Tạp chí Tin học và Điều khiển học, T.27, S.3 (2011), 275–282

TỔ HỢP ĐƯỜNG F0 VÀ VTLN CHO NHẬN DẠNG TÊN TIẾNG VIỆT

NGÔ HOÀNG HUY

Viện Công nghệ thông tin, Viện Khoa học và Công nghệ Việt Nam

Tóm tắt. Bài báo đề xuất và thử nghiệm hiệu ứng của tổ hợp đặc trưng F0 và chuẩn hóa độ dài bộ phận cấu âm (VTLN, vocal tract length normalisation) để nâng cao chất lượng nhận dạng tiếng tên tiếng Việt trong mô hình nhận dạng tiếng nói phát âm liên tục dựa trên HMM. Các kết quả thử nghiệm chứng tỏ hệ nhận dạng tiếng nói độc lập người nói với đặc trưng tiếng nói dựa trên đường F0 và đặc trưng MFCC biến đổi theo VTLN cho các tập tên riêng có độ trùng lặp về âm tiết và từ là khá cao, đã chuẩn hóa tốt biến thiên tần số của người nói mới và cải tiến được kết quả nhận dạng.

Abstract. This paper presents a study and experiment on combination effect of F0 feature and vocal tract length normalisation (VTLN) to improve the HMM-based continuous speech recognition system performance for proper Vietnamese name sets. Our experimental results on the recognizer, independent speaker with speech feature based on F0 contour and MFCC warped by VTLN, for an university name set that has high duplicaton of syllables and words, already well normalized frequency variation of new speakers and improved the recognition performance.

Từ khoá. FO, formant, VTLN, MFCC, ML, hiệu chỉnh tần số, độc lập người nói, thanh điệu.

1. MỞ ĐẦU

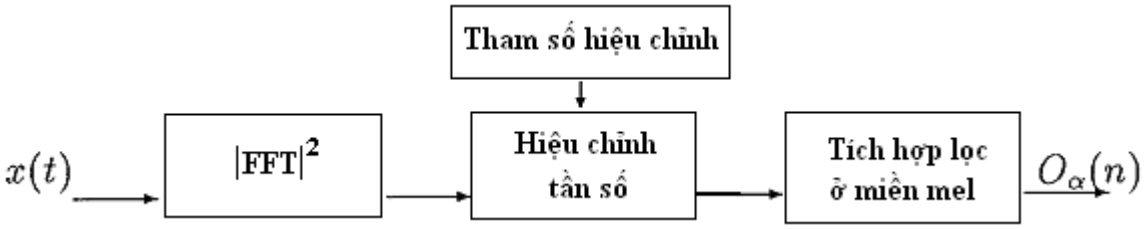
Nghiên cứu các ảnh hưởng của hiện tượng ngôn điệu tiếng Việt như thanh điệu, các tần số formant và trường độ âm tiết trong các hệ thống nhận dạng tiếng Việt là một vấn đề thiết yếu nhưng hiện tại ít được đề cập tới trong các công trình nghiên cứu về xử lý âm thanh tiếng Việt. Các hệ thống nhận dạng tiếng Việt dựa trên HMM thường dựa trên đặc trưng chuẩn MFCC và quy trình huấn luyện, nhận dạng theo thuật toán sau:

Khi ứng dụng thuật toán nhận dạng HMM trên cho việc nhận dạng tiếng nói liên tục không phụ thuộc người nói, hệ thống thường suy giảm độ chính xác với những người nói có giọng nói không phù hợp với những mẫu giọng được sử dụng để huấn luyện mô hình HMM.

