(12) **BẢN MÔ TẢ GIẢI PHÁP HỮU ÍCH THUỘC BẰNG ĐỘC QUYỀN GIẢI PHÁP HỮU ÍCH**

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **2-0001923**

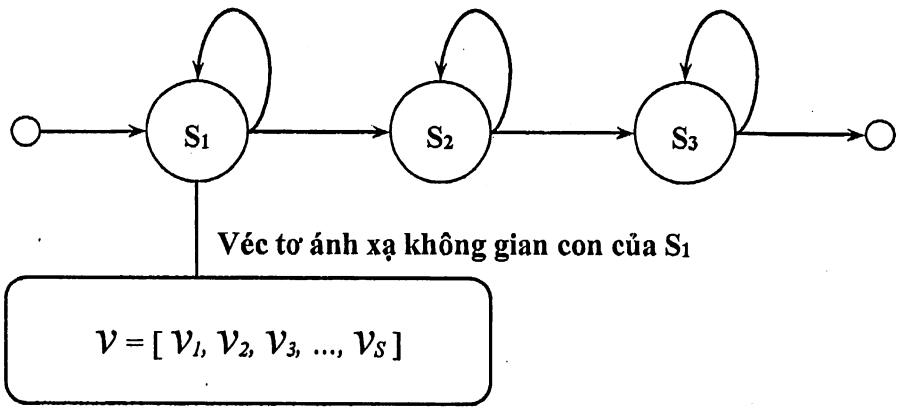
(51)[7] **G10L 15/00**

(13) **Y**

(21) 2-2015-00186
(22) 02.07.2015
(45) 25.12.2018 369
(43) 25.04.2016 337
(73) ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH (VN)
Phường Linh Trung, quận Thủ Đức, thành phố Hồ Chí Minh
(72) Vũ Hải Quân (VN)

(54) PHƯƠNG PHÁP NHẬN DẠNG TIẾNG NÓI TIẾNG VIỆT SỬ DỤNG RÀNG BUỘC KHÔNG GIAN ĐẶC TRƯNG

(57) Giải pháp hữu ích đề cập đến phương pháp xây dựng hệ thống nhận dạng tiếng nói tiếng Việt dựa trên ràng buộc không gian đặc trưng bao gồm các bước sau: bước 1: xác định điều kiện tài nguyên huấn luyện; bước 2: xây dựng mô hình ngữ âm ràng buộc không gian đặc trưng; bước 3: tiến hành mở rộng phân hoạch không gian con cho mô hình ngữ âm; bước 4: lựa chọn phương pháp cho bộ trích đặc trưng, mô hình ngôn ngữ và đồ thị nhận dạng; bước 5: cấu thành hệ thống nhận dạng được tiến hành thông thường chỉ khác trong cách tính tích p(xlj).



Lĩnh vực kỹ thuật được đề cập

Phương pháp thuộc lĩnh vực nhận dạng tiếng nói (ASR – Automatic Speech Reconition), một nhánh của xử lý ngôn ngữ nói (SPL – Spoken Language Processing) đề cập đến việc xây một hệ thống nhận dạng tiếng nói với ngôn ngữ có tài nguyên hạn chế nhằm rút ngắn thời gian, chi phí và công sức trong quá trình thu thập và xử lý dữ liệu. Đối với các ngôn ngữ phổ biến, lượng dữ liệu thu thập được lên đến hàng trăm giờ thu âm với giọng của hàng ngàn người nói khác nhau. Quá trình thu thập dữ liệu tốn rất nhiều thời gian, công sức và chi phí dẫn đến việc nhiều ngôn ngữ chưa thể thực hiện quá trình này.

Tình trạng kỹ thuật của giải pháp hữu ích

Vấn đề cốt yếu trong quá trình xây dựng một hệ thống nhận dạng tiếng nói là kho tài nguyên dữ liệu dưới dạng âm thanh và nội dung văn bản tương ứng của nó. Số lượng dữ liệu càng lớn thì chất lượng của bộ nhận dạng càng cao. Để cải thiện hiệu suất nhận dạng cho các ngôn ngữ có tài nguyên hạn chế, tùy theo số lượng dữ liệu có được, một trong các phương pháp sau đây sẽ được áp dụng để khai thác hệ thống nhận dạng tiếng nói của các ngôn ngữ khác:

- Trường hợp không có tài nguyên dữ liệu hoặc dữ liệu quá ít không đáng kể: áp dụng mô hình xuyên ngữ.
- Trường hợp có tài nguyên dữ liệu nhưng khá hạn chế: áp dụng phương pháp điều chỉnh thích nghi ngôn ngữ.
- Trường hợp tài nguyên dữ liệu tương đối lớn: áp dụng phương pháp khởi tạo nền.

Mô hình xuyên ngữ:

Khi hoàn toàn không có tài nguyên dữ liệu, phương pháp này sử dụng mô hình ngữ âm được huấn luyện trên các ngôn ngữ có đầy đủ tài nguyên để xây dựng bộ nhận dạng tiếng nói cho ngôn ngữ không có tài nguyên. Hai biến thể chính của phương pháp này phụ thuộc vào loại dữ liệu huấn luyện:

- Chuyển thể đơn ngữ: dữ liệu huấn luyện chỉ gồm một ngôn ngữ.
- Chuyển thể đa ngữ: dữ liệu huấn luyện thuộc nhiều ngôn ngữ khác nhau.

Các nghiên cứu trước đây cho thấy dữ liệu đa ngữ cho hiệu quả cao hơn dữ liệu đơn ngữ; đồng thời tương quan giữa ngôn ngữ nguồn (có đầy đủ tài nguyên) và ngôn ngữ đích (không có tài nguyên) cũng quyết định hiệu suất nhận dạng: các ngôn ngữ thuộc cùng một họ (theo cách phát âm) sẽ cho kết quả tốt hơn các ngôn ngữ khác họ. Ví dụ: chuyển đổi giữa tiếng Trung Quốc và tiếng Việt sẽ cho kết quả tốt hơn giữa tiếng Việt với tiếng Anh.

Điều chỉnh thích nghi ngôn ngữ:

Đối với trường hợp tài nguyên dữ liệu khá hạn chế, các kỹ thuật điều chỉnh thích nghi ngôn ngữ sẽ được áp dụng để điều chỉnh bộ nhận dạng của các ngôn ngữ nguồn (ngôn ngữ có tài nguyên đầy đủ) theo hướng thích nghi với tập dữ liệu của ngôn ngữ đích (ngôn ngữ có tài nguyên hạn chế).

Cơ chế điều chỉnh thích nghi ngôn ngữ ở mức tổng quát đòi hỏi phải có đầy đủ hai yếu tố đầu vào là: mô hình ngữ âm của ngôn ngữ nguồn và dữ liệu điều chỉnh. Theo đó, đã có nhiều hướng tiếp cận được đưa ra và nhiều phương pháp được đề xuất, chủ yếu gồm 2 nhóm chính:

- Hướng tiếp cận dựa trên phép biến đổi.
- Hướng tiếp cận dựa trên mô hình.

