Cảm xúc trong tiếng nói và phân tích thống kê ngữ liệu cảm xúc tiếng Việt

Speech Emotions and Statistical Analysis for Vietnamese Emotion Corpus

Lê Xuân Thành, Đào Thị Lệ Thủy, Trịnh Văn Loan, Nguyễn Hồng Quang

Abstract: *Research on emotional speech has been carried out for many languages over the world and for Vietnamese, there was a beginning. This paper describes some research results on main features of four basic emotions: happiness, sadness, anger and neutrality. Our preliminary research on emotions of Vietnamese shows that in general anger and happiness correspond to speech energy and fundamental frequency higher than the one of neutral emotion, the sad emotion has the lowest values for energy and fundamental frequency. These comments come from the statistical methods such as analysis of variance (ANOVA) and Tukey's test applied for our Vietnamese emotion corpus. The classifiers SMO, lBk, trees J48 have been used for preliminary identification of emotions based on BKEmo corpus. The highest recognition rate is 98.17% for the classifier lBk using 384 feature parameters and this rate decreases to 82.59% for the case using only 48 parameters relating to the F0 and intensity.*

Keywords: *Speech, emotions, Vietnamese, corpus, ANOVA, Tukey's test, fundamental frequency, speech energy, recognition, SMO, lBk, trees J48.*

I. GIỚI THIỆU

Tiếng nói ngày càng được sử dụng rộng rãi trong giao tiếp giữa người và máy. Việc trao đổi thông tin tiếng nói cũng chuyển từ việc phải sử dụng các cấu trúc chặt chẽ sang dùng các cách thức giao tiếp linh hoạt hơn, điều này giúp cho ứng dụng tiếng nói được phổ biến đến người dùng phổ thông một cách dễ dàng hơn. Sự linh hoạt này không chỉ thể hiện ở việc sử dụng các cấu trúc câu lệnh linh hoạt mà còn hướng tới thể hiện ở các cung bậc cảm xúc khác nhau trong giao tiếp người máy. Để làm được điều này, các hệ thống tương tác người máy cần được trang bị thêm các tính năng mới. Các tính năng này bao gồm việc phân tích nội dụng của dữ liệu tiếng nói nhận được để lấy ra các thông tin như: cảm xúc trong câu lệnh, nội dung câu lệnh rồi đưa ra các phản hồi với nội dung và cảm xúc phù hợp. Chính vì vậy nghiên cứu về cảm xúc trong tiếng nói trở nên rất quan trọng trong lĩnh vực tương tác người máy.

Hiện nay, các nghiên cứu về tiếng nói tiếng Việt với giọng trần thuật (bình thường) đã có nhiều kết quả rất tốt. Trong khi đó các nghiên cứu về phương diện cảm xúc trong tổng hợp hay nhận dạng tiếng Việt chưa nhiều. 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 [23], [24] đã thử nghiệm mô hình hóa ngôn điệu tiếng Việt với ngữ liệu đa thể thức nhằm tổng hợp tiếng Việt biểu cảm. Các tác giả của [20] đã đề xuất mô hình biến đổi tiếng Việt nói để tạo biểu cảm trong kênh tiếng nói cho nhân vật ảo nói tiếng Việt. Trong nghiên cứu này, ngữ liệu có cảm xúc bao gồm các phát âm tiếng Việt của một nghệ sĩ nam và một nghệ sĩ nữ phát âm 19 câu ở năm trạng thái cơ bản: tự nhiên, vui, buồn, hơi giận, rất giận. Đối với nhận dạng cảm xúc tiếng Việt, nghiên cứu [21] đã sử dụng SVM (Support Vector Machines) để phân lớp với đầu vào là tín hiệu điện não (EEG). 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%. Một số tác giả Trung Quốc [28], [29] có kết hợp với sinh viên Việt Nam xây dựng ngữ liệu cảm xúc tiếng Việt theo cách đóng kịch biểu lộ cảm xúc. Người nói là các sinh viên Việt Nam, trong nghiên cứu [28] có 2 nam, 2 nữ, còn trong [29] có 6 người nói với 6 cảm xúc vui, bình thường, buồn, ngạc nhiên, tức giận, sợ hãi. Các tác giả ban đầu đã xây dựng ngữ liệu 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 [28] còn MRF (Markov Random Fields) được sử dụng trong [29] để nhận dạng cảm xúc.

Những tham số cơ bản nhất để phân biệt các cảm xúc bao gồm tần số cơ bản *F0*, năng lượng tiếng nói [7]. Sự phân biệt này có thể được xác minh thông qua cách sử dụng các phương pháp phân tích và kiểm định giả thuyết thống kê. Bài báo này sẽ trình bày về kết quả nghiên cứu sử dụng phương pháp phân tích ANOVA và kiểm định *T* để giới thiệu phần thử nghiệm phân lớp cảm xúc.

Nội dung tiếp theo của bài báo gồm các phần sau: Phần II trình bày về các tham số cơ bản đặc trưng cho cảm xúc trong tiếng nói; Phần III mô tả phương pháp xây dựng ngữ liệu tiếng Việt có cảm xúc; Phần IV sử dụng phương pháp phân tích phương sai ANOVA và kiểm định *T* để đưa ra kết quả phân tích thống kê sự khác biệt của các cảm xúc theo tần số cơ bản *F0* và năng lượng tiếng nói; Phần V trình bày kết quả thử nghiệm nhận dạng cảm xúc tiếng Việt; Phần VI tổng kết và định hướng nghiên cứu tiếp theo.

II. CÁC THAM SỐ VỀ CẢM XÚC TRONG TIẾNG NÓI

Trong giao tiếp thông thường giữa người với người, ngoài nội dung của thông điệp trao đổi thì người nghe cũng thu được rất nhiều thông tin thông qua các cảm xúc của người nói lúc đó. Vì vậy, trong giao tiếp người máy cần phát triển các hệ thống tiếng nói có thể xử lý các cảm xúc kèm theo nội dung cần truyền tải. Các mục tiêu cơ bản của hệ thống xử lý tiếng nói có cảm xúc là nhận dạng cảm xúc thể hiện trong tiếng nói và tổng hợp cảm xúc mong muốn trong tiếng nói để truyền tải ý định nội dung. Từ góc độ kỹ thuật, để làm được điều này, cần phải tìm được các tham số đặc trưng về cảm xúc trong tiếng nói nói chung và trong tiếng nói tiếng Việt nói riêng. Sau đó đưa ra được các mô hình tổng hợp, nhận dạng tiếng nói có cảm xúc.

Cảm xúc của con người không thể đo lường một cách chính xác bằng các phương tiện đo đạc bình thường. Vì vậy, các phương pháp phân tích nhận dạng và tổng hợp đối với cảm xúc đặt ra các thách thức đối với con người cũng như đối với máy tính. Cowie và Schroder đã chỉ ra rằng không thể phân biệt một cách rõ ràng các loại cảm xúc khác nhau [1]. Tuy nhiên đã có rất nhiều nghiên cứu về phân loại cảm xúc trong tiếng nói và các nhà nghiên cứu hiện đã đưa ra hơn 300 trạng thái cho những cảm xúc khác nhau [2], trong khi đó có tác giả lại thống kê 107 loại cảm xúc [30]. Liên hệ với tiếng Việt cũng dễ thấy đối với chỉ một cảm xúc được coi là buồn lại có thể được phân nhánh thành buồn bã, buồn bực, buồn rười rượi, buồn thiu, buồn tênh, v.v.. [31]. Cũng có nhiều tác giả thống nhất với quan điểm cho rằng một cảm xúc bất kỳ có thể được phân giải thành các cảm xúc cơ bản theo kiểu phân tích màu bất kỳ thành các màu cơ bản. Các cảm xúc cơ bản là: tức giận, chán ghét, sợ hãi, vui, buồn, ngạc nhiên [17]. Miwa và cộng sự [18] đã định nghĩa 6 cảm xúc và gán chúng vào nhóm bốn cảm xúc chủ yếu là: vui, buồn, tức giận, bình thường. Trong khuôn khổ bài báo này, chúng tôi cũng đi theo hướng như vậy bằng cách tập trung vào 4 loại cảm xúc mang tính đại diện là vui, buồn, tức giận và bình thường.

