Kỷ yếu Hội nghị Quốc gia lần thứ VIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 9-10/7/2015
DOI: 10.15625/vap.2015.000190

NHẬN DẠNG PHƯƠNG NGỮ TIẾNG VIỆT SỬ DỤNG MFCC VÀ TẦN SỐ CƠ BẢN

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

pnhung@utehy.edu.vn, loantv@soict.hust.edu.vn, quangnh@soict.hust.edu.vn

TÓM TẮT - *Nhận dạng phương ngữ đã được nghiên cứu cho nhiều ngôn ngữ trên thế giới tuy nhiên với phương ngữ tiếng Việt, nghiên cứu theo phương diện xử lý tín hiệu đến nay vẫn còn hạn chế, chưa có nhiều công trình được công bố. Tiếng Việt là ngôn ngữ có nhiều phương ngữ khác nhau. Ảnh hưởng của yếu tố phương ngữ tới các hệ thống nhận dạng tự động tiếng nói là rất đáng kể. Nếu biết trước tiếng nói cần nhận dạng thuộc phương ngữ nào thì việc nhận dạng nội dung sẽ thuận lợi hơn do ngữ liệu được tổ chức phù hợp cho từng phương ngữ. Bài báo này sẽ trình bày phương pháp sử dụng đặc trưng MFCC kết hợp với khai thác thông tin về tần số cơ bản (F0) của tiếng Việt để thực hiện nhận dạng phương ngữ tiếng Việt dựa trên mô hình GMM. Kết quả thử nghiệm trên bộ ngữ liệu phương ngữ tiếng Việt cho thấy việc kết hợp các tham số F0 và MFCC so với chỉ dùng MFCC đã tăng tỷ lệ nhận dạng đúng phương ngữ từ 64.2% lên 70.3%.*

Từ khóa - *Tần số cơ bản, MFCC, GMM, nhận dạng phương ngữ tiếng Việt.*

I. GIỚI THIỆU

Tiếng Việt là ngôn ngữ có thanh điệu và có nhiều phương ngữ khác nhau. Chính sự đa dạng của các phương ngữ đã tạo nên thách thức đối với các hệ thống nhận dạng tự động tiếng Việt. Chỉ xét về phương diện phát âm, cùng một từ nhưng ở các địa phương khác nhau có thể lại được phát âm theo cách khác nhau. Với hai phương ngữ khác nhau, có những âm nghe như nhau nhưng nội dung lại được hiểu khác nhau theo từng phương ngữ. Chỉ riêng yếu tố này cũng đã có thể gây ra nhầm lẫn, ảnh hưởng đáng kể đến kết quả nhận dạng của các hệ thống nhận dạng tiếng Việt nói. Nếu biết trước nội dung tiếng nói cần nhận dạng được phát âm theo cách nói của vùng miền nào đó, hay nói cách khác, nếu biết tiếng nói đó thuộc phương ngữ nào thì có thể giúp hệ thống nhận dạng giới hạn phạm vi, sử dụng bộ ngữ liệu phù hợp cho tiếng nói cần được nhận dạng, từ đó tăng hiệu quả nhận dạng.

Để xác định tiếng nói thuộc phương ngữ nào, trên thế giới cũng đã có nhiều nghiên cứu và thử nghiệm thành công trên một số ngôn ngữ như tiếng Anh, tiếng Trung, tiếng Nhật,... Nghiên cứu về phương ngữ tiếng Việt cũng đã được thực hiện từ lâu nhưng chủ yếu về phương diện ngôn ngữ; còn về phương diện xử lý tín hiệu còn rất hạn chế. Hầu như chưa có công trình nào được công bố về nghiên cứu nhận dạng phương ngữ tiếng Việt theo phương diện xử lý tín hiệu. Do vậy các nghiên cứu, giải pháp đề xuất cho nhận dạng phương ngữ tiếng Việt là cần thiết và đóng góp đáng kể nhằm nâng cao hiệu quả nhận dạng tiếng Việt nói.

