Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.0005

ẢNH HƯỞNG CỦA ĐẶC TRƯNG PHỔ TÍN HIỆU TIẾNG NÓI ĐẾN NHẬN DẠNG CẢM XÚC TIẾNG VIỆT

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

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

[2] Khoa Công nghệ Thông tin - Đại học Sư phạm Kỹ thuật Hưng Yên

thuydtl@soict.hust.edu.vn, loantv@ soict.hust.edu.vn, quangnh@ soict.hust.edu.vn, thanhlx@ soict.hust.edu.vn

TÓM TẮT: *Một vấn đề quan trọng đối với hệ thống nhận dạng cảm xúc tiếng nói là việc cần phải trích chọn các đặc trưng phù hợp của tín hiệu tiếng nói sao cho các đặc trưng này cho khả năng phân biệt hiệu quả các cảm xúc khác nhau. Bài báo này sử dụng phương pháp ANOVA và kiểm định T đối với một số đặc trưng phổ của bộ ngữ liệu cảm xúc tiếng Việt nói để đánh giá khả năng dựa trên các đặc trưng này để phân biệt 4 cảm xúc cơ bản vui, buồn, tức giận và bình thường. Kết quả thử nghiệm sử dụng mô hình GMM để nhận dạng 4 cảm xúc cho thấy có sự ảnh hưởng khác nhau của từng đặc trưng phổ đến tỷ lệ nhận dạng đúng các cảm xúc đồng thời tỷ lệ nhận dạng đúng tăng đáng kể khi có sự kết hợp các đặc trưng MFCC với các đặc trưng phổ.*

Từ khóa: *nhận dạng cảm xúc, tiếng Việt nói, đặc trưng phổ, MFCC, GMM.*

I. GIỚI THIỆU

Các nghiên cứu về cảm xúc tiếng nói có vai trò quan trọng trong các lĩnh vực tương tác người-máy. Đã có nhiều nghiên cứu nhận dạng cảm xúc tiếng nói cho một số ngôn ngữ khác nhau trên thế giới [1]. Phần lớn các nghiên cứu này thường sử dụng các đặc trưng tín hiệu tiếng nói theo 4 loại [1]: các đặc trưng mang tính liên tục (cao độ, năng lượng, formant), các đặc trưng mang tính chất lượng giọng nói (dễ nghe hay khó nghe, mức độ căng thẳng, mức độ lấy hơi), các đặc trưng phổ (LPC (Linear Prediction Coding), MFCC (Mel Frequency Cepstral Coefficients), LFPC (Log-frequency power coefficients)), các đặc trưng TEO (TEO-Teager-energy-operator) do Teager đề xuất (TEO-FM-Var (TEO-decomposed FM variation), TEO-Auto-Env (normalized TEO autocorrelation envelope area), TEO-CB-Auto-Env (critical bandbased TEO autocorrelation envelope area)). Các nghiên cứu về cảm xúc tiếng Việt hiện nay chủ yếu được thực hiện trên phương diện ngôn ngữ [2]. Theo phương diện xử lý tín hiệu, còn rất ít các công trình nghiên cứu về cảm xúc tiếng Việt nói. Một số nghiên cứu về cảm xúc tiếng Việt đã được công bố thường được thực hiện trên ngữ liệu đa thể thức, kết hợp video biểu hiện khuôn mặt, cử chỉ và tiếng nói với ứng dụng chủ yếu để tổng hợp tiếng Việt. Chẳng hạn nghiên cứu trong [3, 4] đã thử nghiệm mô hình hóa ngôn điệu tiếng Việt với ngữ liệu đa thể thức nhằm tổng hợp tiếng Việt biểu cảm. Nghiên cứu [5] đã sử dụng SVM (Support Vector Machines) để phân lớp với đầu vào là tín hiệu điện não (EEG) và kết quả cho thấy có thể nhận dạng được trên thời gian thực 5 trạng thái cảm xúc cơ bản với độ chính xác trung bình là 70,5%. Ngoài ra, có số ít nghiên cứu về cảm xúc tiếng Việt nói song được thực hiện ở nước ngoài và không phải chủ yếu do người Việt thực hiện [6, 7]. Nhìn chung, các nghiên cứu nhận dạng cảm xúc thường dùng hỗn hợp bốn đặc trưng tín hiệu tiếng nói đã được nêu trên. Tuy vậy, còn hiếm thấy các nghiên cứu xét riêng ảnh hưởng của các đặc trưng trong miền tần số đến kết quả nhận dạng cảm xúc.

Đối với nhận dạng cảm xúc tiếng nói, việc tìm kiếm xác định các tham số đặc trưng của tín hiệu tiếng nói để nhận dạng có hiệu quả là điều rất quan trọng. Trong bài báo này, chúng tôi thực hiện đánh giá ảnh hưởng của các đặc trưng phổ tín hiệu tiếng nói đến nhận dạng cảm xúc và xem xét các lợi ích của việc kết hợp các đặc trưng MFCC để nhận dạng bốn cảm xúc vui, buồn, tức giận và bình thường cho tiếng Việt nói. Bài báo gồm 5 phần. Phần 2 trình bày các đặc trưng phổ của tín hiệu tiếng nói đã được sử dụng trong thử nghiệm của bài báo. Phần 3 nêu phương pháp ANOVA và kiểm định *T* để đánh giá các đặc trưng phổ đã nêu. Các kết quả thử nghiệm dùng mô hình GMM với các đặc trưng phổ để nhận dạng cảm xúc tiếng Việt nói được trình bày trong phần 4. Cuối cùng, các kết luận được rút ra trong phần 5.