Về mặt sinh lý của cơ chế tạo cảm xúc, người ta đã phát hiện ra rằng với biểu hiện của các cảm xúc hưng phấn cao như giận dữ, vui, sợ hãi, hệ thống thần kinh sẽ được kích thích làm cho tim đập nhanh hơn, huyết áp cao hơn, có sự thay đổi trong hơi thở, áp suất không khí trong phổi ứng với phần dưới thanh môn lớn hơn và làm khô miệng. Kết quả là tiếng nói sẽ to hơn, nhanh hơn và năng lượng ở phạm vi tần số cao lớn

hơn, trung bình tần số cơ bản sẽ cao hơn và phạm vi biến thiên cũng rộng hơn [3]. Mặt khác, đối với những cảm xúc hưng phấn thấp như buồn bã, hệ thần kinh được kích thích gây ra sự sụt giảm nhịp tim, huyết áp, dẫn đến tăng tiết nước bọt, nói chậm và tần số cơ bản sẽ giảm với năng lượng tần số cao là nhỏ. Vì vậy, các đặc tính âm học như pitch, năng lượng, nhịp điệu, chất lượng giọng nói, và tín hiệu tiếng nói có độ tương quan lớn với những cảm xúc chính [4].

Về mặt kỹ thuật, có rất nhiều nghiên cứu đưa ra các tham số khác nhau ảnh hưởng đến cảm xúc trong nhận dạng và tổng hợp tiếng nói, các thông số này sẽ được phân tích để tìm ra các quy luật ảnh hưởng đến cảm xúc của từng ngôn ngữ khác nhau.

Đường bao *F0* là một thông số rất quan trọng theo những nghiên cứu của [5], nó được khẳng định lại trong các nghiên cứu về tiếng Đức của Burkhardt và Sendlmeier trong [6] và tiếng Hà Lan của Mozziconacci và Hermes trong [7].

Thời hạn là một trong những tham số ảnh hướng nhiều nhất đến cảm xúc theo Cahn [8] và cùng kết hợp với đường bao *F0* là đủ để phân biệt các cảm xúc bình thường, vui, buồn, giận dữ, chán nản, sợ hãi và phẫn nộ trong tiếng Hà Lan [9]. Nghiên cứu trong [10] cũng tham khảo mối quan hệ giữa đường bao *F0*, tốc độ phát âm, cường độ và cao độ ảnh hưởng đến tiếng nói tổng hợp có cảm xúc trong ngôn ngữ Malayalam.

Đặc tính phổ đã được sử dụng thành công cho các nghiên cứu tiếng nói khác nhau như phát triển hệ thống nhận dạng tiếng nói và nhận dạng người nói. Nghiên cứu cho thấy các đặc tính MFCC (Mel-Frequency Cepstral Coefficients) bậc thấp hơn sẽ mang thông tin về âm vị trong khi đó các đặc tính bậc cao thì chứa các thông tin không phải về tiếng nói. Tổ hợp các hệ số MFCC, LPCC (Linear Predictive Cepstral Coefficients), RASTA PLP (Relative Spectral Transform - Perceptual Linear Prediction) và các hệ số logarit của công suất đối với tần số đã được xem là tập các đặc điểm để phân loại các cảm xúc: tức giận, chán, bình thường, vui, buồn trong tiếng phổ thông Trung Quốc [11]. SVM cũng được dùng để nhận dạng 3 cảm xúc vui, buồn, bình thường của tiếng Trung Quốc [16] sử dụng các tham số như năng lượng, tần số cơ bản, LPCC, MFCC và MEDC (Mel-Energy spectrum Dynamic Coefficients). [17] sử dụng các tham số LPC, MFCC với thuật giải OSALPC (linear prediction of the causal part of the autocorrelation sequence algorithm) cho mô hình GMM (Gaussian Mixture Model) trên ngữ liệu tiếng Đức (Emo-DB) đạt được độ chính xác trung bình 89% cho 7 cảm xúc. Các tham số sử dụng cho mô hình GMM và K-NN (K-Nearest Neighbor) gồm: các hệ số MFCC, đặc trưng sóng con của tiếng nói và tần số cơ bản *F0* cũng được nghiên cứu trong [25] thực hiện đối với ngữ liệu tiếng Đức. Mạng nơ-ron sâu [19] đã được sử dụng với các tham số MFCC, các đặc trưng liên quan cao độ như chu kỳ cơ bản, HNR (Harmonics-to-Noise Ratio) và chênh lệch của các tham số này giữa các khung tiếng nói để nhận dạng cảm xúc trên dữ liệu đa thể thức IEMOCAP (interactive emotional dyadic motion capture database).

Về mặt âm học, nhiều nghiên cứu đã khẳng định có thể nhận thấy và lượng hóa cảm xúc trong tiếng nói bằng cách phân tích các tham số như tần số cơ bản *F0*, cường độ và thời hạn. Ví dụ, các âm tiết có trọng âm có tần số cơ bản cao hơn, biên độ lớn hơn và thời hạn dài hơn so với các âm tiết không có trọng âm. Ở mức cảm thụ, sóng tiếng nói đi vào hệ thống thính giác của người nghe, thông qua ngôn điệu và quá trình xử lý cảm nhận cảm thụ mà sinh ra các thông tin về ngôn ngữ và thông tin đồng hành với ngôn ngữ. Dãy các đặc điểm ngôn điệu theo từng khung được trích rút từ các đoạn tiếng nói dài hơn như từ và câu cũng được dùng để đặc trưng cho các cảm xúc có trong tiếng nói. Thông tin *F0* được phân tích để phân loại cảm xúc và kết quả cho thấy giá trị cực đại, cực tiểu, trung bình của *F0* và đường bao *F0* là các đặc trưng nổi bật cho cảm xúc. Độ chính xác nhận dạng cảm xúc đạt được vào khoảng 80% khi sử dụng các đặc tính *F0* đã nêu cùng với bộ phân lớp láng giềng *K* gần nhất [12].

Các đặc tính ngôn điệu được trích rút từ các đơn vị ngôn ngữ nhỏ hơn như các âm tiết với phụ âm và nguyên âm cũng được dùng để phân tích cảm xúc.

Tầm quan trọng của đường bao ngôn điệu dẫn tới các ngữ cảnh có cảm xúc khác nhau đã được nghiên cứu [13]. Các cực đại và cực tiểu đối với tần số cơ bản, cường độ, thời hạn của khoảng dừng, các đột biến đã được đề xuất để định danh 4 cảm xúc như: sợ hãi, tức giận, buồn và vui [14].

III. XÂY DỰNG NGỮ LIỆU CẢM XÚC TIẾNG VIỆT

Theo thống kê của [22], đã có nhiều dữ liệu cảm xúc được xây dựng cho các ngôn ngữ khác nhau trên thế giới với số lượng dữ liệu tương ứng được đặt trong ngoặc đơn như sau: Anh (43), Pháp (5), Đức (14), Nga (1), Trung Quốc (11), Nhật (6)…Trong số các dữ liệu này, có một số dữ liệu được xây dựng đồng thời cho 2, 3 hoặc 4 ngôn ngữ khác nhau.

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 nhận 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. Trong số các phương pháp này, phương pháp ghi âm giọng các nghệ sĩ biểu đạt cảm xúc cho trước là phương pháp cho phép xây dựng được ngữ liệu thuận lợi hơn theo thiết kế định sẵn [26], dễ đạt được số lớn ngữ liệu đồng nhất, từ đó thuận tiện cho việc phân tích xác định tham số đặc trưng một cách tin cậy. Vì vậy, phương pháp này đã được chúng tôi lựa chọn để xây dựng bộ ngữ liệu cảm xúc tiếng Việt BKEmo. Với mục tiêu chính là phân tích tập trung vào bốn cảm xúc cơ bản vui, buồn, tức giận và bình thường, kịch bản thu âm được xây dựng phù hợp và yêu cầu người nói thể hiện tập trung vào bốn loại cảm xúc này một cách tốt nhất.