Bài báo này đề cập tới phương pháp nhận dạng phương ngữ tiếng Việt sử dụng MFCC và đặc trưng thanh điệu thông qua tham số F0 (tần số cơ bản). Mô hình nhận dạng được triển khai dựa trên mô hình GMM (Gaussian Mixture Model). Các thử nghiệm đã được tiến hành trên bộ ngữ liệu tiếng nói xây dựng công phu cho các nghiên cứu nhận dạng phương ngữ VDSPEC (Vietnamese Dialect Speech Corpus). VDSPEC thực hiện ghi âm trực tiếp từ 100 người nói với tổng thời lượng lên đến 33.79 giờ tiếng nói. Kết quả thử nghiệm cho thấy phương pháp nhận dạng phương ngữ sử dụng MFCC có bổ sung tham số F0 đã làm tăng tỷ lệ nhận dạng phương ngữ tiếng Việt.

Phần II của bài báo giới thiệu tổng quan về phương ngữ tiếng Việt. Phần III trình bày mô hình GMM và các tham số MFCC, tần số cơ bản (F0) được đưa vào mô hình. Các thử nghiệm và kết quả nhận dạng được trình bày ở phần IV. Cuối cùng, phần V là kết luận và hướng phát triển.

II. TỔNG QUAN VỀ PHƯƠNG NGỮ TIẾNG VIỆT

Theo [1]: "Phương ngữ là một thuật ngữ ngôn ngữ học để chỉ sự biểu hiện của ngôn ngữ toàn dân ở một địa phương cụ thể với những nét khác biệt của nó so với ngôn ngữ toàn dân hay với một phương ngữ khác". Tiếng Việt là ngôn ngữ có nhiều phương ngữ. Sự khác biệt giữa các phương ngữ thể hiện trên nhiều yếu tố khác nhau như ngữ âm, ngữ pháp, từ vựng.

Việc phân chia các vùng phương ngữ tiếng Việt đã được các nhà nghiên cứu đề cập đến với nhiều ý kiến khác nhau. Mặc dù chưa có ý kiến thống nhất về cách phân chia song về cơ bản, chiếm số đông các nhà nghiên cứu cho rằng có thể chia phương ngữ tiếng Việt thành 3 vùng chính là phương ngữ Bắc (các tỉnh ở Bắc Bộ), phương ngữ Trung (các tỉnh từ Thanh Hóa vào đến khu vực đèo Hải Vân) và phương ngữ Nam (từ khu vực đèo Hải Vân vào các tỉnh phía Nam) [1]. Việc phân chia các vùng phương ngữ cũng mang tính chất tương đối, không tách biệt hoàn toàn. Giữa các vùng có sự chuyển tiếp. Đôi khi trong một địa phương, một phạm vi địa lý hẹp như giữa các làng, các xã cũng có sự khác biệt rất lớn về phương ngữ.

Khi xem xét những đặc điểm chung nhất của 3 vùng phương ngữ chính (như cách phân chia nêu trên), ngoài sự khác biệt đáng kể về từ vựng thì điều khiến người nghe dễ dàng cảm nhận, phân biệt giữa các phương ngữ đó chính là ngữ âm.

Ngữ âm của ba phương ngữ chính có sự khác biệt đáng kể. Trước hết, khi xem xét về hệ thống thanh điệu. Phương ngữ Bắc có đủ 6 thanh điệu (huyền, sắc, nặng, hỏi, ngã và thanh ngang). Các thanh điệu đối lập từng đôi về âm vực và âm điệu.

Trong khi đó phương ngữ Trung, hệ thống thanh điệu chỉ có 5 thanh điệu. Có khu vực thanh hỏi và thanh ngã không phân biệt (như Thanh Hóa). Có vùng thanh ngã và thanh nặng lại trùng nhau như Nghệ An, Hà Tĩnh. Trong khi đó khu vực Bình-Trị-Thiên không phân biệt thanh ngã và thanh hỏi. Phương ngữ Nam cũng chỉ có 5 thanh điệu. Thanh ngã và thanh hỏi trùng nhau. Xét về mặt điệu tính, hệ thống thanh điệu phương ngữ Nam khác với hệ thống thanh điệu phương ngữ Bắc và phương ngữ Trung [1].

