Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.0009

CẢI THIỆN HIỆU NĂNG HỆ THỐNG NHẬN DẠNG TIẾNG VIỆT VỚI THÔNG TIN VỀ PHƯƠNG NGỮ

Phạm Ngọc Hưng [1,2], Trịnh Văn Loan[1,2], Nguyễn Hồng Quang[2], Trần Vũ Duy[2]

[1] Khoa Công nghệ thông tin, Trường Đại học Sư phạm Kỹ thuật Hưng Yên

[2] Viện Công nghệ Thông tin và Truyền thông, Trường Đại học Bách khoa Hà Nội

phamngochung@gmail.com, loantv@soict.hust.edu.vn, quangnh@soict.hust.edu.vn, tranvuduy14@gmail.com

TÓM TẮT— *Tiếng Việt là ngôn ngữ có thanh điệu và có nhiều phương ngữ khác nhau. Ảnh hưởng của yếu tố phương ngữ tới các hệ thống nhận dạng tự động tiếng Việt nói là đáng kể. Có nhiều phương pháp khác nhau đã được nghiên cứu và áp dụng cho nhận dạng phương ngữ như GMM (Gaussian Mixture Model), Supervector,... Bài báo trình bày kết quả thử nghiệm nhận dạng phương ngữ tiếng Việt và việc cải thiện hiệu năng của hệ thống nhận dạng tiếng Việt khi có thông tin về phương ngữ. Ngữ liệu dùng cho nhận dạng là phương ngữ của ba giọng Hà Nội, Huế, Thành phố Hồ Chí Minh đại diện cho ba miền Bắc, Trung, Nam. Mô hình GMM đã được dùng để nhận dạng phương ngữ. Kết quả thử nghiệm cho thấy, tỷ lệ nhận dạng đúng phương ngữ tiếng Việt đạt 71% khi sử dụng các tham số MFCC kết hợp với F0 chuẩn hóa theo trung bình của F0, tăng 12% so với trường hợp chỉ sử dụng MFCC. Tỷ lệ nhận dạng tiếng Việt nói sử dụng HMM được nâng cao khi bổ sung thông tin về phương ngữ với lỗi từ là 6,76%, giảm 2,61% so với trường hợp chưa có thông tin phương ngữ.*

Từ khóa— *Nhận dạng phương ngữ, tiếng Việt, GMM, HMM, MFCC, tần số cơ bản, nhận dạng tiếng Việt nói.*

I. GIỚI THIỆU

Hệ thống nhận dạng tự động tiếng nói đã được nghiên cứu từ sớm và đạt được nhiều tiến bộ. Chất lượng nhận dạng đã được nâng cao tuy nhiên hiệu năng nhận dạng, tỷ lệ nhận dạng chưa đúng vẫn còn đáng kể. Có nhiều yếu tố tác động và là nguyên nhân ảnh hưởng đến hiệu năng của hệ thống nhận dạng tự động tiếng nói. Trong số đó có thể kể đến yếu tố về phương ngữ. Với cùng nội dung nhưng phương thức phát âm khác nhau giữa các vùng miền có thể khiến cho hệ thống nhận dạng có kết quả khác nhau. Tiếng Việt là ngôn ngữ có thanh điệu với nhiều phương ngữ khác nhau và đặc biệt phương thức phát âm của các phương ngữ có thể rất khác nhau. Chính vì vậy, các hệ thống nhận dạng tự động tiếng Việt nói cũng chịu ảnh hưởng nhiều bởi yếu tố phương ngữ của tiếng Việt. Nếu biết trước tiếng nói cần nhận dạng thuộc phương ngữ nào thì hệ thống nhận dạng có thể tổ chức cho phù hợp với phương ngữ tương ứng nhằm đạt được kết quả nhận dạng đúng với nội dung thực sự cần nhận dạng. Hay nói khác đi, hiệu năng hệ thống nhận dạng sẽ được cải thiện nếu biết trước phương ngữ của tiếng nói cần nhận dạng.

Để cải thiện hiệu năng của hệ thống nhận dạng tiếng Việt, trước khi nhận dạng nội dung cần tiến hành định danh phương ngữ của tiếng nói cần nhận dạng. Hệ thống định danh phương ngữ được nghiên cứu trong bài báo này dựa trên phương thức phát âm mà không sử dụng các từ địa phương của phương ngữ đó. Điều này cho phép thực hiện linh hoạt hệ thống định danh phương ngữ không phụ thuộc nội dung nói. Sau khi xác định được phương ngữ của tiếng Việt cần nhận dạng, bước tiếp theo là thực hiện nhận dạng nội dung sử dụng mô hình phù hợp với phương ngữ tiếng Việt tương ứng đã được huấn luyện.

Kết quả thử nghiệm cho thấy hiệu năng hệ thống nhận dạng tiếng Việt nói được cải thiện khi biết trước phương ngữ tiếng nói cần nhận dạng.