Một số nghiên cứu gần đây đã khảo sát đường thanh điệu tiếng Việt trong ngữ cảnh để nhận dạng thanh điệu và cải tiến kết quả nhận dạng từ và câu tiếng Việt [7,8,10,12] . Các tiếp cận này chủ yếu vẫn ghép trực tiếp đặc trưng thanh điệu vào các kiểu đặc trưng tiếng nói như MFCC, PLP [10,12]. Có một tiếp cận khác sử dụng giá trị đường F0, các tần số

Bảng 1. Thuật toán nhận dạng tiếng nói dựa trên HMM với đặc trưng MFCC

Huấn luyện HMM	Nhận dạng với HMM
Đầu vào gồm T frame các đặc trưng MFCC Bước 1: Xác định dãy trạng thái tối ưu bằng thuật toán Viterbi: $S = \left(\{S_t\}_{t=1}^{T}\right) = argmax \sum_{t=1}^{T} logp\,(O_t\|\lambda, W)$ Bước 2: Hiệu chỉnh lại tham số mô hình HMM: $\lambda^* = argmax \sum_{t=1}^{T} logp\,(O_t\|\lambda, s_t, W)$ Bước 3: Đặt $\lambda = \lambda^*$, lặp lại tới khi mô hình hội tụ	Đầu vào gồm T frame các đặc trưng MFCC Giải mã theo Viterbi để xác định tập nhãn và dãy trạng thái tối ưu ứng với bộ tham số mô hình HMM đã cho: $\left(W, S = \{s_t\}_{t=1}^{T}\right) = argmax \sum_{t=1}^{T} logp\,(O_t\|\lambda, s_t)$



Hình 2.1. Hiệu chỉnh tần số và trích chọn đặc trưng MFCC

formant để xây dựng một phép hiệu chỉnh lại các đặc trưng MFCC, dẫn đến kiểu đặc trưng không phụ thuộc người nói trong cả quá trình huấn luyện và nhận dạng, qua đó kết quả nhận dạng của các hệ thống không phụ thuộc người nói được cải thiện đáng kể [5,6,9,13,14].

Bài báo đề xuất phương pháp ghép trực tiếp giá trị đường F0 vào các vector đặc trưng MFCC đã được hiệu chỉnh theo VTLN trong các hệ thống nhận dạng tên riêng tiếng Việt phát âm liên tục và độc lập người nói.

2. PHÉP CHUẨN HÓA VTLN

Các hệ nhận dạng tiếng nói thường trích chọn đặc trưng của mỗi khung tiếng nói (độ dài 10ms-25ms) theo kiểu MFCC ở thang tần số mel dựa trên đặc điểm cảm thụ tần số âm của tai người, tuy vậy các hệ số MFCC không thể hiện được các biến thể bên trong của mỗi người nói như VTL (vocal tract length, độ dài bộ cấu âm), dẫn đến việc suy giảm chất lượng nhận dạng trong các hệ thống nhận dạng độc lập người nói.

VTLN là phép chuẩn hóa tín hiệu tiếng nói để VTL đạt được mức trung bình nhờ các tham số hiệu chỉnh tần số cho mỗi người nói hoặc một phát âm. Có hai tiếp cận chính cho VTLN , một là ước lượng hệ số hiệu chỉnh tần số dựa vào đặc điểm âm học của người nói như các tần số formant, hai là cách duyệt trên lưới của tham số hiệu chỉnh để tối ưu hóa hàm mục tiêu của mô hình nhận dạng.

Biểu diễn tiếng nói đầu vào đã qua tiền xử lý $x(t)$ trong miền tần số bằng biến đổi FFT:

Bảng 2. Một số dạng của phép hiệu chỉnh tần số

Dạng biến đổi	Công thức biến đổi
Phi tuyến trong miền tần số	$\varphi_\alpha(\omega) = \omega + 2tan^{-1}\left(\frac{(1-\alpha)sin\omega}{1-(1-\alpha)cos\omega}\right)$
Tuyến tính từng đoạn trong miền tần số	$(**)\ \varphi_\alpha(\omega) = \begin{cases}\alpha^{-1}\omega, \omega \le \omega_0 \\ b\omega + c, \omega \ge \omega_0\end{cases}$; [HTK] $\varphi_\alpha(\omega) = \begin{cases}a\omega + b, \omega < \frac{2\omega_l}{1+\alpha^{-1}} \\ \alpha^{-1}\omega, \omega \in \left[\frac{2\omega_l}{1+\alpha^{-1}}, \frac{2\omega_u}{1+\alpha^{-1}}\right] \\ c\omega + d, \omega > \frac{2\omega_u}{1+\alpha^{-1}}\end{cases}$
Dịch chuyển trong miền mel	$\varphi_\alpha(\omega) = e^{\frac{\alpha}{1127}}.\omega + 700\left(e^{\frac{\alpha}{1127}} - 1\right)$ (ứng với $W_\alpha(z) = z + \alpha$)
Tuyến tính từng đoạn trong miền mel	$\varphi_\alpha(\omega) = 700\left(e^{\frac{W_\alpha(z)}{1127}}\right)$, $W_\alpha(z) = \begin{cases}z_{min} + (z - z_l)\frac{z_l + \alpha - z_{min}}{z_l - z_{min}}, z < z_l \\ z + \alpha, z \in [z_l, z_u] \\ z_u + \alpha + (z - z_u)\frac{z_{max} - z_u - \alpha}{z_{max} - z_u}, z > z_u\end{cases}$

