MỞ ĐẦU

1. Lý do chọn đề tài

Ngày nay, đã có những thay đổi rất lớn về cách thức con người trao đổi thông tin với hệ thống. Sự thay đổi này biểu hiện ở chỗ, các cách thức trao đổi thông tin đã được định dạng và có cấu trúc chặt chẽ được chuyển sang các cách thức linh hoạt và tự nhiên hơn. Trong đó, tiếng nói là cách thức trao đổi thông tin tự nhiên nhất, cho phép tương tác giữa con người với hệ thống nhanh và dễ dàng. Đối thoại dùng ngôn ngữ nói không chỉ đơn giản, thuận tiện và tiết kiệm thời gian mà còn góp phần đảm bảo khía cạnh an toàn trong những môi trường có tính rủi ro.

Để có thể thiết lập hệ thống tương tác có tính linh hoạt cao, kiến trúc của các hệ thống đối thoại người - máy cần được trang bị thêm các chức năng mới. Các chức năng này bao gồm nhận dạng cảm xúc tiếng nói, phát hiện các tham biến dựa trên tình huống cũng như trạng thái của người dùng và quản lý tình huống để đưa ra các mô hình dựa trên các tham biến đã được phát hiện làm cho quá trình đối thoại phù hợp. Chính vì vậy, trong nhiều năm qua, các nghiên cứu về cảm xúc tiếng nói đã thu hút mối quan tâm mạnh mẽ trong lĩnh vực tương tác người - máy và mong muốn tìm ra cách làm thế nào có thể tích hợp trạng thái cảm xúc của người nói vào hệ thống đối thoại người - máy dùng tiếng nói.

Trên thế giới đã có nhiều nghiên cứu về cảm xúc và nhận dạng cảm xúc tiếng nói với các ngôn ngữ khác nhau nhưng kết quả ứng dụng trên thực tế còn nhiều khó khăn vì cảm xúc được thể hiện rất đa dạng trong mỗi con người. Do đó, việc phát hiện chính xác cảm xúc còn phải được tiếp tục nghiên cứu. Riêng về nhận dạng cảm xúc cho tiếng Việt nói, còn rất ít các công trình nghiên cứu, mặc dù cũng đã có những nghiên cứu và đã đạt được những thành công nhất định nhưng để triển khai thành các sản phẩm ứng dụng thực tế vẫn còn nhiều mặt hạn chế, đặc biệt là độ chính xác, chất lượng nhận dạng. Chính vì vậy, cần thiết phải nghiên cứu nhận dạng cảm xúc cho tiếng Việt nói để tăng cường hiệu quả và ứng dụng được cho các hệ thống tương tác dùng tiếng Việt nói.

Từ những lý do nêu trên, tác giả lựa chọn đề tài nghiên cứu "Nhận dạng cảm xúc cho tiếng Việt nói" nhằm nghiên cứu sâu hơn về vấn đề xử lý nhận dạng cảm xúc, đặc biệt đối với tiếng Việt nói để tìm ra các tham số cũng như mô hình nhận dạng cảm xúc phù hợp cho tiếng Việt, góp phần phát triển các ứng dụng công nghệ thông tin cho người Việt cũng như các sản phẩm ứng dụng công nghệ thông tin sử dụng tiếng Việt nói trong giao tiếp và tương tác người-máy.

2. Mục tiêu nghiên cứu của luận án

Với tính thiết thực của cảm xúc trong tiếng nói được áp dụng trong thực tế đang rất được quan tâm, mục tiêu chính của đề tài là nghiên cứu nhận dạng cảm xúc cho tiếng Việt nói dựa trên phương diện xử lý tín hiệu tiếng nói. Đề tài nghiên cứu thử nghiệm và đề xuất mô hình nhận dạng cảm xúc cho tiếng Việt nói dựa trên việc nghiên cứu đánh giá các tham số và so sánh một số mô hình nhận dạng. Bốn cảm xúc cơ bản sẽ được nghiên cứu bao gồm cảm xúc: vui, buồn, tức và bình thường. Ngữ liệu tiếng Việt dùng cho nhận dạng là giọng phổ thông miền Bắc có cả giọng nam và giọng nữ.