Để phân biệt được các phương ngữ có thể dựa trên một hoặc nhiều yếu tố khác biệt giữa các phương ngữ. Trong phạm vi nghiên cứu của bài báo này, khác biệt về mặt ngữ âm giữa các phương ngữ được tập trung khai thác và làm cơ sở cho nhận dạng phương ngữ.

III. MÔ HÌNH GMM VỚI CÁC THAM SỐ MFCC VÀ F0

Mô hình hỗn hợp Gauss đa biến vào (Gaussian Mixture Model: GMM) đã được sử dụng trong các nghiên cứu về nhận dạng người nói [7], định danh phương ngữ tiếng Anh [3], tiếng Trung [5], nhận dạng ngôn ngữ [2][6]. Supervectors cũng được sử dụng trong nghiên cứu nhận dạng phương ngữ và cho kết quả khả quan [4]. 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 hàm tương đồng sẽ tính toán không tốn kém, dựa trên mô hình thống kê đã được biết rõ.

Một mô hình hỗn hợp Gauss đa biến vào là tổng có trọng số của M thành phần mật độ Gauss như biểu thức (1):

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

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 D×D.

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

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)

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

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:

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_T\} \quad (4)$$

Tương đồng GMM là :

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

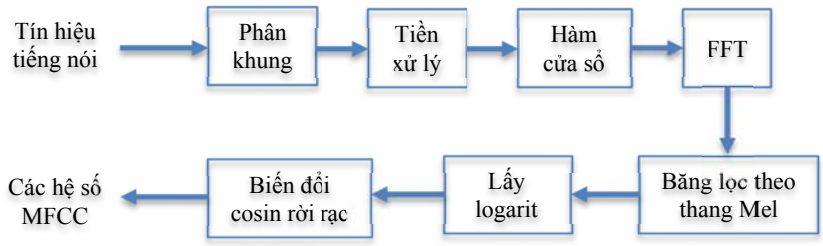
Biểu thức (5) là hàm phi tuyến đối với λ nên không thể trực tiếp cực đại hóa mà các tham số tương đồng cực đại có thể nhận được bằng cách dùng thuật giải cực đại hóa kỳ vọng EM (EM: *expectation-maximization*).

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

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

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 nghiên cứu được công bố tại [8], mô hình GMM được sử dụng chỉ với bộ tham số MFCC (Mel Frequency Cepstral Coefficients: các hệ số phổ theo thang tần số mel). Giải thuật thực hiện tính bộ tham số này được mô tả ở hình 1.



Hình 1. Sơ đồ khối giải thuật tính bộ tham số MFCC

Trong hình 1, đầu tiên tín hiệu tiếng nói sẽ được chia thành các khung với độ dài từng khung là 0,1 giây và độ dịch của khung là 0,01 giây. Sau đó mỗi khung tín hiệu tiếng nói sẽ được thực hiện tiền xử lý theo (7).

$$y(n) = x(n) - 0{,}96x(n-1) \quad (7)$$

Tín hiệu sau khi đã thực hiện tiền xử lý sẽ được nhân với hàm cửa sổ Hamming biểu diễn theo (8) trong đó N là số mẫu của một khung tín hiệu tiếng nói.

$$w(n) = 0{,}54 - 0{,}46\cos(2\pi n/(N-1)) \text{ với } 0 \le n < N \quad (8)$$

Tiếp theo thực hiện tính phép biến đổi Fourier nhanh (FFT: Fast Fourier Transform) trên tín hiệu đã cho qua hàm cửa sổ. Phổ tín hiệu thu được sẽ cho qua băng lọc theo thang Mel. Số bộ lọc chính là số các hệ số MFCC cần tính. Cuối cùng tính logarit trên các giá trị đầu ra băng lọc và tiến hành thực hiện phép biến đổi cosin rời rạc sẽ thu được các hệ số MFCC.