X(ω)= H(ω)S(ω) + N(ω), ở đó H(ω) là biến dạng kênh và N(ω) nhiễu cộng của tín hiệu. Sử dụng M bộ lọc tam giác với khoảng cách giữa các vị trí ω_k trong thang tần số mel:

$$B_k(\omega) = \begin{cases}\frac{\omega}{\omega_k - \omega_{k-1}}, \omega \in [\omega_{k-1}, \omega_k] \\ \frac{\omega^{k+1} - \omega}{\omega_{k+1} - \omega_k}, \omega \in [\omega_k, \omega_{k+1}]\end{cases},$$

$$Y(m) = \sum_{\omega \in [\omega_{k-1}, \omega_{k+1}]} B_k(\omega)\,|X(\omega)|^2, 0 \le m \le M-1\ (*)$$

$$MFCC(n) = \sum_{m=0}^{M-1} cos\frac{n\pi\left(m - \frac{1}{2}\right)}{M} logY(m), 0 \le n \le N-1$$

Khi đó với phép hiệu chỉnh tần số có dạng $\omega' = \omega_\alpha(\omega)$ thì công thức (*) trở thành

$$Y(m) = \sum_{\omega \in [\omega_{k-1}, \omega_{k+1}]} B_k(\omega)\,|X(\varphi_\alpha(\omega))|^2$$

Bảng dưới đây cho ta một số dạng biến đổi tuyến tính và phi tuyến của hàm $\varphi_\alpha(\omega)$, sử dụng hàm $z = mel(\omega) = 1127 * ln\left(1 + \frac{\omega f_s}{2\pi.700}\right), \omega \in [0, \pi]$ và f_s là tần số lấy mẫu.

Tham số α đặc trưng cho mỗi người nói có thể được ước lượng tự động từ các giá trị trung bình của formant F3 của tiếng nói đầu vào và của tập huấn luyện [15]. Trong [6] các tác giả ước lượng dựa trên giá trị trung bình của đường F0 của câu phát âm. Kiểm nghiệm nhận dạng trên tập tên riêng các trường học chúng tôi thấy phương pháp ước lượng này không tăng được đáng kể độ chính xác nhận dạng, do mới chỉ nhấn mạnh vào việc thích ứng với lớp giọng có tần số cơ bản cao và thấp. Ngoài ra, việc chuẩn hóa VTL cho từng người nói theo phương pháp này tuy đạt được hiệu quả về tốc độ xử lý nhưng không thích ứng được với hình dạng đường F0 của các câu tiếng Việt biến đổi mạnh theo kiểu thanh điệu của âm tiết.

Sử dụng chính hàm mục tiêu của các mô hình HMM (π, λ), tham số hiệu chỉnh tần số $\alpha \in [\alpha_{min}, \alpha_{max}]$ có thể ước lượng trên từng phát âm tiếng nói đầu vào [2,3,14] theo công thức sau:

$$\begin{cases} \left(W, S = \{s_t\}_{t=1}^{T}\right) = argmax \sum_{t=1}^{T} logp\left(O_t|\lambda, s_t\right) \\ \alpha = argmax_{\alpha\in[\alpha_{min},\alpha_{max}]} \sum_{t=1}^{T} logp\left(O_t^{\alpha}|\lambda, s_t\right), O_t^{\alpha} = O_t\left(\varphi_{\alpha}\right) \end{cases}$$