3. Nhiệm vụ nghiên cứu của luận án

Để đạt được những mục tiêu đã đề ra, luận án cần thực hiện các nhiệm vụ chính sau:

- Nghiên cứu tổng quan về cảm xúc và nhận dạng cảm xúc tiếng nói.
- Nghiên cứu một số mô hình nhận dạng dùng cho nhận dạng cảm xúc tiếng nói như mô hình GMM, ANN, …
- Phân tích đánh giá và đề xuất bộ ngữ liệu cảm xúc tiếng Việt dùng cho nhận dạng bốn cảm xúc cơ bản vui, buồn, tức và bình thường.
- Nghiên cứu đề xuất và phân tích ảnh hưởng của các tham số đặc trưng tín hiệu tiếng nói đến cảm xúc tiếng Việt.
- Thử nghiệm nhận dạng cảm xúc tiếng Việt dựa trên các mô hình đã nghiên cứu có tính đến các đặc trưng của tiếng Việt nói.
- Phân tích đánh giá kết quả nhận dạng cảm xúc của các mô hình dựa trên các kết quả thử nghiệm.

4. Đối tượng và phạm vi nghiên cứu của luận án

Đối tượng nghiên cứu của luận án là nhận dạng cảm xúc cho tiếng Việt nói theo phương diện xử lý tín hiệu tiếng nói. Từ kết quả nhận dạng cảm xúc, xây dựng mô hình nhận dạng cảm xúc cho tiếng Việt nói. Các hình thái cảm xúc rất đa dạng và ở những vùng miền khác nhau thì ngôn điệu đối với biểu hiện cảm xúc cũng khác nhau. Trong khuôn khổ có hạn, luận án tập trung thực hiện nghiên cứu nhận dạng 4 cảm xúc cơ bản: vui, buồn, tức và bình thường với giọng phổ thông miền Bắc gồm cả giọng nam và nữ.

Nghiên cứu của luận án nhằm nhận dạng cảm xúc chỉ qua diễn đạt câu nói mà tín hiệu tiếng nói đã thu thập được tương ứng và cũng không xét đến các từ biểu lộ cảm xúc, hoặc biểu lộ cảm xúc qua khuôn mặt cũng như chưa thể xét đến suy nghĩ thực tế trong bộ não của con người liên quan đến cảm xúc.

5. Ý nghĩa khoa học và thực tiễn của luận án

Về mặt lý thuyết, luận án góp phần làm sáng tỏ các mô hình nhận dạng tiếng nói và nhận dạng cảm xúc đối với tiếng Việt nói, đánh giá kết quả thử nghiệm với các mô hình nhận dạng cảm xúc tiếng Việt nói và tạo tiền đề cho các nghiên cứu tiếp theo về cảm xúc tiếng Việt.

Về mặt thực tiễn, kết quả nghiên cứu của luận án có thể được ứng dụng đa dạng trong các lĩnh vực khoa học, công nghệ, đặc biệt trong lĩnh vực tương tác người-hệ thống sử dụng tiếng nói với việc tổng hợp và nhận dạng tiếng Việt có cảm xúc.

6. Phương pháp nghiên cứu

Phương pháp nghiên cứu thực hiện trong luận án là nghiên cứu lý thuyết kết hợp với thực nghiệm.

Về mặt lý thuyết, luận án tìm hiểu tổng quan về cảm xúc trong tiếng nói, các phương pháp nhận dạng cảm xúc, các tham số đặc trưng của tín hiệu tiếng nói có ảnh hưởng đến cảm xúc xét theo phương diện tín hiệu tiếng nói đồng thời cũng trình bày một số mô hình nhận dạng cảm xúc tiếng nói được tổng hợp từ các tài liệu, bài báo khoa học.

Về mặt thực nghiệm, lựa chọn và đánh giá bộ ngữ liệu cảm xúc tiếng Việt, sử dụng các bộ công cụ để tính toán, phân tích, thống kê và đánh giá các tham số đặc trưng, tiến hành nghiên cứu và thực hiện các thử nghiệm nhận dạng cảm xúc dựa trên các mô hình nhận dạng cảm xúc cho ngữ liệu tiếng Việt với bốn cảm xúc vui, buồn, tức, bình thường từ đó đánh giá kết quả đạt được để xác nhận giá trị của các mô hình và các tham số sử dụng.

7. Kết quả mới của luận án

Kết quả nghiên cứu mới của luận án có thể được tóm tắt tập trung vào các điểm chính sau:

- Sử dụng các phương pháp thích hợp để đánh giá bộ ngữ liệu cảm xúc tiếng Việt từ đó đề xuất được bộ ngữ liệu cảm xúc tiếng Việt dùng cho thử nghiệm nhận dạng cảm xúc tiếng Việt nói.
- Nghiên cứu, khai thác và đề xuất được các mô hình GMM, DCNN và các tham số đặc trưng phù hợp cho nhận dạng cảm xúc tiếng Việt nói đồng thời đánh giá được ảnh hưởng của các tham số đặc trưng đến kết quả nhận dạng cảm xúc tiếng Việt với bốn cảm xúc vui, buồn, tức và bình thường.