Khởi tạo nền:

Khi tài nguyên dữ liệu có được tương đối lớn, phương pháp khởi tạo nền thường được áp dụng. Ý tưởng chính của phương pháp này là sử dụng mô hình ngữ âm của các ngôn ngữ nguồn (ngôn ngữ có tài nguyên đầy đủ) để khởi tạo tham số cho mô hình ngữ âm của ngôn ngữ đích (ngôn ngữ có tài nguyên tương đối lớn). Sau khi đã được khởi tạo, mô hình ngữ âm của ngôn ngữ đích sẽ được huấn luyện lại từng bước một dựa trên kho tài nguyên sẵn có của ngôn ngữ này.

Nghiên cứu cũng cho thấy cơ chế khởi tạo nền đem lại kết quả tốt hơn so với việc khởi tạo phẳng hay khởi tạo ngẫu nhiên cho các tham số ngữ âm ban đầu. Đồng thời, cũng giống như trong phương pháp xuyên ngữ, mô hình ngữ âm nguồn nếu được huấn luyện từ dữ liệu đa ngữ sẽ cho kết quả tốt hơn dữ liệu đơn ngữ.

Nhìn chung, cả ba phương pháp này có ưu điểm là nhanh chóng đưa ra được mô hình nhận dạng tiếng nói có chất lượng trung bình để áp dụng trong các ứng dụng qui mô nhỏ hoặc được dùng như hệ thống ban đầu để phát triển thêm. Tuy nhiên, chất lượng của chúng thường không đạt yêu cầu với bài toán nhận dạng tiếng nói có bộ từ vựng lớn.

Ở trong nước, tài nguyên để sử dụng cho bài toán nhận dạng tiếng nói tiếng Việt hiện nay vẫn còn rất hạn chế. Tiếng Việt không nằm trong danh sách đầu tư phát triển của các công ty hàng đầu thế giới về nhận dạng tiếng nói (Nuance, Microsoft, IBM...) mặc dù tiếng Việt được xếp vào nhóm 20 ngôn ngữ được sử dụng nhiều nhất trên thế giới với dân số Việt Nam lên đến 85 triệu người. Một số nghiên cứu đã được thực hiện trong các Viện nghiên cứu, trường Đại học trên thế giới cũng như ở trong nước trong vòng 5 năm trở lại đây. Các kho ngữ liệu cho nhận dạng tiếng nói tiếng Việt hiện tại vẫn còn hạn chế ở mức hàng chục giờ thu âm với số lượng người nói cũng chỉ ở mức hàng chục người khác nhau. Ngoài ra, các kho ngữ liệu này lại sử dụng tiếng nói của các vùng miền khác nhau trên khắp Việt Nam, và do đó, mang những đặc trưng riêng về giọng nói và phong cách nói của từng vùng miền. Hiện tại vẫn chưa có một phương pháp nào về nhận dạng tiếng nói cho toàn bộ các vùng miền khác nhau của Việt Nam và nguồn dữ liệu còn hạn chế dẫn đến việc chưa thể xây dựng một bộ nhận dạng tiếng nói chất lượng cao sử dụng các phương pháp truyền thống.

Các nguyên nhân dẫn đến khó khăn này bao gồm:

- Các kho ngữ liệu cho nhận dạng tiếng nói tiếng Việt hiện tại vẫn còn hạn chế ở mức hàng chục giờ thu âm với số lượng người nói cũng chỉ ở mức hàng chục người khác nhau. Trong khi đó, để có thể áp dụng các phương pháp truyền thống cho việc xây dựng bộ nhận dạng đòi hỏi lượng dữ liệu thu thập được phải lên đến hàng trăm giờ thu âm với giọng của hàng ngàn người nói khác nhau.
- Các kỹ thuật cải thiện hiệu suất nhận dạng cho ngôn ngữ có tài nguyên hạn chế có ưu điểm là nhanh chóng đưa ra được mô hình nhận dạng tiếng nói có chất lượng trung bình để áp dụng trong các ứng dụng qui mô nhỏ hoặc được dùng như hệ thống ban đầu để phát triển thêm. Tuy nhiên, chất lượng của các phương pháp này thường không đạt yêu cầu với bài toán nhận dạng tiếng nói có bộ từ vựng lớn.

Bản chất kỹ thuật giải pháp hữu ích

Trong phương pháp này, chúng tôi đề xuất xây dựng mô hình ngữ âm dựa trên ràng buộc không gian đặc trưng. Toàn cảnh mà phương pháp nhắm tới là một mô hình ngữ âm mới và phương pháp huấn luyện tương ứng; song song đó là một hệ thống nhận dạng tiếng nói tiếng Việt hoàn chỉnh cho hiệu suất cao với tiếng nói ba miền. Cụ thể là xây dựng các mô hình Markov ẩn hỗ trợ đa ngôn ngữ và sử dụng các tham số ràng buộc đã được huấn luyện cho các ngôn ngữ có nhiều tài nguyên hơn để ngoại suy giá trị của các tham số ràng buộc cho các ngôn ngữ có tài nguyên hạn chế. Giải pháp đề xuất phương pháp nhận dạng tiếng nói tiếng Việt sử dụng ràng buộc không gian đặc trưng bao gồm các bước sau:

bước 1: xác định điều kiện tài nguyên huấn luyện;

bước 2: xây dựng mô hình ngữ âm ràng buộc không gian đặc trưng;

bước 3: tiến hành mở rộng phân hoạch không gian con cho mô hình ngữ âm;

bước 4: lựa chọn phương pháp cho bộ trích đặc trưng, mô hình ngôn ngữ và đồ thị nhận dạng;

bước 5: cấu thành hệ thống nhận dạng được tiến hành thông thường chỉ khác trong cách tính tích p(x|j).

Mô tả vắn tắt các hình vẽ

Hình 1 là hình vẽ mô tả kiến trúc tổng quát của hệ thống nhận dạng tiếng nói tiếng Việt.

Hình 2 là hình vẽ mô tả cơ cấu mô hình ngữ âm truyền thống.

Hình 3 là hình vẽ mô tả mô hình ngữ âm ràng buộc không gian đặc trưng.

Hình 4 là hình vẽ mô tả phương pháp huấn luyện mô hình ngữ âm ràng buộc không gian đặc trưng.

Hình 5 là hình vẽ biểu diễn đa phân hoạch cho mỗi trạng thái của mô hình ngữ âm ràng buộc không gian đặc trưng.

Mô tả chi tiết giải pháp hữu ích

Nhận dạng tiếng nói là tiến trình chuyển giọng nói của con người (biểu diễn dưới dạng sóng âm) sang văn bản (text) một cách tự động. Nói cách khác, công nghệ nhận dạng tiếng nói cho phép máy tính biết được nội dung truyền đạt trong lời nói của con người. Cần phân biệt nhận dạng tiếng nói (speech recognition) với nhận dạng người nói (speaker recognition) – công nghệ nhận dạng chủ thể của mẫu tiếng nói (trả lời cho câu hỏi "giọng nói đó là của ai?").