CÁC ĐẶC TRƯNG PHỔ CỦA TÍN HIỆU TIẾNG NÓI VÀ NGỮ LIỆU CẢM XÚC DÙNG CHO THỬ NGHIỆM

Với tín hiệu tiếng nói, các hệ số MFCC và LPCC (Linear Predictive Cepstral Coefficients) là các tham số phổ biến được dẫn xuất từ miền cepstral đặc trưng cho thông tin tuyến âm. Những đặc điểm này của tuyến âm cũng được biết đến như là các đặc trưng phổ hoặc đặc trưng hệ thống. Nói chung, các đặc trưng phổ được xem như có độ tương quan rất lớn giữa cấu hình thay đổi của tuyến âm và tốc độ di chuyển của các thành phần tham gia phát âm. Thông tin chuyên biệt về cảm xúc hiện diện trong chuỗi cấu hình của tuyến âm đóng vai trò quan trọng trong việc tạo ra các đơn vị âm khác nhau với các cảm xúc khác nhau. Các hệ số MFCC, LPCC và các đặc trưng formant đã được biết đến một cách rộng rãi như là các đặc trưng của hệ thống. Trong bài báo này, các đặc trưng thống kê của phổ được trích chọn từ ngữ liệu cảm xúc bằng công cụ Praat gồm các đặc trưng như sau (trong ngoặc đơn là các ký hiệu sẽ được dùng trong bài báo để chỉ các đặc trưng này): các thành phần hài (harmonicity), trọng tâm phổ (central gravity), mômen trung tâm (central moment), độ bất đối xứng (skewness), độ nhọn (kurtosis), độ lệch chuẩn (stand_dev), giá trị trung bình (mean), độ dốc (slope) và độ lệch chuẩn (stddevi) của phổ trung bình dài hạn LTAS (Long Term Average Spectrum).

Các thành phần hài đại diện cho mức độ tuần hoàn còn được gọi là tỷ lệ sóng hài-nhiễu HNR (Harmonics-to-Noise Ratio). Harmonicity được biểu diễn bằng thang đo dB. Nếu 99% năng lượng của tín hiệu nằm trong chu kỳ và 1% là nhiễu thì HNR là 10 * log10 (99/1) = 20dB. Nếu HNR bằng 0dB có nghĩa là năng lượng trong sóng hài và trong nhiễu bằng nhau [18].

Nếu $S(f)$ là phổ phức, trong đó f là tần số thì trọng tâm phổ được cho bởi công thức (1):

$$\frac{\int_0^\infty f|S(f)|^p df}{\int_0^\infty |S(f)|^p df} \quad (1)$$

Trong đó $\int_0^\infty |S(f)|^p df$ là năng lượng. Như vậy, trọng tâm phổ là trung bình của tần số trên toàn bộ miền tần số với trọng số là $|S(f)|^p$. Khi $p = 2$, trọng số là phổ công suất, còn $p = 1$ trọng số là trị tuyệt đối của phổ. Giá trị thường được dùng là $p = 2/3$. Trọng tâm phổ là phép đo tần số trung bình của tần số trong phổ. Đối với tín hiệu hình sin ở tần số 377 Hz thì trọng tâm phổ là 377 Hz. Đối với nhiễu trắng ở tần số 22050 Hz thì trọng tâm phổ là 5512,5 Hz, tức là bằng nửa tần số Nyquist.

Nếu $S(f)$ là phổ phức thì mômen phổ trung tâm thứ n được cho bởi công thức (2) với f_c là trọng tâm phổ.

$$\frac{\int_0^\infty (f - f_c)^n |S(f)|^p df}{\int_0^\infty |S(f)|^p df} \quad (2)$$

Mômen trung tâm thứ n là giá trị trung bình của $(f - fc)^n$ trên toàn bộ miền tần số với trọng số là $|S(f)|^p$. Mômen liên quan đến bậc n trong công thức (2). Nếu $n = 2$ ta có phương sai của các tần số trong phổ. Độ lệch chuẩn tần số chính là căn bậc hai của phương sai này.

Nếu $n = 3$ ta sẽ có mômen phổ trung tâm bậc 3, đó cũng chính là độ bất đối xứng skewness không chuẩn hóa của phổ. Để chuẩn hóa, cần chia cho 1,5 công suất của mômen bậc hai. Skewness cho biết độ lệch của tập dữ liệu so với phân bố chuẩn, nếu độ lệch nằm dưới giá trị trung bình thì dữ liệu tập trung hơn so với độ lệch nằm trên giá trị trung bình. Độ bất đối xứng skewness của một phân bố xác suất là độ đo sự bất đối xứng của phân bố đó. Giá trị tuyệt đối của skewness càng cao thì phân bố đó càng bất đối xứng. Một phân bố đối xứng có skewness bằng 0.

Với $n = 4$, ta có kurtosis của phổ không chuẩn hóa. Để chuẩn hóa cần chia cho bình phương của mômen bậc hai và trừ đi 3. Kurtosis là một chỉ số để đánh giá đặc điểm hình dáng của một phân bố xác suất. Cụ thể, kurtosis so sánh độ cao phần trung tâm của một phân bố so với phân bố chuẩn. Phần trung tâm của phân bố càng cao và nhọn thì chỉ số kurtosis của phân bố đó càng lớn. Phân bố chuẩn có kurtosis bằng 3.

Giá trị trung bình của phổ liên quan đến độ lệch chuẩn của phổ. Với bài toán phân lớp, khi một tập các giá trị của dữ liệu có xu hướng phân bố gần giá trị trung tâm thì mức độ tập trung của dữ liệu tốt hơn so với tập dữ liệu có xu hướng phân bố xa giá trị trung tâm. Như vậy, giá trị trung bình có thể là hữu ích để mô tả tập các giá trị của dữ liệu có mối tương quan với nhau. Trung bình của các giá trị $x_1, \ldots, x_N$ là:

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j \quad (3)$$

Các đặc trưng phổ trung bình dài hạn LTAS, độ dốc (slope) của LTAS và độ lệch chuẩn (stddevi) của LTAS cũng được xem xét trong bài báo này.

Ngữ liệu cảm xúc tiếng Việt nói dùng cho thử nghiệm trong bài báo gồm 5584 file của bộ ngữ liệu cảm xúc tiếng Việt BKEmo [8] với bốn cảm xúc vui, buồn, tức giận, bình thường của 8 giọng nam và 8 giọng nữ. Các file cảm xúc được lựa chọn với 22 câu có nội dung khác nhau. Trong số các câu đó, có câu ngắn, câu dài, câu cảm thán như "*Có lương rồi*", "*Ôi dào, người như vậy không thay đổi được đâu*" để phân tích các tham số đặc trưng của cảm xúc. Mỗi câu được nói 4 lần. Số lượng file cảm xúc của mỗi giọng nam và nữ là 2792 file, mỗi cảm xúc có 698 file. Tập ngữ liệu được chia đôi, một nửa dùng để huấn luyện (2792 file), nửa còn lại để thử nghiệm (2792 file).

