toán đánh giá để chọn ra được mô hình mô tả đúng nhất từ đưa vào (Hình 2).

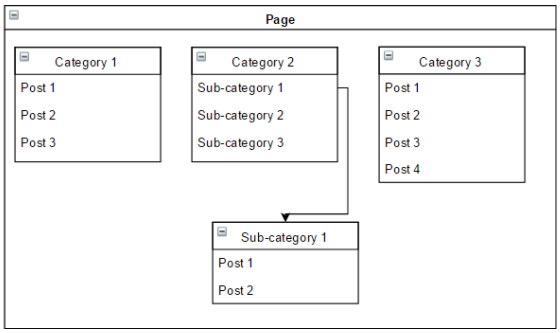
3 Thiết kế tập lệnh

Tập lệnh được thiết kế phục vụ điều khiển ứng dụng đọc báo điện tử, hướng tới sự dễ sử dụng cho người khiếm thị, nghĩa là giúp một người không có khả năng nhìn nhưng vẫn có thể truy cập được toàn bộ nội dung trang web. Thêm vào đó, hệ thống chúng tôi

sử dụng là hệ thống nhận dạng tiếng nói rời rạc, nên các lệnh chỉ gồm một từ. Do đó cần thiết kế sao cho chỉ với một từ nhưng vẫn mô tả câu lệnh thực hiện điều gì, và vẫn đảm bảo sự dễ đọc cho người dùng. Dựa vào mục đích điều khiển mà các lệnh được phân nhóm như sau.

3.1 Nhóm điều hướng (Navigation)

Đây là nhóm lệnh giúp hệ thống trỏ đến những trạng thái của một trang web nhằm điều hướng truy cập. Mỗi trang web thường được cấu thành từ danh sách các mục, mỗi mục sẽ chứa các mục con (nếu có) hoặc danh sách các bài, nơi chứa nội dung.



Hình 4. Hình minh họa một trang web

Có 3 cấp điều hướng: trang (page), mục (category) và bài (post).

Cấp điều hướng trang. Những lệnh điều hướng nào có thực hiện tải một đường dẫn trang web thì được nhóm vào cấp này. Một ví dụ dễ thấy là các lệnh thuộc về trình duyệt như "Về" để lui lại trang trước, hoặc "Chủ" ra lệnh trở về trang chủ. Ngoài ra còn có lệnh "Vào" giúp tải đường dẫn đến mục hoặc bài đang được chỉ đến, tương tự như việc click chuột vào một mục hay bài trên trang web.

Cấp điều hướng mục. Việc điều hướng mục không làm thay đổi trang hiện tại, chỉ thực hiện trỏ đến phần tử trong danh sách mục, giống như việc di chuyển chuột đến các mục trong trang. Chỉ khi nào đã trỏ đến mục mong muốn, thực hiện mở mục bằng các lệnh ở cấp điều hướng trang. Các lệnh như "Đầu", "Cuối", "Kế", "Trước" lần lượt dùng để thao tác chỉ đến mục đầu tiên, mục cuối cùng, mục kế sau, mục kế trước. Mỗi khi các lệnh này được thực hiện, hệ thống sẽ đọc ra tên mục đang được chỉ đến để người dùng biết được mình đang ở mục nào và lựa chọn "Vào" hoặc không.

Cấp điều hướng bài. Giống như cấp điều hướng mục, nhưng con trỏ chỉ đến danh sách bài. Vì có sự giống nhau giữa "Đầu", "Cuối", "Kế", "Trước" nên cần lệnh hướng con trỏ đến đúng danh sách cần chỉ, "Mục" và "Bài" giúp báo ứng dụng chuẩn bị nhận lệnh tiếp theo trong mục hay bài.

3.2 Nhóm điều khiển (Control)

Nhóm điều khiển thực hiện chức năng giống như một trình phát nhạc, dùng để kiểm soát âm thanh phát ra từ ứng dụng. Ví dụ: các lệnh như “Dừng”/“Tiếp” hay “To”/”Nhỏ” có chức năng như nút Play/Pause và điều chỉnh âm lượng.