8. Cấu trúc của luận án

Luận án được trình bày trong 4 chương với nội dung tóm tắt như sau:

Chương 1: Tổng quan về cảm xúc và nhận dạng cảm xúc tiếng nói.

Chương này trình bày các nghiên cứu về cảm xúc, phân loại cảm xúc và các cảm xúc cơ bản. Đồng thời, các nghiên cứu về nhận dạng cảm xúc tiếng nói trong và ngoài nước, các mô hình được thực hiện để nhận dạng cảm xúc tiếng nói cũng được nêu rõ.

Chương 2: Ngữ liệu cảm xúc và các tham số đặc trưng cho cảm xúc tiếng Việt nói. Nội dung của chương trình bày các phương pháp xây dựng ngữ liệu cảm xúc nói chung, các bộ ngữ liệu cảm xúc có sẵn với các ngôn ngữ khác nhau. Chương này sẽ tập trung vào việc lựa chọn đề xuất bộ ngữ liệu cảm xúc tiếng Việt dùng cho thử nghiệm của luận án, đề xuất và đánh giá các tham số đặc trưng của tín hiệu tiếng nói ảnh hưởng đến cảm xúc. Phần cuối của chương đánh giá bộ ngữ liệu cảm xúc tiếng Việt dùng cho thử nghiệm dựa trên một số bộ phân lớp LDA, IBk, SVM, Tree-J48.

Chương 3: Nhận dạng cảm xúc tiếng Việt nói với mô hình GMM. Các kết quả nhận dạng cảm xúc tiếng Việt với mô hình GMM được thử nghiệm chi tiết với nhiều bộ tham số khác nhau. Các tham số dùng cho thử nghiệm bao gồm các tham số đặc trưng MFCC, năng lượng, đặc trưng phổ, tần số cơ bản F0 và các biến thể của nó. Từ các kết quả này, luận án đưa ra những nhận xét, đánh giá và đề xuất bộ tham số để nhận dạng cảm xúc cho tiếng Việt nói sử dụng mô hình GMM.

Chương 4: Nhận dạng cảm xúc tiếng Việt nói sử dụng mô hình DCNN. Chương này trình bày nghiên cứu về mạng nơron lấy chập CNN, nghiên cứu và đề xuất mô hình DCNN cho nhận dạng cảm xúc tiếng Việt. Các tham số sử dụng bao gồm các đặc trưng về phổ mel, các tham số liên quan đến tuyến âm và các tham số liên quan đến nguồn âm như tần số cơ bản. Kết quả thử nghiệm nhận dạng cảm xúc với mô hình này cũng được thống kê chi tiết với từng tập ngữ liệu cảm xúc tiếng Việt và bộ tham số sử dụng.

Cuối cùng, phần Kết luận tổng hợp các kết quả nghiên cứu đã đạt được, những đóng góp mới và hướng mở rộng nghiên cứu phát triển của luận án.

Chương 1. TỔNG QUAN VỀ CẢM XÚC VÀ NHẬN DẠNG CẢM XÚC TIẾNG NÓI

1.1 Cảm xúc tiếng nói và phân loại cảm xúc

Phần này của luận án trình bày về cảm xúc tiếng nói và phân loại cảm xúc. Đã có các nghiên cứu đưa ra hơn 300 trạng thái cho những cảm xúc khác nhau. Tuy nhiên, không phải toàn bộ những cảm xúc đó đều được trải nghiệm trong đời sống hàng ngày. Về mặt này, hầu hết các nhà nghiên cứu đồng ý với lý thuyết Palette cho rằng, bất kỳ cảm xúc nào cũng đều được cấu thành từ sáu loại cảm xúc cơ bản giống như bất kỳ màu sắc nào đó đều là sự tổ hợp của 3 màu cơ bản [6]. Các nhà nghiên cứu cũng cho

rằng các cảm xúc giận dữ, ghê tởm, sợ hãi, vui, buồn và ngạc nhiên được coi là những cảm xúc chính yếu hoặc cơ bản hiển nhiên nhất [7]. Đây cũng được gọi là cảm xúc nguyên mẫu [8].