II. PHÂN TÍCH CÁC ĐẶC TRƯNG PHỔ BẰNG PHƯƠNG PHÁP ANOVA VÀ KIỂM ĐỊNH *T*

A. Phương pháp phân tích ANOVA và kiểm định T

1. Phân tích phương sai one-way ANOVA

Các phân tích ANOVA [10], thường được xem như là tập hợp của các tình huống thực nghiệm và các thủ tục thống kê để phân tích các đáp ứng có tính định lượng từ các đơn vị thử nghiệm. Bài toán ANOVA đơn giản được gọi với các tên khác nhau như nhân tố đơn (single-factor) hoặc one-way ANOVA. Bài toán ANOVA đơn giản liên quan đến việc phân tích trong trường hợp các dữ liệu được lấy mẫu từ hai quần thể trở lên hoặc khi dữ liệu được lấy từ các thử nghiệm trong đó phương pháp xử lý được sử dụng là nhiều hơn 2. Các đặc tính phân biệt các phương pháp xử lý

hoặc các quần thể với nhau gọi là các nhân tố (factor) được dùng trong nghiên cứu, còn các phương pháp xử lý khác nhau hoặc quần thể khác nhau được gọi là các mức độ của nhân tố. Phân tích one-way ANOVA được sử dụng trong trường hợp chỉ có một yếu tố nào đó được xem xét nhằm xác định ảnh hưởng của nó đến một yếu tố khác. Yếu tố được xem xét ảnh hưởng sẽ được dùng để phân loại các quan sát thành các nhóm khác nhau.

Phương pháp one-way ANOVA thực hiện so sánh các giá trị thống kê (giá trị trung bình) của nhiều tập dữ liệu. Giả sử I là số tập dữ liệu cần so sánh, $\mu_1, \dots, \mu_I$ là các giá trị trung bình của từng tập dữ liệu, giả thuyết cần kiểm định là:

H_0: $\mu_1 = \mu_2 = \cdots \mu_I$ (giá trị trung bình của các tập dữ liệu bằng nhau),

H_a: ít nhất 1 trong 2 giá trị μ_i khác nhau.

Nếu H_0 là đúng, các J quan sát trong mỗi cá thể từ một quần thể phân bố chuẩn thông thường có cùng một giá trị trung bình μ. Trong trường hợp đó, giá trị trung bình của các cá thể $\overline{x_{1.}}, \overline{x_{2.}}, \dots, \overline{x_{I.}}$ là gần với nhau. Các thủ tục kiểm tra dựa trên việc so sánh phép đo độ chênh lệch giữa các $\overline{x_{i.}}$ so với phép đo độ biến đổi được tính toán từ mỗi mẫu. Để kiểm định các giả thuyết trên, cần tính giá trị trung bình bình phương $MSTr$ (Mean Square for Treatments) và trung bình bình phương lỗi MSE (Mean Square for Error) theo các công thức (4) và (5).

$$MSTr = \frac{J}{I-1}[(\bar{X}_{1.} - \bar{X}_{..})^2 + (\bar{X}_{2.} - \bar{X}_{..})^2] + \cdots + (\bar{X}_{I.} - \bar{X}_{..})^2$$
$$= \frac{J}{I-1}\sum_i (\bar{X}_{i.} - \bar{X}_{..})^2 \quad (4)$$

$$MSE = \frac{S_1^2 + S_2^2 + \cdots + S_I^2}{I} \quad (5)$$

Trong công thức (4), I là số tập dữ liệu và J là số giá trị đo cho mỗi tập dữ liệu. $\bar{X}_{i.}$là giá trị trung bình trên mẫu thứ i, $\bar{X}_{..}$là giá trị trung bình trên toàn bộ dữ liệu. Trong công thức (5), S_i^2 là phương sai mẫu thứ i. Thử nghiệm thống kê cho one-way ANOVA là $F = MSTr/MSE$.

2. Kiểm định *T*

Kết quả phân tích phương sai ANOVA loại bỏ giả thuyết H_0 và chấp nhận H_a, như vậy sẽ có các cặp giá trị $\mu_i - \mu_j$ của các tập dữ liệu khác nhau. Khi đó, cần biết chính xác những cặp giá trị nào có sự khác biệt đáng kể. Để kiểm định điều này, một trong những phương pháp được sử dụng phổ biến là kiểm định *T* (Tukey's test [10]). Kiểm định *T* sử dụng phân bố Student để đánh giá các cặp giá trị $\mu_i - \mu_j$. Các khoảng tin cậy của các cặp giá trị $\mu_i - \mu_j$ được tính để so sánh. Khoảng tin cậy của giá trị này được mô tả ở phương trình (6) với $Q_{(\alpha,I,I(J-1))}$ là giá trị của phân bố Student tại mức ý nghĩa α.

$$\bar{X}_i - \bar{X}_j - Q_{(\alpha,I,I(J-1))}\sqrt{MSE/J} \leq \mu_i - \mu_j \leq \bar{X}_i - \bar{X}_j + Q_{(\alpha,I,I(J-1))}\sqrt{MSE/J} \quad (6)$$

với mọi i, j và ($i = 1, \dots, I$ và $j = 1, \dots, I$) với ($i < j$).

Giả sử $I = 4$, ta cần tính khoảng tin cậy cho 6 cặp: $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, $\mu_1 - \mu_4$, $\mu_2 - \mu_3$, $\mu_2 - \mu_4$, $\mu_3 - \mu_4$. Ngoài ra $P - value$ cũng được tính cho các trường hợp này. Gọi F là tỷ lệ trung bình bình phương lỗi, P là xác suất thống kê thử nghiệm, có thể có một giá trị lớn hơn hoặc bằng giá trị thống kê kiểm định ($P > F$).