Môt hệ thống nhận dạng tiếng nói điển hình nhận đầu vào là các tín hiệu tiếng nói dưới dạng các file/luồng âm thanh. Tín hiệu âm thanh này sẽ được đưa qua bước rút trích đặc trưng, rồi thực hiện tiến trình nhận dạng và trả về kết quả ở dạng văn bản. Nói nôm na, nhận dạng tiếng nói chính là tác vụ chuyển tiếng nói thành văn bản.

Cách tiếp cận của nhận dạng tiếng nói (ASR) là dựa vào thống kê, cụ thể là các mô hình Markov ẩn, mô hình ngôn ngữ N-gram. Ở mức tổng quát nhất, một hệ thống ASR được mô tả trong Hình 1, bao gồm các thành phần:

bộ trích đặc trưng: thực hiện rút trích đặc trưng từ tín hiệu âm thanh trước khi đưa vào nhận dạng;

mô hình ngữ âm: liên quan đến việc biểu diễn tri thức cho tín hiệu ngữ âm, âm vị, ngữ điệu…

mô hình ngôn ngữ: liên quan đến việc biểu diễn tri thức của các từ, chuỗi từ, hình thành nên câu;

tiến trình tìm kiếm trên đồ thị nhận dạng: chọn lựa chuỗi từ ứng với tín hiệu ngữ âm. Về mặt bản chất đây là việc tìm kiếm tối ưu trên đồ thị được xây dựng bằng cách kết-ghép các mô hình ngôn ngữ, mô hình ngữ âm và từ điển phát âm.

Với kiến trúc này, tất cả các tri thức về ngữ âm, ngôn ngữ, và thống kê đều được tích hợp trong đó. Bài toán nhận dạng tiếng nói trở thành bài toán tổ chức và tìm kiếm trên đồ thị, bao gồm các thách thức:

- Kích thước: đồ thị cần được tổ chức và tối ưu sao cho có kích thước càng nhỏ (thông qua số đỉnh và số cạnh) càng tốt. Tuy nhiên vẫn phải đảm bảo tính đầy đủ cho toàn bộ thông tin tri thức được tích hợp trong đó.
- Độ chính xác: phép tìm kiếm trên đồ thị tuân theo ràng buộc cơ bản, đó là phải đảm bảo sao cho kết quả tìm kiếm gần giống nhất với chuỗi từ đã được phát âm.

Như vậy, với các thành phần trên, việc xây một hệ thống nhận dạng tiếng nói điển hình đòi hỏi phải có một kho tài nguyên dữ liệu rất lớn. Đối với các ngôn ngữ phổ biến, lượng dữ liệu thu thập được lên đến hàng trăm giờ thu âm với giọng của hàng ngàn người nói khác

nhau. Quá trình thu thập dữ liệu tốn rất nhiều thời gian, công sức và chi phí dẫn đến việc nhiều ngôn ngữ chưa thể thực hiện quá trình này.

Phương pháp huấn luyện bộ nhận dạng tiếng nói cổ điển hoạt động không hiệu quả đối với các ngôn ngữ có tài nguyên hạn chế (thuật ngữ under-resourced languages – ngôn ngữ mà kho ngữ liệu cho huấn luyện/nghiên cứu khoa học chưa được xây dựng hay thu thập nhiều). Tuy nhiên, trong trường hợp này kết quả nhận dạng có thể được cải thiện bằng cách sử dụng phương pháp khởi tạo nền, học thích nghi và chuyển thể mô hình từ ngôn ngữ khác.

Kiến trúc tổng quát của hệ thống nhận dạng tiếng nói tiếng Việt được mô tả trong Hình 1, gồm các thành phần chính là: bộ trích đặc trưng, mô hình ngữ âm, mô hình ngôn ngữ, đồ thị nhận dạng.

Phương pháp này xoáy sâu vào xây dựng mô hình ngữ âm với điều kiện tài nguyên hạn chế, từ đó tích hợp và liên kết vào hệ thống nhận dạng tiếng nói tiếng Việt, đảm bảo hiệu suất đầu ra. Các mục con sau sẽ đi vào mô tả chi tiết cho phương pháp này.

Mô hình ngữ âm cho điều kiện tài nguyên hạn chế:

Các hệ thống nhận dạng tiếng nói dựa trên mô hình Markov ẩn sử dụng hỗn hợp phân bố Gauss để mô hình hóa xác suất của các mẫu quan sát được. Một không gian đặc trưng cho nhận dạng tiếng nói được định nghĩa bởi Mel Cepstra, bao gồm đạo hàm bậc nhất và đạo hàm bậc hai, thường được biến đổi tiếp (sử dụng phân tích biệt lập tuyến tính -LDA hoặc các phép biến đổi khác) về một không gian phù hợp hơn với mô hình hỗn hợp Gauss. Số chiều của các không gian đặc trưng này (từ 30 đến 60) lớn hơn rất nhiều so với số chiều mà những nhà nghiên cứu về lý thuyết ngữ âm đề nghị. Tuy nhiên, không có một phép biến đổi đơn giản nào để chuyển từ không gian Mel Cepstra về một không gian có số chiều ít như vậy.

Thử khảo sát vấn đề tham số của một hệ thống nhận dạng tiếng nói điển hình theo cơ cấu Mô hình Markov ẩn (Hidden Markov Model-HMM) - Mô hình hợp Gauss (Gaussian Mixture Model-GMM) như trong Hình 2, với các input (đầu vào) chuẩn như sau:

- Mỗi đơn vị ngữ âm (phone) có 3 states (trạng thái).
- Mỗi state có I (~20) phân bố Gauss (Gaussian mixtures).
- Véc tơ đặc trưng có số chiều D = 39.

Khi đó, tổng lượng tham số cho 100 đơn vị ngữ âm không ngữ cảnh (monophone) sẽ là:

$$\sum|\text{params}| = |\text{ngữ âm}| \,.\, |\text{trạng thái}| \,.\, |\text{phân bố trên trạng thái}|$$

$$\sum|\text{params}| = 100 \,.\, 3 \,.\, |(\mu_i, \sigma_i, w_i)|$$

$$\sum|\text{params}| = 100 \,.\, 3 \,.\, I \,.\, (D + D + 1)$$

$$\sum|\text{params}| = 100 \,.\, 3 \,.\, 20 \,.\, (39 + 39 + 1)$$

$$\sum|\text{params}| = 474000 \text{ tham số.}$$

Và đối với trường hợp đơn vị ngữ âm với ngữ cảnh lân cận (triphones) (~5341 tied-triphones), tổng lượng tham số sẽ là:

$\sum|params| \sim 25$ triệu tham số.

Như vậy, có thể thấy rằng lượng tham số này là quá lớn để có thể huấn luyện chính xác với một lượng nhỏ tài nguyên.