3.3 Nhóm tương tác (Interaction)

Nhận lệnh “Có”/”Không” hoặc “Đúng”/”Sai” từ người dùng để ra quyết định thực hiện lệnh kế trước.

3.4 Nhóm lựa chọn (Selection)

Đánh số các mục và bài để tiện cho việc lựa chọn và nhận lệnh bằng các lệnh “Một” đến “Chín”.

4 Thực nghiệm và kết quả

4.1 Chuẩn bị dữ liệu

Điều kiện thu âm. Tần số lấy mẫu của dữ liệu được thiết lập ở mức 16kHz, tín hiệu ở định dạng WAV (Microsoft) 16-bit PCM. Người nói được thu trong điều kiện phòng tương đối yên tĩnh (có nhiễu nhẹ từ những tạp âm như tiếng quạt, tiếng chim hót, …), và thu âm sử dụng cùng một thiết bị micro.
Phụ thuộc người nói. Bộ dữ liệu gồm 20 set thu âm trên cùng một người nói là nữ, mỗi set gồm 46 lệnh như ở phần 3.

Độc lập người nói. Thu âm dữ liệu trên 42 người nói thuộc các vùng khác nhau (đa số giọng miền Trung), mỗi người thu âm 2 set với mỗi set gồm 46 câu lệnh như mô tả ở phần 3.

Từ bộ dữ liệu, chúng tôi phân nhóm dữ liệu huấn luyện và kiểm thử như sau:

Dữ liệu huấn luyện:	2756 câu thu âm từ 31 người nói giọng Quảng Nam hoặc Đà Nẵng (trong đó có 5 nữ)
Dữ liệu kiểm thử:	1012 câu thu âm từ 11 người nói giọng địa phương khác (4 nữ)

4.2 Huấn luyện mô hình

Việc lựa chọn các tham số phù hợp với hệ thống đang sử dụng để thiết lập cho thực nghiệm cụ thể như sau:

Loại bộ mô hình:	đơn giản, không chia sẻ tham số
Loại ma trận phương sai:	ma trận đường chéo
Số luồng:	1

Ngữ cảnh: không phụ thuộc ngữ cảnh

Loại tham số phổ: MFCC

Kích thước vector tham số: 39 chiều

Các tham số thay đổi:

Số trạng thái HMM (nStates): từ 3 đến 7

Số hỗn hợp của phân bố Gauss (nMixes): từ 1 đến 5

Với những mô hình HMM liên tục, có nghiên cứu cho rằng việc dùng ma trận phương sai đường chéo thì thuận tiện và thích hợp hơn việc dùng toàn bộ ma trận phương sai. Lý do chính là vì việc tính toán các thành phần nằm ngoài đường chéo mà chỉ dựa trên kích thước dữ liệu nhỏ thì không đáng tin cậy [1][2]. Ngữ cảnh được thiết lập là mono (không phụ thuộc ngữ cảnh) vì sử dụng hệ thống nhận dạng từ rời rạc nên các tính chất âm học của một từ không phụ thuộc vào các từ lân cận. Kích thước của vector đặc trưng là 39 được chọn như sau: phương pháp trích xuất vector đặc trưng là Mel Frequency Cepstral Coefficients (MFCCs), vector tĩnh được tham số hóa có 12 chiều, có chứa thành phần năng lượng (energy component), và có các hệ số delta (đạo hàm cấp một), các hệ số acceleration (đạo hàm cấp hai). Số 39 được tính từ chiều dài của vector tĩnh là 13 (12 chiều và thêm 1 thành phần năng lượng), cộng các hệ số delta (+13), cộng các hệ số acceleration (+13) [2].

4.3 Kết quả thực nghiệm

Bảng 1 và 2 lần lượt trình bày kết quả nhận dạng phụ thuộc và độc lập người nói ở chế độ offline, trong đó việc huấn luyện và nhận dạng thực hiện trên các tệp lưu tín hiệu tiếng nói ở định dạng WAV của Microsoft dùng máy tính laptop và thư viện HTK [4].