Phần II của bài báo sẽ trình bày tổng quan về phương ngữ tiếng Việt, ngữ liệu và thử nghiệm nhận dạng phương ngữ tiếng Việt. Phần III trình bày kết quả cải thiện hiệu năng nhận dạng tiếng Việt nói khi có thông tin về phương ngữ. Cuối cùng, phần IV là kết luận.

II. PHƯƠNG NGỮ TIẾNG VIỆT, NGỮ LIỆU VÀ NHẬN DẠNG PHƯƠNG NGỮ TIẾNG VIỆT

A. Phương ngữ và ngữ liệu phương ngữ tiếng Việt

Như đã biết, phương ngữ là sự khác biệt của ngôn ngữ nói giữa các vùng miền ở mỗi quốc gia. Sự khác biệt này thể hiện ở nhiều yếu tố như từ vựng, ngữ pháp và phương thức phát âm. Tiếng Việt là ngôn ngữ có nhiều phương ngữ. Sự phân chia phương ngữ tiếng Việt đã được nhiều nhà nghiên cứu đề cập tới và cũng có nhiều cách phân chia khác nhau. Tuy nhiên, phần lớn các nhà nghiên cứu đều cho rằng phương ngữ tiếng Việt có thể được chia làm ba phương ngữ chính đó là: phương ngữ Bắc tương ứng với khu vực Bắc Bộ, phương ngữ Trung tương ứng với khu vực các tỉnh từ Thanh Hóa đến đèo Hải Vân và phương ngữ Nam tương ứng các tỉnh từ đèo Hải Vân đến các tỉnh thành phía Nam [1]. Sự phân chia này chỉ là tương đối vì các ranh giới địa lý để phân chia các phương ngữ không phải là hoàn toàn rõ ràng. Trong thực tế, ở cùng một khu vực, phương ngữ có thể khác nhau ngay cả giữa các làng, xã với nhau. Đối với ba phương ngữ chính trên, ngoài sự khác biệt đáng kể trong vốn từ vựng, điều khiến cho người nghe dễ dàng nhận biết, phân biệt giữa các phương ngữ đó là phương thức phát âm. Ngữ âm của ba phương ngữ chính có sự khác biệt đáng kể. Đối với hệ thống thanh điệu tiếng Việt, phương ngữ Bắc có đủ sáu thanh bao gồm thanh bằng ("level tone"), thanh huyền ("low-falling tone"), thanh hỏi ("asking tone"), thanh sắc ("rising tone"), thanh ngã ("broken tone") và thanh nặng ("heavy tone"), trong khi phương ngữ Trung chỉ có năm thanh. Đối với giọng các tỉnh Thanh Hóa, Quảng Bình,

Quảng Trị, Thừa Thiên và giọng miền Nam nói chung, không có sự phân biệt giữa thanh hỏi và thanh ngã. Đối với giọng Nghệ An và Hà Tĩnh, thanh ngã và thanh nặng đều giống nhau. Xét về ngôn điệu, ba phương ngữ chính là hoàn toàn khác nhau. Trong nghiên cứu này, sự khác nhau về phương thức phát âm được khai thác để nhận dạng phương ngữ mà không sử dụng đến yếu tố khác biệt của từ địa phương.

Để thực hiện các thử nghiệm, bộ ngữ liệu mới đã được nhóm tác giả đã tiến hành xây dựng và đặt tên là VDSPEC [2]. Bộ ngữ liệu này không chỉ dùng cho nghiên cứu nhận dạng tiếng Việt nói nói chung mà được xây dựng đặc biệt dành cho nghiên cứu nhận dạng phương ngữ tiếng Việt.

Bộ ngữ liệu VDSPEC được ghi âm trực tiếp từ người nói thông qua việc đọc các đoạn văn bản đã được chuẩn bị sẵn. Văn bản này được tổ chức theo các chủ đề khác nhau và cân bằng về thanh điệu (số lượng các từ cho mỗi thanh là xấp xỉ như nhau, khoảng 717 từ). Tiếng nói được ghi âm với tần số lấy mẫu là 16000 Hz, 16 bit cho mỗi mẫu. Độ tuổi của người nói trung bình là 21 tuổi. Ở độ tuổi này, tiếng nói đã ổn định và thể hiện rõ được tiếng địa phương. Mỗi phương ngữ có 50 người nói bao gồm 25 nữ và 25 nam. Giọng Hà Nội được chọn đại diện cho phương ngữ Bắc, Huế cho phương ngữ Trung và giọng Thành phố Hồ Chí Minh đại diện cho phương ngữ Nam. Mỗi chủ đề, người nói đọc 25 câu với mỗi câu có độ dài ghi âm khoảng 10 giây. Tổng thời lượng tiếng nói đã ghi âm của VDSPEC là 45,12 giờ, chiếm dung lượng 4,84 GB bộ nhớ.