3. TỔ HỢP GIÁ TRỊ F0 VÀ CHUẨN HÓA VTLN

Phương pháp ghép giá trị F0 vào các hệ số MFCC đã được nắn lại sau phép hiệu chỉnh tần số đề xuất trong bài báo này được thực hiện gồm 4 bước chính sau:

Bước 1. Xác định tham số α và hiệu chỉnh lại các vector đặc trưng MFCC nhờ phương pháp huấn luyện hợp lý cực đại ML (xem quy trình nêu ở bước 4)

Bước 2. Tính F0 theo thuật toán RAPT, nội suy xác định giá trị liên tục của đường F0 trên cả đoạn vô thanh. Làm trơn và chuẩn hóa các giá trị F0.

Bước 3. Kết hợp F0 với các hệ số MFCC đã hiệu chỉnh
Thuật toán xác định tham số hiệu chỉnh α dựa trên phương pháp huấn luyện hợp lý cực đại ML (maximum likelihood) được cài đặt như sau:

Bước 4. Huấn luyện và giải mã.
i) Giai đoạn huấn luyện : Với mỗi phát âm tiếng nói đã gán nhãn W, gồm T frame

Bước 1: Khởi tạo α=1.0 và xác định dãy trạng thái tối ưu bằng thuật toán Viterbi:

$$S = \left(\{s_t\}_{t=1}^{T}\right) = argmax \sum_{t=1}^{T} logp\left(O_t^{\alpha}|\lambda, s_t\right), O_t^{\alpha} = O_t\left(\varphi_{\alpha}\right)$$

Bước 2: Duyệt tìm giá trị tối ưu trên lưới giá trị của tham số α:

$$\alpha^* = argmax_{\alpha\in[\alpha_{min},\alpha_{max}]} \sum_{t=1}^{T} logp\left(O_t^{\alpha}|\lambda, s_t\right), O_t^{\alpha} = O_t\left(\varphi_{\alpha}\right)$$

Bước 3: Thực hiện phân đoạn cưỡng bức (forced alignment) dựa trên bộ nhãn W và tham số hiệu chỉnh α^* và hiệu chỉnh lại tham số mô hình HMM:

$$\lambda^* = argmax_{\alpha\in[\alpha_{min},\alpha_{max}]} \sum_{t=1}^{T} logp\left(O_t^{\alpha^*}|\lambda, s_t\right), O_t^{\alpha^*} = O_t\left(\varphi_{\alpha^*}\right)$$

Bước 4: Đặt $\alpha = \alpha^*$, và $\lambda = \lambda^*$, lặp lại tới khi mô hình hội tụ.
ii) Giai đoạn giải mã (nhận dạng): Với một phát âm đầu vào gồm T frame

Bước 1: Giải mã theo Viterbi để xác định tập nhãn, và dãy trạng thái tối ưu ứng với bộ tham số mô hình HMM đã cho:

$$\left(W, S = \{s_t\}_{t=1}^{T}\right) = argmax \sum_{t=1}^{T} logp\left(O_t|\lambda, s_t\right)$$

Bước 2: Duyệt tìm giá trị tối ưu trên lưới giá trị của tham số α:

$$\alpha^* = argmax_{\alpha\in[\alpha_{min},\alpha_{max}]} \sum_{t=1}^{T} logp\left(O_t^{\alpha}|\lambda, s_t\right), O_t^{\alpha} = O_t\left(\varphi_{\alpha}\right)$$

Bước 3: Giải mã Viterbi lần nữa với tham số α^* để xác định tập nhãn đầu ra:

$$Ws^* = argmax \sum_{t=1}^{T} logp\left(O_t^{\alpha^*}|\lambda, s_t\right)$$

Do thuật toán HMM chuẩn có độ phức tạp là O(VN^2T), ở đây N là số trạng thái của mô hình HMM (N = 5 trong thử nghiệm này), V số phần tử từ vựng (V = 117 trong thử nghiệm này, là số âm vị với các nguyên âm mang thanh điệu) và T là số frame đầu vào, nên độ phức tạp của thuật toán HMM có kết hợp với VTLN là O(KVN^2T), trong đó K là số giá trị rời rạc hóa trên lưới giá trị của tham số $\alpha \in [\alpha_{min}, \alpha_{max}]$ (ở đây $\alpha_{min} = 0.85, \alpha_{max} = 1.15$, mức rời rạc hóa theo bước 0.0001)

4. THỬ NGHIỆM TRÊN TẬP TÊN RIÊNG PHỨC TẠP

Tập thử nghiệm của ứng dụng là tên của khoảng 300 trường Đại học và Cao đẳng trong nước (một số tên trường có thể không còn trong thực tế).