Bảng 1. Thống kê kết quả nhận dạng phụ thuộc người nói

nStates **nMixes**	**3**	**4**	**5**	**6**	**7**
1	99.13%	99.13%	99.13%	**99.42%**	99.28%
2	99.28%	99.13%	99.13%	99.13%	99.13%
3	99.13%	99.13%	99.13%	99.13%	99.13%
4	99.13%	99.13%	99.13%	99.13%	99.13%
5	99.13%	99.13%	99.13%	99.13%	99.13%

Bảng 2. Thống kê kết quả nhận dạng độc lập người nói

nStates **nMixes**	**3**	**4**	**5**	**6**	**7**
1	88.04%	88.11%	89.49%	90.74%	90.65%
2	90.09%	90.18%	-	-	-
3	90.61%	-	-	-	-
4	**91.14%**	-	-	-	-
5	90.78%	-	-	-	-

Hệ thống nhận dạng phụ thuộc người nói đạt kết quả rất cao, trên 99%, ở tất cả các cấu hình số trạng thái HMM và số hỗn hợp của phân bố Gauss, và đạt độ chính xác cao nhất tại số trạng thái bằng 6 và số hỗn hợp bằng 1 với 99,42%. Trong khi đó, hệ thống độc lập người nói đạt kết quả cao nhất 91,14% ở số trạng thái bằng 3 và số hỗn hợp bằng 4. Một số ô của kết quả nhận dạng độc lập người nói không có vì dữ liệu huấn luyện không đủ để huấn luyện mô hình phức tạp (mô hình có số lượng tham số lớn).

Chúng tôi cũng tiến hành thử nghiệm nhận dạng tiếng nói online trên máy tính nhúng có tài nguyên hạn chế Raspberry Pi 3 [5] kết hợp với một card âm thanh có giá thành rẻ. Ở chế độ online, việc nhận dạng được thực hiện trực tiếp trên tín hiệu tiếng nói thu được từ micro và chuyển đến card âm thanh. Chúng tôi dùng toàn bộ dữ liệu tiếng nói của 42 người mô tả ở Phần 4.1 cho việc huấn luyện mô hình, và dùng giọng nói của các thành viên trong nhóm nghiên cứu để kiểm thử. Do HTK không hỗ trợ chế độ online, chúng tôi đã dùng thư viện pocketsphinx [6] để thử nghiệm. Hệ thống đạt độ chính xác trên 80% khi nhận dạng độc lập người nói trong điều kiện phòng tương đối yên tĩnh.

5 Kết luận

Bài báo trình bày khái quát mô hình Markov ẩn và thiết kế tập lệnh giúp người khiếm thị điều khiển ứng dụng đọc báo điện tử bằng giọng nói. Chúng tôi đã thu âm bộ dữ liệu của nhiều người nói, sau đó tiến hành thực nghiệm huấn luyện mô hình và nhận dạng tiếng nói trên bộ dữ liệu nhằm đánh giá sự phụ thuộc của hệ thống vào các tham số khác nhau của mô hình. Kết quả khả quan khi thử nghiệm nhận dạng không phụ thuộc người nói trên hệ thống máy tính có tài nguyên hạn chế cho phép chúng tôi phát triển một thiết bị cho phép người dùng ra lệnh điều khiển bằng giọng nói theo thời gian thực trong tương lai.

Tài liệu tham khảo

1. Lawrence R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, Vol. 77, No. 2, 1989.
2. Mark Gales, Steve Young, The Application of Hidden Markov Models in Speech Recognition, Foundations and Trends in Signal Processing, Vol. 1, No. 3, 2008.
3. https://play.google.com/store/apps/details?id=com.vnspeak.newsreader4blind&hl=vi. Last accessed: 30/11/2017.
4. Steve Young et al., The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, 2009.
5. https://www.raspberrypi.org/products/raspberry-pi-3-model-b/. Last accessed: 30/11/2017.
6. https://sourceforge.net/projects/cmusphinx/. Last accessed: 30/11/2017.