B. Nhận dạng phương ngữ tiếng Việt dùng mô hình GMM với MFCC và F0

Mô hình hỗn hợp Gauss đa thể hiện (Gaussian Mixture Model: GMM) đã được sử dụng trong các nghiên cứu về nhận dạng người nói [3], định danh phương ngữ tiếng Anh [4], tiếng Trung [5], tiếng Thái [6], tiếng Hindi [7], tiếng Việt [8], nhận dạng ngôn ngữ [9], [10]. Supervectors cũng được sử dụng trong nghiên cứu nhận dạng phương ngữ và cho kết quả khả quan [11]. Để giải thích lý do tại sao GMM thường được dùng trong nhận dạng người nói, định danh ngôn ngữ và định danh phương ngữ,... có thể suy diễn như sau. Ngay cả trong trường hợp không nghe rõ nội dung câu nói, con người vẫn có khả năng cảm nhận đang nghe giọng người, ngôn ngữ, phương ngữ nào,... mà mình đã biết. Trong trường hợp đó, thông tin tổng quát hay đường bao thông tin về ngữ âm đã giúp con người nhận ra giọng, ngôn ngữ, phương ngữ mà chưa cần dùng đến các thông tin chi tiết khác về nội dung cũng như về ngữ âm mà người nói truyền tải. Bằng cách lấy số các thành phần phân bố Gauss đủ lớn, điều chỉnh trung bình và phương sai của chúng cũng như các trọng số trong tổ hợp tuyến tính, GMM có thể xấp xỉ phần lớn các mật độ phân bố liên tục với độ chính xác tùy chọn. Cũng chính vì vậy, GMM cho phép mô hình hóa chỉ các phân bố cơ bản của cảm nhận về ngữ âm của người nói hay cũng là cảm nhận đường bao thông tin ngữ âm đã nói ở trên. Yếu tố của phép trung bình trong khi xác định mô hình GMM có thể loại đi các nhân tố ảnh hưởng đến đặc trưng âm học như biến thiên ngữ âm theo thời gian của người nói khác nhau và chỉ giữ lại những gì là đặc trưng cơ bản cho giọng vùng, miền như trong trường hợp định danh phương ngữ. Mặt khác, về mặt tính toán, việc sử dụng GMM như là khả hiện sẽ tính toán không tốn kém, dựa trên mô hình thống kê đã được biết rõ.

Mô hình hỗn hợp Gauss đa thể hiện là tổng có trọng số của M thành phần mật độ Gauss như biểu thức (1):

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

Trong (1), $\boldsymbol{X}$ là véctơ dữ liệu (chứa các tham số của đối tượng cần biểu diễn), π_i, i=1,..., M là các trọng số của hỗn hợp và $g_i(\boldsymbol{X}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$ là các hàm mật độ Gauss thành phần theo biểu thức (2) với véctơ trung bình $\boldsymbol{\mu_i}$ của véctơ D chiều và ma trận hiệp phương sai $\boldsymbol{\Sigma_i}$ kích thước DxD.

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

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

Một GMM đầy đủ được tham số hóa bởi véctơ trung bình, ma trận hiệp phương sai và các trọng số hỗn hợp từ tất cả các thành phần Gauss. Các tham số này có thể được biểu diễn gọn lại theo (3)

$$\lambda = \{\boldsymbol{\pi_i}, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\}, i = 1,2, \dots, M \tag{3}$$

Để định danh phương ngữ, mỗi phương ngữ được biểu diễn bằng một GMM và được tham chiếu bởi mô hình λ của phương ngữ đó. Trong trường hợp dùng MFCC như là véctơ đặc trưng, đường bao phổ của lớp âm học thứ i được biểu diễn bằng trung bình $\boldsymbol{\mu_i}$ của thành phần thứ i, còn biến thiên của đường bao phổ trung bình được biểu diễn bằng ma trận hiệp phương sai $\boldsymbol{\Sigma_i}$.

Giả thiết T là số lượng véctơ đặc trưng hay cũng là toàn bộ số lượng khung (frame) tiếng nói, M là số thành phần Gauss:

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_T\} \tag{4}$$

Khả hiện GMM là:

$$p(\boldsymbol{X}|\lambda) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\lambda) \tag{5}$$

Biểu thức (5) là hàm phi tuyến đối với λ nên không thể trực tiếp cực đại hóa mà các tham số khả hiện cực đại có thể nhận được bằng cách dùng thuật giải cực đại hóa kỳ vọng EM (EM - *Expectation-Maximization*).