Phương pháp vẫn sử dụng đặc trưng dựa trên Mel Cepstra và một danh sách các phân bố thống kê (Gauss) chung cho tất cả các âm của tất cả các ngôn ngữ. Trong quá trình huấn luyện, vấn đề dữ liệu sẽ được giải quyết bằng cách xem các phân bố Gauss này như các hàm cơ bản. Sự phụ thuộc giữa các Gauss sẽ tạo ra một không gian biểu diễn âm thanh có số chiều ít hơn. Làm việc trên không gian ít chiều hơn sẽ yêu cầu ít dữ liệu huấn luyện hơn. Cụ thể, phương pháp đề xuất cách biểu diễn mô hình ngữ âm như ở Hình 3, trong đó mỗi trạng thái chỉ chứa 1 véc tơ ánh xạ không gian con thay cho mô hình hợp (Gauss Gaussian Mixture Model – GMM) của mô hình truyền thống. Véc tơ này có số chiều là S (~20, hoặc có thể tùy biến từ 20 đến D) nhỏ hơn rất nhiều so với số chiều của véc tơ đặc trưng (D = 39).

Với cách biểu diễn này, ở bước nhận dạng ta chỉ cần tái thiết lập cấu trúc mô hình hợp Gauss (Gaussian Mixture Model – GMM) truyền thống cho mỗi trạng thái (gồm μ_i, $\sum_i$ và w_i) từ véc tơ v của trạng thái đó bằng phép nhân ma trận như sau:

- $\mu_i = M_i.v$
- $\omega_i = \dfrac{W_i.v}{\sum_{i'=1}^{I} W_{i'}.v}$
- $\sum_i$ dùng chung cho tất cả các trạng thái.

Bước tái thiết lập mô hình hợp Gauss (Gaussian Mixture Model – GMM) này cần sự hiện diện của các ma trận M_i, W_i và $\sum_i$. Bộ tham số (M_i, W_i, $\sum_i$) được gọi là bộ tham số toàn cục, dùng chung cho tất cả các trạng thái của tất cả các ngữ âm. Còn tập các véc tơ {v} được gọi là bộ tham số riêng (mỗi véc tơ v biểu diễn riêng cho 1 trạng thái).

Trong mô hình ngữ âm mới này, tổng lượng tham số cho 100 đơn vị ngữ âm không ngữ cảnh - monophones sẽ là:

$\sum|params| = |\text{ngữ âm}| \,.\, |\text{trạng thái}| \,.\, |\text{params trên trạng thái}| + |(M_i, W_i, \sum_i)|$

$\sum|params| = 100 \,.\, 3 \,.\, S + I \,.\, (D.S + S + D.D)$

$\sum|params| = 100 \,.\, 3 \,.\, 20 + 20\ (39.20 + 20 + 39.39)$

$\sum|params| = 6000 + 46420$

$\sum|params| = 52420$ nhỏ hơn rất nhiều so với 474000 tham số của mô hình truyền thống.

Tương tự cho trường hợp đơn vị ngữ âm với ngữ cảnh lân cận (triphones) (~ 5341 tied-triphones):

$\sum|params| = 320460 + 46420 = 366880 << 25$ triệu tham số của mô hình cũ.

Như vậy, mô hình ngữ âm mới này đã bó không gian ngữ âm toàn cục thành một không gian con mô tả phạm vi các âm sắc mà con người có khả năng phát âm được. Ít lượng tham số hơn cũng đồng nghĩa với việc yêu cầu lượng tài nguyên huấn luyện ít hơn.

Trong bước huấn luyện mô hình, bộ tham số toàn cục (M_i, W_i, $\sum_i$) sẽ được huấn luyện trên kho ngữ liệu đa ngữ, còn bộ tham số riêng {v} sẽ được huấn luyện trên kho tài nguyên hạn chế của tiếng Việt. Sau đó, 2 bộ tham số này sẽ được dùng để xây dựng mô hình ngữ âm hoàn chỉnh cuối cùng thông qua phép nhân ma trận. Hình 4 minh họa các bước của quá trình này.

Ngoài ra, nếu lượng tài nguyên tiếng Việt thu thập được vượt hơn định mức, ta có thể tăng số lượng véc tơ biểu diễn cho mỗi trạng thái. Việc tăng số véc tơ sẽ làm tăng tổng lượng tham số của mô hình, kéo theo một hiệu suất nhận dạng tốt hơn toàn cục. Ta gọi khái niệm này là phân hoạch trạng thái (Hình 5), trong đó mỗi trạng thái con sẽ bao gồm một véc tơ v và trọng c tương ứng của nó.

Đặc tả hình thức

Mô hình ngữ âm

Trước hết, ta quy ước một số khái niệm sau:

v^+ là véc tơ v thêm thành phần mang giá trí 1 ở cuối

(vd: $v = [v_1, v_2, \ldots, v_n, 1]$)

v^- là véc tơ v bỏ đi thành phần cuối

M^+ là ma trận M thêm vào dòng cuối với giá trị [0, 0, …, 0, 1]

M^{+0} là ma trận M thêm vào dòng cuối với giá trị [0, 0, …, 0, 0]

M^- là ma trận M bỏ đi dòng cuối

M^{--} là ma trận M bỏ đi dòng cuối và cột cuối

M^{-C} là ma trận M bỏ đi cột cuối

I là số lượng phân bố Gauss trong không gian đặc trưng

J là tổng số trạng thái của toàn bộ mô hình Markov ẩn(Hidden Markov Model-HMM) đại diện cho hệ ngữ âm.

D là số chiều của véc tơ đặc trưng

S là số chiều của véc tơ không gian con.

Khi đó, xác suất quan sát được đặc trưng x ở trạng thái j sẽ là:

$$p(\mathbf{x}|j) = \sum_{i=1}^{I} w_{ji}\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_{ji},\boldsymbol{\Sigma}_i)$$

Với:

$$\boldsymbol{\mu}_{ji} = \mathbf{M}_i\mathbf{v}_j^+$$

$$w_{ji} = \frac{\exp(\mathbf{w}_i^T\mathbf{v}_j^+)}{\sum_{i'=1}^{I}\exp(\mathbf{w}_{i'}{}^T\mathbf{v}_j^+)}$$

Xác suất p(x|j) cùng với ma trận chuyển trạng thái a_{ij} tạo thành bộ mô hình Markov ẩn mô hình hóa cho hệ ngữ âm của ngôn ngữ tương ứng, đây chính là mô hình ngữ âm ràng buộc không gian đặc trưng, đại diện bởi bộ tham số:

- Ma trận chiếu đặc trưng $M_i \in R^{D \times (S+1)}$
- Véc tơ chiếu trọng $w_i \in R^{(S+1)}$
- Ma trận hiệp phương sai $\sum_i \in R^{D \times D}$
- Véc tơ không gian con $v_j \in R^S$
- Ma trận chuyển trạng thái a_{ij}
- Bộ ngữ âm (phoneset) của ngôn ngữ

Nếu ta mở rộng số lượng véc tơ không gian con trong mỗi trạng thái, xác suất p(x|j) sẽ trở thành:

$$p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm}\sum_{i=1}^{I} w_{jmi}\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_{jmi},\boldsymbol{\Sigma}_i)$$

Với:

$$\boldsymbol{\mu}_{jmi} = \mathbf{M}_i\mathbf{v}_{jm}^+$$

$$w_{jmi} = \frac{\exp(\mathbf{w}_i^T\mathbf{v}_{jm}^+)}{\sum_{i'=1}^{I}\exp(\mathbf{w}_{i'}^T\mathbf{v}_{jm}^+)}.$$

M_j ~ số lượng phân hoạch trạng thái con của mô hình Markov ẩn (sub-state) của trạng thái j

c_{jm} ~ trọng số của phân hoạch trạng thái con của mô hình Markov ẩn (sub-state) thứ m trong trạng thái j,

thỏa điều kiện $\sum c_{jm} = 1$

Trong trường hợp này, mô hình mở rộng phân hoạch không gian con thành M cụm con theo cách như mô hình hợp Gauss (Gaussian Mixture Model-GMM) phân hoạch các phân bố Guass, với cùng giả thuyết xấp xỉ của định lý giới hạn trung tâm. Ta gọi mô hình mở rộng này là mô hình ngữ âm ràng buộc đa không gian con.