B. Kết quả phân tích và kiểm định các đặc trưng phổ

1. Kết quả phân tích ANOVA

Với nhận dạng cảm xúc, thử nghiệm trong bài báo là xem xét yếu tố đặc trưng phổ có ảnh hưởng đến sự phân loại bốn cảm xúc cơ bản hay không, nếu có ảnh hưởng thì giá trị trung bình của đặc trưng phổ của các tập cảm xúc sẽ là khác nhau. Phương pháp one-way ANOVA được sử dụng để phân tích sự ảnh hưởng của các đặc trưng phổ đến sự phân biệt các cảm xúc với nhau. Ngữ liệu cảm xúc đã giới thiệu ở phần trên được đưa vào thử nghiệm. Giả thuyết cần kiểm định là các giá trị kỳ vọng của bốn tập cảm xúc vui, buồn, tức giận và bình thường là như nhau, nghĩa là không có sự khác biệt giữa bốn cảm xúc. Giả thuyết đối lập: có ít nhất hai cặp cảm xúc có sự khác biệt với nhau.

Giả sử cần kiểm định xem tham số đặc trưng skewness của phổ có ảnh hưởng đến sự phân biệt các cảm xúc hay không, các giá trị MSTr và MSE sẽ được tính toán dựa trên các giá trị trung bình và phương sai của tham số này. Tương tự với các tham số đặc trưng khác, các kết quả phân tích thống kê F và P-value của các tham số đặc trưng phổ được trình bày trong Bảng 1.

Bảng 1 cho thấy giá trị $P - value$ của các tham số rất nhỏ, có nghĩa là xác suất sẽ rất thấp để cho $\mu_1 = \mu_2 = \mu_3 = \mu_4$. Như vậy, giả thuyết H_0 bị bác bỏ và có thể khẳng định rằng có sự phân biệt các cảm xúc dựa vào các tham số trên.

Bảng 1. Giá trị thống kê F và $P-value$ của phân tích ANOVA với các tham số đặc trưng phổ

Thứ tự	Tham số	Giá trị F	Giá trị $P-value$
1	Harmonicity	218,4038	5,7727E-134
2	Centre_gravity	1317,4812	0,0
3	Stand_dev	1397,2756	0,0
4	Skewness	1114,7564	0,0
5	Kurtosis	594,1112	0,0
6	Cmoment	1664,5398	0,0
7	Mean	783,2338	0,0
8	Slope	1218,3402	0,0
9	Stddevi	751,0346	0,0

2. Kết quả kiểm định *T*

Kết quả phân tích ANOVA với bốn cảm xúc cho thấy có thể phân biệt được các cảm xúc dựa trên các tham số đặc trưng phổ của tín hiệu. Vậy làm sao để biết những cặp cảm xúc nào có thể phân biệt được với nhau dựa vào các tham số đó? Để tìm ra các cặp cảm xúc này, chúng tôi tiến hành thử nghiệm bằng phương pháp kiểm định *T*, kết quả thống kê trong Bảng 2.

Bảng 2. Giá trị $P-value$ của kiểm định *T* với các tham số đặc trưng phổ cho các cặp cảm xúc

Thứ tự	Tham số	$P-value$					
		BT-Buồn	BT-Tức	BT-Vui	Buồn-Tức	Buồn-Vui	Tức-Vui
1	Harmonicity	3,7683E-09	3,7683E-09	2,8624E-06	3,7683E-09	3,7683E-09	3,7683E-09
2	Centre_gravity	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	1,3024E-06
3	Stand_dev	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
4	Skewness	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
5	Kurtosis	3,7683E-09	3,7683E-09	6,4688E-09	3,7683E-09	3,7683E-09	3,7683E-09
6	Central monent	0,010095	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
7	Mean	3,7683E-09	0,20606	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
8	Slope	3,7683E-09	3,8234E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
9	Stddevi	3,7683E-09	0,61003	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09

Bảng 2 cho thấy giá trị $P-value$ là rất nhỏ khi đánh giá cho từng tham số đối với từng cặp cảm xúc. Điều này cho thấy các tham số trên có ảnh hưởng đến sự phân biệt các cặp cảm xúc với nhau. Nhìn chung, $P-value$ của các cặp cảm xúc Buồn-Tức, Buồn-Vui đều có giá trị nhỏ khá đồng đều hơn so với $P-value$ của các cặp cảm xúc còn lại. Các cặp cảm xúc còn lại cũng được phân biệt rõ, tuy nhiên các tham số mean, stddevi có ảnh hưởng ít hơn với cặp cảm xúc BT-Tức. Điều này cũng được phản ánh thông qua kết quả nhận dạng cảm xúc được trình bày ở phần sau trong đó tỷ lệ nhận dạng cảm xúc tăng lên khi bổ sung các đặc trưng phổ.

III. THỬ NGHIỆM NHẬN DẠNG BẰNG MÔ HÌNH GMM

A. Mô hình hỗn hợp Gauss (GMM- Gauss Mixture Model)

Để nhận dạng cảm xúc tiếng nói, đã có nhiều bộ phân lớp được dùng như HMM (Hiden Markov Model), GMM, SVM (Support Vector Machine), ANN (Artificial Neural Network), KNN (K-Nearest Neighbors) và nhiều bộ phân lớp khác [1]. Trên thực tế, không có sự thỏa thuận nào về một bộ phân lớp nào đó là thích hợp nhất cho nhận dạng cảm xúc. Bởi vì mỗi bộ phân lớp có ưu thế và hạn chế riêng của nó.

Theo khía cạnh thống kê để nhận dạng mẫu, mỗi lớp được mô hình hóa bằng phân bố xác suất dựa trên ngữ liệu huấn luyện sẵn có. Các bộ phân lớp thống kê đã được sử dụng trong nhiều ứng dụng nhận dạng tiếng nói như HMM, GMM. Mô hình GMM là mô hình xác suất để đánh giá mật độ bằng cách sử dụng tổ hợp lồi của các mật độ chuẩn đa thể hiện. GMM có thể được xem như HMM liên tục đặc biệt chỉ chứa một trạng thái [14]. GMM rất hiệu quả khi mô hình hóa các phân bố đa thể thức và các yêu cầu về việc huấn luyện ít hơn nhiều so với yêu cầu của HMM liên tục tổng quát. Do vậy, GMM là thích hợp hơn so với HMM cho nhận dạng cảm xúc tiếng nói khi chỉ có đặc trưng tổng quan được trích rút từ tiếng nói dùng cho huấn luyện. Tuy nhiên, GMM không thể mô hình hóa cấu trúc thời gian của ngữ liệu huấn luyện bởi vì các phương trình huấn luyện và nhận dạng đều dựa trên giả thiết rằng tất cả các vectơ là độc lập. Trên thực tế, GMM đã được dùng khá phổ biến cho các trường hợp định danh người nói [11], định danh ngôn ngữ [12], định danh phương ngữ [13] hoặc phân lớp thể loại âm nhạc [15]. Trong trường hợp nhận dạng cảm xúc, mỗi cảm xúc sẽ được mô hình hóa bằng một mô hình GMM và bộ các tham số sẽ được xác định thông qua việc huấn luyện trên tập mẫu học.