Tiếp theo, bài báo trình bày phương pháp định danh phương ngữ dựa trên mô hình GMM trong đó sử dụng kết hợp cả bộ tham số MFCC với tần số cơ bản F0, LogF0 và các giá trị chuẩn hóa của F0 và LogF0. Các cài đặt thử nghiệm trong bài báo sử dụng bộ công cụ mã nguồn mở ALIZE [7]. Bộ tham số MFCC sau khi được trích rút được bổ sung tham số F0 cùng các giá trị chuẩn hóa từ F0 vào từng véctơ đặc trưng. Mỗi véctơ đặc trưng tương ứng với khung thời gian 0,01 giây. Do vậy, các giá trị F0 được tính cho mỗi câu (tương ứng mỗi file wav chứa nội dung tiếng nói cần nhận dạng) cũng theo khung thời gian 0,01 giây. Các tham số F0 được bổ sung vào cuối mỗi véctơ đặc trưng. Các véctơ này sau đó được sử dụng để huấn luyện mô hình và dùng cho nhận dạng ở pha thử nghiệm.

IV. KẾT QUẢ THỬ NGHIỆM

A. Dữ liệu tiếng nói dùng cho thử nghiệm

Dữ liệu tiếng nói dùng cho thử nghiệm được xây dựng dành cho nghiên cứu nhận dạng phương ngữ. Tiếng nói được ghi âm trực tiếp trong đó nội dung văn bản dùng để đọc được tổ chức theo chủ đề và có sự cân bằng về thanh điệu (trung bình 717 từ cho mỗi thanh điệu).

Tín hiệu tiếng nói được ghi âm với tần số lấy mẫu là 16000Hz, ghi một kênh (mono) và 16 bit cho một mẫu. Ngữ liệu gồm có 50 giọng nam và 50 giọng nữ với tuổi trung bình là 21. Các giọng đã ghi âm được chọn đại diện cho 2 phương ngữ chính của tiếng Việt. Phương ngữ Bắc có 50 giọng (25 nam, 25 nữ). Phương ngữ Trung có 50 giọng (25 nam, 25 nữ). Phương ngữ Bắc được lựa chọn là giọng Hà Nội, còn phương ngữ Trung là giọng Huế. Đối với mỗi chủ đề, mỗi người nói đọc 25 câu, mỗi câu là một đoạn văn ngắn. Trung bình thời lượng ghi âm một câu là 10 giây. Số câu đã ghi âm là 15000 câu (100 người nói, mỗi người nói 150 câu) với dung lượng 3,62GB. Tổng cộng thời lượng là 33,79 giờ tiếng nói (Bảng 1, Bảng 2).

Bảng 1. Một số đặc điểm bộ dữ liệu tiếng nói thử nghiệm

STT	Phương ngữ	Số câu	Thời lượng (giờ)
1	Bắc	7500	16,82
2	Trung	7500	16,97
	Tổng	**15000**	**33,79**

Bảng 2. Phân bố theo chủ đề trong bộ dữ liệu tiếng nói thử nghiệm

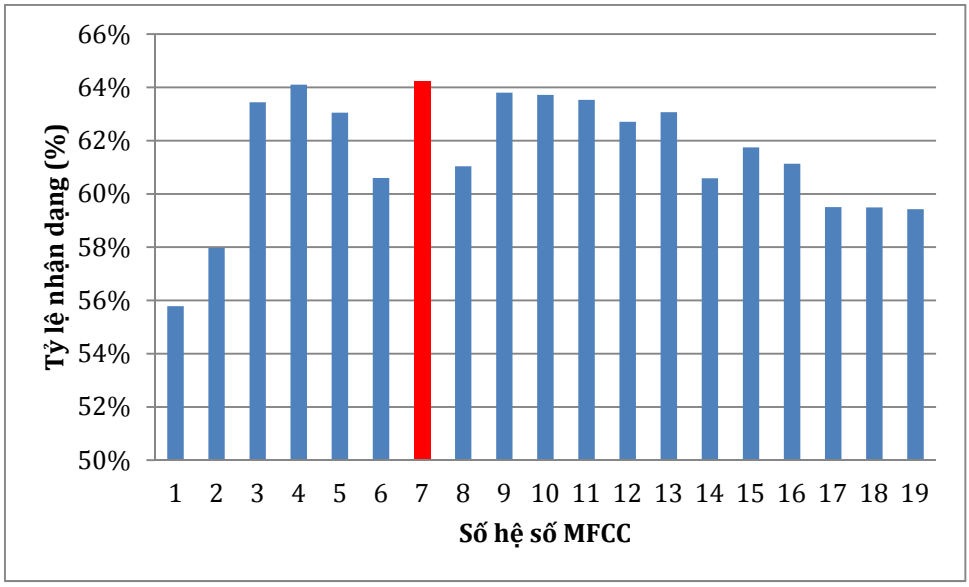
Chủ đề	Số câu	Số âm tiết	Thời lượng (phút)
Cơ bản	25	349	283,84
Đời sống	25	855	386,13
Khoa học	25	893	310,85
Kinh doanh	25	729	388,63
Ô tô-xe máy	25	652	282,23
Pháp luật	25	855	375,76
Tổng	**150**	**4333**	**2027,4**