Tập tên riêng này có các đặc điểm sau:

- Tên trường chứa hơn 200 từ đa âm tiết tiếng Việt bao gồm tên địa danh cổ, tỉnh thành, tên các danh nhân, ngành nghề, phiên âm tiếng nước ngoài, số đếm chỉ chi nhánh trường.
- Độ dài của một tên trường : ngắn nhất 4 âm tiết, dài nhất 15 âm tiết.
- Không có 2 âm tiết nào có cùng âm tiết gốc (âm không mang thanh điệu).
- Các tên trường có sự trùng lặp âm tiết lớn như các cặp tên trường sau
  - Học viện hành chính quốc gia, Học viện hành chính quốc gia Hồ Chí Minh
  - ĐH dân lập Văn Lang, Đại học dân lập Văn hiến
  - Trường sỹ quan lục quân một, Trường sỹ quan lục quân hai
  - ĐH dân lập Thăng Long, ĐH dân lập Cửu Long
  - ĐH lâm nghiệp, ĐH nông nghiệp
  - ĐH dân lập Bình Dương, ĐH dân lập Hùng Vương
  - ĐH Sư phạm kỹ thuật, ĐH Sư phạm mỹ thuật

Dữ liệu huấn luyện nhanh các mô hình HMM của các âm vị tiếng Việt là các bài đọc truyện và tin tức của 2 phát thanh viên (một nam, một nữ giọng Hà Nội) của đài tiếng nói Việt Nam, dung lượng dữ liệu hơn 1GB, và chưa được gán nhãn ngữ âm.

Dữ liệu huấn luyện thích ứng được ghi bởi giọng đọc của 12 nam, 10 nữ sinh viên, môi trường tín hiệu trong lớp học trên giảng đường, mỗi sinh viên đọc 300 tên trường đúng một lượt. Tín hiệu thu có tần số lấy mẫu 11025Hz, đơn kênh 16 bit và có nhiễu.

Dữ liệu kiểm tra hệ thống nhận dạng là giọng đọc của 3 nam, 6 nữ sinh viên và một giọng nữ tiếng miền Nam, được thu trong cùng môi trường tín hiệu như với dữ liệu huấn luyện thích ứng.

Bảng 3. Bảng kết quả thực nghiệm kiểm tra trên tập 10 người nói

Người đọc	Đặc trưng chuẩn MFCC		Đặc trưng MFCC + F0		VTLN-MFCC + F0	
	mức từ	mức câu	mức từ	mức câu	mức từ	mức câu
Nu1	93,8	82,2	92,7	81,2	96,7	88,5
Nu2	97,0	86,3	97,3	88,4	97,6	89,4
Nu3	95,5	82,8	93,4	78,7	96,4	87,8
Nu4	897,4	890,9	897,0	888,8	897,6	891,9
Nu5	95,5	85,1	97,1	91,0	98,9	97,0
Nu6	86,7	67,7	88,1	72,9	91,8	81,2
Nu7	95,7	84,6	96,9	89,9	98,1	93,8
Nam1	98,6	92,9	98,5	92,9	98,9	94,9
Nam3	97,1	88,8	97,7	93,9	98,6	96,2
Trung bình	95,3	84,6	95,6	86,4	97,1	90,6

Bảng 4. Bảng kết quả giải mã tên trường của người đọc “Nam2”

Phát âm	Kết quả giải mã cho tên đúng
ĐH dân lập kỹ thuật công nghiệp	ĐH dân lập kỹ thuật công nghệ
ĐH Thái Bình	ĐH Y Thái Bình
Học viện bưu chính viễn TP. HCM	Học viện bưu chính viễn thông TP. HCM
ĐH ngoại thương Đà Nẵng	ĐH ngoại ngữ Đà Nẵng

Khi tiến hành thử nghiệm, chúng tôi sử dụng phiên bản HTK 3.4 để huấn luyện, kiểm thử và trích chọn đặc trưng MFCC thông thường.

