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.2016.00082

SO SÁNH HIỆU NĂNG MỘT SỐ PHƯƠNG PHÁP NHẬN DẠNG CẢM XÚC TIẾNG VIỆT NÓI

Lê Xuân Thành[1], Đào Thị Lệ Thủy[2], Nguyễn Hồng Quang[1], Trịnh Văn Loan[1],

[1] Viện Công nghệ Thông tin và Truyền thông, Trường Đại học Bách khoa Hà Nội
[2] Khoa Công nghệ Thông tin, Trường Cao đẳng nghề Công nghệ cao Hà Nội

thanhlx@soict.hust.edu.vn, thuydt@hht.edu.vn, quangnh@soict.hust.edu.vn, loantv@soict.hust.edu.vn

TÓM TẮT— *Nhận dạng cảm xúc là hướng nghiên cứu được quan tâm trong thời gian gần đây. Những kết quả đã công bố hầu như mới chỉ tập trung vào một số ngôn ngữ thông dụng trên thế giới. Trong khi đó, các nghiên cứu trên tiếng Việt được thực hiện còn rất ít. Phần đầu bài báo sẽ mô tả phương pháp mới để xây dựng bộ ngữ liệu cảm xúc cho tiếng Việt nói với bốn cảm xúc cơ bản: bình thường, vui, buồn và tức giận. Dựa trên bộ ngữ liệu này, việc phân tích ảnh hưởng của các cảm xúc đến hai tham số cơ bản của tiếng nói là tần số cơ bản F0 và cường độ tiếng nói đã được thực hiện. Kết quả phân tích cho thấy, có sự phân nhóm rõ ràng giữa cảm xúc bình thường/buồn với cảm xúc vui/tức giận. Quy luật biến thiên tần số cơ bản F0 đóng vai trò rất quan trọng đối với tiếng Việt nói vì quy luật này quyết định 6 thanh điệu khác nhau của tiếng Việt đồng thời tham gia biểu hiện các cảm xúc khác nhau. Tần số cơ bản F0 cùng với cường độ tiếng nói đã được bước đầu sử dụng làm các tham số đặc trưng thử nghiệm cho các bộ nhận dạng cảm xúc bao gồm: K láng giềng gần nhất (KNN: K-Nearest Neighbor), phân tích phân biệt tuyến tính (LDA: Linear Discriminant Analysis), phân tích phân biệt toàn phương (QDA: Quadratic Discriminant Analysis), bộ phân lớp các véc tơ hỗ trợ (SVC: Support Vector Classifier) và máy véc tơ hỗ trợ (SVM: Support Vector Machine). Chỉ riêng với các tham số đặc trưng nêu trên, phương pháp SVC cho kết quả tốt nhất với giọng nam, tỷ lệ nhận dạng cảm xúc đúng đạt 56,9%. Với giọng nữ, kết quả tốt nhất là 57,7% khi sử dụng phương pháp SVM.*

Từ khóa— *Tiếng Việt nói, nhận dạng cảm xúc, F0, cường độ tiếng nói, K láng giềng gần nhất KNN, phân tích phân biệt tuyến tính LDA, phân tích phân biệt toàn phương QDA, máy véc tơ hỗ trợ SVM.*

I. GIỚI THIỆU

Cảm xúc của người nói là một hiện tượng tự nhiên, tồn tại vốn có trong tiếng nói con người. Việc xác minh cảm xúc của người nói sẽ giúp hệ thống hiểu rõ hơn về trạng thái của người nói, từ đó có thể đưa ra những trợ giúp quyết định cho con người. Hệ thống nhận dạng cảm xúc được thực hiện để xác định trạng thái cảm xúc của người nói. Những hệ thống này đã và đang được áp dụng hiệu quả trong một số lĩnh vực như trợ giúp lái xe thông minh, trợ giúp bệnh nhân trong bệnh viện, các hệ thống trả lời thông tin tự động v.v…

Những kết quả nghiên cứu về nhận dạng cảm xúc đã công bố hầu như mới chỉ tập trung vào một số ngôn ngữ thông dụng trên thế giới. Trong khi đó, các nghiên cứu trên tiếng Việt được thực hiện còn rất ít [3], [2], [17], [18]. Một số tác giả Trung Quốc [9], [13] có kết hợp với sinh viên Việt Nam xây dựng ngữ liệu cảm xúc tiếng Việt theo cách đóng kịch biểu lộ cảm xúc. Trong nghiên cứu [9] có 2 giọng nam và 2 giọng nữ, còn trong [13] có 6 người nói với 6 cảm xúc vui, bình thường, buồn, ngạc nhiên, tức giận, sợ hãi. Người thể hiện cảm xúc đều là các sinh viên Việt Nam. Các tác giả ban đầu đã xây dựng ngữ liệu này với ý định nghiên cứu chéo ngôn ngữ Việt Nam và Trung Quốc. Các tham số của ngữ liệu được phân tích phục vụ nhận dạng cảm xúc bao gồm cao độ (pitch), các formant F1, F2, F3 và năng lượng tín hiệu. GMM (Gaussian Mixture Model) đã được sủ dụng trong [9], [15], còn MRF (Markov Random Fields) được sử dụng trong [13] để nhận dạng cảm xúc.