1.2 Nghiên cứu về nhận dạng cảm xúc

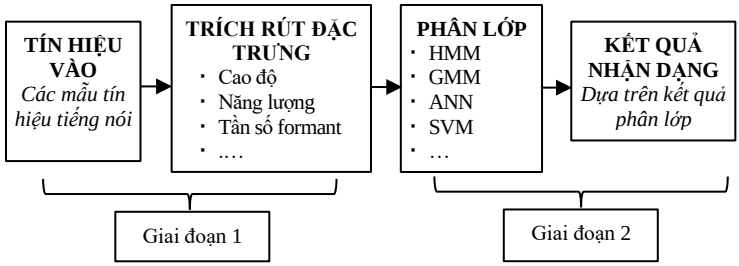
- Những kết quả nghiên cứu về nhận dạng cảm xúc hầu như chỉ mới tập trung vào một số ngôn ngữ thông dụng trên thế giới.
- Có nhiều bộ phân lớp được sử dụng nhưng khó đánh giá bộ phân lớp nào là tốt nhất
- Các nghiên cứu về cảm xúc tiếng Việt theo phương diện xử lý tín hiệu được thực hiện còn rất ít

1.3 Sơ đồ chung cho hệ thống nhận dạng cảm xúc tiếng nói

Các hệ thống nhận dạng cảm xúc tiếng nói thường gồm 2 giai đoạn:

Giai đoạn 1: Xử lý tín hiệu vào để trích rút các đặc trưng

Giai đoạn 2: Phân lớp dựa trên các mô hình nhận dạng



Hình 1.2 *Sơ đồ chung cho hệ thống nhận dạng cảm xúc tiếng nói*

Trên thực tế, phần lớn các nghiên cứu hiện tại trong nhận dạng cảm xúc đều tập trung vào giai đoạn 2 bởi vì giai đoạn này là kết nối giữa kết quả nhận dạng và các kỹ thuật phân lớp. Luận án sẽ tập trung vào các bộ phân lớp thống kê vì các bộ phân lớp này được dùng rộng rãi nhất trong bối cảnh nhận dạng cảm xúc tiếng nói.

1.4 Một số bộ phân lớp thường dùng cho nhận dạng cảm xúc

1.4.1 Bộ phân lớp phân tích phân biệt tuyến tính LDA

1.4.2 Bộ phân lớp phân tích khác biệt toàn phương QDA

1.4.3 Bộ phân lớp k láng giềng gần nhất k-NN

1.4.4 Bộ phân lớp hỗ trợ véctơ SVC

1.4.6 Bộ phân lớp HMM

1.4.7 Bộ phân lớp GMM [64]

1.4.8 Bộ phân lớp ANN

1.5 Một số kết quả nhận dạng cảm xúc được thực hiện trong và ngoài nước

Mục 1.5 trình bày một số kết quả nghiên cứu nhận dạng cảm xúc trong và ngoài nước. Hiện đã có nhiều kết quả nghiên cứu nhận dạng cảm xúc với các ngôn ngữ và mô hình nhận dạng cùng bộ tham số khác nhau. Tuy nhiên, với tiếng Việt còn rất ít các công trình nghiên cứu về nhận dạng cảm xúc tiếng Việt dựa trên phương diện xử lý tín hiệu tiếng nói. Một số nghiên cứu chủ yếu tập trung dựa vào ngôn ngữ hoặc kết hợp đa thể thức.

1.6 Kết chương 1

Chương 1 đã trình bày tổng quan nghiên cứu về phân loại cảm xúc và một số nghiên cứu mới về nhận dạng cảm xúc đã được tiến hành trong và ngoài nước. Các kỹ thuật nhận dạng đã liên tục được cải tiến nhằm cải thiện độ chính xác nhận dạng và đây vẫn là thách thức đối với các nhà nghiên cứu. Các kết quả cũng cho thấy, đối với tiếng Việt chưa có nhiều nghiên cứu được công bố, do đó cần có những nghiên cứu về nhận dạng cảm xúc của tiếng Việt nói để góp phần cải thiện các ứng dụng cho tiếng Việt có liên quan đến xử lý tiếng nói.

Chương 2. NGỮ LIỆU CẢM XÚC VÀ CÁC THAM SỐ ĐẶC TRƯNG CHO CẢM XÚC TIẾNG VIỆT NÓI

2.1 Phương pháp xây dựng ngữ liệu cảm xúc

Ngữ liệu tiếng nói được xây dựng dùng cho phát triển hệ thống tiếng nói có cảm xúc có thể được chia thành ba loại:

- Ngữ liệu tiếng nói có cảm xúc được xây dựng dựa trên đóng kịch
- Ngữ liệu tiếng nói có cảm xúc được xây dựng dựa trên suy diễn
- Ngữ liệu tiếng nói được xây dựng dựa trên cảm xúc tự nhiên

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.