Phương pháp huấn luyện

Phương pháp huấn luyện mô hình ngữ âm ràng buộc không gian đặc trưng gồm 3 pha chính:

pha 1: xây dựng không gian đặc trưng

pha 2: ước lượng tương quan giữa các hàm cơ sở trong không gian con

pha 3: huấn luyện không gian con

Pha 1 xây dựng không gian đặc trưng

a) phân cụm các véc tơ đặc trưng thành I nhóm, mỗi nhóm đại diện bởi một phân bố Gauss;

b) lặp: với mỗi cụm:

tính toán lại giá trị ràng buộc hồi qui tuyến tính theo hướng cực đại hóa khả xuất (Maximum Likelihood Linear Regression- MLLR) cho từng cụm;

tích lũy các biến ẩn của mô hình hợp Gauss (Gaussian Mixture Model-GMM)

cập nhật lại tham số của mô hình hợp Gauss (Gaussian Mixture Model-GMM)

chuẩn hóa (Normalize) lại trọng số của từng phân bố Gauss trong cụm mô hình hợp Gauss (Gaussian Mixture Model-GMM)

trích xuất ma trận phương sai đường chéo từ mô hình hợp Gauss (Gaussian Mixture Model-GMM);

c) phân hoạch I phân bố Gauss thành K nhóm cha (mô hình hợp Gauss (Gaussian Mixture Model-GMM));

Pha 2 ước lượng tương quan giữa các hàm cơ sở trong không gian con

Với k = 1…K

Với i = 1…I, tính:

$$\log p^{diag}(x(t), k, i) = \log\det_k^{(s)} + \log w_{ki} + \log N(x_{ki}(t)|u_{ki}^{diag}, \sum_{ki}^{diag})$$

tinh giản (prune) ra các cặp $p^{diag}(k,i)$ có $\log p^{diag}(x(t), k, i)$ cao nhất;

với mỗi cặp như vậy, tính:

$$\log P(x(t), k, i) = \log\det_k^{(s)} + \log w_{ki} + \log N(x_{ki}(t)|u_{ki}, \sum_{ki})$$

tinh giản (prune) ra các cặp P(k, i) có logP(x(t), k, i) cao nhất;

Pha 3 huấn luyện không gian con

a) khởi tạo không gian đặc trưng với cấu trúc mô hình hợp Gauss (Gaussian Mixture Model-GMM);

b) khởi tạo mô hình Markov ẩn (Hidden Markov Models –HMMs);

c) lặp: Chọn từng phân hoạch không gian con để huấn luyện

với mỗi mẫu huấn luyện:

tính lại giá trị của các biến ẩn (i, j, k) và biến tích lũy (α, β, γ, δ, ζ)

chọn ra các phân bố Gauss có p(k, i) cao nhất như trong Pha 2

xác định ánh xạ không gian con từ các phân bố Gauss chọn được và bộ mô hình hợp Gauss (Gaussian Mixture Model-GMM) của Pha 1 (không gian đặc trưng)

cập nhật lại giá trị của các véc tơ không gian con

gia giảm số lượng trạng thái con theo phương pháp nhân đôi và kỹ thuật trộn tham số (mixing) của mô hình hợp Gauss (Gaussian Mixture Model-GMM)

tùy chọn gia giảm kích thước không gian con cho khớp nhất với kho ngữ liệu.

Phương pháp mở rộng phân hoạch không gian con, thông qua 3 bước:

Bước 1 xác định số phân hoạch dựa trên điều kiện tài nguyên hiện hữu

i) chọn ra bộ ngữ liệu kiểm định hiệu suất nhận dạng D_T

gọi M là số phân hoạch con cần xác định

ii) lặp:

tăng dần M từ 1, với mỗi giá trị của M thực hiện:

tính $\log P(x(t), k, i) = \log \det_k^{(s)} + \log w_{ki} + \log N(x_{ki}(t)|u_{ki}, \sum_{ki})$ với $x \in D_T$

tính độ tăng $\Delta = \log_M P - \log_{M-1} P$

cho đến khi $\Delta < 0$

iii) chọn M lớn nhất mà tại đó $\Delta >= 0$

Bước 2 mở rộng cấu trúc mô hình ngữ âm theo số phân hoạch đã định

i) chuyển đổi hàm mật độ xác suất trong mỗi trạng thái thành

$$p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{jmi}, \boldsymbol{\Sigma}_i)$$

ii) từ 1 phân hoạch của mỗi trạng thái, tiến hành lặp:

sao chép tham số, nhân bản ra *m* phân hoạch con, theo độ tăng 1 phân hoạch mới ở mỗi bước lặp: *m = 1 ... M*

thực hiện lại Pha 3 (huấn luyện không gian con) của mô hình

cho đến khi tổng số phân hoạch *m* đạt đến giá trị của *M* đã định

Bước 3 huấn luyện phân hoạch không gian con

mục tiêu là hiệu chỉnh lại giá trị của bộ tham số c_{jm} và v_{jm} cho phù hợp nhất với dữ liệu học, theo 2 biểu thức:

$$\hat{c}_{jkm} = \frac{\gamma_{jkm} + \tau^{(w)}}{\sum_{k=1}^{K}\sum_{m=1}^{M_{jk}}(\gamma_{jkm} + \tau^{(w)})}$$

$$\hat{\mathbf{v}}_{jkm'} = \mathbf{v}_{jkm} - 0.1\mathbf{H}_k^{(\mathrm{sm})^{-0.5}}\mathbf{r}$$

trong đó,

$$\mathbf{H}_k^{(\mathrm{sm})} = \frac{1}{\sum_i \gamma_{ki}}\sum_i \gamma_{ki}$$

và *r* là giá trị phát sinh ngẫu nhiên theo phân phối chuẩn.