Phần đầu bài báo sẽ mô tả vắn tắt phương pháp mới để xây dựng bộ ngữ liệu cảm xúc cho tiếng Việt nói với bốn cảm xúc cơ bản: bình thường, vui, buồn, tức giận. Để xây dựng ngữ liệu cảm xúc, có thể thực hiện theo các phương pháp như: ghi âm trực tiếp các đối thoại tự nhiên, xây dựng kịch bản sao cho các đối thoại được các nhân vật tùy biến cảm xúc theo tình huống, ghi âm trực tiếp giọng các nghệ sĩ diễn đạt các nội dung theo yêu cầu biểu đạt cảm xúc cho trước [20]. Phương pháp sau cùng đã được áp dụng để xây dựng ngữ liệu cảm xúc cho tiếng Đức [1] và cũng là phương pháp đã được chúng tôi chọn lựa để xây dựng ngữ liệu cảm xúc cho tiếng Việt. Đây là phương pháp cho phép chủ động xây dựng được ngữ liệu một cách hiệu quả.

Tiếp theo, thử nghiệm nhận dạng cảm xúc được thực hiện trên bộ ngữ liệu cảm xúc tiếng Việt đã xây dựng. Để nhận dạng cảm xúc cho tiếng nói thu âm từ một tổng đài trả lời tự động, Laurence Vidrascu [5] sử dụng máy hỗ trợ véc tơ SVM và mô hình cây logic (LMT: Logistic Model Tree). Kalyana Kumar Inakollu [11], sử dụng mô hình hỗn hợp Gauss đa thể hiện (GMM: Gaussian Mixture Model) với tiếng nói được mô hình hóa bởi các hệ số theo thang tần số Mel (MFCC: Mel Frequency Cepstral Coefficients) [12]. Thurid [16] sử dụng thông tin về giới tính để cải thiện hiệu năng của hệ thống nhận dạng cảm xúc.

Phần đầu của bài báo sẽ trình bày kết quả phân tích ảnh hưởng của các cảm xúc đến hai tham số cơ bản của tiếng nói là tần số cơ bản *F*0 [6], [4] và cường độ tiếng nói. Sau đó, bài báo trình bày việc thực hiện nhận dạng cảm xúc dựa trên một số bộ nhận dạng, bao gồm: *K* láng giềng gần nhất [14], phân tích phân biệt tuyến tính LDA [8], phân tích phân biệt toàn phương QDA, bộ phân lớp các véc tơ hỗ trợ SVC và máy véc tơ hỗ trợ SVM [19].

Nội dung tiếp theo của bài báo bao gồm:

Phần 2 trình bày phương pháp xây dựng bộ ngữ liệu cho tiếng Việt nói có cảm xúc.

Phần 3 trình bày các phương pháp nhận dạng cảm xúc và đánh giá, so sánh các phương pháp này.

Phần 4 phân tích ảnh hưởng của các cảm xúc đến hai tham số cơ bản của tiếng nói là tần số cơ bản *F*0 và cường độ tiếng nói.

Phần 5 đưa ra kết quả nhận dạng cảm xúc.

Cuối cùng phần 6 tổng kết và mô tả hướng nghiên cứu tiếp theo.

II. XÂY DỰNG NGỮ LIỆU CHO TIẾNG VIỆT NÓI CÓ CẢM XÚC

Bộ ngữ liệu này được xây dựng cho 4 cảm xúc: bình thường, vui, buồn, tức giận. Đầu tiên, chúng tôi chọn lựa kịch bản để diễn viên thể hiện được 4 cảm xúc một cách tự nhiên nhất. Kịch bản này được xây dựng với sự giúp đỡ của các nhà ngôn ngữ của Viện Ngôn ngữ Việt Nam. Kịch bản thu âm được xây dựng gồm 55 câu theo các tiêu chí sau:

Các câu cần được biểu lộ cả 4 cảm xúc khi nói, không chứa các từ ngữ cảm thán, biểu cảm về mặt cảm xúc. Với các câu không có từ cảm thán (ví dụ: *"Vườn hoa trước nhà", "Trường Đại học Bách khoa Hà Nội"*…) người nói sẽ tập trung vào việc biểu lộ cảm xúc mà không bị ảnh hưởng bởi nội dung của câu nói.

Kịch bản có các tổ hợp từ (ví dụ: *"Thật á"*) và các câu ngắn (ví dụ: *"Vườn hoa trước nhà"*), câu dài (ví dụ: *"À anh dám ăn nói với bố thế à"*) nhằm mục đích phân tích được ảnh hưởng của các tham số trên một từ riêng lẻ hay trên cả câu;

Kịch bản cố gắng lựa chọn các câu sao cho có càng nhiều âm tiết cơ bản của tiếng Việt càng tốt.

Có 56 giọng được thu âm, gồm 28 nữ và 28 nam là các diễn viên, nghệ sĩ lồng tiếng chuyên nghiệp, được lựa chọn theo các tiêu chí: có độ tuổi trải đều từ 18 đến 60 tuổi, có phân bố cân bằng giữa giọng nam và giọng nữ, có kinh nghiệm biểu đạt tốt, rõ ràng cảm xúc khi nói. Với mỗi cảm xúc, một câu sẽ được diễn đạt lặp lại 4 lần, được sắp xếp sao cho xuất hiện ngẫu nhiên để người nói có thể biểu lộ cảm xúc tốt nhất. Người nói được huấn luyện biểu diễn mỗi cảm xúc theo một cách thống nhất (cùng một kiểu vui, cùng một kiểu buồn..) dễ nhận ra hay dễ biểu lộ nhất để tránh tình trạng dữ liệu gồm rất nhiều cách biểu lộ khác nhau nhưng mỗi loại lại chỉ có vài câu gây khó khăn trong việc tìm quy luật.

Dữ liệu thu xong được xử lý trước bằng cách sử dụng công cụ cắt bỏ hết khoảng lặng ở đầu và cuối câu, được nghe nhanh một lượt để loại bỏ các câu bị lỗi trong quá trình thu hoặc cắt tự động.