2.2 Một số bộ ngữ liệu cảm xúc hiện có trên thế giới

Trong luận án đã thống kê 14 bộ ngữ liệu hiện có trên thế giới. Hầu hết các bộ ngữ liệu đều không được phổ biến rộng rãi nên khó có thể lấy để dùng chung cho các nghiên cứu. Nhìn chung, số lượng giọng nói và nội dung nói chưa nhiều, số lượng các phát ngôn cho các cảm xúc không đều nhau. Vì vậy, các nhà nghiên cứu sẽ khó so sánh kết quả trong quá trình đánh giá khi thử nghiệm.

2.3 Ngữ liệu cảm xúc tiếng Việt

Bộ ngữ liệu cảm xúc tiếng Việt dùng cho các nghiên cứu trong luận án được lựa chọn từ bộ ngữ liệu BKEmo [128]. Bộ ngữ liệu được sử dụng để nhận dạng trong luận án là ngữ liệu được chọn ra từ bộ ngữ liệu cảm xúc tiếng Việt BKEmo gồm 5584 file. Trong đó, số lượng file cảm xúc của mỗi giọng nam và nữ là 2792 file. Mỗi cảm xúc có 1396 file. Bộ ngữ liệu dùng để thử nghiệm nhận dạng cảm xúc tiếng Việt trong luận án được chia thành bốn tập ngữ liệu (Bảng 2.2).

Bảng 2.2 *Ngữ liệu cảm xúc tiếng Việt dùng cho thử nghiệm*

Tập ngữ liệu	Ngữ liệu thử nghiệm	Tổng số file	Số file huấn luyện	Số file thử nghiệm
Test1	Phụ thuộc cả người nói và nội dung	5584	2792	2792
Test2	Phụ thuộc người nói, độc lập nội dung	5584	2793	2791
Test3	Độc lập người nói, phụ thuộc nội dung	5584	2794	2790
Test4	Độc lập cả người nói và nội dung	2803	1403	1400

Bốn tập ngữ liệu trên sẽ dùng các ký hiệu như sau: Test1 được ký hiệu T1, Test2 được ký hiệu T2, Test3 được ký hiệu T3, Test4 được ký hiệu T4.

2.4 Tham số đặc trưng của tín hiệu tiếng nói dùng cho nhận dạng cảm xúc

2.4.1 Đặc trưng của nguồn âm và tuyến âm

Là các đặc trưng được trích rút từ nguồn âm và tuyến âm như các hệ số cepstrum tiên đoán tuyến tính (LPCC), các hệ số cepstrum theo thang tần số mel (MFCC), các hệ số tiên đoán tuyến tính cảm thụ (PLPC), formant, …

2.4.2 Đặc trưng ngôn điệu

Các đặc trưng của tiếng nói được trích chọn từ các đoạn tín hiệu tiếng nói dài hơn như âm tiết, từ và câu chính là các đặc trưng ngôn điệu. Bao gồm chu kỳ cơ bản, thời hạn, năng lượng, cao độ, tốc độ nói,… và các dẫn xuất tương ứng của chúng như cực đại, cực tiểu, trung bình, phương sai, phạm vi giá trị và độ lệch chuẩn.

2.5 Tham số đặc trưng dùng cho nhận dạng cảm xúc tiếng Việt

2.5.1 Các hệ số MFCC

2.5.2 Năng lượng tiếng nói

2.5.3 Cường độ tiếng nói

2.5.4 Tần số cơ bản F0 và các biến thể của F0

Tiếng Việt là ngôn ngữ có thanh điệu, các thanh điệu trong tiếng Việt nói được thể hiện qua qui luật biến thiên tần số cơ bản $F0$. Vì vậy,

đặc trưng tần số cơ bản $F0$ và các biến thể của $F0$ sẽ là những tham số hữu ích cho nhận dạng cảm xúc tiếng Việt. Bao gồm: Đạo hàm $F0$, chuẩn hóa $F0$ theo giá trị trung bình của $F0$, chuẩn hóa $F0$ theo giá trị min $F0$ và max $F0$, chuẩn hóa $F0$ theo trung bình và độ lệch chuẩn của $F0$, đạo hàm $LogF0$, chuẩn hóa $LogF0$ theo giá trị min $LogF0$ và max $LogF0$, chuẩn hóa $LogF0$ theo trung bình *LogF0*, chuẩn hóa $LogF0$ theo trung bình và độ lệch chuẩn của $LogF0$.