Kịch bản thu âm được xây dựng gồm 55 câu theo các tiêu chí sau:

- Nội dung gồm các câu cảm thán biểu lộ được cả 4 cảm xúc khi nói, các câu bình thường không có các từ ngữ cảm thán, biểu cảm mặt cảm xúc. Với các câu không có từ ngữ cảm thán (ví dụ: "*Vườn hoa trước nhà*", "*Trường Đại học Bách khoa Hà Nội*"…) người nói sẽ tập trung được vào việc biểu lộ cảm xúc mà không bị ảnh hưởng bởi nội dung của câu nói. Với loại câu có cảm thán (ví dụ: "*Thật á!*", "*Có lương rồi!*"….) sẽ giúp phân tích được nhiều tham số cảm xúc và các tham số phụ ảnh hưởng đến cảm xúc đó;
- Kịch bản có các tổ hợp từ (ví dụ: "*Thật á!*") và các câu 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.

Ngữ liệu được thu trong phòng thu âm, lồng tiếng chuyên nghiệp với 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 câu được nói lặp lại 4 lần cho mỗi cảm xúc. 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.

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 và biểu đạt tốt, rõ ràng cảm xúc khi nói. Kịch bản thu được sắp xếp không xuất hiện theo quy luật cụ thể 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, nghe nhanh một lượt để loại bỏ các câu bị lỗi trong quá trình thu hoặc cắt tự động.

IV. PHÂN TÍCH VÀ ĐÁNH GIÁ MỘT SỐ THAM SỐ VỀ CẢM XÚC TRONG TIẾNG VIỆT NÓI

Bài báo sử dụng phân tích phương sai ANOVA và kiểm định *T* (Tukey's test) để đánh giá sự biến thiên về tần số cơ bản *F0* trung bình và năng lượng trung bình của các cảm xúc trong ngữ liệu cảm xúc tiếng Việt đã được xây dựng. Mặt khác, để lấy các mẫu tham gia phân tích thống kê, chúng tôi dùng 2 phương pháp: phương pháp kinh nghiệm chủ quan trong đó chủ động lựa chọn các mẫu là các nghệ sĩ được biết nổi tiếng, rất có kinh nghiệm lồng tiếng cho phim và phương pháp cảm nhận thực tế trong đó dùng người nghe để lựa chọn các mẫu đã được phát âm phù hợp với cảm xúc quy định.

IV.1. Phân tích phương sai ANOVA và kiểm định T

IV.1.1. Phân tích phương sai ANOVA

Phương pháp này thực hiện so sánh các giá trị thống kê (giá trị trung bình) của nhiều tập hợp dữ liệu. Giả sử I là số tập hợp dữ liệu cần so sánh. $\mu_1, \ldots, \mu_I$ là các giá trị kỳ vọng của từng tập hợp. Khi đó giả thuyết cần kiểm định H_0: $\mu_1 = \mu_2 = \ldots \mu_I$ (1). Giả thuyết đối lập sẽ là H_a: ít nhất 1 trong 2 giá trị μ_i khác nhau.

Phương pháp ANOVA [15] để kiểm định giả thuyết này bao gồm:

- Tính trung bình bình phương giữa các tập hợp *MSTr* (Phương trình 1). Trong phương trình 1, I là số tập hợp và J là số giá trị đo cho mỗi tập hợp. $\bar{X}_i$ là giá trị trung bình trên mẫu i, $\bar{X}_{..}$ là giá trị trung bình trên toàn bộ dữ liệu.

$$MSTr = \frac{J}{I-1} \sum_i (\bar{X}_i - \bar{X}_{..})^2 \quad (1)$$

- Tính trung bình bình phương lỗi *MSE* (Phương trình 2). Trong phương trình 2, S_i^2 là phương sai mẫu thứ i.

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

- Giá trị thống kê cho kiểm định: $F = MSTr/MSE$. Giá trị này có phân bố F với $(I-1)$ bậc tự do ở tử số và $I(J-1)$ bậc tự do ở mẫu số. Khi đó với mức ý nghĩa α, vùng loại bỏ sẽ là: $f \geq F\alpha,\ I-1,\ I(J-1)$. *P*-value chính là phần diện tích ở phía dưới đường cong F nằm bên phải giá trị trên.

IV.1.2. Kiểm định T

Khi phân tích phương sai ANOVA đã cho kết quả là loại bỏ giả thuyết H_0, tức là sẽ có các cặp giá trị kỳ vọng của các tập hợp khác nhau; khi đó chúng ta sẽ cần biết chính xác đây là những cặp giá trị nào. Một trong những phương pháp được sử dụng phổ biến là kiểm định *T* (Tukey's test [15]). Phương pháp này sử dụng phân phối Student để đánh giá các giá trị μ_i - μ_j. Khoảng tin cậy của giá trị này được mô tả ở phương trình 3 với $Q_{\alpha,I,I(J-1)}$ là giá trị của phân phối Student tại mức ý nghĩa α.

$$\bar{X}_i - \bar{X}_j - Q_{\alpha,I,I(J-1)} \leq \mu_i - \mu_j \leq \bar{X}_i - \bar{X}_j + Q_{\alpha,I,I(J-1)} \quad (3)$$

Ngoài ra *P*-value cũng được tính cho các trường hợp này.

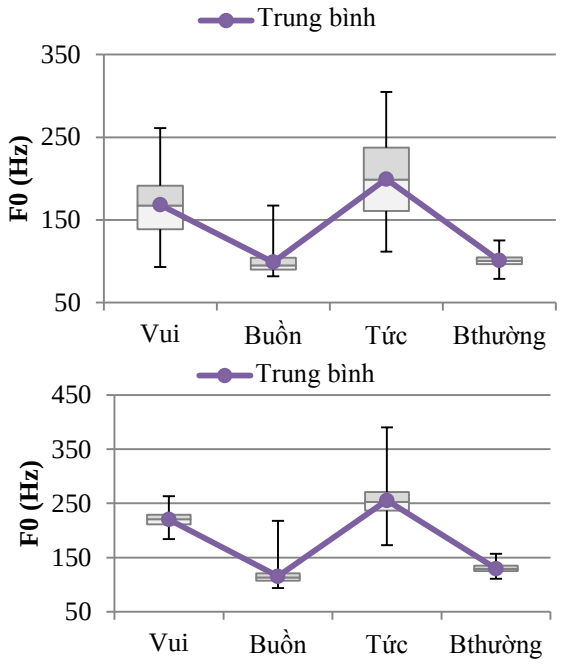
IV.2. Phân tích sự biến thiên F0 giữa các cảm xúc

Praat [27] đã được sử dụng để tính *F0*. Giá trị *F0* được tính trung bình trên từng câu tiếng nói (mỗi câu được người nói thể hiện bằng một cảm xúc cụ thể). Các giá trị *F0* trung bình này sẽ được thể hiện bằng đồ thị box-plot, và được phân tích thống kê bằng phương pháp phân tích phương sai ANOVA sau đó được kiểm định lại bằng phương pháp kiểm định *T*.

Theo kinh nghiệm chủ quan, bốn nghệ sĩ rất nổi tiếng gồm hai nghệ sĩ nam Đ.K (50 tuổi), H.P (40 tuổi) và hai nghệ sĩ nữ T.T.H (34 tuổi), B.H.G. (38 tuổi) đã được lựa chọn để đánh giá. Các nghệ sĩ này cũng trong số 56 nghệ sĩ tham gia thu âm. Mỗi cảm xúc được từng nghệ sĩ thể hiện trong 55 câu, 4 lần (220 file dữ liệu cho từng cảm xúc). Hình 1 mô tả đồ thị box-plot phân bố của các giá trị *F0* trung bình theo 4 cảm xúc.