Ý tưởng của thuật giải EM là bắt đầu với mô hình khởi đầu λ, đánh giá mô hình mới $\bar{\lambda}$ sao cho:

$$p(\boldsymbol{X}|\bar{\lambda}) \geq p(\boldsymbol{X}|\lambda) \tag{6}$$

Mô hình mới lại là mô hình khởi đầu cho bước lặp tiếp theo và quá trình lặp lại cho đến khi ngưỡng hội tụ đạt được.

Trong trường hợp nhận dạng phương ngữ tiếng Việt, véctơ $\boldsymbol{X}$ sẽ chứa các hệ số MFCC và các tham số liên quan đến *F*0. Để tìm ra số tốt nhất các hệ số MFCC dùng để nhận dạng phương ngữ không phân biệt giới tính, số lượng các hệ số MFCC lựa chọn thử nghiệm từ 5 đến 19. Các thí nghiệm được thực hiện đối với từng phương ngữ và lấy giá trị trung bình. Kết quả cho thấy, nếu chọn số hệ số MFCC bằng 13, cả 3 phương ngữ cùng đạt tỷ lệ nhận dạng cao như nhau. Còn nếu chọn số hệ số MFCC bằng 11, tỉ lệ nhận dạng trung bình sẽ cao hơn so với trường hợp số hệ số MFCC bằng 13 song mỗi phương ngữ lại có tỷ lệ nhận dạng khác nhau. Do vậy, số hệ số MFCC bằng 11 và 13 được chọn cho các thử nghiệm nhận dạng phương ngữ.

Trong các thử nghiệm này, bộ tham số MFCC được kết hợp với tần số cơ bản *F*0, *LogF0(t)* và các dạng chuẩn hóa *F*0, *LogF0(t)*. Chuẩn hóa *F*0 và *LogF0(t)* dùng các công thức sau:

- Đạo hàm *F*0 (*diffF0(t)*):

$$diffF0(t) = dF0(t)/dt \tag{9}$$

- Chuẩn hóa *F*0 theo xu hướng đi lên hoặc đi xuống của *F*0 mỗi câu (*cdF0(t)*):

$$cdF0(t) = \begin{cases} -1 & nếu\ ((F0_i - F0_{i-1}) \leq -3) \\ 0 & nếu\ (-3 < (F0_i - F0_{i-1}) < 3) \\ 1 & nếu\ ((F0_i - F0_{i-1}) \geq 3) \end{cases} \tag{10}$$

- Chuẩn hóa *F*0 theo giá trị trung bình *F*0 cho mỗi câu (*F0sbM(t)*):

$$F0sbM(t) = F0(t)/\overline{F0(t)} \tag{11}$$

- Chuẩn hóa *F*0 theo trung bình và độ lệch chuẩn của *F*0 (*F0sbMSD(t)*):

$$F0sbMSD(t) = \frac{F0(t) - \overline{F0(t)}}{\sigma F0(t)} \tag{12}$$

- Đạo hàm *LogF0(t)* (*diffLogF0(t)*):

$$diffLogF0(t) = dLogF0(t)/dt \tag{13}$$

- Chuẩn hóa *LogF*0*(t)* theo giá trị min *LogF0(t)* và max *LogF0(t)* cho mỗi câu (*LogF0sbMM(t)*):

$$LogF0sbMM(t) = \frac{LogF0(t) - \min LogF0(t)}{\max\ LogF0(t) - \min\ LogF0(t)} \tag{14}$$

- Chuẩn hóa *LogF*0*(t)* theo trung bình *LogF*0*(t)* mỗi câu (*LogF0sbM(t)*):

$$LogF0sbM(t) = logF0(t)/\overline{logF0(t)} \tag{15}$$

- Chuẩn hóa theo *LogF*0*(t)* theo trung bình và độ lệch chuẩn của *LogF*0*(t)* (*LogF0sbMSD(t)*):

$$LogF0sbMSD(t) = \frac{logF0(t) - \overline{logF0(t)}}{\sigma logF0(t)} \tag{16}$$

Praat [12] được sử dụng để xác định tần số cơ bản *F*0 của tiếng nói trong ngữ liệu VDSPEC. Kết quả cho thấy, với 11 hệ số MFCC, tỉ lệ nhận dạng cao nhất là 70% đối với hai trường hợp: MFCC được kết hợp với *F0sbM(t)* và MFCC được kết hợp với *LogF0sbM(t)*. Nếu số hệ số MFCC = 13, tỷ lệ nhận dạng đạt cao nhất là 71% đối với trường hợp MFCC được kết hợp với *F0sbM(t)*. Điều này cũng phù hợp với các trường hợp MFCC = 11. Với sự kết hợp của MFCC và *F0*, tỷ lệ nhận được cải thiện đáng kể (tăng 12%) so với trường hợp không có thông tin *F*0.