Tập âm vị tiếng Việt gồm khoảng 52 âm vị bao gồm các phụ âm đầu, âm đệm, nguyên âm chính với thanh điệu và âm cuối tương ứng với 52 mô hình HMM cần được huấn luyện.

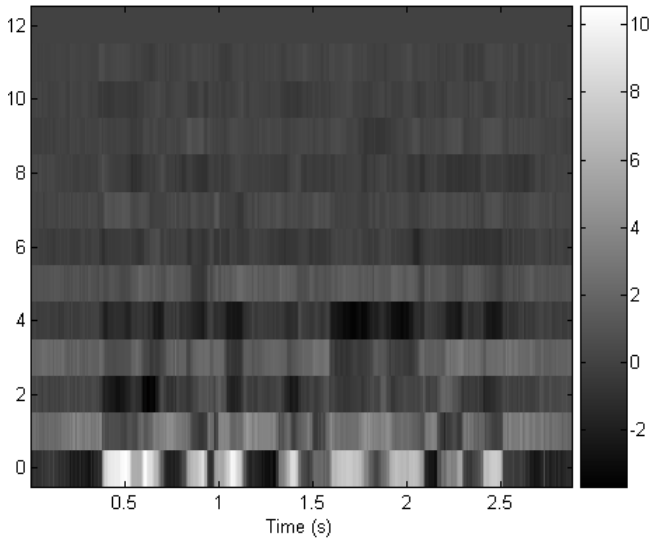
Giá trị đường F0 và các vector đặc trưng MFCC được trích chọn với các frame độ dài 25 mili giây, phần trùng nhau giữa 2 frame liên tiếp là 15 mili giây. Các vector đặc trưng gồm 12 hệ số MFCC và 1 hệ số năng lượng và các sai phân bậc 1 và bậc 2 của 13 hệ số này. Các HMM có hàm mật độ xác suất liên tục. Việc hiệu chỉnh các vector MFCC và ghép giá trị F0 sử dụng phương pháp đã trình bày ở trên.

Các thực nghiệm cho thấy các biến đổi tần số kiểu “tịnh tiến” trong miền mel cho kết quả thấp hơn một chút so với phép biến đổi theo hệ số tỉ lệ. Bảng 3 được cho với phép biến đổi tần số theo phương pháp của HTK (xem bảng 1).

Quan sát bảng thực nghiệm 2, kết quả nhận dạng của “Nu6” thấp do đây là một giọng nói tiếng miền Nam, thanh điệu và các tham số VTL hoàn toàn khác so với các giọng đọc trong tập huấn luyện (chỉ có giọng miền Bắc).

Người nói “Nu6", câu "ĐH mỹ thuật TP.HCM", nhận dạng nhầm thành: “ĐH luật TP. HCM”. Áp dụng phép hiệu chỉnh VTLN cho giọng “Nu6”, cho kết quả nhận dạng đúng.

Người nói “Nam2” thực tế kết quả nhận dạng sẽ cao hơn nhiều, lỗi xảy ra do người đọc đã nhầm một số âm tiết của tên trường, hệ thống giải mã cho HMM đã nhận dạng phát âm này về tên trường có trong tập từ vựng, chẳng hạn như:



Hình 4.2. MFCC chuẩn, người nói "Nu6", câu "ĐH mỹ thuật TP.HCM"

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Theo nội dung đã trình bày ở trên, các giá trị của đường F0 và độ dài bộ phận cấu âm của mỗi giọng nói đã ảnh hưởng đáng kể đến kết quả nhận dạng tiếng Việt. Để tích hợp các giá trị này vào hệ thống nhận dạng, đầu tiên áp dụng phép chuẩn hóa kiểu VTLN cho tiếng nói đầu vào để hiệu chỉnh lại tần số trước khi tính các hệ số MFCC như thông thường, sau đó ghép giá trị đường F0 đã được tiền xử lý (làm đầy trên các đoạn vô thanh và làm trơn) của phát âm và chuyển tới bộ huấn luyện hoặc giải mã của các HMM.