Hình 1 cho thấy tần số cơ bản *F0* trung bình cho cảm xúc buồn là thấp nhất, tiếp theo là cảm xúc bình thường. Cảm xúc tức giận và cảm xúc vui có *F0* lớn hơn so với cảm xúc buồn và cảm xúc bình thường. Cảm xúc tức giận có giá trị *F0* trung bình lớn nhất.

Phương pháp phân tích phương sai ANOVA đã được sử dụng để kiểm định lại nhận xét trên, giá trị *F* và *P*-value được cho trong Bảng 1.



Hình 1. Đồ thị box-plot phân bố của các giá trị *F0* trung bình theo 4 cảm xúc của nghệ sĩ Đ.K. (bên trên) và H.P. (bên dưới)

Bảng 1. Giá trị *F* và *P*-value của phân tích phương sai ANOVA cho các giọng nam và nữ với tần số cơ bản *F0* trung bình và năng lượng trung bình

Người nói	*F0* Trung bình		Năng lượng trung bình	
	Giá trị *F*	*P*-value : Pr(>*F*)	Giá trị *F*	*P*-value : Pr(>*F*)
Đ.K.	586,93	$< 2{,}2.10^{-16}$	111,2	$< 2{,}2.10^{-16}$
H.P.	2931,7	$< 2{,}2.10^{-16}$	188,25	$< 2{,}2.10^{-16}$
T.T.H.	2681,1	$< 2{,}2.10^{-16}$	223,43	$< 2{,}2.10^{-16}$
B.H.G.	2543,4	$< 2{,}2.10^{-16}$	100,05	$< 2{,}2.10^{-16}$

Bảng 1 cho thấy giá trị *P*-value rất nhỏ, như vậy giả thuyết H_0 bị loại bỏ với tất cả các mức ý nghĩa quan trọng.

Để đánh giá sự khác biệt giữa các giá trị *F0* trung bình của các cảm xúc khác nhau, kiểm định *T* với mức ý nghĩa 95% đã được sử dụng. Kết quả được cho ở bảng 2.

Bảng 2 cho thấy có sự khác biệt về giá trị *F0* trung bình giữa tất cả các cảm xúc với nhau ngoại trừ giữa cảm xúc buồn và cảm xúc bình thường (P-value = 0,9). Điều này cũng phù hợp với Hình 1. Cảm xúc tức giận và cảm xúc buồn có độ chênh lệch *F0* cao nhất, khoảng tin cậy cho sự sai lệch là (92,9 Hz, 107,9 Hz).

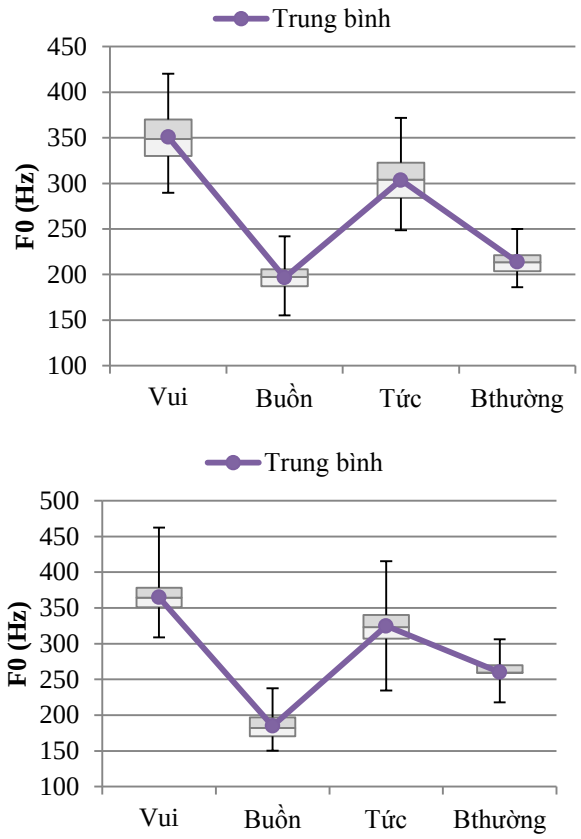
Bảng 2. Kết quả phân tích kiểm định *T* về tần số cơ bản *F0* cho giọng của người nói T.T.H và Đ.K.

$\mu_i - \mu_j$	***F0* trung bình của T.T.H.**			
	Giá trị trung bình	**Giá trị dưới của khoảng tin cậy**	**Giá trị trên của khoảng tin cậy**	***P*-value**
Buồn – BT	-75,2	-80,7	-69,3	≅0
Tức – BT	64,7	59,1	70,3	≅0
Vui – BT	104,8	99,3	110,3	≅0
Tức – Buồn	139,9	134,4	145,4	≅0
Vui – Buồn	179,9	174,4	185,5	≅0
Vui – Tức	40,1	34,6	45,6	≅0
$\mu_i - \mu_j$	***F0* trung bình của Đ.K.**			
	Giá trị trung bình	**Giá trị dưới của khoảng tin cậy**	**Giá trị trên của khoảng tin cậy**	***P*-value**
Buồn – BT	-2,0	-9,5	5,5	0,9
Tức – BT	98,3	90,9	105,9	≅0
Vui – BT	67,2	59,7	74,8	≅0
Tức – Buồn	100,4	92,9	107,9	≅0
Vui – Buồn	69,3	61,7	76,8	≅0
Vui – Tức	-31,2	-38,7	-23,6	≅0

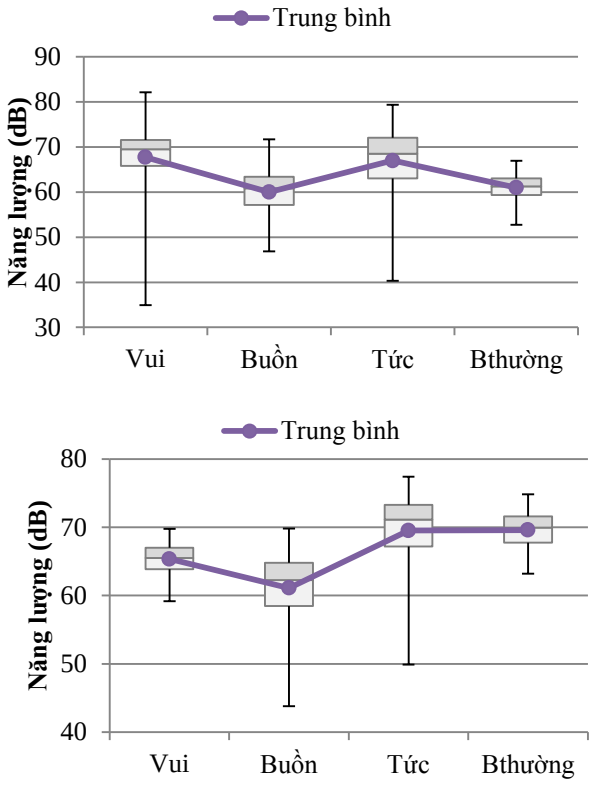
Hình 2 mô tả đồ thị box-plot phân bố của các giá trị *F0* trung bình theo 4 cảm xúc của 2 giọng nữ đã chọn.

Hình 2 cho thấy cũng như với giọng nam, cảm xúc tức giận và cảm xúc vui của giọng nữ cũng có *F0* lớn hơn so với cảm xúc buồn và cảm xúc bình thường. Tuy nhiên với giọng nữ, cảm xúc vui lại có *F0* lớn hơn so với cảm xúc tức giận.

Để đánh giá sự khác biệt giữa các giá trị *F0* trung bình của các cảm xúc khác nhau, kiểm định T với mức ý nghĩa 95% đã được sử dụng. Từ Bảng 2 có thể thấy có sự khác biệt về giá trị *F0* trung bình giữa tất cả các cảm xúc với nhau. Điều này cũng phù hợp với Hình 2. Cảm xúc vui và cảm xúc buồn có độ chênh lệch *F0* cao nhất, khoảng tin cậy cho sự sai lệch là (174,4 Hz, 185,5 Hz).