Các ma trận nhầm lẫn trong nhận dạng phương ngữ không phân biệt giới tính với sự kết hợp của MFCC và tham số F0 được trình bày ở Bảng 1. Nhìn chung, Bảng 1 cho thấy phương ngữ Trung có xu hướng nhận dạng thành phương ngữ Bắc nhiều hơn và phương ngữ Nam có xu hướng nhầm sang phương ngữ Trung hơn. Điều này phù hợp với thực tế là phương ngữ Bắc và phương ngữ Trung có nhiều điểm tương đồng và phương thức phát âm là gần như giống nhau ở hầu hết các thanh điệu. Khoảng cách địa lý càng xa thì mức độ khác biệt giữa các phương ngữ càng lớn.

Bảng 1. Ma trận nhầm lẫn nhận dạng phương ngữ không phụ thuộc giới tính với sự kết hợp sử dụng hệ số MFCC và tham số F0; a) MFCC=11, b) MFCC=13

	PNB	**PNT**	**PNN**	**Tỷ lệ nhận dạng đúng**
PNB	824	220	206	66%
PNT	178	932	140	75%
PNN	140	258	852	68%

a)

	PNB	**PNT**	**PNN**	**Tỷ lệ nhận dạng đúng**
PNB	826	226	198	66%
PNT	152	965	133	77%
SD	158	229	863	69%

b)

Trong thực tế, dựa trên sự khác biệt của phương thức phát âm đặc biệt là đối với biến thiên của *F*0, người ta có thể phân biệt ba phương ngữ chính của tiếng Việt là Bắc, Trung và Nam. Vì vậy, bằng sự kết hợp của MFCC và tham số *F*0 trong mô hình GMM, tỉ lệ nhận dạng của các phương ngữ tiếng Việt được cải thiện đáng kể. Các thử nghiệm cho thấy điểm số tốt nhất để có được các mô hình GMM thích hợp dùng cho nhận dạng phương ngữ khi số lượng các hệ số MFCC chọn bằng 13.

Phần tiếp theo của bài báo trình bày ứng dụng kết quả nhận dạng phương ngữ vào hệ thống nhận dạng tiếng Việt nói giúp cải thiện hiệu năng nhận dạng.

III. CẢI THIỆN HIỆU NĂNG NHẬN DẠNG TIẾNG VIỆT NÓI KHI CÓ THÔNG TIN PHƯƠNG NGỮ

A. Hệ thống nhận dạng tự động tiếng nói

Nhận dạng tiếng nói là quá trình tìm ra chuỗi các từ trong dữ liệu tiếng nói dưới dạng sóng. Giả sử, tín hiệu đầu vào được tham số hóa thành các véctơ âm học "a". Trong các hệ thống nhận dạng tự động tiếng nói, nhận dạng mẫu được dùng làm phương tiện giải mã. Bộ giải mã tìm kiếm chuỗi các từ "w" có nhiều khả năng tương ứng với các đặc tính âm học này.

$$w = \arg\max_{w} P(w|a) = \arg\max_{w} P(a|w)\,P(w) \tag{17}$$

Xác suất của các đặc trưng âm học $P(a)$ được loại bỏ khỏi phương trình bởi vì nó không có liên quan đến việc tìm kiếm chuỗi từ tốt nhất "w". Xác suất có điều kiện $P(a|w)$ của véctơ âm "a" cho bởi chuỗi các từ "w" được xác định bởi một mô hình âm học còn các xác suất $P(w)$ của chuỗi được tính toán bằng mô hình ngôn ngữ.

B. Bộ công cụ nhận dạng Kaldi

Kaldi là bộ công cụ nhận dạng tiếng nói mã nguồn mở [13]. Như đã mô tả ở trên, mô hình âm học và mô hình ngôn ngữ là những thành phần quan trọng của hệ thống nhận dạng tiếng nói. Sau đây mô tả các thành phần này trong Kaldi.

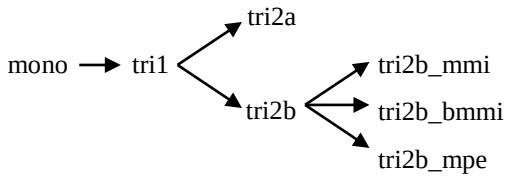
1. Mô hình âm học

HMMs (Hidden Markov Models) được sử dụng để huấn luyện mô hình âm học. Các thông số của HMMs được ước lượng bằng huấn luyện Viterbi. Các HMM có thể biểu diễn cho âm đơn (monophone) và âm ba (triphone). Hình 1 mô tả quá trình huấn luyện mô hình âm học AM (Acoustic Model). Mô hình AM đầu tiên được huấn luyện với monophone (mono) sử dụng các đặc trưng MFCC và Delta-Deltas. Sau đó, huấn luyện bằng các triphone "tri1". Mô hình "tri2a" được tạo ra sau khi huấn luyện lại triphone.