Ngữ liệu được thu trong phòng thu âm, lồng tiếng chuyên nghiệp có hệ thống cách âm, lọc nhiễu tốt. Mỗi câu được lưu thành một file wav, tín hiệu thu được lấy mẫu ở tần số 16000Hz và 16 bit cho một mẫu. Mỗi giọng nói sẽ thu được 220 file cho một cảm xúc. Dữ liệu thu được gồm có 52800 file với tổng dung lượng là 2,68Gb.

III. CÁC PHƯƠNG PHÁP NHẬN DẠNG CẢM XÚC TIẾNG VIỆT NÓI

Trong phần này, bài báo trình bày các bộ phân lớp được thử nghiệm để nhận dạng cảm xúc cho tiếng Việt nói: *K* láng giềng gần nhất KNN, phân tích phân biệt tuyến tính LDA, phân tích phân biệt toàn phương QDA, bộ phân lớp các véc tơ hỗ trợ SVC và máy véc tơ hỗ trợ SVM [10].

3.1. Phương pháp phân tích phân biệt tuyến tính LDA

Giả sử các đối tượng thuộc vào *N* lớp. π_n là xác suất tiên nghiệm để một đối tượng đến từ lớp thứ *n*. $f_n(x) = P_r(X = x|Y = n)$ là hàm mật độ xác suất để đối tượng *X* lấy giá trị *x* khi đang ở lớp thứ *n*, giả định $f_n(x)$ là hàm chuẩn Gauss đa thể hiện (phương trình (1)).

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \quad (1)$$

Định lý Bayes [7] cho phép tính xác suất hậu nghiệm đối tượng thuộc vào lớp *n* khi có giá trị bằng *x* được mô tả ở phương trình (2).

$$Pr(Y = n|X = x) = \frac{\pi_n f_n(x)}{\sum_{l=1}^{N}\pi_l f_l(x)} \quad (2)$$

Đối tượng được nhận dạng vào lớp có giá trị xác suất hậu nghiệm lớn nhất (phương trình (2)) sẽ tương ứng với lớp này.

Với phương pháp phân tích phân biệt tuyến tính LDA, giả sử mỗi lớp có riêng giá trị kỳ vọng μ_n song tất cả các lớp đều có chung ma trận hiệp phương sai Σ. Thực hiện lấy logarit phương trình (4) sẽ thu được phương trình (3).

$$\delta_n(x) = x^T\Sigma^{-1}\mu_n - \frac{1}{2}\mu_n^T\Sigma^{-1}\mu_n + log\pi_n \quad (3)$$

Trong phương trình (3), $\delta_n(x)$ được gọi là hàm phân biệt (discriminant function). Vì $\delta_n(x)$ là hàm tuyến tính của x nên phương pháp này được gọi là phương pháp phân biệt tuyến tính. Các tham số μ_n và Σ_n được xác định dựa trên sự ước lượng tham số từ bộ dữ liệu huấn luyện.

3.2. Phân tích khác biệt toàn phương QDA

Với phương pháp này, giả sử mỗi lớp sẽ có một ma trận hiệp phương sai riêng Σ_n, khi đó hàm phân biệt sẽ được biểu diễn bằng phương trình (4).

$$\delta_n(x) = -\frac{1}{2}x^T\Sigma_n^{-1}x + x^T\Sigma_n^{-1}\mu_n - \frac{1}{2}\mu_n^T\Sigma_n^{-1}\mu_n - \frac{1}{2}log|\Sigma_n| + log\pi_n \quad (4)$$

Các tham số μ_n and Σ_n trong các phương trình (3) và (4) sẽ được xác định trong quá trình huấn luyện dựa vào các dữ liệu huấn luyện.

3.3. K láng giềng gần nhất KNN

Với mỗi đối tượng x trong tập thử nghiệm, tính giá trị $Y_l(x)$ theo phương trình (5).

$$Y_l(x) = \frac{1}{K}\Sigma_{x_i \in N_K(x)} y_i \quad (5)$$

Trong phương trình (5), $N_K(x)$ là láng giềng của x, bao gồm K điểm gần x nhất trong tập huấn luyện, y_i là trọng số của điểm trong tập huấn luyện y_i. Đối tượng x được nhận dạng vào lớp L nếu $Y_l(x)$ đạt giá trị lớn nhất khi so sánh với các giá trị $Y_l(x)$.

3.4. Bộ phân lớp phân biệt tuyến tính với lề cực đại (maximal margin classifier)

Lề cực đại được xác định như sau: với mỗi mẫu trong tập huấn luyện, tính khoảng cách trực giao đến biên giới phân lớp; lề là khoảng cách trực giao tối thiểu tìm được. Bộ phân lớp này chọn biên giới phân lớp có lề đạt giá trị lớn nhất, nghĩa là biên giới phân lớp phân biệt tốt nhất các mẫu trong tập huấn luyện. Các véc tơ nằm trên lề được gọi là các véc tơ hỗ trợ (support vector).

3.5. Bộ phân lớp hỗ trợ véc tơ SVC

Phương pháp này là sự mở rộng của bộ phân lớp phân biệt tuyến tính với lề cực đại (maximal margin classifier), cho phép phân lớp với các lớp không thể phân tách bằng một biên giới tuyến tính [21]. Phương pháp này sẽ tìm biên giới phân lớp phù hợp nhất với đa số các mẫu, và chấp nhận một số mẫu huấn luyện bị phân lớp sai (được điều chỉnh bằng tham số C – phương trình (7)). Phiên bản mở rộng của phương pháp này là máy hỗ trợ véc tơ SVM.