Nội dung ghi âm được phân bổ tương đối đều theo phương ngữ cũng như theo giới tính. Phương ngữ Bắc gồm 7500 câu với thời lượng 16,82 giờ tiếng nói. Phương ngữ Trung bao gồm 7500 câu tương ứng 16,97 giờ tiếng nói. Giọng nam gồm 16,85 giờ tiếng nói. Giọng nữ gồm 16,94 giờ tiếng nói.

Để thử nghiệm, dữ liệu tiếng nói nêu trên được chia làm 2 nhóm tách biệt. Nhóm thứ nhất chiếm 50% dữ liệu dành riêng cho huấn luyện (7500 câu). Nhóm thứ hai là phần dữ liệu còn lại dùng cho thử nghiệm.

B. Thử nghiệm trong trường hợp chỉ sử dụng MFCC

Đây là trường hợp trong đó chỉ các tham số MFCC được sử dụng cho huấn luyện và thử nghiệm. Để tìm được số tham số MFCC tốt nhất cho nhận dạng, các thử nghiệm được tiến hành lần lượt với số hệ số MFCC tăng dần từ 1 đến 19 trên tổng số 7500 câu cần nhận dạng. Kết quả thử nghiệm thể hiện ở hình 2.



Hình 2. Kết quả thử nghiệm nhận dạng phương ngữ chỉ sử dụng tham số MFCC

Thử nghiệm cho thấy số hệ số MFCC=7 ứng với kết quả nhận dạng cao nhất là 64,2%. Vì vậy, trong các thử nghiệm sau, số hệ số MFCC sẽ lấy bằng 7 để kết hợp với tham số F0 và các dạng chuẩn hóa trên cơ sở F0.

C. Thử nghiệm trong trường hợp kết hợp MFCC với tham số F0

Trong trường hợp này, bộ tham số MFCC được kết hợp với tần số cơ bản F0, LogF0 và các dạng chuẩn hóa F0, LogF0. Chuẩn hóa F0 và LogF0 dùng các công thức sau:

- Đạo hàm F0 (dF0):

$$f_0(t) = dF0 \tag{9}$$

- Chuẩn hóa F0 theo xu hướng đi lên hoặc đi xuống của F0 mỗi câu (cdF0):

$$f_0(t) = \begin{cases} -1 & nếu\ ((F0_i - F0_{i-1}) \le -3) \\ 0 & nếu\ (-3 < (F0_i - F0_{i-1}) < 3) \\ 1 & nếu\ ((F0_i - F0_{i-1}) \ge -3) \end{cases} \tag{10}$$