Mặt khác, mô hình "tri2b" được huấn luyện bằng cách sử dụng biến đổi tuyến tính LDA + MLLT. Từ mô hình này, hệ thống tiếp tục huấn luyện dựa trên các đặc trưng LDA + MLLT bằng các phương pháp huấn luyện phân biệt. Các phương pháp đó là:

- Thông tin tương hỗ cực đại (MMI - Maximum Mutual Information): tối ưu hóa đúng đắn của một mô hình bằng cách xây dựng một hàm mục tiêu có xu hướng tối đa hóa xác suất kết hợp $P(a, q)$ và thông tin tương hỗ [14].
- Thông tin tương hỗ cực đại tăng cường (BMMI - Boosted Maximum Mutual Information): sử dụng biến thể của hàm giá MMI với hệ số tăng cường để làm tăng các mô hình có thể trộn được [15].
- Lỗi âm cực tiểu hóa (MPE - Minimum Phone Error): nhằm cực tiểu hóa lỗi âm có thể có [16].

Các phương pháp huấn luyện nêu trên cũng được mô tả trên Hình 1. Ngoài ra còn bổ sung phương pháp thích nghi người nói ký hiệu là "tri3b".



Tên phương pháp huấn luyện	Viết tắt
Monophone	Mono
Triphone	tri1
Δ + ΔΔ	tri2a
LDA+MLLT	tri2b
LDA+MLLT+NMI	tri2b_mmi
LDA+MLLT+bMMI	tri2b_bmmi
MPE	tri2b_mpe

Hình 1. Phương pháp huấn luyện của Kaldi [17]

2. Mô hình ngôn ngữ

Mô hình ngôn ngữ tính toán xác suất của chuỗi từ theo công thức:

$$P(W) = P(w_1, w_2, w_3 \dots w_k) = \prod_{i=1}^{k} P(w_i | w_{i-1} w_{i-2} \dots w_1) \quad (18)$$

Kaldi cung cấp công cụ cho phép tạo mô hình ngôn ngữ theo định dạng ARPA (Advanced Research Projects Agency) từ ngữ liệu văn bản và cả công cụ cho phép chuyển đổi từ định dạng ARPA sang định dạng máy chuyển trạng thái hữu hạn (FST - Finite-state-transducer).

3. Giải mã

Bộ giải mã của hệ thống nhận dạng tự động tiếng nói tìm chuỗi từ giống nhất với chuỗi từ được cho thông qua véctơ đặc trưng. Thuật toán tìm kiếm Viterbi được sử dụng để tìm ra chuỗi như vậy [18].

Lưới từ (Word lattice) là kiểu đầu ra của nhận dạng, chỉ ra các phần chung với các giả thiết khác. Bởi vì các giả thiết được gắn với một xác suất rất nhỏ, các xác suất được tính bằng phép toán lôgarit. Vì thế, dữ liệu đầu ra chứa thông tin về chất lượng của mỗi giả thiết.

C. Thử nghiệm nhận dạng sử dụng bộ công cụ Kaldi

1. Xây dựng mô hình ngôn ngữ

Từ vựng: từ điển phát âm bao gồm 1072 từ đơn được xây dựng bằng 2 phương pháp:

- Phương pháp 1: các âm tiết của một từ đơn không chứa thông tin thanh điệu (có 47 âm vị).
- Phương pháp 2: bao gồm các âm tiết của từ đơn và thông tin thanh điệu trên nguyên âm chính (có 126 âm vị).

Dữ liệu văn bản được dùng để tạo mô hình ngôn ngữ thống kê. Dữ liệu này bao gồm 4 triệu câu với 90 triệu âm tiết được thu thập từ các tài liệu điện tử tiếng Việt. Các ký tự được chuyển đổi theo định dạng Bach Khoa Text Code (BKTC) [19]. Độ phức tạp của mô hình ngôn ngữ bigram là 108,57 và mô hình trigram là 62,43.

Bộ công cụ SRILM [20] được sử dụng để tạo mô hình ngôn ngữ theo định dạng ARPA. Mô hình ngôn ngữ bigram chứa 8.925 unigrams và 3.742.980 bigrams. Mô hình trigram bao gồm nội dung như như mô hình bigram và 11.593.319 trigram. Các file này sau đó được dùng để tạo ra mô hình ngôn ngữ theo định dạng file FST.

2. Kết quả thử nghiệm

Ngữ liệu tiếng nói VDSPEC được sử dụng cho thử nghiệm. Bộ ngữ liệu được chia thành 5 tập trong đó 4 tập dùng huấn luyện và 1 tập dùng cho thử nghiệm như trình bày ở Bảng 2.

Bảng 2. Phân chia tập dữ liệu dùng cho huấn luyện và thử nghiệm

STT	Tên tập dữ liệu	Số giọng nam		Số giọng nữ	
		Huấn luyện	Thử nghiệm	Huấn luyện	Thử nghiệm
1	Phương ngữ Bắc	20	5	20	5
2	Phương ngữ Trung	20	5	20	5
3	Phương ngữ Nam	20	5	20	5
4	Chung cả 3 phương ngữ	60	15	60	15

Thử nghiệm nhận dạng được tiến hành cho hai trường hợp: không có và có thông tin phương ngữ.