3.6. Máy hỗ trợ véc tơ SVM

Phương pháp SVC chỉ có khả năng tìm được biên giới phân lớp tuyến tính. Trong khi đó, biên giới phân lớp tuyến tính lại không phù hợp với một số dữ liệu cụ thể. Để vẫn có thể sử dụng biên giới phân lớp tuyến tính, một phương pháp được đề xuất là mở rộng số tham số biểu diễn đối tượng dựa trên các tham số đã có. SVM là phương pháp cho phép thực hiện hiệu quả sự mở rộng này với mức độ tính toán hợp lý.

Xét bài toán sử dụng SVM để phân chia các mẫu thành 2 lớp. Giả sử tập huấn luyện bao gồm N mẫu x_i, $i = 1,2,\dots,N$. Các mẫu này được phân vào lớp y_i, $i = 1,2,\dots,N$; y chỉ lấy các giá trị -1 hoặc 1. Biên giới phân lớp được biểu diễn bằng vế trái của phương trình (6).

$$f(x) = \beta_0 + \sum_{i=1}^{N}\alpha_i k(x, x_i) \quad (6)$$

Thực chất đa phần các giá trị α_i đều bằng 0, chỉ trừ những giá trị α_i của các véc tơ hỗ trợ. Các giá trị này bị giới hạn theo phương trình (7).

$$0 \le \alpha_i \le C, i = 1,2,\dots,N \quad (7)$$

C là giá trị cho phép các mẫu bị vi phạm. Khi C càng nhỏ thì lề sẽ càng rộng, và ngược lại khi C càng lớn thì lề sẽ càng hẹp.

k là hàm kernel của hệ thống, u và v là hai vec tơ của tập huấn luyện, với bộ phân lớp hỗ trợ véc tơ SVC thì k được tính theo phương trình (8).

$$k(u, v) = u^T v \quad (8)$$

Với SVM, hàm k được sử dụng để biến đổi không gian tham số, và được tính theo phương trình (9), trong đó γ là hệ số biến đổi của hàm k.

$$k(u, v) = \exp\{-\gamma|u - v|^2\} \quad (9)$$

Khi đó giải thuật thực hiện tìm các giá trị β_0 và α_i theo phương trình (10).

$$\min_{\beta_0,\alpha}\sum_{i=1}^{N}\big(1 - y_i f(x_i)\big) + \frac{1}{2}\alpha^T k\alpha \quad (10)$$

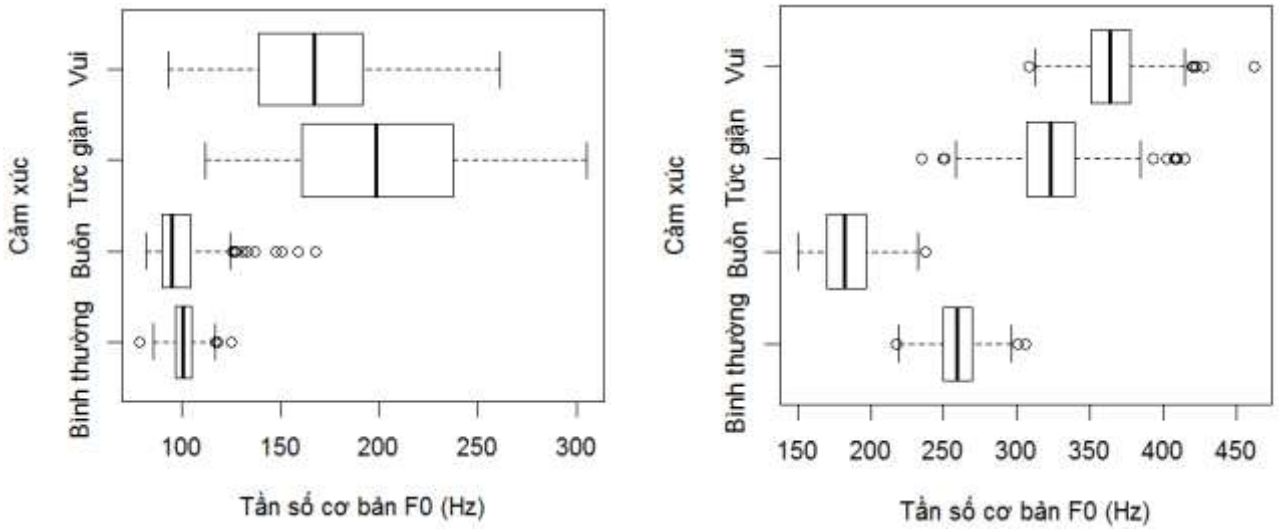
với k là ma trận $N \times N$ tính trên tất cả các cặp mẫu sử dụng trong quá trình huấn luyện.

Quá trình phân lớp được thực hiện tính hàm f (phương trình (6)) trên mẫu cần thử nghiệm. Tùy vào dấu của hàm f mà mẫu thử nghiệm sẽ được phân vào 1 trong 2 lớp.

Để áp dụng SVM cho bài toán phân lớp nhiều mẫu, phương pháp được sử dụng là one-versus-one: xây dựng $\binom{k}{2}$ bộ phân lớp cho từng cặp lớp. Mỗi mẫu thử nghiệm sẽ được đưa qua tất cả các bộ phân lớp này. Lớp nào chiếm đa số sẽ được coi là kết quả nhận dạng.