2.5.5 Các formant và dải thông tương ứng

2.5.6 Các đặc trưng phổ

Bảng 2.6 thống kê các tham số đặc trưng sẽ được sử dụng cho các thử nghiệm nhận dạng bốn cảm xúc vui, buồn, tức, bình thường trong nghiên cứu của luận án.

Bảng 2.6 Các tham số đặc trưng được dùng cho nhận dạng cảm xúc tiếng Việt.

Chỉ số	Tham số đặc trưng	Số lượng
(1)	*Các hệ số MFCC*	19
(2)	*Đạo hàm bậc nhất MFCC*	19
(3)	*Đạo hàm bậc hai MFCC*	19
(4)	*Năng lượng, đạo hàm bậc nhất, bậc hai của năng lượng*	3
(5)	*Tần số cơ bản F0*	1
(6)	*Cường độ tiếng nói*	1
(7)	*Các formant và dải thông tương ứng*	8
(8)	*Các thành phần hài*	1
(9)	*Trọng tâm phổ*	1
(10)	*Mômen trung tâm*	1
(11)	*Skewness*	1
(12)	*Kurtosis*	1
(13)	*Độ lệch chuẩn tần số*	1
(14)	*Giá trị trung bình của phổ*	1
(15)	*Độ dốc và độ lệch chuẩn của phổ trung bình dài hạn LTAS (Long Term Average Spectrum)*	2
(16)	*dF0*	1
(17)	*F0NormAver*	1
(18)	*F0NormMinMax*	1
(19)	*F0NormAverStd*	1
(20)	*dLogF0*	1
(21)	*LogF0NormMinMax*	1
(22)	*LogF0NormAver*	1
(23)	*LogF0NormAverStd*	1

2.6 Phân tích ảnh hưởng của một số tham số đến khả năng phân biệt các cảm xúc của bộ ngữ liệu cảm xúc tiếng Việt

2.6.1 Phân tích phương sai ANOVA và kiểm định T

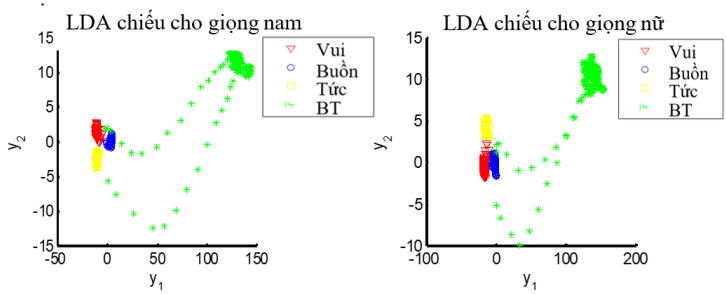
2.6.2 Ảnh hưởng của tham số đặc trưng đến phân biệt các cảm xúc

Kết quả phân tích ANOVA và kiểm định T cho thấy có thể phân biệt được bốn cảm xúc với nhau dựa trên các tham số đặc trưng về tần số, cường độ, formant và dải thông tương ứng, các đặc trưng phổ.

2.7 Đánh giá sự phân lớp của bộ ngữ liệu cảm xúc tiếng Việt

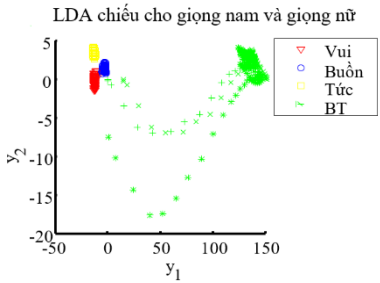
2.7.1 Kết quả phân lớp với LDA

Kết quả phân lớp bằng phương pháp LDA trên Hình 2.5 cho thấy, 4 cảm xúc vui, buồn, tức, bình thường được phân lớp tương đối rõ ràng cho cả giọng nam và giọng nữ. Trong 4 cảm xúc, cảm xúc bình thường được phân biệt rõ nhất so với 3 cảm xúc còn lại.



Hình 2.5 *Kết quả phân lớp cảm xúc giọng nam và nữ bằng LDA*

Hình 2.6 là kết quả phân lớp cảm xúc cho cả giọng nam và nữ. Cả bốn cảm xúc được quan sát phân biệt rõ ràng, việc phân cụm các cảm xúc của bộ ngữ liệu khá tốt trong đó cảm xúc bình thường được phân lớp khá tách biệt so với 3 cảm xúc còn lại.