Kết quả thực nghiệm được áp dụng vào vấn đề nhận dạng tập tên riêng tiếng Việt có độ lặp lại cao về âm tiết và đa dạng như tập tên của khoảng 300 trường Đại học và Cao đẳng đã chứng tỏ phương pháp đề xuất cải tiến được đáng kể kết quả nhận dạng của hệ thống nhận dạng tiếng Việt độc lập người nói với tiếng nói đầu vào được phát âm liên tục.

Trong các nghiên cứu tiếp theo chúng tôi sẽ tập trung vào vấn đề ước lượng nhanh hệ số của phép hiệu chỉnh tần số dựa trên các tần số formant.

TÀI LIỆU THAM KHẢO

[1]. Daniel Elenius, Mats Blomberg, Dynamic vocal tract length normalization in speech recognition, *Proceedings from Fonetik 2010 Lund*, 2010, ISSN 0280-526X, 29-34.

[2]. Tadashi Emori, Koichi Shinoda, "Rapid vocal tract length normalization using maximum likelihood estimation", Eurospeech – Scandinavia, 2001.

[3]. Ramalingam Hariharan, Olli Viikki, "On combining vocal tract length normalisation and speaker adaptation for noise robust speech recognition", Eurospeech, 1999.

[4]. Chin-Hui Lee, Haizhou Li, Lin-shan Lee, Ren-Hua Wang, Qiang Huo, *Advances In Chinese Spoken Language Processing*, World Scientific Publishing Co.Pte.Ltd, 2007 (ISBN-13 978-981-256-904-2, 25–31).

[5]. Li Lee, Richard C. Rose, Speaker normalization using efficient frequency warping procedures, *ICASSP*, 1996.

[6]. Jian Liu, Thomas Fang Zheng, and Wenhu Wu, *Pitch Mean Based Frequency Warping*, Springer-Verlag Berlin Heidelberg, ISCSLP, LNAI 4274, 2006, 87–94.

[7]. Ngô Hoàng Huy, Nguyễn Thị Thanh Mai, Phân lớp các đường thanh điệu trong ngữ cảnh câu, *Kỷ yếu Hội thảo Quốc gia*, NXB KHKT, 2006 (279–284).

[8]. Ngô Hoàng Huy, Nguyễn Thị Thanh Mai, Nhận dạng thanh điệu tiếng Việt trên tiếng nói rời rạc phụ thuộc người nói, *Kỷ yếu Hội thảo Quốc gia*, NXB KHKT, 2006 (443–449).

[9]. Sankaran Panchapagesan, Abeer Alwan, Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC, *Computer Speech and Language 23* , 2009, 42–64.

[10]. Nguyen Hong Quang, N. Pascal, C.Ericy, and Trinh Van Loan, Tone recognition of vietnamese continuous speech using hidden markov model, *HUT-ICCE 2008*, Hoian, Vietnam, 2008.

[11]. William R. Rodrıguez, Oscar Saz, Antonio Miguel and Eduardo Lleida, On Line Vocal Tract Length Estimation for Speaker Normalization in Speech Recognition, *VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop*, FALA 2010.

[12]. Thang Tat Vu, Dung Tien Nguyen, Mai Chi Luong, John-Paul Hosom, Vietnamese Large Vocabulary Continuous Speech Recognition, *Proc. of Eurospeech Conference 2005*, Lisbon, Septemper, 4-8, 2005 (1172–1175).

[13]. Puming Zhan and Alex Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition" Technical report, CMU-LTI-97-150, 1997.

[14]. Puming Zhan, Martin Westphal, Speaker normalization based on frequency warping, *ICASSP 1997*, 1997.

[15]. Shizhen Wang, Yi-Hui Lee, Abeer Alwan, Bark-shift based nonlinear speaker normalization using the second subglottal resonance, *INTERSPEECH 2009*, 2009 (1619–1622).

Ngày nhận bài 10 - 3 - 2011

Nhận lại sau sửa 21 - 7 - 2011