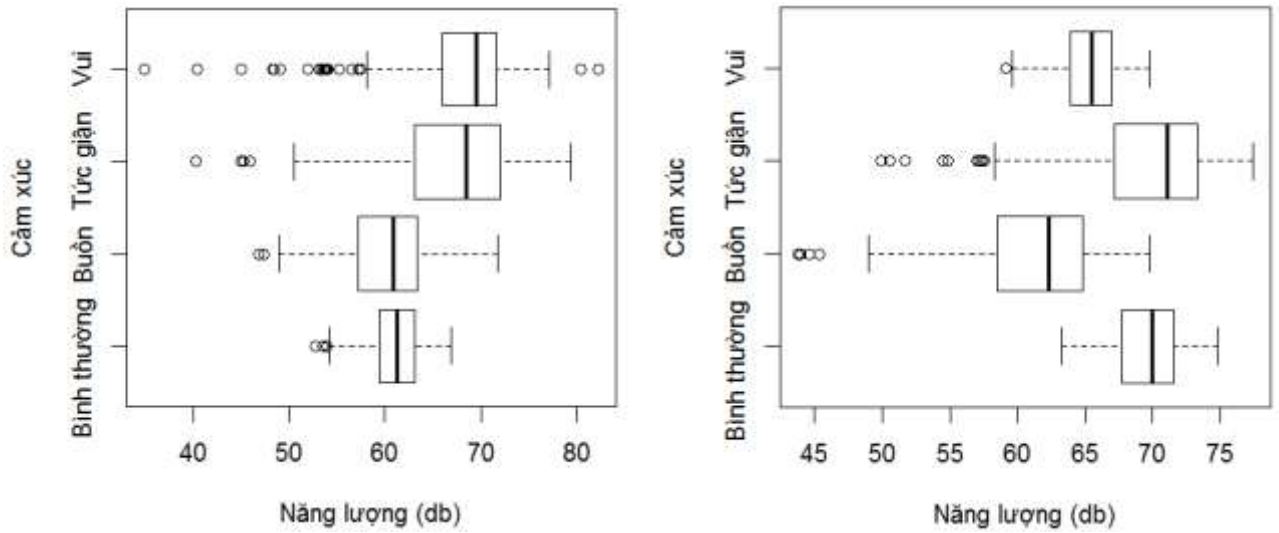
3.7. Nhận xét

Trong ba phương pháp đầu, phương pháp QDA thực hiện phân biệt giữa các lớp thông qua biên giới phân lớp tuyến tính, như vậy là biên giới phân lớp tương đối thô với các bộ dữ liệu phức tạp. Trong khi đó với phương pháp KNN, kết quả nhận dạng lại quá phụ thuộc vào một số mẫu nhất định (K mẫu) xung quanh mẫu cần nhận dạng. Vì thế, phương pháp KNN cho kết quả rất dao động theo bộ dữ liệu. Là một cải tiến của phương pháp LDA, phương pháp QDA cho phép tạo ra biên giới phân lớp phi tuyến, như vậy cho phép nhận dạng các mẫu mềm dẻo hơn.



Hình 1. Phân bố tần số cơ bản $F0$ trung bình theo 4 cảm xúc của nam nghệ sĩ Đ.K (hình trái) và nữ nghệ sĩ T.T.H (hình phải).

Các phương pháp trên đã sử dụng toàn bộ dữ liệu huấn luyện để xây dựng biên giới phân lớp. Trong khi đó, phương pháp SVM chỉ sử dụng các véc tơ hỗ trợ để quyết định biên giới phân lớp. Phương pháp sử dụng bộ phân lớp hỗ trợ véc tơ SVC chỉ sử dụng biên giới phân lớp tuyến tính, còn phương pháp SVM lại cho phép xây dựng biên giới phi tuyến với sự mở rộng số lượng tham số lớn. Về mặt thực chất, phương pháp SVC có thể coi là phương pháp SVM với hàm nhân tuyến tính (được tính theo phương trình 8). Trên cơ sở nhận xét trên, nhóm nghiên cứu đánh giá phương pháp QDA và SVM sẽ cho kết quả nhận dạng tốt nhất.



Hình 2. Phân bố cường độ tiếng nói trung bình theo 4 cảm xúc của nam nghệ sĩ Đ.K (hình trái) và nữ nghệ sĩ T.T.H (hình phải).

IV. ẢNH HƯỞNG CỦA CẢM XÚC ĐẾN TẦN SỐ CƠ BẢN F0 VÀ CƯỜNG ĐỘ TIẾNG NÓI

Thông thường, trong các hệ thống nhận dạng tiếng nói, các hệ số MFCC thường được sử dụng như là tham số đặc trưng. Tiếng Việt là ngôn ngữ có thanh điệu. Quy luật biến thiên tần số cơ bản *F0* khác nhau dẫn đến 6 thanh điệu khác nhau trong tiếng Việt. Từ đó có thể thấy tần số cơ bản đóng vai trò rất quan trọng đối với tiếng Việt nói. Mặt khác quy luật biến thiên của tần số cơ bản khác nhau cũng dẫn đến thể hiện các cảm xúc phân biệt đối với tiếng Việt nói như phân tích ở trên. Vì vậy, trong bài bài báo này chúng tôi mong muốn trước hết khảo sát ảnh hưởng của tham số *F0* kết hợp với cường độ tiếng nói để nhận dạng cảm xúc tiếng Việt.

Dựa trên cảm nhận chủ quan, hai nghệ sĩ nổi tiếng của Việt Nam là nghệ sĩ nam Đ.K (50 tuổi) và nữ nghệ sĩ T.T.H (34 tuổi) thể hiện các cảm xúc rất chân thật. Mỗi nghệ sĩ này thể hiện 55 câu, mỗi câu lặp lại 4 lần cho một cảm xúc. Như vậy, mỗi nghệ sĩ ghi âm 880 file tiếng nói. Giá trị *F*0 và cường độ tiếng nói được lấy trung bình trên từng file wav.