Đối với trường hợp thử nghiệm nhận dạng không có thông tin phương ngữ, dữ liệu huấn luyện là tập dữ liệu chung cả 3 phương ngữ tương ứng dòng 4 của Bảng 2. Kết quả thử nghiệm được cho ở bảng 3.

Bảng 3. Kết quả nhận dạng khi chưa biết thông tin phương ngữ

Phương pháp huấn luyện	WER %
mono	39,77
tri1	16,78
tri2a	16,48
tri2b	13,57
tri2b_mmi	11,00
tri2b_bmmi	10,81
tri2b_mpe	10,48
tri3b	9,37

Từ Bảng 3 có thể thấy, kết quả nhận dạng tốt nhất ứng với phương pháp huấn luyện tri3b cho tỷ lệ lỗi từ là 9,37%.

Khi có thông tin về phương ngữ, dữ liệu huấn luyện và thử nghiệm lần lượt là là các tập dữ liệu xây dựng cho từng phương ngữ như đã mô tả trên Bảng 2. Kết quả thử nghiệm được cho trên Bảng 4.

Bảng 4. Kết quả nhận dạng khi đã biết thông tin phương ngữ

Phương pháp huấn luyện	WER %			
	Bắc	Trung	Nam	Trung bình
mono	25,02	21,84	36,20	27,69
tri1	9,26	8,88	18,05	12,06
tri2a	8,95	8,58	18,06	11,86
tri2b	6,99	7,06	14,40	9,48
tri2b_mmi	6,34	6,60	13,94	8,96
tri2b_bmmi	6,21	6,48	13,74	8,81
tri2b_mpe	5,87	6,21	13,06	8,38
tri3b	5,02	5,21	10,05	6,76

Bảng 4 cho thấy kết quả nhận dạng tốt nhất với tỷ lệ lỗi từ trung bình là 6,76% cho phương pháp huấn luyện tri3b cũng là phương pháp cho kết quả nhận dạng tốt nhất khi không biết thông tin phương ngữ.

Kết quả của Bảng 3 và Bảng 4 cho thấy khi có thông tin phương ngữ tỷ lệ nhận dạng chính xác tăng lên 2,61%.

IV. KẾT LUẬN

Bài báo đã trình bày kết quả nhận dạng cho ba phương ngữ chính Bắc, Trung, Nam của tiếng Việt và kết quả thử nghiệm nhận dạng tiếng Việt khi không có và có thông tin về phương ngữ. Việc bổ sung thông tin F0 giúp tỷ lệ nhận dạng đúng phương ngữ tiếng Việt tăng 12% so với trường hợp chỉ dùng MFCC. Thực nghiệm cho thấy hiệu năng của hệ thống nhận dạng tiếng Việt được cải thiện đáng kể khi có thông tin phương ngữ. Tỷ lệ lỗi từ trong trường hợp nhận dạng khi đã có thông tin phương ngữ giảm xuống 2,61% so với trường hợp nhận dạng khi chưa có thông tin phương ngữ. Tiếng Việt là ngôn ngữ có phương ngữ rất đa dạng. Vì vậy, muốn có hệ thống hoàn thiện nhận dạng tự động tiếng Việt nói cần có hệ thống tiền xử lý để định danh phương ngữ. Hệ thống tiền xử lý để định danh phương ngữ này chỉ dựa trên phương thức phát âm đặc trưng cho phương ngữ mà không sử dụng đến các từ địa phương. Điều này cho phép định danh linh hoạt, không phụ thuộc nội dung nói. Trong khuôn khổ thời gian nghiên cứu còn hạn chế, việc định danh tự động phương ngữ tiếng Việt mới tập trung vào ba phương ngữ đại diện bao gồm các giọng Hà Nội, Huế và Thành phố Hồ Chí Minh. Hướng nghiên cứu tiếp theo sẽ là mở rộng nghiên cứu các vùng phương ngữ khác của tiếng Việt. Điều này sẽ góp phần xây dựng hệ thống nhận dạng tự động tiếng Việt ngày càng hoàn thiện.

TÀI LIỆU THAM KHẢO

[1] Hoàng Thị Châu. Phương ngữ học tiếng Việt. NXB Đại học Quốc gia Hà Nội, 2009.

[2] Pham Ngoc Hung, Trinh Van Loan, Nguyen Hong Quang, "Building of corpus for Vietnamese dialect identification", Journal of Science and Technology Technical Universities, No.109-2015. ISSN 2354-1083, pp.49-55.