Hình 2. Đồ thị box-plot phân bố các giá trị *F0* trung bình theo 4 cảm xúc của người nói T.T.H. (dưới) và B.H.G. (trên)



Hình 3. Đồ thị box-plot phân bố của các giá trị năng lượng trung bình theo 4 cảm xúc của người nói Đ.K. (trên: giọng nam) và T.T.H. (dưới: giọng nữ)

IV.3. Phân tích sự biến thiên năng lượng giữa các cảm xúc

Giá trị năng lượng được tính trung bình trên từng câu nói, được thể hiện bằng đồ thị box-plot và được kiểm định bằng phương pháp phân tích phương sai ANOVA và kiểm định *T*.

Đồ thị box-plot phân bố năng lượng của người nói Đ.K. và T.T.H.cho trên Hình 3.

Hình 3 cho thấy với giọng nam vẫn có sự phân biệt rõ rệt về mặt năng lượng giữa các cảm xúc vui/tức giận so với các cảm xúc bình thường/buồn.

Kết quả phân tích ANOVA trong Bảng 2 cho thấy vẫn có sự khác biệt về mặt năng lượng trung bình giữa các cảm xúc này. Tuy nhiên, dải biến thiên của năng lượng của từng cảm xúc khá rộng. Do đó, không thể hiện được sự tách biệt giữa các cảm xúc như trong trường hợp tần số cơ bản *F0*.

Kiểm định T với mức ý nghĩa 95% được sử dụng để đánh giá sự khác biệt giữa các giá trị năng lượng trung bình của các cảm xúc khác nhau. Kết quả được cho ở Bảng 3.

Bảng 3 cho thấy có sự khác biệt về giá trị năng lượng trung bình giữa tất cả các cảm xúc với nhau ngoại trừ giữa cảm xúc buồn và cảm xúc bình thường (P-value = 0,22) và giữa cảm xúc vui và cảm xúc tức (P-value = 0,47). Điều này cũng phù hợp với Hình 5 và nhận định ở trên. Cảm xúc vui và cảm xúc bình thường có độ chênh lệch năng lượng cao nhất, khoảng tin cậy cho sự sai lệch là (5,34 dB, 8,09 dB).

Từ Hình 3 cũng có thể thấy với nữ giới, các cảm xúc không được thể hiện rõ ràng qua giá trị năng lượng trung bình. Chẳng hạn, cảm xúc bình thường lại có năng lượng trung bình cao hơn so với cảm xúc vui. Phân tích ANOVA (Bảng 4) vẫn cho thấy có thể phân

biệt giữa các cảm xúc với nhau dựa trên giá trị năng lượng.

Bảng 3. Kết quả phân tích kiểm định *T* về năng lượng trung bình cho giọng của Đ.K. (nam) và T.T.H. (nữ)

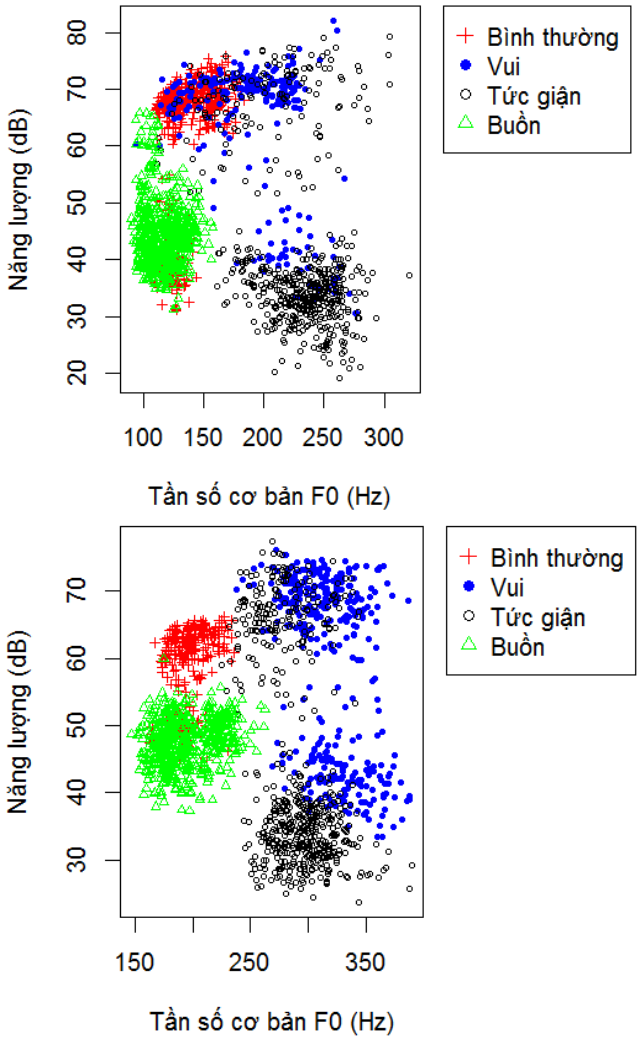
$\mu_i - \mu_j$	**Năng lượng trung bình của T.T.H**			
	Giá trị trung bình	**Giá trị dưới của khoảng tin cậy**	**Giá trị trên của khoảng tin cậy**	***P*-value**
Buồn – BT	-8,49	-9,48	-7,50	≅0
Tức – BT	-0,06	-1,04	0,93	0,99
Vui – BT	-4,25	-5,23	-3,26	≅0
Tức – Buồn	8,43	7,45	9,42	≅0
Vui – Buồn	4,24	3,26	5,23	≅0
Vui – Tức	-4,19	-5,17	-3,20	≅0
$\mu_i - \mu_j$	**Năng lượng trung bình của Đ.K.**			
	Giá trị trung bình	**Giá trị dưới của khoảng tin cậy**	**Giá trị trên của khoảng tin cậy**	***P*-value**
Buồn – BT	-1,02	-2,39	0,35	0,22
Tức – BT	5,94	4,56	7,31	≅0
Vui – BT	6,71	5,34	8,09	≅0
Tức – Buồn	6,96	5,59	8,33	≅0
Vui – Buồn	7,74	6,36	9,11	≅0
Vui – Tức	0,77	-0,61	2,14	0,47

Từ Bảng 3 ta thấy có sự khác biệt về giá trị năng lượng trung bình giữa tất cả các cảm xúc với nhau ngoại trừ giữa cảm xúc tức và cảm xúc bình thường (*P*-value = 0,99). Điều này cũng phù hợp với Hình 3. Cảm xúc buồn và cảm xúc tức giận có độ chênh lệch năng lượng cao nhất, khoảng tin cậy cho sự sai lệch là (7,45 dB, 9,42 dB).

IV.4. Phương pháp cảm nhận thực tế

Phần này trình bày các kết quả kiểm định theo phương pháp cảm nhận thực tế bằng cách thực hiện nghe lại và đánh giá trực tiếp để xác định những câu nói thể hiện được đúng cảm xúc theo yêu cầu. Trung bình mỗi cảm xúc cho mỗi giới tính có khoảng 500 câu được đánh giá với 5 người nói cho mỗi giới tính được lấy ngẫu nhiên.