Giả sử với một phát ngôn của cảm xúc j tương ứng có K khung tiếng nói và mỗi khung tiếng nói trích chọn được véctơ đặc trưng $\boldsymbol{x}_i$ có D chiều. Như vậy, một phát ngôn của cảm xúc j sẽ tương ứng với tập $\boldsymbol{X}$ chứa K véctơ đặc trưng $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_K\}$. Giả thiết các véctơ đặc trưng phù hợp với phân bố Gauss trong đó phân bố này được xác

định bởi trung bình và độ lệch so với giá trị trung bình. Từ đó, phân bố các đặc trưng của cảm xúc j có thể được mô hình hóa bằng hỗn hợp các phân bố Gauss. Mô hình hỗn hợp các phân bố Gauss λ_j của cảm xúc j sẽ bằng tổng có trọng số của M phân bố thành phần được xác định bởi xác suất:

$$p(\boldsymbol{X}|\lambda_j) = \sum_{m=1}^{M} g_m N(\boldsymbol{X}; \boldsymbol{\mu}_m, \Sigma_m) \quad (7)$$

Trong (7), g_m là các trọng số của hỗn hợp thỏa mãn điều kiện $\sum_{m=1}^{M} g_m = 1$, $N(\boldsymbol{X}; \boldsymbol{\mu}_m, \Sigma_m)$ là các hàm mật độ thành phần với phân bố Gauss D thể hiện có dạng:

$$N(\boldsymbol{X}; \boldsymbol{\mu}_m, \Sigma_m) = \frac{1}{(2\pi)^{D/2}|\Sigma_m|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_m)^T \Sigma_m{}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_m)} \quad (8)$$

Trong (8), $\boldsymbol{\mu}_m$ là véctơ trung bình $\boldsymbol{\mu}_m \in \mathbb{R}^D$ còn Σ_m là ma trận hiệp phương sai $\boldsymbol{\Sigma}_m \in \mathbb{R}^{D\times D}$. Như vậy, mô hình GMM λ_j cho cảm xúc j được xác định bởi bộ ba: các véctơ trung bình, các ma trận hiệp phương sai và các trọng số cho M thành phần: $\lambda_j = \{\boldsymbol{\mu}_m, \Sigma_m, g_m\}_j, m = 1,2,\dots,M$. Trên thực tế, việc xác định mô hình GMM λ_j của cảm xúc j sẽ được thực hiện theo thuật giải cực đại kỳ vọng (EM: Expectation-Maximization). Thuật giải này sẽ xác định cực đại khả hiện (ML: Maximum- Likelihood) của log khả hiện $\log(p(\boldsymbol{X}|\lambda_j)$ [17].

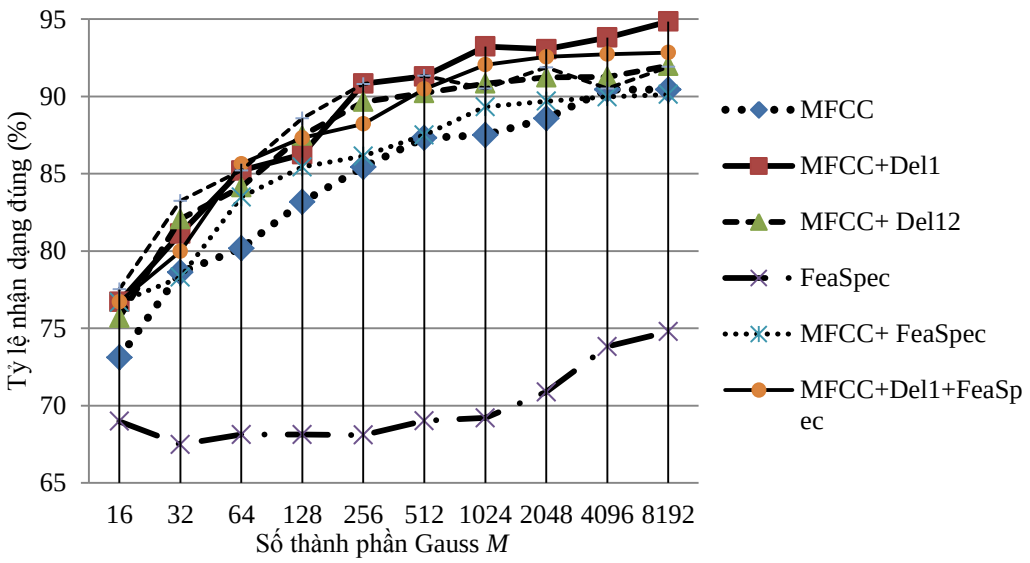
B. Nhận dạng kết hợp MFCC với các đặc trưng phổ

Phần này sẽ trình bày các kết quả thử nghiệm sử dụng MFCC, các đạo hàm bậc một và đạo hàm bậc hai của MFCC và các tham số đặc trưng phổ để nhận dạng bốn cảm xúc. Các hệ số MFCC (19 hệ số) cùng đạo hàm của các hệ số này được trích chọn bằng bộ công cụ Alize [11], còn các đặc trưng phổ được trích chọn bằng bộ công cụ Praat [18]. Mỗi thử nghiệm được tiến hành với số thành phần Gauss M tăng từ 16 đến 8192 theo lũy thừa 2 ($M = 2^n$, $n = 4,5,\dots,13$).