Phương pháp xây dựng hệ thống nhận dạng tiếng nói tiếng Việt sử dụng ràng buộc không gian đặc trưng

Như đã đề cập ở nội dung trước, một hệ thống nhận dạng tiếng nói gồm có 4 thành phần chính là: bộ trích đặc trưng, mô hình ngữ âm, mô hình ngôn ngữ, đồ thị nhận dạng. Cách thức xây dựng và liên kết các thành phần này đã đi vào kinh điển. Điểm khác biệt ở đây thể hiện qua việc đổi mới thành phần mô hình ngữ âm, kéo theo một số thay đổi khác. Cụ thể, phương pháp xây dựng được thực hiện thông qua gồm 5 bước:

Bước 1: xác định điều kiện tài nguyên huấn luyện

Tùy theo lượng dự liệu tiếng nói có được, ta phân ra 3 ngạch sau:

- Ngạch 1: Tổng thời lượng ít hơn 20 giờ hoặc tổng số người nói ít hơn 100 người. Đối với ngạch này, tiếp tục thực hiện lần lượt các bước 2, 3, 4, và 5.
- Ngạch 2: không thuộc các trường hợp của ngạch 1 và, tổng thời lượng ít hơn 60 giờ hoặc tổng số người nói ít hơn 300 người. Đối với ngạch này, tiếp tục thực hiện các bước 2, 4, và 5, bỏ qua bước 3.
- Ngạch 3: Các trường hợp còn lại. Đối với ngạch này, phương pháp xây dựng hệ nhận dạng tiếng nói kinh điển được áp dụng.

Bước 2: xây dựng mô hình ngữ âm ràng buộc không gian đặc trưng

Từ kho dữ liệu tiếng nói, tiến hành huấn luyện mô hình ngữ âm ràng buộc không gian đặc trưng như đã mô tả ở mục "Phương pháp huấn luyện".

Bước 3: tiến hành mở rộng phân hoạch không gian con cho mô hình ngữ âm

Nếu thuộc ngạch 2, dữ liệu có được ở mức tương đối, ta thực hiện mở rộng phân hoạch con theo như mô tả ở mục "Phương pháp mở rộng phân hoạch không gian con".

Bước 4: lựa chọn phương pháp cho bộ trích đặc trưng, mô hình ngôn ngữ, và đồ thị nhận dạng

Các thành phần còn lại của hệ nhận dạng có thể được lựa chọn tự do theo kinh điển. Tuy nhiên, khuyến khích dùng:

- Đặc trưng Các hệ số phổ Mel (Mel Frequency Cepstral Coefficient –MFCC), một đặc trưng phổ biến dùng trong nhận dạng tiếng nói.
- Mô hình N-gram, một dạng mô hình thống kê, cho mô hình ngôn ngữ.
- Máy chuyển đổi trạng thái hữu hạn (Finite State Transducer – FST) cho đồ thị nhận dạng.

Bước 5: cấu thành hệ thống nhận dạng tiếng nói tiếng Việt tiến hành theo thông thường, chỉ khác trong cách tính tích p(x|j):

$$p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \mu_{jmi}, \Sigma_i)$$

Ví dụ thực hiện phương pháp

Phương pháp nhận dạng tiếng nói tiếng Việt sử dụng ràng buộc không gian đặc trưng đã được triển khai ứng dụng trong nhiều lĩnh vực khác nhau như:

- Hệ thống call-center phục vụ giao tiếp trả lời thông tin tự động cho khách hàng qua điện thoại. Cụ thể là tổng đài chuyển mạch tự động của Trường ĐH.Khoa học Tự nhiên.
- Bộ giao tiếp thoại của ứng dụng báo nói iNghe và trợ lý du lịch iSago trên iPhone.
- Phương pháp cũng đóng vai trò hạt nhân nhận dạng tiếng nói trong sản phẩm VIS - đoạt giải 2 giải thưởng Nhân tài Đất Việt 2012.

Những lợi ích có thể đạt được của phương pháp

Tận dụng ràng buộc không gian đặc trưng trong nhận dạng tiếng nói, phương pháp cung cấp những lợi ích tiềm tàng như sau:

- Khả năng xây dựng một hệ thống nhận dạng tiếng nói với chi phí thấp, đáp ứng hiệu suất cao.
- Đem lại độ chính xác cao cho nhận dạng tiếng nói tiếng Việt trong điều kiện ở thời điểm hiện tại: điều kiện tài nguyên ngữ liệu ít của tiếng Việt.
- Có khả năng áp dụng được cho các ngôn ngữ nói của người dân tộc thiểu số đang sinh sống tại Việt Nam. Các ngôn ngữ này có rất ít mẫu ngữ liệu, hoặc hoàn toàn không.
- Việc chuyển đổi lĩnh vực ứng dụng của một hệ thống nhận dạng tiếng nói cũng trở nên dễ dàng hơn, tiết kiệm được nhiều thời gian, công sức và kinh phí.

YÊU CẦU BẢO HỘ

1. Phương pháp nhận dạng tiếng nói tiếng Việt sử dụng ràng buộc không gian đặc trưng bao gồm các bước sau:

bước 1: xác định điều kiện tài nguyên huấn luyện: tùy theo lượng dự liệu tiếng nói có được, phân ra 3 ngạch :

ngạch 1: tổng thời lượng ít hơn 20 giờ hoặc tổng số người nói ít hơn 100 người; tiếp tục thực hiện lần lượt các bước 2, 3, 4, và 5;

ngạch 2: không thuộc các trường hợp của ngạch 1 và tổng thời lượng ít hơn 60 giờ hoặc tổng số người nói ít hơn 300 người; tiếp tục thực hiện các bước 2, 4, và 5, bỏ qua bước 3;

ngạch 3: các trường hợp còn lại, phương pháp xây dựng hệ nhận dạng tiếng nói kinh điển được áp dụng;

bước 2: xây dựng mô hình ngữ âm ràng buộc không gian đặc trưng :từ kho dữ liệu tiếng nói, tiến hành huấn luyện mô hình ngữ âm ràng buộc không gian đặc trưng gồm 3 pha chính:

pha 1: xây dựng không gian đặc trưng;

pha 2: ước lượng tương quan giữa các hàm cơ sở trong không gian con;

pha 3: huấn luyện không gian con;

bước 3: tiến hành mở rộng phân hoạch không gian con cho mô hình ngữ âm gồm các bước:

i) xác định số phân hoạch dựa trên điều kiện tài nguyên hiện hữu;

ii) mở rộng cấu trúc mô hình ngữ âm theo số phân hoạch đã định;

iii) huấn luyện phân hoạch không gian con;

bước 4: lựa chọn phương pháp cho bộ trích đặc trưng, mô hình ngôn ngữ, và đồ thị nhận dạng: các thành phần còn lại của hệ nhận dạng có thể được lựa chọn tự do theo kinh điển; có thể dùng:

đặc trưng các hệ số phổ Mel (Mel Frequency Cepstral Coefficient –MFCC), một đặc trưng phổ biến trong nhận dạng tiếng nói;

mô hình N-gram, một dạng mô hình thống kê, cho mô hình ngôn ngữ;

máy chuyển đổi trạng thái hữu hạn (Finite State Transducer – FST) cho đồ thị nhận dạng;

bước 5: cấu thành hệ thống nhận dạng được tiến hành thông thường chỉ khác trong cách tính tích p(x|j):

$$p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \mu_{jmi}, \Sigma_i)$$

2. Phương pháp theo điểm 1, trong đó ở bước 2 xây dựng mô hình ngữ âm ràng buộc không gian đặc trưng gồm 3 pha chính như sau:

pha 1: xây dựng không gian đặc trưng:

a) phân cụm các véc tơ đặc trưng thành I nhóm, mỗi nhóm đại diện bởi một phân bố Gauss;

b) lặp:

với mỗi cụm:

tính toán lại giá trị ràng buộc hồi qui tuyến tính theo hướng cực đại hóa khả xuất (Maximum Likelihood Linear Regression- MLLR) cho từng cụm;

tích lũy các biến ẩn của mô hình hợp Gauss (Gaussian Mixture Model-GMM);

cập nhật lại tham số của mô hình hợp Gauss (Gaussian Mixture Model-GMM);

chuẩn hóa (normalize) lại trọng số của từng phân bố Gauss trong cụm mô hình hợp Gauss (Gaussian Mixture Model-GMM);

trích xuất ma trận phương sai đường chéo từ mô hình hợp (Gaussian Mixture Model-GMM);

c) phân hoạch I phân bố Gauss thành K nhóm cha (mô hình hợp Gauss (Gaussian Mixture Model-GMM));

pha 2: ước lượng tương quan giữa các hàm cơ sở trong không gian con:

Với k = 1…K

Với i = 1…I, tính:

$$\text{logp}^{\text{diag}}(x(t), k, i) = \text{logdet}_k^{(s)} + \text{logw}_{ki} + \text{logN}(x_{ki}(t)|u_{ki}^{\text{diag}}, \sum_{ki}^{\text{diag}})$$

tinh giản (prune) ra các cặp $p^{diag}(k,i)$ có $logp^{diag}(x(t), k, i)$ cao nhất;

với mỗi cặp như vậy, tính:

$$\text{logP}(x(t), k, i) = \text{logdet}_k^{(s)} + \text{logw}_{ki} + \text{logN}(x_{ki}(t)|u_{ki}, \sum_{ki})$$

tinh giản (prune) ra các cặp P(k, i) có logP(x(t), k, i) cao nhất;

pha 3: huấn luyện không gian con

a) khởi tạo không gian đặc trưng với cấu trúc mô hình hợp Gauss (Gaussian Mixture Model-GMM);

b) khởi tạo mô hình Markov ẩn (Hidden Markov Models -HMMs);

c) lặp: chọn từng phân hoạch không gian con để huấn luyện;

với mỗi mẫu huấn luyện:

tính lại giá trị của các biến ẩn (i, j, k) và biến tích lũy (α, β, γ, δ, ζ);

chọn ra các phân bố Gauss có p(k, i) cao nhất như trong pha 2;

xác định ánh xạ không gian con từ các phân bố Gauss chọn được và bộ mô hình hợp Gauss (Gaussian Mixture Model-GMM) của pha 1 (không gian đặc trưng);

cập nhật lại giá trị của các véc tơ không gian con;

gia giảm số lượng trạng thái con theo phương pháp nhân đôi và kỹ thuật trộn tham số (mixing) của mô hình hợp Gauss (Gaussian Mixture Model-GMM);

tùy chọn gia giảm kích thước không gian con cho khớp nhất với kho ngữ liệu.

3. Phương pháp theo điểm 1, trong đó ở bước 3 tiến hành mở rộng phân hoạch không gian con cho mô hình ngữ âm gồm các bước như sau:

bước 1: xác định số phân hoạch dựa trên điều kiện tài nguyên hiện hữu:

i) chọn ra bộ ngữ liệu kiểm định hiệu suất nhận dạng D_T ;

gọi M là số phân hoạch con cần xác định;

ii) lặp:

tăng dần M từ 1, với mỗi giá trị của M thực hiện:

tính $\log P(x(t), k, i) = \log\det_k^{(s)} + \log w_{ki} + \log N(x_{ki}(t)|u_{ki}, \sum_{ki})$ với $x \in D_T$;

tính độ tăng $\Delta = \log_M P - \log_{M-1} P$;

cho đến khi $\Delta < 0$;

iii) chọn M lớn nhất mà tại đó $\Delta >= 0$;

bước 2: mở rộng cấu trúc mô hình ngữ âm theo số phân hoạch đã định:

i) chuyển đổi hàm mật độ xác suất trong mỗi trạng thái thành:

$$p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{jmi}, \boldsymbol{\Sigma}_i)$$

ii) từ 1 phân hoạch của mỗi trạng thái, tiến hành lặp:

sao chép tham số, nhân bản ra *m* phân hoạch con, theo độ tăng 1 phân hoạch mới ở mỗi bước lặp: *m = 1 ... M*;

thực hiện lại Pha 3 (huấn luyện không gian con) của mô hình;

cho đến khi tổng số phân hoạch *m* đạt đến giá trị của *M* đã định;

bước 3: huấn luyện phân hoạch không gian con:

mục tiêu là hiệu chỉnh lại giá trị của bộ tham số c_{jm} và v_{jm} cho phù hợp nhất với dữ liệu học, theo 2 biểu thức:

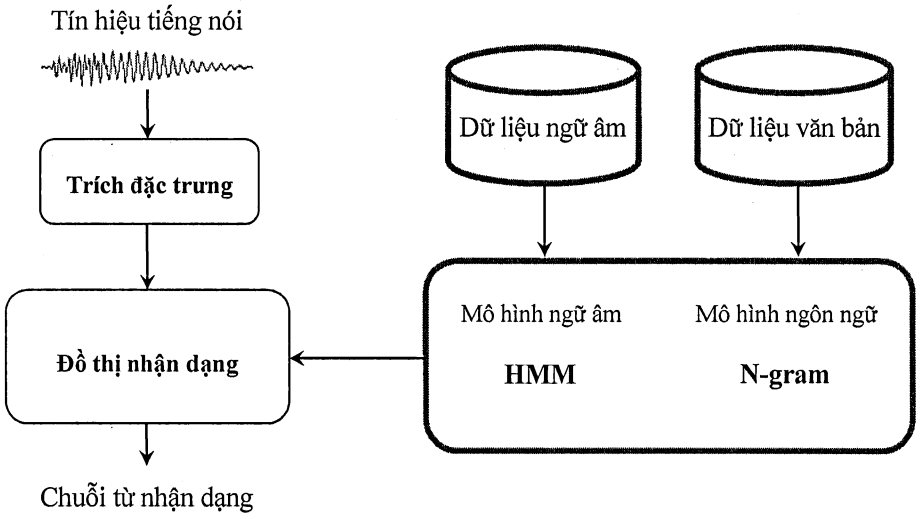
$$\hat{c}_{jkm} = \frac{\gamma_{jkm} + \tau^{(w)}}{\sum_{k=1}^{K} \sum_{m=1}^{M_{jk}} (\gamma_{jkm} + \tau^{(w)})}$$

$$\hat{\mathbf{v}}_{jkm'} = \mathbf{v}_{jkm} - 0.1\mathbf{H}_k^{(\mathrm{sm})^{-0.5}}\mathbf{r}$$