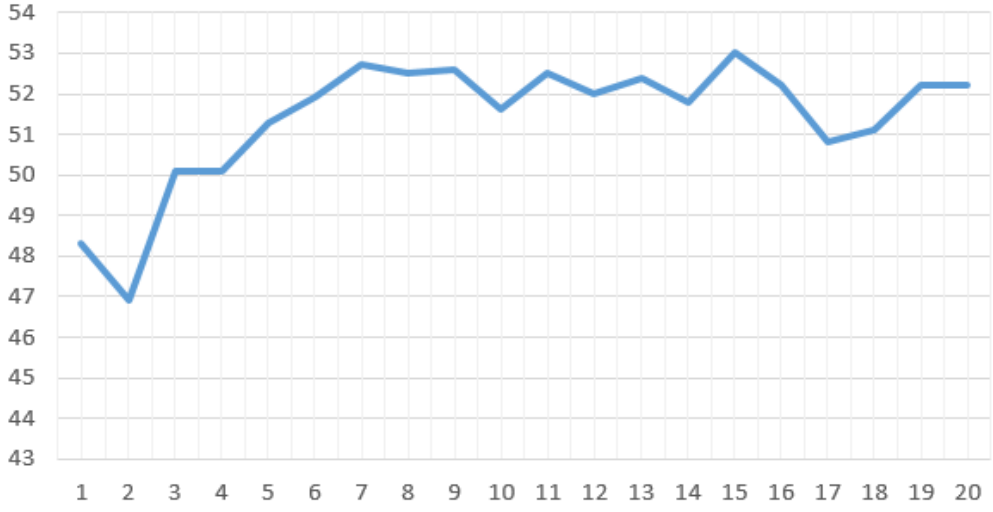
Hình 1 mô tả sự phân bố *F*0 và hình 2 mô tả sự phân bố cường độ tiếng nói theo từng cảm xúc dưới dạng đồ thị box-plot.

Hình 1 cho thấy tần số cơ bản *F*0 trung bình của cảm xúc buồn là thấp nhất, tiếp theo là của cảm xúc bình thường. Tần số *F*0 của cảm xúc vui và tức giận thì cao hơn. Trong 4 cảm xúc, tần số *F*0 của cảm xúc tức giận là lớn nhất với giọng nam và cảm xúc vui với giọng nữ.

Kết quả phân tích sự biến thiên của cường độ tiếng nói theo từng cảm xúc được mô tả ở hình 2. Hình 2 cho thấy có sự phân biệt rõ rệt về cường độ giữa cảm xúc vui/tức giận và cảm xúc buồn/bình thường. Ngoài ra, không có sự phân biệt rõ ràng về cường độ tiếng nói giữa cảm xúc buồn và cảm xúc bình thường, giữa cảm xúc vui và cảm xúc tức giận. Hơn nữa, với giọng nữ thì cảm xúc không được thể hiện rõ rệt qua cường độ tiếng nói. Chẳng hạn, cường độ trung bình của cảm xúc bình thường lại cao hơn so với cảm xúc vui.

V. THỬ NGHIỆM NHẬN DẠNG CẢM XÚC TIẾNG VIỆT NÓI

Ngữ liệu của 4 nghệ sĩ nam (Đ.A.N, Đ.K, H.P, L.V.H) và 4 nghệ sĩ nữ (B.H.G, Đ.T.H, N.B.T, T.T.H) đã được sử dụng để thử nghiệm nhận dạng. Hai thử nghiệm đã được thực hiện cho giọng nữ và cho giọng nam. Mỗi thử nghiệm được thực hiện theo phương pháp đánh giá chéo (cross-validation): 3 người nói được chọn để huấn luyện mô hình, số người nói còn lại được chọn để thử nghiệm nhận dạng; kết quả nhận dạng được tính trung bình cho 4 lần thực hiện. Mỗi file cảm xúc được biểu diễn bằng 2 tham số: tần số cơ bản *F*0 trung bình và cường độ tiếng nói trung bình.



Hình 3. Tỷ lệ nhận dạng cảm xúc đúng của thử nghiệm sử dụng bộ phân lớp KNN với giá trị *K* biến thiên từ 1 đến 20.

Đối với phương pháp nhận dạng cảm xúc sử dụng bộ phân lớp KNN, cần xác định giá trị *K* tối ưu (xem mục 3.3). Giá trị *K* này được xác định dựa trên thử nghiệm với giọng nam. Tập huấn luyện bao gồm 3 nghệ sĩ Đ.K, H.P, L.V.H. Tập thử nghiệm bao gồm ngữ liệu của nghệ sĩ Đ.A.N. Các giá trị *K* được thử nghiệm từ 1 đến 20. Kết quả của các thử nghiệm này được mô tả ở hình 3. Hình 3 cho thấy kết quả tốt nhất đạt được khi *K*=15. Giá trị này được sử dụng trong các thử nghiệm nhận dạng cảm xúc với phương pháp KNN.

Các kết quả thử nghiệm được trình bày ở bảng 1 cho thấy phương pháp KNN cho tỉ lệ nhận dạng thấp nhất (tuy nhiên có nhiều ngoại lệ). Trong khi đó, phương pháp QDA cho kết quả nhận dạng tốt hơn phương pháp LDA. Như vậy, có thể kết luận rằng biên giới phân lớp toàn phương cho kết quả nhận dạng chính xác hơn so với phương pháp sử dụng biên giới phân lớp tuyến tính (khi chỉ sử dụng bộ tham số gồm 2 thành phần là tần số cơ bản *F*0 và cường độ tiếng nói).