Hình 2.6 *Kết quả phân lớp cảm xúc cả giọng nam và nữ bằng LDA*

2.7.2 Thử nghiệm nhận dạng cảm xúc tiếng Việt dựa trên bộ phân lớp IBk, SMO và Trees J48

2.7.2.1 Công cụ, ngữ liệu và tham số sử dụng

Phần này sử dụng các bộ phân lớp IBk, Trees J48, SMO thuộc bộ công cụ Weka để nhận dạng cảm xúc. Ngữ liệu dùng cho các thử nghiệm là tập

ngữ liệu T1 đã được trình bày trong Chương 2. Tham số được trích chọn gồm 384 tham số bằng công cụ OpenSmile.

2.7.2.2 Kết quả thử nghiệm

Kết quả thử nghiệm nhận dạng trên công cụ Weka với 3 bộ phân lớp trên cho thấy bộ ngữ liệu cảm xúc tiếng Việt đã đề xuất có chất lượng đảm bảo để thực hiện các thử nghiệm nhận dạng cảm xúc trong luận án.

***Bảng 2.9** Tỷ lệ (%) nhận dạng cảm xúc với 384 tham số*

Bộ phân lớp	**Cảm xúc** / **Cảm xúc**	Tức	Vui	Bình thường	Buồn	Trung bình
IBk	Tức	**99,07**	0,64	0,14	0,14	**98,17**
	Vui	0,93	**98,85**	0,07	0,14	
	Bình thường	0	0	**97,92**	2,08	
	Buồn	0	0,07	3,08	**96,85**	
SMO	Tức	**96,06**	3,65	0,29	0	**94,73**
	Vui	2,94	**96,13**	0,93	0	
	Bình thường	0,29	0,57	**93,12**	6,02	
	Buồn	0,21	0,79	5,37	**93,62**	
Trees J48	Tức	**77,65**	16,12	4,44	1,79	**80,64**
	Vui	15,47	**79,01**	3,87	1,65	
	Bình thường	4,37	4,15	**80,8**	10,67	
	Buồn	1,36	1,79	11,75	**85,1**	

***Bảng 2.10** Tỷ lệ (%) nhận dạng cảm xúc chỉ dùng 228 tham số liên quan đến MFCC*

Bộ phân lớp	**Cảm xúc** / **Cảm xúc**	Tức	Vui	Bình thường	Buồn	Trung bình
IBk	Tức	**98,28**	1,29	0,29	0,14	**98,17**
	Vui	0,93	**98,93**	0,07	0,07	
	Bình thường	0	0	**98,85**	1,15	
	Buồn	0	0	2,51	**97,49**	
SMO	Tức	**93,34**	5,80	0,72	0,14	**94,73**
	Vui	5,23	**93,34**	1,36	0,07	
	Bình thường	0,36	0,86	**92,34**	6,45	
	Buồn	0,14	1,72	6,09	**92,05**	
Trees J48	Tức	**77,36**	17,62	3,65	1,36	**80,64**
	Vui	16,48	**77,29**	3,94	2,29	
	Bình thường	3,65	2,58	**80,30**	13,47	
	Buồn	1,5	2,22	13,97	**82,31**	

***Bảng 2.11** Tỷ lệ (%) nhận dạng cảm xúc chỉ dùng 48 tham số liên quan đến F0 và năng lượng*

Bộ phân lớp	**Cảm xúc** / **Cảm xúc**	Tức	Vui	Bình thường	Buồn	Trung bình
IBk	Tức	**84,96**	10,32	3,22	1,50	**82,59**
	Vui	9,96	**84,1**	4,51	1,43	

	Bình thường	2,15	3,58	**78,3**	15,97	
	Buồn	1,50	0,93	14,54	**83,02**	
SMO	Tức	**81,95**	12,75	3,80	1,50	**77,73**
	Vui	13,04	**79,01**	7,16	0,79	
	Bình thường	2,22	7,09	**64,68**	26	
	Buồn	1,00	2,36	11,17	**85,46**	
Trees J48	Tức	**77,65**	15,62	5,01	1,72	**75,25**
	Vui	16,26	**75,36**	7,09	1,29	
	Bình thường	5,52	6,59	**69,41**	18,48	
	Buồn	1,22	2,36	17,84	**78,58**	

2.8 Kết chương 2

Chương 2 đã trình bày các phương pháp xây dựng ngữ liệu tiếng nói có cảm xúc để thực hiện các nghiên cứu về nhận dạng cảm xúc và cách lựa chọn, phân tích đánh giá bộ ngữ liệu cảm xúc tiếng Việt.