Bảng 3. Kết quả thử nghiệm nhận dạng sử dụng bộ tham số MFCC và tham số F0

Test case	dF0	cdF0	F0sbM	F0sbMSD	LogF0	dLogF0	LogF0sbMM	LogF0sbM	LogF0sbMSD	Tỷ lệ nhận dạng
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	0	0	0	0	1	1	1	1	1	64,5%
2	0	0	0	0	0	0	0	1	1	64,5%
3	0	0	0	0	0	1	0	1	0	64,5%
4	0	0	0	0	1	0	0	1	1	65,1%
5	0	0	0	0	1	1	0	1	0	65,1%
6	0	0	0	0	1	1	1	0	1	65,3%
7	0	1	0	0	0	0	0	0	0	65,4%
8	0	0	0	0	1	0	1	0	0	65,5%
9	0	0	0	0	0	1	0	1	1	65,7%
10	0	0	0	0	0	0	1	0	0	66,0%
11	0	0	0	0	0	1	1	1	1	66,2%
12	0	0	0	1	0	0	1	0	0	66,2%

Test case	dF0	cdF0	F0sbM	F0sbMSD	LogF0	dLogF0	LogF0sbMM	LogF0sbM	LogF0sbMSD	Tỷ lệ nhận dạng
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
13	0	0	0	1	0	0	0	1	0	66.4%
14	0	0	0	0	1	1	0	0	1	66.6%
15	0	0	0	1	0	0	1	0	1	66.7%
16	0	0	0	1	0	1	1	0	0	66.7%
17	0	0	0	1	0	0	0	0	0	66.9%
18	1	0	0	0	0	0	0	0	0	66,9%
19	0	0	0	0	1	0	1	1	0	67,0%
20	0	0	0	1	0	0	1	1	0	67,1%
21	0	0	0	0	1	0	0	0	1	67,2%
22	0	0	0	0	1	1	0	0	0	67,2%
23	0	0	0	0	0	0	1	1	1	67,3%
24	0	0	0	0	0	1	1	1	0	67,3%
25	0	0	0	1	0	1	0	0	1	67,3%
26	0	0	0	0	1	0	1	0	1	67,4%
27	0	0	0	0	1	1	1	0	0	67,4%
28	0	0	0	1	1	0	0	1	0	67,4%
29	0	0	0	1	1	0	0	0	1	67,6%
30	0	0	0	1	0	1	1	0	1	67,7%
31	0	0	0	1	0	0	0	0	1	67,8%
32	0	0	0	1	0	1	0	0	0	67,8%
33	0	0	0	1	0	0	0	1	1	67,9%
34	0	0	0	1	0	1	0	1	0	67,9%
35	0	0	0	0	0	0	0	1	0	68,3%
36	0	0	0	0	0	0	1	1	0	68,6%
37	0	0	0	1	0	0	1	1	1	69,0%
38	0	0	0	1	0	1	1	1	0	69,0%
39	0	0	0	0	1	0	0	1	0	69,2%
40	0	0	0	0	1	0	0	0	0	69,3%
41	0	0	0	1	1	0	0	0	0	69,4%
42	0	0	0	1	0	1	1	1	1	69,6%
43	0	0	0	1	0	1	0	1	1	69,8%
44	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**70,3%**

- Chuẩn hóa F0 theo giá trị trung bình F0 cho mỗi câu (F0sbM):

$$f_0(t) = F_0(t)/\overline{F_0(t)} \tag{11}$$

- Chuẩn hóa F0 theo trung bình và độ lệch chuẩn của F0 (F0sbMSD):

$$f_0(t) = \frac{F_0(t)-\overline{F_0(t)}}{\sigma F_0(t)} \tag{12}$$

- Đạo hàm LogF0 (dLogF0):

$$f_0(t) = d\ LogF0 \tag{13}$$

- Chuẩn hóa LogF0 theo giá trị MinLogF0 và MaxLogF0 cho mỗi câu (LogF0sbMM):

$$logf_0(t) = \frac{LogF_0(t)-\min LogF_0(t)}{\max LogF_0(t)-\min LogF_0(t)} \tag{14}$$