[3] Jean-François Bonastre, Frédéric Wils, "ALIZE, A FREE TOOLKIT FOR SPEAKER RECOGNITION", IEEE International Conference , pp. I 737 - I 740, 2005.

[4] Torres-Carrasquillo, P. A., Gleason, T. P., and Reynolds, D. A., "Dialect Identification Using Gaussian Mixture Models", In Proc. Odyssey: The Speaker and Language Recognition Workshop in Toledo, Spain, ISCA, pp. 297-300, 31 May - 3 June 2004

[5] Bin MA, Donglai ZHU and Rong TONG, "Chinese Dialect Identification Using Tone Features Based On Pitch", ICASSP 2006.

[6] Sittichok Aunkaew, Montri Karnjanadecha, Chai Wutiwiwatchai, "Development of a Corpus for Southern Thai Dialect Speech Recognition: Design and Text Preparation", The 10th International Symposium on Natural Language Processing, October 28-30, 2013, Phuket, Thailand .

[7] Shweta Sinha, Aruna Jain, S. S. Agrawal, "Acoustic-Phonetic Feature Based Dialect Identification in Hindi Speech", International Journal on Smart Sensing and Intelligent Systems Vol. 8, No. 1, March 2015, pp 235-254.

[8] Phạm Ngọc Hưng, Trịnh Văn Loan, Nguyễn Hồng Quang, Phạm Quốc Hùng , "Nhận dạng phương ngữ tiếng Việt sử dụng mô hình Gauss hỗn hợp", Kỷ yếu Hội nghị Khoa học Công nghệ Quốc gia lần thứ 6 FAIR, 20-21 tháng 6, 2014, ISBN 978-604-913-165-3, pp 449-452

[9] Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., and Deller Jr., J. R., "Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features". In Proc. International Conference on Spoken Language Processing in Denver, CO, ISCA, pp. 33-36, 82-92 September 2002.

[10] Campbell, W. M., Singer, E., Torres-Carrasquillo, P. A., and Reynolds, D. A., "Language Recognition with Support Vector Machines". In Proc. Odyssey: The Speaker and Language Recognition Workshop in Toledo, Spain, ISCA, pp. 41-44, 31 May - 3 June 2004.

[11] Fadi Biadsy, Julia Hirschberg, Daniel P. W. Ellis, "Dialect and Accent Recognition using Phonetic-Segmentation Supervectors", Interspeech 2011.

[12] www.praat.org.

[13] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Silovsky, J. . The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding (No. EPFL-CONF-192584). IEEE Signal Processing Society.

[14] George Saon and Daniel Povey, "Penalty Function Maximization for Large Margin HMM Training", Interspeech 2008.

[15] Daniel Povey, Dimitri Kanevsky, Brian Kingsbury, Bhuvana Ramabhadran, George Saon & Karthik Visweswariah, "Boosted MMI for Model and Feature Space Discriminative Training", ICASSP 2008.

[16] Daniel Povey, "Minimum Phone Error - Better than MMI," talk given at IBM, 2003

[17] Ondřej Plátek, Speech recognition using KALDI, MASTER THESIS, Charles University in Prague Faculty of Mathematics and Physics, 2014

[18] Viterbi AJ , "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". IEEE Transactions on Information Theory 13 (2): 260–269, April 1967.

[19] Nguyen Quoc Cuong, Pham Thi Ngoc and Castelli, E. "Shape vector characterization of Vietnamese tones and application to automatic recognition". Automatic Speech Recognition and Understanding – ASRU. Italy: IEEE, 2001. 437-440

[20] Stolcke, A., Zheng, J., Wang, W., & Abrash, V. . SRILM at sixteen: Update and outlook. In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (p. 5), December 2011.

IMPROVEMENT OF VIETNAMESE RECOGNITION SYSTEM PERFORMANCE USING INFORMATION ABOUT DIALECTS

Pham Ngoc Hung, Trinh Van Loan, Nguyen Hong Quang, Tran Vu Duy

ABSTRACT— *Vietnamese is a tonal language with many different dialects. The influence of dialectal features on Vietnamese speech recognition systems is significant. There are many different methods which have been studied and applied for dialect recognition such as GMM , Supervector ... This paper presents the experimental results of Vietnamese dialect identification and the improving of the performance of the Vietnamese recognition system using information about Vietnamese dialects. The corpus used for identification contain the voices of Hanoi, Hue, and Ho Chi Minh City considered as the representable voices for Northern, Central, and Southern dialects. GMM model has been used for dialect identification. Test results showed that Vietnamese dialect recognition rate is 71% in the case using MFCC combined with normalized F0 according to average F0 for each sentence and this rate increases 12% in comparison with the case using MFCC only. The performance of Vietnamese speech recognition using HMM is considerably improved with the additional dialectal information, word error rate is 6.76% and this rate decreases 2.61% in comparison with the case without dialect information.*