Bảng 1. Tỷ lệ phần trăm nhận dạng cảm xúc đúng

Phương pháp	Giọng nam	Giọng nữ
KNN : K=15	47,4	53,0
LDA	51,3	56,4
QDA	55,1	57,0
SVC : C=0.1	56,3	56,2
SVC : C=1	56,8	55,5
SVC : C=10	56,9	55,6
SVM : γ=0,5, C=0,1	53,4	58,1
SVM : γ=0,5, C=1	53,9	57,2
SVM : γ=0,5, C=10	53,0	56,8
SVM : γ=1, C=0,1	53,3	57,7
SVM : γ=1, C=1	53,0	57,1
SVM : γ=1, C=10	53,1	57,2

Trong các phương pháp thử nghiệm, phương pháp SVC cho kết quả nhận dạng tốt nhất với giọng nam và phương pháp SVM cho kết quả tốt nhất với giọng nữ (mặc dù không có sự cải thiện đáng kể khi so sánh với phương pháp QDA và SVC).

Bảng 2. Ma trận nhầm lẫn (tỷ lệ %) giữa các cảm xúc khi sử dụng phương pháp QDA trên giọng nam.

Kết quả nhận dạng của hệ thống	Tỉ lệ nhận dạng			
	Bình thường	Buồn	Tức giận	Vui
Bình thường	59,7	39,0	8,6	17,0
Buồn	38,6	60,7	3,0	3,0
Tức giận	0,0	0,3	41,4	36,8
Vui	1,7	0,0	47,0	43,2

Bảng 3. Ma trận nhầm lẫn (tỷ lệ %) giữa các cảm xúc khi sử dụng phương pháp QDA trên giọng nữ.

Kết quả nhận dạng của hệ thống	Tỉ lệ nhận dạng			
	Bình thường	Buồn	Tức giận	Vui
Bình thường	33,8	36,8	9,4	0,1
Buồn	47,8	62,6	0,6	0,0
Tức giận	18,4	0,6	56,7	27,6
Vui	0,0	0,0	33,3	72,3

Ma trận nhầm lẫn giữa các cảm xúc được thể hiện ở bảng 2 (cho giọng nam) và bảng 3 (cho giọng nữ), số liệu được cho trong hai bảng này là tỉ lệ nhận dạng đúng tính theo phần trăm. Với cùng một câu được thể hiện theo cảm xúc nào đó, hệ thống có thể nhận dạng nhầm sang các cảm xúc khác. Do đó, lấy tổng theo một hàng không nhất thiết phải bằng tổng số câu được dùng để nhận dạng hoặc không nhất thiết phải bằng 100% nếu tính theo tỷ lệ nhận dạng. Kết quả trên bảng 2 và bảng 3 cho thấy hầu hết các lỗi nhận dạng nhầm xảy ra giữa cảm xúc bình thường và cảm xúc buồn, giữa cảm xúc vui và cảm xúc tức giận. Điều này phù hợp với phân tích đã đưa ra ở mục 4.

VI. KẾT LUẬN

Bài báo đã mô tả phương pháp xây dựng ngữ liệu có cảm xúc cho tiếng Việt nói và việc phân tích tần số cơ bản *F0*, cường độ tiếng nói của ngữ liệu này cho thấy có thể phân biệt được hai nhóm cảm xúc bình thường/buồn và vui/tức giận. Việc thử nghiệm một số phương pháp phân lớp để nhận dạng cảm xúc tiếng Việt cũng đã được thực hiện. Biến thiên tần số cơ bản *F0* đóng vai trò quan trọng đối với tiếng Việt nói và đã được sử dụng kết hợp với cường độ tiếng nói như là tham số đặc trưng cho các bộ phân lớp. Kết quả cho thấy chỉ riêng tham số *F0* và cường độ tiếng nói đã cho tỉ lệ nhận dạng tốt nhất là 56,9% đối với giọng nam khi sử dụng phương pháp SVC còn đối với giọng nữ tỉ lệ này là 57,7% khi sử dụng phương pháp SVM.

Trong nghiên cứu tiếp theo, để có thể đề xuất mô hình đầy đủ cho nhận dạng cảm xúc của tiếng Việt nói, các tham số chi tiết hơn về nguồn âm, tuyến âm và các kỹ thuật nhận dạng tiên tiến khác sẽ được sử dụng nhằm tăng tỷ lệ nhận dạng cảm xúc đúng cho hệ thống như mô hình hỗn hợp Gauss đa thể hiện hay mạng nơ ron sâu.

VII. LỜI CẢM ƠN

Bài báo này được thực hiện trong khuôn khổ đề tài nghiên cứu khoa học cấp trường "Xây dựng bộ ngữ liệu cảm xúc tiếng Việt" của Trường Đại học Bách khoa Hà Nội. Các tác giả chân thành cảm ơn Trường Đại học Bách khoa Hà Nội, Phòng Khoa học Công nghệ, Viện Công nghệ Thông tin và Truyền thông đã hỗ trợ để chúng tôi có thể thực hiện thành công đề tài.

TÀI LIỆU THAM KHẢO

[1] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter Sendlmeier and Benjamin Weiss, "A Database of German Emotional Speech", In Proceeding of 9th European Conference on Speech Communication and Technology (INTERSPEECH 2005), pp. 1-4, 2005.

[2] Viet Hoang Anh, Manh Ngo Van, Bang Ban Ha, Thang Huynh Quyet, "A real-time model based Support Vector Machine for emotion recognition through EEG", In Processding of International Conference on Control, Automation and Information Sciences (ICCAIS), Ho Chi Minh City, pp. 191-196, 2012.

[3] Thi Duyen Ngo, The Duy Bui, "A study on prosody of Vietnamese emotional speech", In Proceedings of the Fourth International Conference on Knowledge and Systems Engineering, pp. 151-155, 2012.