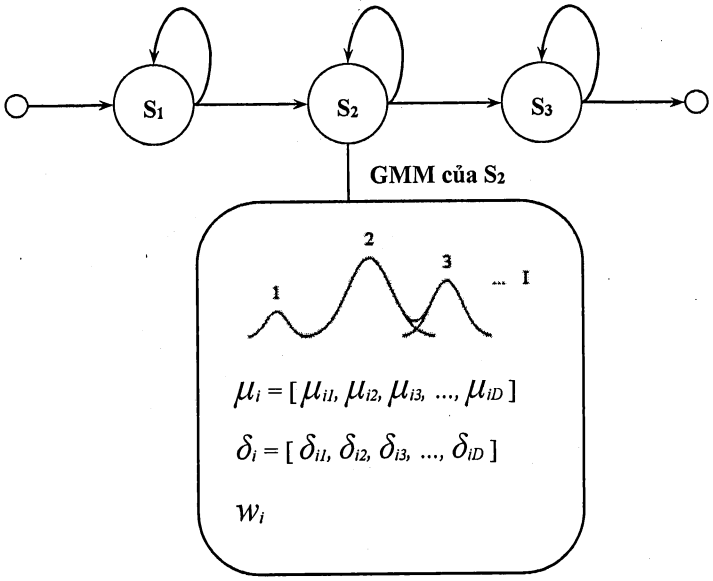
trong đó,

$$\mathbf{H}_k^{(\mathrm{sm})} = \frac{1}{\sum_i \gamma_{ki}} \sum_i \gamma_{ki}$$

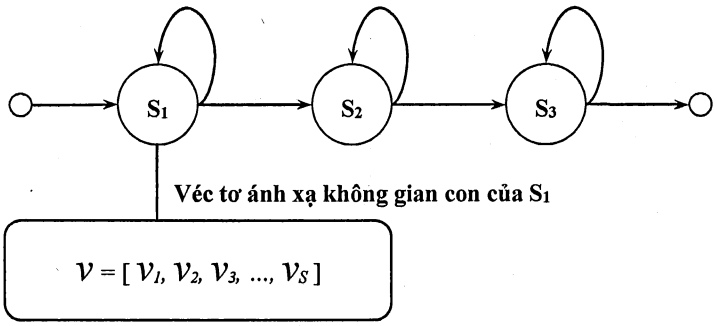
và *r* là giá trị phát sinh ngẫu nhiên theo phân phối chuẩn.


Tín hiệu tiếng nói
Trích đặc trưng
Đồ thị nhận dạng
Chuỗi từ nhận dạng
Dữ liệu ngữ âm
Dữ liệu văn bản
Mô hình ngữ âm
HMM
Mô hình ngôn ngữ
N-gram

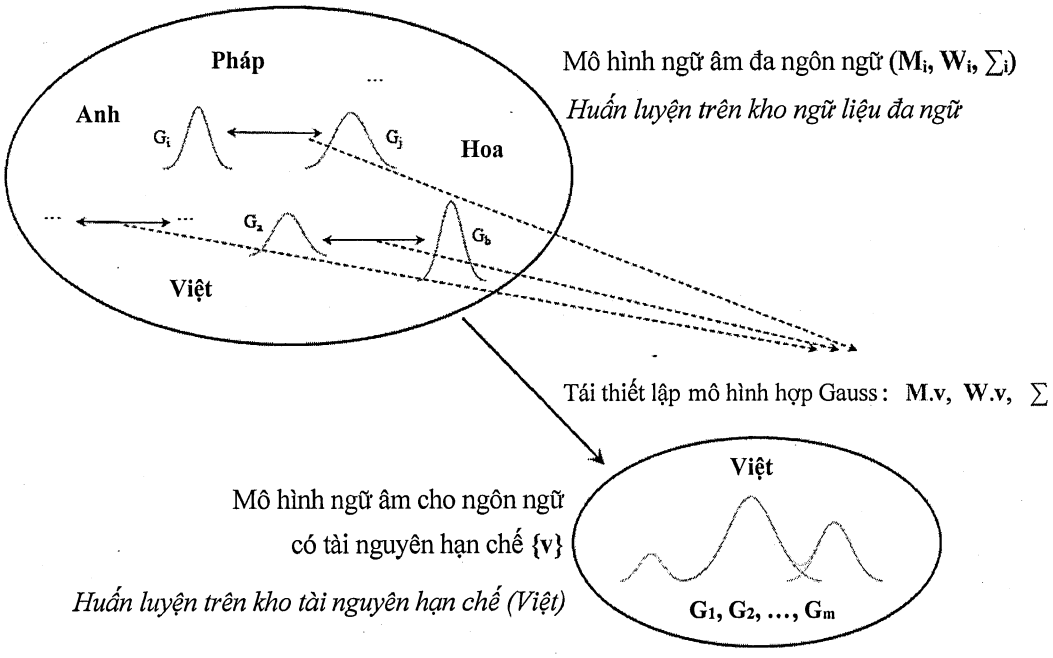
Hình 1


S1
S2
S3
GMM của S2
1
2
3
... I
$\mu_i = [\mu_{i1}, \mu_{i2}, \mu_{i3}, ..., \mu_{iD}]$
$\delta_i = [\delta_{i1}, \delta_{i2}, \delta_{i3}, ..., \delta_{iD}]$
w_i

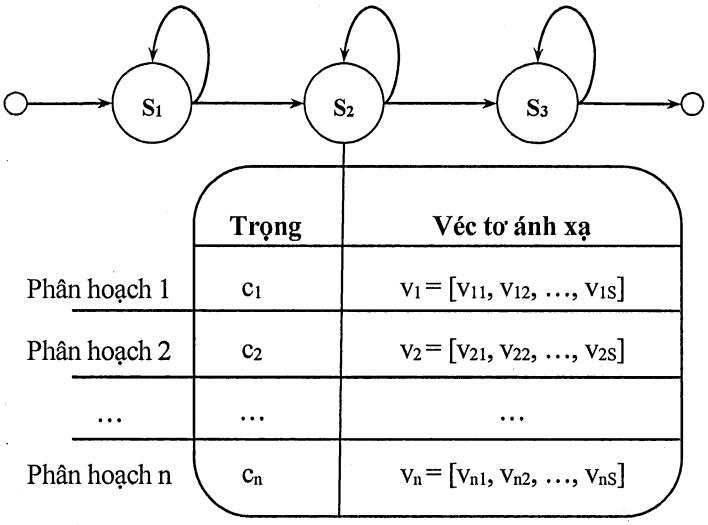
Hình 2



Hình 3


Pháp
Anh
Hoa
Việt
Mô hình ngữ âm đa ngôn ngữ (M_i, W_i, $\sum_i$)
Huấn luyện trên kho ngữ liệu đa ngữ
Tái thiết lập mô hình hợp Gauss : M.v, W.v, ∑
Việt
Mô hình ngữ âm cho ngôn ngữ
có tài nguyên hạn chế {v}
Huấn luyện trên kho tài nguyên hạn chế (Việt)
$G_1, G_2, \ldots, G_m$

Hình 4


S1
S2
S3
Trọng
Véc tơ ánh xạ
Phân hoạch 1
c1
v1 = [v11, v12, …, v1S]
Phân hoạch 2
c2
v2 = [v21, v22, …, v2S]
…
…
…
Phân hoạch n
cn
vn = [vn1, vn2, …, vnS]

Hình 5