Có 7 trường hợp thử nghiệm đã được thực hiện và có thể chia 7 trường hợp này thành 2 nhóm như sau. Nhóm thứ nhất gồm 4 thử nghiệm: sử dụng chỉ MFCC, MFCC + Del1 (đạo hàm bậc nhất của MFCC), MFCC+Del12 (đạo hàm bậc nhất và bậc hai của MFCC), sử dụng chỉ các đặc trưng phổ FeaSpec. Nhóm thứ 2 gồm 3 thử nghiệm: sử dụng MFCC+FeatSpec, MFCC+Del1+FeaSpec, MFCC+Del12+FeaSpec.

Kết quả đạt được của nhóm thứ nhất như sau. Đối với trường hợp chỉ sử dụng MFCC, tỷ lệ nhận dạng chính xác đạt được trong khoảng 73,1% - 90,44%. Khi kết hợp MFCC với đạo hàm bậc nhất (Del1), tỷ lệ nhận dạng chính xác tăng lên và nằm trong khoảng 76,72% - 93,8%. Với trường hợp dùng MFCC và có cả đạo hàm bậc nhất, đạo hàm bậc hai (Del12), tỷ lệ nhận dạng chính xác thuộc khoảng 75,68% - 91,26%. Nếu chỉ dùng các tham số đặc trưng phổ (FeaSpec) thì tỷ lệ nhận dạng sẽ thấp hơn và có giá trị từ 67,48% - 73,82%.

Đối với nhóm thứ hai, kết quả thử nghiệm đạt được như sau. Tỷ lệ nhận dạng chính xác từ 76,68% - 89,97% khi sử dụng MFCC+FeaSpec, tỷ lệ này là từ 76,72% - 92,73% khi sử dụng MFCC+Del1+FeaSpec và từ 77,51% - 91,87% khi sử dụng MFCC+Del12+FeaSpec.

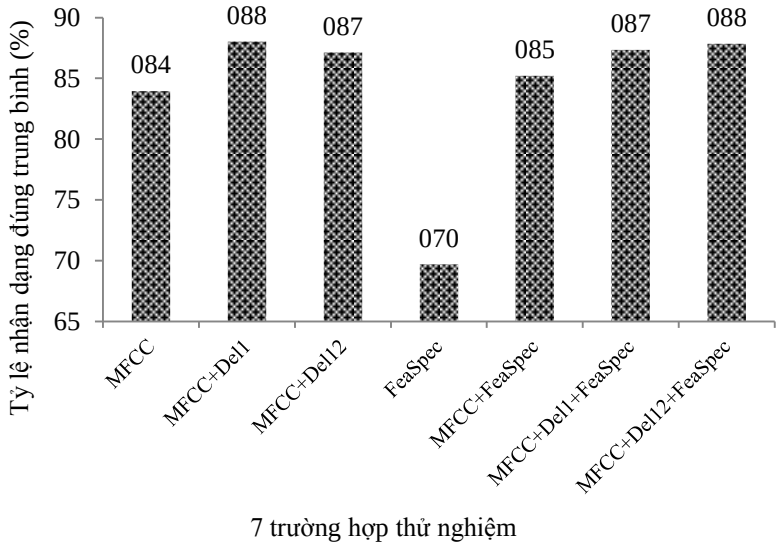


Hình 1. Tỷ lệ nhận dạng đúng cảm xúc với tham số MFCC và các đặc trưng phổ

Hình 1 cho thấy, nhìn chung khi số thành phần Gauss M tăng thì tỷ lệ nhận dạng cũng tăng lên. Với giá trị của M từ 16 đến 256, tỷ lệ nhận dạng đúng trung bình của bốn cảm xúc đạt được khi sử dụng đầy đủ bộ tham số gồm MFCC+Del12+FeaSpec nói chung đều cao hơn so với các trường hợp không sử dụng đầy đủ bộ tham số này. Còn với

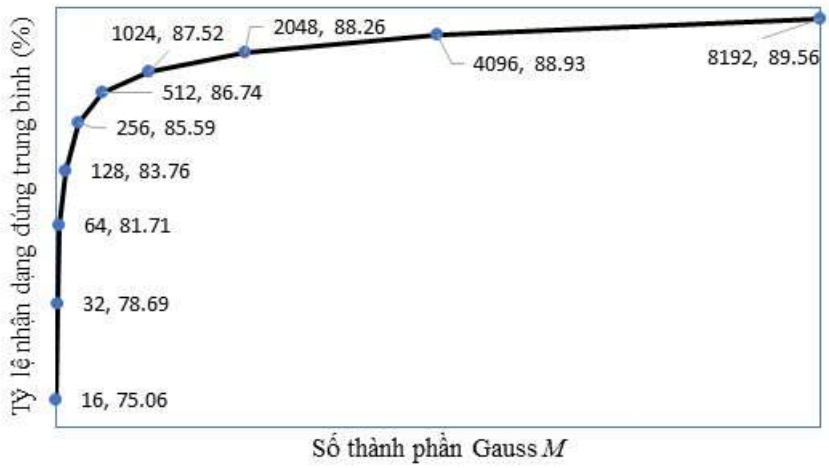
giá trị *M* từ 256 đến 8192, tỷ lệ nhận dạng đúng trung bình bốn cảm xúc khi sử dụng MFCC+Del1 cao hơn so với các trường hợp còn lại.

Hình 2 thống kê tỷ lệ nhận dạng đúng trung bình cho 7 thử nghiệm. Tỷ lệ nhận dạng đúng trung bình khi chỉ dùng đặc trưng phổ là thấp nhất và bằng 69,71%. Tỷ lệ nhận dạng đúng trung bình đạt cao nhất bằng 88,03% khi dùng MFCC+Del1. Nếu dùng MFCC+Del12 thì tỷ lệ nhận dạng là 87,16% và tỷ lệ này tăng 0,71% khi có kết hợp với đặc trưng phổ FeaSpec. Việc kết hợp với đặc trưng phổ đều làm tăng tỷ lệ nhận dạng trong 2 trường hợp MFCC+FeaSpec và MFCC+Del12+FeaSpec.



Hình 2. Tỷ lệ nhận dạng đúng trung bình cho 7 trường hợp thử nghiệm

Hình 3 là quan hệ giữa số thành phần Gauss *M* và tỷ lệ nhận dạng đúng trung bình cho 7 thử nghiệm đã nêu trên. Hình 3 cho thấy, với giá trị *M* thấp thì tỷ lệ nhận dạng đúng tăng lên đáng kể. Khi *M* thay đổi từ 16 đến 256, tỷ lệ nhận dạng đúng trung bình tăng 10,53%. Khi *M* tăng từ 512 đến 8192, tỷ lệ nhận dạng đúng trung bình tăng chỉ 2,82%.