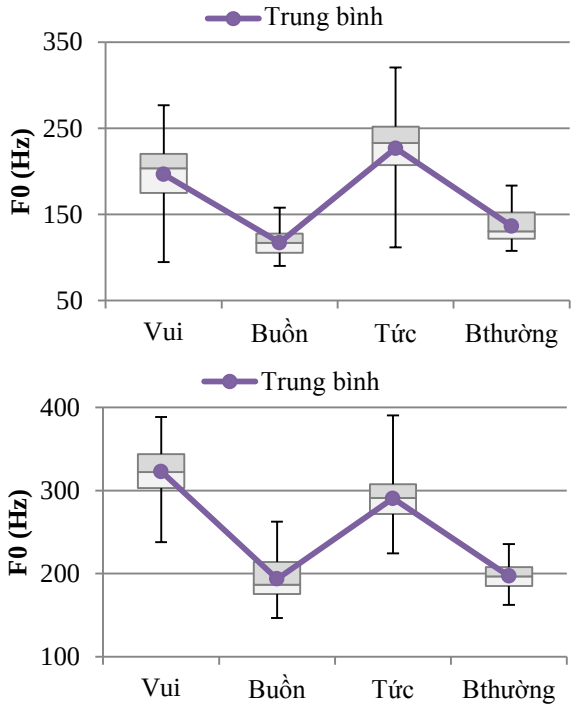
Từ Hình 4 có thể nhận thấy các cảm xúc có sự tập trung tốt tại một vùng nhất định: năng lượng là bộ tham số rất tốt để phân biệt giữa cảm xúc buồn và cảm xúc bình thường, giữa cảm xúc vui và cảm xúc tức giận. Ngoài ra cũng có sự phân biệt rất rõ về tần số *F0* giữa cảm xúc buồn/bình thường so với cảm xúc vui/tức giận.



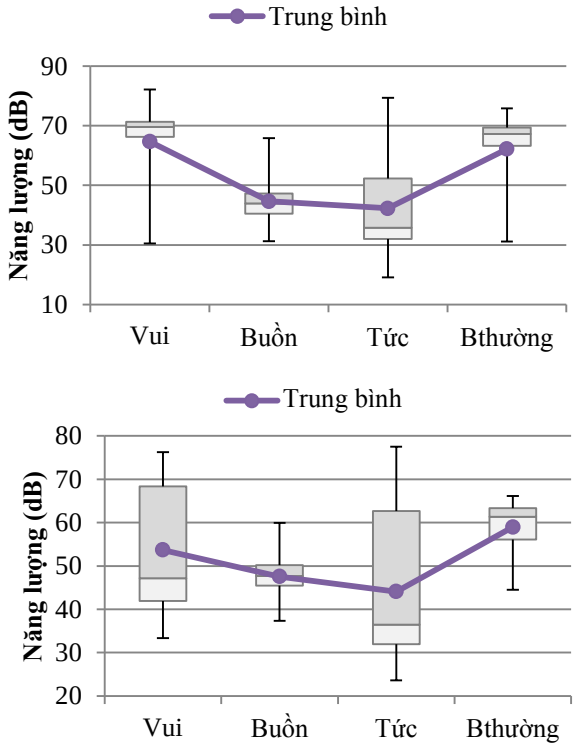
Hình 4. Đồ thị phân bố điểm của các giá trị *F0* trung bình so với năng lượng trung bình theo 4 cảm xúc của giọng nam (trái) và giọng nữ (phải)

Từ Hình 5, tần số *F0* trung bình của cảm xúc bình thường và cảm xúc buồn có xu hướng nhỏ hơn so với cảm xúc tức giận và cảm xúc vui. Ở giọng nam, *F0* trung bình của cảm xúc tức giận lớn hơn so với cảm xúc vui, và ngược lại ở giọng nữ.

Phương pháp phân tích phương sai ANOVA đã được thực hiện trên tần số *F0* trung bình và năng lượng trung bình. Kết quả trong Bảng 4 cho thấy có sự khác biệt của các tham số này trên các cảm xúc.



Hình 5. Đồ thị box-plot phân bố của các giá trị *F0* trung bình theo 4 cảm xúc của giọng nam (trên) và giọng nữ (dưới)



Hình 6. Đồ thị box-plot phân bố các giá trị năng lượng trung bình theo 4 cảm xúc, giọng nam (trên) và giọng nữ (dưới)

Bảng 4. Giá trị *F* và *P*-value của phân tích phương sai ANOVA cho các giọng nam và nữ với *F0* trung bình và năng lượng trung bình

Giới tính	*F0* trung bình	Năng lượng trung bình		
	Giá trị *F*	*P*-value : Pr(>*F*)	Giá trị *F*	*P*-value: Pr(>*F*)
Nam	2049	< 2,2e-16	427,94	< 2,2e-16
Nữ	3277,7	< 2,2e-16	132,65	< 2,2e-16

Bảng 5. Kết quả phân tích kiểm định *T* về *F0* trung bình và năng lượng trung bình cho giọng của các giọng nam

$\mu_i - \mu_j$	Năng lượng trung bình			
	Giá trị trung bình	Giá trị dưới của khoảng tin cậy	Giá trị trên của khoảng tin cậy	*P*-value
Buồn – BT	-17,6	-19,4	-15,7	≅0
Tức – BT	-19,9	-21,8	-18,0	≅0
Vui – BT	2,49	0,23	4,77	0,0242
Tức –Buồn	-2,35	-4,17	-0,54	0,0048
Vui –Buồn	20,1	17,9	22,3	≅0
Vui –Tức	22,4	20,2	24,6	≅0
$\mu_i - \mu_j$	*F0* trung bình			
	Giá trị trung bình	Giá trị dưới của khoảng tin cậy	Giá trị trên của khoảng tin cậy	*P*-value
Buồn – BT	-19,1	-23,2	-14,9	≅0
Tức –BT	90,4	86,3	94,5	≅0
Vui – BT	60,2	55,2	65,1	≅0
Tức –Buồn	109,5	105,5	113,4	≅0
Vui –Buồn	79,2	74,4	84,0	≅0
Vui –Tức	-30,2	-35,1	-25,4	≅0

Kiểm định T được thực hiện để đánh giá sự khác nhau của các tham số trên giữa các cảm xúc. Kết quả của giọng nam được mô tả ở Bảng 5 và của giọng nữ được mô tả ở Bảng 6.

Kết quả trong Bảng 5 cho thấy có sự phân biệt rất rõ rệt về *F0* giữa các cảm xúc cho cả giọng nam (*P*-value ≅0). *F0* trung bình giữa cảm xúc tức-buồn cao nhất với khoảng tin cậy (105,5Hz, 113,4Hz). Như vậy, lựa chọn mẫu theo đánh giá cảm nhận cho kết quả phân biệt cảm xúc chính xác hơn so với lựa chọn mẫu theo kinh nghiệm chủ quan. Tuy nhiên, với năng

lượng thì vẫn có những giá trị *P*-value đáng kể (ví dụ 0,0242), như vậy sẽ không thể phân biệt được 2 cảm xúc này với mức ý nghĩa 0,01.

Bảng 6. Kết quả phân tích kiểm định *T* về *F0* trung bình và năng lượng trung bình cho giọng của các giọng nữ

$\mu_i - \mu_j$	**Năng lượng trung bình**			
	Giá trị trung bình	**Giá trị dưới của khoảng tin cậy**	**Giá trị trên của khoảng tin cậy**	***P*-value**
Buồn – BT	-11,4	-13,6	-9,2	≅0
Tức – BT	-14,9	-17,1	-12,7	≅0
Vui – BT	-5,3	-7,5	-3,1	≅0
Tức –Buồn	-3,5	-5,3	-1,7	≅0
Vui –Buồn	6,1	4,3	7,9	≅0
Vui –Tức	9,6	7,8	11,4	≅0
$\mu_i - \mu_j$	***F0* trung bình**			
	Giá trị trung bình	**Giá trị dưới của khoảng tin cậy**	**Giá trị trên của khoảng tin cậy**	***P*-value**
Buồn – BT	-3,5	-8,2	1,2	0,22
Tức –BT	93,4	88,7	98,2	≅0
Vui –BT	125,6	120,9	130,4	≅0
Tức –Buồn	96,9	93,1	100,7	≅0
Vui –Buồn	129,1	125,2	133,1	≅0
Vui –Tức	32,2	28,3	36,1	≅0

Với giọng nữ, kết quả ở Bảng 6 cho thấy không có sự phân biệt rõ rệt về *F0* trung bình giữa cảm xúc buồn và cảm xúc bình thường (*P*-value = 0,22). *F0* trung bình giữa cảm xúc vui và buồn cao nhất với độ tin cậy (125,2Hz, 133,1Hz).