[4] Indranil Chatterjee, Hindol Halder, Sayani Bari, Suman Kumar, Amitabha Roychoudhury, "An Analytical Study of Age and Gender Effects on Voice Range Profile in Bengali Adult Speakers using Phonetogram", Jaypee Journals, pp.65-70, 2011.

[5] Laurence Vidrascu, Laurence Devillers, "Detection of real-life emotions in call centers", In Proceeding of 9th European Conference on Speech Communication and Technology (INTERSPEECH 2005), pp. 1841-1844, 2005.

[6] S. Mwangi, Werner Spiegl, Florian Hoenig, T Haderlein, A. Maier, Elmar Noeth, "Effects of vocal aging on fundamental frequency and formants", In Proceedings of the International Conference on Acoustics (NAG/DAGA) , pp.1761-1764, 2009.

[7] Jay L. Devore, Probability and Statistics for Engineering and the Sciences, Eighth Edition, Brooks/Cole Edition, USA, 2010.

[8] Felix Burkhardt, Markus van Ballegooy, Klaus-Peter Engelbrecht, Tim Polzehl, Joachim Stegmann, "Emotion Detection in Dialog Systems: Applications, Strategies and Challenges", In Proceeding of 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009), pp. 1-6, 2009.

[9] La Vutuan, Huang Cheng-Wei, Ha Cheng, Zhao Li, "Emotional Feature Analysis and Recognition from Vietnamese Speech", Journal of Signal Processing, vol 29, issue 10, pp. 1423-1432, 2013.

[10] Prasad Reddy P. V. G. D, Prasad A, Srinivas Y, Brahmaiah P, "Gender Based Emotion Recognition System for Telugu Rural Dialects Using Hidden Markov Models", Journal of Computing, vol 2, issue 6, pp. 94-98, 2010.

[11] Kalyana Kumar Inakollu, Sreenath Kocharla, "Gender Dependent and Independent Emotion Recognition System for Telugu Speeches Using Gaussian Mixture Models", International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, issue 11, pp. 4172-4175, 2013.

[12] Igor Bisio, Alessandro Delfino, Fabio Lavagetto, Mario Marchese, And Andrea Sciarrone, "Gender-Driven Emotion Recognition Through Speech Signals for Ambient Intelligence Applications", IEEE transactions on Emerging topics in computing, vol. 1, no. 2, pp. 244-257, 2013.

[13] Jiang Zhipeng, Huang Chengwei, "High-Order Markov Random Fields and Their Applications in Cross-Language Speech Recognition", Cybernetics and Information Technologies, vol 15, no 4, pp. 50-57, 2015.

[14] Rahul B. Lanewar, Swarup Mathurkar, Nilesh Patel, "Implementation and Comparison of Speech Emotion Recognition System using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques", Procedia Computer Science, vol 49, pp. 50-57, 2015.

[15] Elif Bozkurt, Engin Erzin, Çidem Erolu Erdem, A. Tanju Erdem, "Improving Automatic Emotion Recognition from Speech Signals", In Proceeding of 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009), pp. 324-327, 2009.

[16] Thurid Vogt, Elisabeth André, "Improving Automatic Emotion Recognition from Speech via Gender Differentiation", In Proceedings of Language Resources and Evaluation Conference LREC, pp. 1123-1126, 2006.

[17] Dang-Khoa_Mac, Eric Castelli, Véronique Aubergé, "Modeling the Prosody of Vietnamese Attitudes for Expressive Speech Synthesis", In Processding of International workshop on Spoken Language Technologies for Under-resourced languages (SLTU 2012), pp. 114-118, 2012.

[18] Dang-Khoa Mac, Do-Dat Tran, "Modeling Vietnamese Speech Prosody: A Step-by-Step Approach Towards an Expressive Speech Synthesis System", Trends and Applications in Knowledge Discovery and Data Mining, pp. 273-287, 2015.

[19] Kun Han, Dong Yu, Ivan Tashev, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine", In Processding of International Speech Communication Association 2014, pp 223-227, 2014.

[20] Moataz El Ayadi, Mohamed S. Kamel, Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition Journal, vol 44, issue 3, pp. 572-587, 2011.

[21] Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning, 10th Edition, Springer, USA, 2013.

COMPARING PERFORMANCE OF SOME RECOGNITION METHODS FOR EMOTION RECOGNITION OF VIETNAMESE

Le Xuan Thanh, Dao Thi Le Thuy, Nguyen Hong Quang, Trinh Van Loan

ABSTRACT— *Emotional identification is an issue receiving the most interest in recent times. Recent studies have focused on a number of popular languages in the world. However, there is very little research on Vietnamese. In this paper, we describe the method to build a corpus of Vietnamese emotional speech and the preliminary evaluation of the distribution of F0 fundamental frequency and intensity for the corpus are also described. The variation of F0 plays an important role because this variation decides the six different tones of Vietnamese and takes part in the emotion expression. The fundamental frequency and intensity have been used firstly as feature parameters for different classifiers to perform the identification of Vietnamese emotions: KNN (K-Nearest Neighbor), LDA (Linear Discriminant Analysis), QDA (Quadratic Discriminant Analysis), SVC (Support Vector Classifier), and SVM (Support Vector Machine). The recognition results showed a significant proximity between the neutral emotion and the sad emotion, between the happy emotion and angry emotion. With only the feature parameters mentionned above, SVC method gave the best results for male voices; the correct emotion recognition rate is 56.9%. For female voices, SVM method gave the best result with the correct emotion recognition rate 57.7%.*

Keywords— *Vietnamese speech, Emotion recognition, F0, intensity, K-Nearest Neighbors, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Support Vector Machine.*