Hình 3. Quan hệ giữa số thành phần Gauss *M* và tỷ lệ nhận dạng đúng trung bình

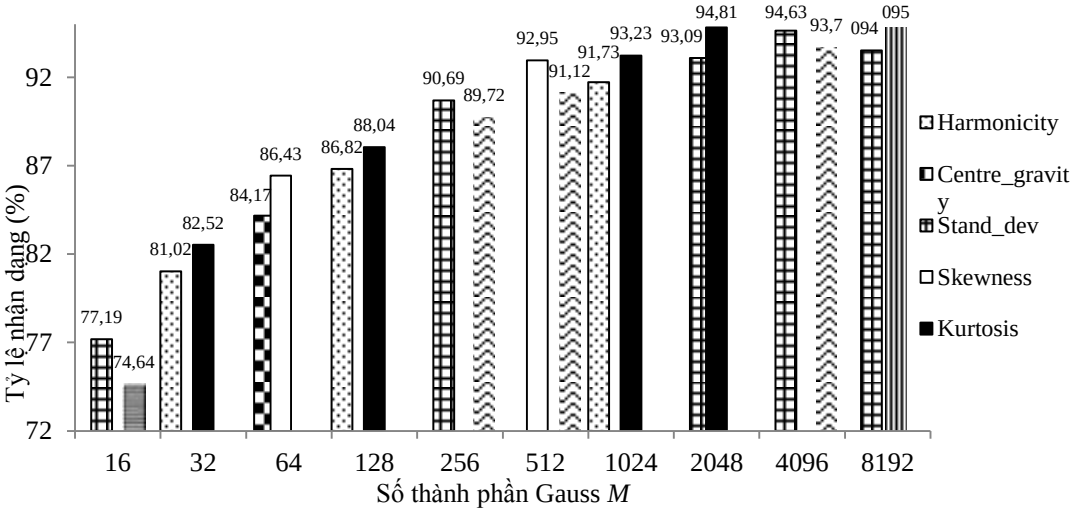
Có thể thấy rằng, khi *M* tăng đủ lớn (khoảng trên 512), mô hình GMM hầu như đã đạt tới mức xấp xỉ việc mô hình hóa các cảm xúc nên tỷ lệ nhận dạng đúng trung bình tăng theo dạng bão hòa khi tăng *M*.

Việc xác định tối ưu các thành phần Gauss *M* là quan trọng nhưng đó cũng lại là bài toán khó [16]. *M* càng tăng thì thời gian tính toán cũng tăng theo. Tùy từng bộ tham số đưa vào nhận dạng mà giá trị tối ưu của *M* cần được lựa chọn thích hợp tùy theo thời gian tính toán cần thiết và độ chính xác nhận dạng theo yêu cầu.

C. Nhận dạng kết hợp MFCC với từng đặc trưng phổ

Để đánh giá ảnh hưởng của riêng từng đặc trưng phổ, lần lượt từng đặc trưng phổ đã được đưa vào mô hình GMM với số thành phần Gauss *M* thay đổi từ 16 đến 8192. Kết quả cho thấy, tỷ lệ nhận dạng đúng trung bình đạt cao nhất là 88,61% đối với đặc trưng skewness cho các giá trị của *M* từ 16 đến 8192. Tỷ lệ nhận dạng đúng trung bình thấp

nhất là 87,77% đối với đặc trưng harmonicity. Hình 4 là tỷ lệ nhận dạng đúng cao nhất và thấp nhất tương ứng với đặc trưng phổ và các giá trị khác nhau của số thành phần Gauss M. Với từng giá trị của M, biểu đồ trên Hình 4 chỉ biểu diễn tỷ lệ nhận dạng đúng cao nhất và thấp nhất tương ứng với đặc trưng phổ.



Hình 4. Tỷ lệ nhận dạng đúng cao nhất và thấp nhất tương ứng với đặc trưng phổ cho các giá trị của M

Từ Hình 4 có thể nhận xét: đặc trưng độ lệch chuẩn của LTAS không xuất hiện trên biểu đồ, nghĩa là tỷ lệ nhận dạng cao nhất hoặc thấp nhất cho từng giá trị của M không thuộc về đặc trưng này. Trong khi đó, đặc trưng kurtosis có bốn lần xuất hiện tỷ lệ nhận dạng đúng cao nhất ứng với các giá trị của $M = 32, 128, 1024, 2048$. Đặc trưng độ lệch chuẩn của phổ có ba lần xuất hiện tỷ lệ nhận dạng đúng cao nhất ứng với các giá trị của $M = 16, 256, 4096$. Đặc trưng skewness có hai lần xuất hiện tỷ lệ nhận dạng đúng cao nhất ứng với các giá trị của $M = 64, 512$. Còn lại, đặc trưng mômen trung tâm của phổ chỉ xuất hiện một lần có tỷ lệ nhận dạng cao nhất ứng với $M = 8192$. Có thể suy diễn lý do để đặc trưng kurtosis có số lần xuất hiện nhiều nhất ứng với tỷ lệ nhận dạng đúng cao nhất như sau. Bản chất của đặc trưng kurtosis là đánh giá độ nhọn phần trung tâm của phân bố phổ so với phân bố chuẩn. Trong khi đó, GMM là mô hình gồm tổ hợp tuyến tính các phân bố chuẩn. Chính vì vậy, phương thức xác định đặc trưng kurtosis khá tương đồng với phương thức mô hình hóa của GMM.

Kết quả đánh giá ảnh hưởng của từng đặc trưng phổ khi được kết hợp với MFCC+Del1 được trình bày ở Bảng 3. Bảng này cho thấy, khi đặc trưng kurtosis của phổ được kết hợp, tỷ lệ nhận dạng trung bình cao nhất đối với cảm xúc vui là 88,80% và cảm xúc bình thường là 86,26%. Khi kết hợp với đặc trưng skewness, cả cảm xúc buồn và cảm xúc tức giận đều cho tỷ lệ nhận dạng trung bình cao nhất lần lượt là 91,49% và 90,82%.