V. THỬ NGHIỆM NHẬN DẠNG CẢM XÚC TIẾNG VIỆT

Với bộ ngữ liệu cảm xúc tiếng Việt BKEmo, các bộ phân lớp SMO, lBk, trees J48 đã được thử nghiệm để nhận dạng cảm xúc. Các bộ phân lớp này thuộc công cụ Weka gồm tập hợp các thuật giải học máy dùng cho khai phá dữ liệu do Đại học Waikato, NewZealand phát triển [34]. SMO (Sequential Minimal Optimization) [32] là thuật giải tối ưu hóa cực tiểu lần lượt để huấn luyện bộ phân lớp hỗ trợ véc-tơ dùng kernel đa thức hoặc Gauss. lBk là bộ phân lớp *k* láng giềng gần nhất sử dụng độ đo khoảng cách Ơclit [34]. Bộ phân lớp trees J48 [33] được dùng để có các luật từ các cây quyết định riêng phần đã được xây dựng bằng cách sử dụng J48. J48 là cài đặt mã nguồn mở Java của thuật giải C4.5 và thuật giải này được dùng để tạo cây quyết định do Ross Quinlan phát triển Ngữ liệu dùng cho thử nghiệm gồm 5584 file tương ứng với 4 cảm xúc được 16 nghệ sĩ (8 giọng nam và 8 giọng nữ) thể hiện. Số file này được chia làm 2 phần bằng nhau, một phần dùng để huấn luyện và phần còn lại dùng cho nhận dạng. Thử nghiệm nhận dạng được thực hiện theo phương pháp đánh giá chéo (cross-validation). Bộ tham số đặc trưng được trích rút nhờ công cụ OpenSMILE [35] với 384 tham số bao gồm: năng lượng, MFCC, tỉ lệ biến thiên qua trục không, tần số cơ bản F0, xác suất xuất hiện âm hữu thanh. Các tham số này lại được đánh giá theo giá trị cực đại, cực tiểu, vị trí xuất hiện cực đại, vị trí xuất hiện cực tiểu, dải giá trị, giá trị trung bình, độ lệch chuẩn, độ lệch phổ so với tần số trung bình (Skewness), độ khác biệt phổ quanh tâm phổ so với phân bố Gauss (Kurtosis).

Bảng 7. Ma trận nhầm lẫn nhận dạng cảm xúc với 384 tham số

Bộ phân lớp		**Tức**	**Vui**	**BT**	**Buồn**
SMO	**Tức**	1341	51	4	0
	Vui	41	1342	13	0
	BT	4	8	1300	84
	Buồn	3	11	75	1307
lBk	**Tức**	1383	9	2	2
	Vui	13	1380	1	2
	BT	0	0	1367	29
	Buồn	0	1	43	1352
Trees J48	**Tức**	1084	225	62	25
	Vui	216	1103	54	23
	BT	61	58	1128	149
	Buồn	19	25	164	1188

Bảng 7 là ma trận nhầm lẫn nhận dạng cảm xúc dùng bộ 384 tham số còn Bảng 8 là ma trận nhầm lẫn nhận dạng cảm xúc chỉ dùng các tham số liên quan đến *F0* và năng lượng. Kết quả trên cả hai bảng đều dùng các bộ phân lớp SMO, lBk, trees J48. Bảng 7 cho thấy tỉ lệ nhận dạng đúng trung bình cao nhất cho cả 4

cảm xúc đạt 98,17% với bộ phân lớp lBk còn tỉ lệ nhận dạng đúng trung bình thấp nhất là 80,64% với bộ phân lớp trees J48. Đối với Bảng 8, khi số tham số giảm xuống chỉ còn 48 tham số liên quan đến *F0* và năng lượng, tỉ lệ nhận dạng đúng đều giảm so với Bảng 7 tuy nhiên vẫn giữ quy luật tỉ lệ nhận dạng đúng cao nhất cho bộ phân lớp lBk và thấp nhất cho bộ phân lớp trees J48. Trường hợp chỉ sử dụng các tham số liên quan đến *F0* và năng lượng, tỉ lệ nhận dạng đúng trung bình cao nhất giảm xuống còn 82,59% và tỉ lệ nhận dạng đúng trung bình thấp nhất giảm xuống còn 75,25%. Nhìn chung, các kết quả này đều khả quan so với một số kết quả nhận dạng cảm xúc tiếng Việt đã được công bố [28], [29] hoặc kết quả nhận dạng cảm xúc của một số ngôn ngữ khác [36-39].

Bảng 8. Ma trận nhầm lẫn nhận dạng cảm xúc với 48 tham số liên quan đến *F0* và năng lượng

Bộ phân lớp		**Tức**	**Vui**	**BT**	**Buồn**
SMO	**Tức**	1144	178	53	21
	Vui	182	1103	100	11
	BT	31	99	903	363
	Buồn	14	33	156	1193
lBk	**Tức**	1186	144	45	21
	Vui	139	1174	63	20
	BT	30	50	1093	223
	Buồn	21	13	203	1159
trees J48	**Tức**	1084	218	70	24
	Vui	227	1052	99	18
	BT	77	92	969	258
	Buồn	17	33	249	1097

VI. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong bài báo này, các tham số cơ bản của các cảm xúc, việc xây dựng ngữ liệu cảm xúc cho tiếng Việt, sử dụng phân tích phương sai ANOVA và kiểm định *T* để đánh giá sự biến thiên *F0* và năng lượng trung bình giữa các cảm xúc đã được trình bày. Kết quả phân tích cho thấy tần số cơ bản *F0* là một tham số đáng tin cậy để phân biệt giữa các cảm xúc. Năng lượng cũng là một tham số hiệu quả về phân biệt cảm xúc, phản ánh rõ nét trên nam giới hơn so với trên nữ giới. Trong số các bộ phân lớp được sử dụng để thử nghiệm bước đầu nhận dạng cảm xúc theo bộ ngữ liệu BKEmo, bộ phân lớp lBK cho kết quả nhận dạng tốt nhất. Hướng nghiên cứu tiếp theo của chúng tôi là tập trung vào phân tích ảnh hưởng đến cảm xúc của các tham số như trường độ, tốc độ nói cũng như một số tham số khác liên quan đến nguồn âm và tiến hành nhận dạng cảm xúc tiếng Việt dùng các mô hình nhận dạng khác nhau sử dụng ngữ liệu đã được xây dựng. Bên cạnh đó sẽ mở rộng nghiên cứu cho các hình thái cảm xúc đa dạng hơn.

LỜI CẢM ƠN

Bài báo này được thực hiện trong khuôn khổ đề tài nghiên cứu "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] RODDY COWIE, MARC SCHRÖDER, *"Piecing together the emotion jigsaw"*, Workshop on Machine Learning for Multimodal Interaction (MLMI04), Martigny, Switzerland, June 21-23, 2004.

[2] MARIA SCHUBIGER, *"English intonation: its form and function"*. Language Vol. 36, No. 4, 1960, pp. 544-548.

[3] KLAUS. R. SCHERER, *"Vocal communication of emotion: A review of research paradigms"*, Speech Communication, vol. 40, 2003, pp. 227–256.

[4] JANET CAHN, *"The generation of affect in synthesized speech"*. Journal of American Voice Input/Output Society, vol. 8, 1990, pp. 1–19.

[5] CARL E. WILLIAMS, KENNETH N. STEVENS, *"Emotions and speech: Some acoustical correlates"*. The Journal of the Acoustical Society of America Vol. 52 (4), 1972, pp. 1238-1250.

[6] FELIX BURKHARDT, WALTER F. SENDLMEIER, *"Verification of acoustical correlates of emotional speech using formant-synthesis"*. In Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, UK, 2000.

[7] SYLVIE MOZZICONACCI, DIK J. HERMES, *"Role of intonation patterns in conveying emotion in speech"*. In Proceedings of ICPhS 1999 , San Francisco 1999, pp. 2001-2004.

[8] JANET E. CAHN, *"Generating expression in synthesized speech"*, Master's Thesis, Massachusetts Institute of Technology, May 1989.