- Chuẩn hóa LogF0 theo trung bình LogF0 mỗi câu (LogF0sbM):

$$logf_0(t) = logF_0(t)/\overline{logF_0(t)} \tag{15}$$

- Chuẩn hóa theo LogF0 theo trung bình và độ lệch chuẩn của LogF0 (LogF0sbMSD):

$$logf_0(t) = \frac{logF_0(t)-\overline{logF_0(t)}}{\sigma logF_0(t)} \tag{16}$$

Các thử nghiệm đã được tiến hành bằng cách kết hợp các tham số MFCC với F0, LogF0 và các dạng chuẩn hóa tương ứng. Kết quả của các thử nghiệm được cho ở Bảng 3. Từ cột 2 đến cột 10 là giá trị F0, LogF0 cùng các giá trị chuẩn hóa tương ứng. Mỗi hàng tương ứng với một thử nghiệm, giá trị nào được dùng thì vị trí tương ứng cột có giá trị 1, không dùng có giá trị là 0. Cột 11 là tỷ lệ nhận dạng. Số liệu trên Bảng 3 đã được sắp xếp theo thứ tự tăng dần của tỷ lệ nhận dạng.

Số liệu Bảng 3 cho thấy, việc bổ sung tham số F0 vào nhận dạng nhìn chung cho kết quả cao hơn so với trường hợp chỉ sử dụng bộ tham số MFCC. Điều này hoàn toàn xác đáng vì hai yếu tố quan trọng sau đây đối với tiếng Việt và phương ngữ tiếng Việt. Thứ nhất, tần số cơ bản đóng vai trò vô cùng quan trọng với tiếng Việt do tần số cơ bản quyết định các thanh điệu.

Thứ hai, việc phân biệt các phương ngữ tiếng Việt theo ngữ âm có thể cơ bản dựa trên quy luật biến thiên F0 trong quá trình phát âm của các phương ngữ. Với các thử nghiệm chỉ sử dụng bộ tham số MFCC, kết quả nhận dạng cao nhất đạt 64,2% (Hình 2). Thử nghiệm bổ sung tham số F0 được chuẩn hóa theo giá trị trung bình F0 cho mỗi câu (F0sbM) có kết quả nhận dạng cao nhất đạt 70,3%.

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Kết quả thử nghiệm cho thấy việc kết hợp sử dụng tham số F0 với bộ tham số MFCC cho kết quả nhận dạng phương ngữ tốt hơn khi không sử dụng tham số F0. Bên cạnh đó, nghiên cứu cũng cho thấy khả năng ứng dụng mô hình hỗn hợp Gauss đa biến vào (GMM) trong nhận dạng phương ngữ tiếng Việt. Các thử nghiệm trên bộ ngữ liệu phương ngữ tiếng Việt VDSPEC cũng chỉ ra bộ tham số MFCC với số hệ số bằng 7 cho kết quả nhận dạng phương ngữ tiếng Việt tốt nhất. Các kết quả nghiên cứu này có thể tiếp tục phát triển để ứng dụng trong các hệ thống nhận dạng tự động tiếng Việt nói nhằm nâng cao hiệu năng nhận dạng.

VI. TÀI LIỆU THAM KHẢO

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

[2] 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.

[3] 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.

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

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

[6] Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., and Deller Jr., J. R. (2002), "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.

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

[8] Phạm Ngọc Hưng, Trịnh Văn Loan, Nguyễn Hồng Quang, Phạm Quốc Hùng (2014), "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.

VIETNAMESE DIALECT IDENTIFICATION USING MFCC AND FUNDAMENTAL FREQUENCY

Pham Ngoc Hung, Trinh Van Loan, Nguyen Hong Quang

ABSTRACT *- The dialect identification was studied for many languages over the world nevertheless the research on signal processing for Vietnamese dialects is still limited and there were not many published works. Il exists many different dialects for Vietnamese. The influence of dialectal features on speech recognition systems is important. If the information about dialects is known during speech recognition process, the performance of recognition systems will be better because the corpus of these systems is normally organized according to different dialects. This paper will present the combination of MFCC coefficients and fundamental frequency features of Vietnamese for dialectal identification based on GMM. The experiment result for the dialect corpus of Vietnamese shows that the performance of dialectal identification is increased from 64.2% for the case using only MFCC coefficients to 70.3% for the case using MFCC coefficients and the information of fundamental frequency.*

Keywords *- Fundamental frequency, MFCC, GMM, identification of Vietnamese dialects.*

.