Bảng 3. Tỷ lệ nhận dạng trung bình của M khi kết hợp MFCC+Del1 với từng đặc trưng phổ cho các cảm xúc

Thứ tự	Tham số	Tỷ lệ (%) nhận dạng đúng cho từng cảm xúc			
		Vui	Buồn	Tức giận	Bình thường
1	Harmonicity	88,41	90.43	89.41	85.20
2	Centre_gravity	88,78	90.76	89.31	85.09
3	Stand_dev	88,73	90.26	90.30	85.86
4	Skewness	89.14	**91.49**	**90.82**	85.13
5	Kurtosis	**88.80**	91.12	90.37	**86.26**
6	Cmonent	88.44	90.99	89.70	84.89
7	Mean	89.17	91.10	89.11	84.67
8	Slope	88.74	91.06	88.87	85.53
9	Stddevi	88.48	90.46	90.13	85.65

IV. KẾT LUẬN

Phương pháp thống kê ANOVA và kiểm định T đã được sử dụng và cho thấy các tham số đặc trưng phổ đều cho khả năng phân biệt 4 cảm xúc khác nhau của tiếng Việt nói. Điều này cũng được thể hiện thông qua kết quả nhận dạng các cảm xúc dựa trên mô hình GMM trong đó các tham số của mô hình là MFCC kết hợp với các đặc trưng phổ. Trong số các đặc trưng phổ harmonicity, centre_gravity, stand_dev, skewness, kurtosis, mean, slope và stddevi, đặc trưng kurtosis của phổ tỏ ra có ảnh hưởng quan trọng hơn đến tỷ lệ nhận dạng đúng các cảm xúc. Đối với 2 cảm xúc Buồn và Tức giận, đặc trưng skewness cho tỷ lệ nhận dạng đúng cao hơn cả. Tỷ lệ nhận dạng đúng cũng cao hơn đối với 2 cảm xúc Vui và Bình thường khi sử dụng đặc trưng kurtosis. Kết quả thử nghiệm cũng cho thấy việc lựa chọn số thành phần Gauss M cho mô hình GMM cần phải được cân nhắc dựa trên bộ tham số đặc trưng của mô hình và yêu cầu cụ thể của bài toán nhận dạng cảm xúc.

Được sự hỗ trợ của Trung tâm Nghiên cứu khoa học và Ứng dụng công nghệ - Trường Đại học Sư phạm Kỹ thuật Hưng Yên, bài báo này đã được hoàn thành. Nhóm tác giả xin bày tỏ sự cảm ơn đối với Trung tâm về sự hỗ trợ đó.

TÀI LIỆU THAM KHẢO

[1] Moataz El Ayadi, Mohamed S. Kamel, Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition 44 (2011) 572–587, 2011.

[2] Đỗ Tiến Thắng, "Ngữ điệu tiếng Việt sơ khảo", NXB Đại học Quốc gia Hà Nội, 2009.

[3] Dang-Khoa_Mac, Eric Castelli, Véronique Aubergé, "Modeling the Prosody of Vietnamese Attitudes for Expressive Speech Synthesis", Workshop of Spoken Languages Technologies for Under-resourced Languages (SLTU 2012), Cape Town, South Africa, May 7-9, 2012.

[4] Dang-Khoa Mac, Do-Dat Tran, "Modeling Vietnamese Speech Prosody: A Step-by-Step Approach Towards an Expressive Speech Synthesis System", Springer, Trends and Applications in Knowledge Discovery and Data Mining, vol 9441, Springer, pp. 273-287, 2015.

[5] 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", International Conference on Control, Automation and Information Sciences (ICCAIS), Ho Chi Minh city, Vietnam, Nov 26-29, 2012.

[6] La Vutuan, Huang Cheng-Wei, Ha Cheng, Zhao Li, "Emotional Feature Analysis and Recognition from Vietnamese Speech", Journal of Signal Processing, China, 2013.

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

[8] Lê Xuân Thành, Đào Thị Lệ Thủy, Trịnh Văn Loan, Nguyễn Hồng Quang, "Cảm xúc trong tiếng nói và phân tích thống kê ngữ liệu cảm xúc tiếng Việt", Tạp chí Công nghệ Thông tin và Truyền thông, trang 86-98, tập V-1 số 15 (35), 2016.

[9] Lê Xuân Thành, Đào Thị Lệ Thủy, Trịnh Văn Loan, Nguyễn Hồng Quang, "So sánh hiệu năng một số phương pháp nhận dạng cảm xúc tiếng Việt nói", Kỷ yếu Hội nghị Quốc gia lần thứ IX về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR) – Cần Thơ, ISBN: 978-604-913-472-2, trang 656-662, 2016.

[10] Jay L. Devor, "Probability and Statistics for Engineering and the Sciences", Eighth Edition, Brooks/Cole Edition, 2010.

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

[12] 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.

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

[14] D. Reynolds, C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Trans, Speech Audio Process, 3 (1) 72–83, 1995.

[15] Bağcı U., Erzin E., "Boosting Classifiers for Music Genre Classification", In: Yolum., Güngör T., Gürgen F., Özturan C. (eds) Computer and Information Sciences – ISCIS, Lecture Notes in Computer Science, vol 3733. Springer, Berlin, Heidelberg, 2005.

[16] Ayadi M. E. , Kamel M. S., and Karray F., "Survey on speech emotion recognition: Features,classification schemes, and databases", Pattern Recognition, vol. 44, pp. 572–587, 2011.

[17] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models", International Computer Science Institute, 1998.

[18] www.praat.org.

INFLUENCE OF SPECTRAL FEATURES OF SPEECH SIGNAL ON EMOTION RECOGNITION OF VIETNAMESE

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

ABSTRACT: *An important issue for speech emotion recognition systems is the need to extract appropriate features of speech signals so that these features give the ability to differentiate efficiently different emotions. This paper uses the ANOVA method and the T test for some of the spectral features of the Vietnamese emotion corpus to assess the ability based on these features to distinguish the four basic emotions: happiness, sadness, anger and neutrality. Using the GMM model to identify the four emotions, the test results showed that there were a different effect of each spectral feature on the exact recognition rates and the considerable increase of exact recognition rates with the combination of MFCC and spectral features.*