[9] JEAN VROOMEN, RENÉ COLLIER, SYLVIE MOZZICONACCI, *"Duration and intonation in emotional speech"*, Proceedings of the Third European Conference on Speech Communication and Technology, Berlin, Germany, September 21-23, 1993.

[10] DEEPA P. GOPINATH, SHEEBA P.S, ACHUTHSANKAR S. NAIR, *"Emotional Analysis for Malayalam Text to Speech Synthesis Systems",* Proceedings of the Setit 2007 - 4th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, Tunisia, March 25-29, 2007.

[11] TSANG-LONG PAO, YU-TE CHEN, JUN-HENG YEH, WEN_YUAN LIAO, *"Combining acoustic features for improved emotion recognition in mandarin speech"*, in ACII (Affective Computing and Intelligent Interaction), Beijing, China, October 22-24, 2005.

[12] FRANK DELLERT, THOMAS POLZIN, ALEX WAIBEL, *"Recognising emotions in speech"*, ICSLP 96, Philadelphia, USA, Oct 03-06, 1996.

[13] IAIN R. MURRAY, JOHN L. ARNOTT, ELIZABETH A. ROHWER, *"Emotional stress in synthetic speech: Progress and future directions"*, Speech Communication, vol. 20, Nov 1996, pp. 85-91.

[14] SINÉAD MCGILLOWAY, RODDY COWIE, ELLEN DOUGLAS-COWIE, STAN GIELEN, MACHIEL WESTERDIJK, SYBERT STROEVE *"Approaching automatic recognition of emotion from voice: A rough benchmark",* Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, UK, Sep 5-9, 2000.

[15] JAY L. DEVORE, *"Probability and Statistics for Engineering and the Sciences"*, Eighth Edition, Brooks/Cole Edition, 2010.

[16] YIXIONG PAN, PEIPEI SHEN, LIPING SHEN, *"Speech Emotion Recognition Using Support Vector Machine"*, International Journal of Smart Home Vol. 6, No. 2, April, 2012, pp 101-108.

[17] R. SUBHASHREE1, G. N. RATHNA, *"Speech Emotion Recognition: Performance Analysis based on Fused Algorithms and GMM Modelling"*, Indian Journal of Science and Technology, Vol 9(11), March 2016, pp. 1-8.

[18] H. MIWA, T. UMETSU, A. TAKANISHI, H. TAKANOBU, *"Robot personalization based on the mental dynamics",* IEEE/RSJ Conference on Intelligent Robots and Systems, vol 1, Takamatsu, Oct 31-Nov 5, 2000.

[19] KUN HAN, DONG YU, IVAN TASHEV, *"Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine"*, INTERSPEECH 2014, Singapore, September 14-18, 2014

[20] THI DUYEN NGO, THE DUY BUI, *"A study on prosody of Vietnamese emotional speech"*, Proceedings of the Fourth International Conference on Knowledge and Systems Engineering (KSE 2012), IEEE, Danang city, Vietnam, Aug 17-19, 2012

[21] 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.

[22] JOHANNES PITTERMANN, ANGELA PITTERMANN, WOLFGANG MINKER, *"Handling Emotions in Human-Computer Dialogues"*, Springer, 2010.

[23] 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.

[24] 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, 2015, pp. 273-287.

[25] 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, Elsevier, 2015, pp. 50-57.

[26] 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, Elsevier, March 2011, pp 572–587.

[27] www.praat.org, last visited 20/02/2016.

[28] LA VUTUAN, HUANG CHENG-WEI, HA CHENG, ZHAO LI, "*Emotional Feature Analysis and Recognition from Vietnamese Speech*", Journal of Signal Processing, China, 2013.

[29] 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, 2015, pp 50-57.

[30] ROBERT PLUTCHIK, HENRY KELLERMAN, *"Emotion: Theory, research and experience",* vol 4. Academic Press, New York, USA, 1989.

[31] NGUYỄN TÔN NHAN, PHÚ VĂN HẴN, "Từ điển tiếng Việt", Nhà xuất bản Từ điển Bách Khoa, 2013.

[32] JOHN C. PLATT, *"Technical Report MSR-TR-98-14"*, Microsoft Research, April 21, 1998

[33] QUINLAN, J. R. *"C4.5: Programs for Machine Learning"*, Morgan Kaufmann Publishers, 1993.

[34] WITTEN, IAN H., AND EIBE FRANK, *"Data Mining: Practical machine learning tools and techniques"*, Morgan Kaufmann Publishers, 2005.

[35] EYBEN, FLORIAN, MARTIN WÖLLMER, AND BJÖRN SCHULLER, *"Opensmile: the munich versatile and fast open-source audio feature extractor"*, Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italia, Oct 25-29, 2010.

[36] SIQING WUA, TIAGO H. FALKB, WAI-YIP CHAN, *"Automatic speech emotion recognition using modulation spectral features"*, Speech Communication, Volume 53, Issue 5, 2011, pp. 768–785.

[37] S. LALITHA, ABHISHEK MADHAVAN, BHARATH BHUSHAN, SRINIVAS SAKETH, *"Speech emotion recognition"*, Proceedings of the International Conference on Advances in Electronics, Computers and Communications, Bangalore, India, Oct 10-11, 2014.

[38] MARTIN GJORESKI, HRISTIJAN GJORESKI, ANDREA KULAKOV, "*Machine Learning Approach for Emotion Recognition in Speech*", Informatica, vol 38, no 4, 2014, pp. 377-384.

[39] ANKUSH CHAUDHARY,ASHISH KUMAR SHARMA, JYOTI DALAL, LEENA CHOUKIKER, *"Speech Emotion Recognition"*, Journal of Emerging Technologies and Innovative Research, vol. 2, issue 4, 2015, pp 1169-1171.

Nhận bài ngày: 26/02/2016

SƠ LƯỢC VỀ TÁC GIẢ

LÊ XUÂN THÀNH

Sinh năm 1982.

Tốt nghiệp ĐH Bách khoa Hà Nội năm 2006.

Hiện tại là giảng viên và nghiên cứu sinh tại Bộ môn Kỹ thuật Máy tính, Trường ĐH Bách khoa Hà Nội.

Lĩnh vực nghiên cứu: Xử lý tín hiệu, Xử lý tiếng nói, Hệ nhúng.

Email: thanhlx@soict.hust.edu.vn

Điện thoại : 0906755789

ĐÀO THỊ LỆ THỦY

Sinh năm 1976.

Tốt nghiệp Học viện Kỹ thuật Quân sự năm 2008.

Hiện đang là nghiên cứu sinh tại Viện CNTT và Truyền thông, Trường ĐH Bách khoa Hà Nội.

Lĩnh vực nghiên cứu: Xử lý tín hiệu, Xử lý tiếng nói, công nghệ phần mềm.

Email: thuydt@hht.edu.vn

TRỊNH VĂN LOAN

Sinh năm 1956.

Tốt nghiệp ĐH Bách khoa Hà Nội năm 1978. Nhận bằng DEA năm 1988 và nhận bằng Docteur năm 1992 tại Viện ĐH Bách khoa Quốc gia Grenoble (INPG) Pháp.

Hiện công tác tại Viện CNTT và Truyền thông, Trường ĐH Bách khoa Hà Nội.

Lĩnh vực nghiên cứu: Xử lý tín hiệu, Xử lý tiếng nói, Hệ nhúng.

Email: loantv@soict.hust.edu.vn

NGUYỄN HỒNG QUANG

Sinh năm 1978.

Tốt nghiệp ĐH Bách khoa Hà Nội năm 2000.

Nhận bằng tiến sỹ tại Trường ĐH Avignon, CH Pháp năm 2008.

Hiện tại là giảng viên Viện CNTT và Truyền thông, Trường ĐH Bách khoa Hà Nội.

Lĩnh vực nghiên cứu: Xử lý tiếng nói, Học máy thống kê.

Email: quangnh@soict.hust.edu.vn