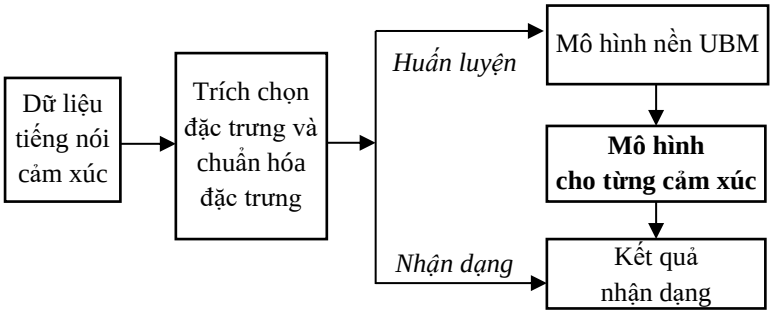
Bộ ngữ liệu này đã được nghe và đánh giá mức độ phân lớp bằng phương pháp LDA, đánh giá tỷ lệ nhận dạng đúng bằng mô hình SMO, IBk, Trees J48 của bộ công cụ Weka. Kết quả cho thấy bộ ngữ liệu có sự phân lớp rõ ràng các cảm xúc với nhau và đáng tin cậy để thực hiện các thử nghiệm nhận cảm xúc đối với tiếng Việt.

Kết quả phân tích phương sai ANOVA và kiểm định *T* cho thấy các tham số liên quan đến tần số cơ bản $F0$, năng lượng và các đặc trưng phổ của tín hiệu tiếng nói đều có ảnh hưởng đến sự phân biệt các cảm xúc vui, buồn, tức và bình thường. Những kết quả này là cơ sở để tiến hành nghiên cứu thử nghiệm các mô hình nhận dạng cảm xúc cho tiếng Việt nói được trình bày trong các chương tiếp theo của luận án dựa trên bộ ngữ liệu và các tham số đã được đánh giá trong Chương 2.

Chương 3. NHẬN DẠNG CẢM XÚC TIẾNG VIỆT NÓI VỚI MÔ HÌNH GMM

3.1 Mô hình GMM cho nhận dạng cảm xúc

GMM là thích hợp cho nhận dạng cảm xúc tiếng nói bởi chỉ có đặc trưng tổng quan được trích rút từ tiếng nói dùng cho huấn luyện. Trên thực tế, GMM đã được dùng khá phổ biến cho các trường hợp định danh người nói, định danh ngôn ngữ, định danh phương ngữ, hoặc phân lớp thể loại âm nhạc. 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.



Hình 3.1 *Sơ đồ mô hình GMM tổng quát cho nhận dạng cảm xúc*

3.2 Công cụ, tham số và ngữ liệu sử dụng

Bộ công cụ Alize được sử dụng để đánh giá mô hình GMM và thực hiện nhận dạng cảm xúc. Matlab là ngôn ngữ lập trình trung gian dùng để kết nối, phối hợp, tính toán và thiết lập các cấu hình tương ứng. Vì vậy việc nhận dạng cảm xúc tiếng Việt trong nghiên cứu của luận án đã được thực hiện hoàn toàn tự động. Ngữ liệu dùng cho các thử nghiệm trong mục 3.3 sau đây gồm 4 tập ngữ liệu T1, T2, T3 và T4 và đã được trình bày trong Bảng 2.2 của Chương 2. Tham số sử dụng trong phần thử nghiệm này gồm các tham số đã được trình bày chi tiết ở mục 2.5 của Chương 2. Mỗi thử nghiệm được thực hiện với số thành phần Gauss *M* tăng từ 16 đến 8192 theo lũy thừa 2.

3.3 Các thử nghiệm nhận dạng

Luận án đã tiến hành 13 thử nghiệm nhận dạng với mô hình GMM.

Bảng 3.1 *Các thử nghiệm nhận dạng cảm xúc với GMM*

Các thử nghiệm	Tập tham số	Ghi chú	Số lượng tham số
Thử nghiệm 1	*MFCC*	19 MFCC	19
Thử nghiệm 2	*MFCC+Delta1*	19 MFCC + 19 Delta1của MFCC	38
Thử nghiệm 3	*MFCC+Delta12*	19 MFCC + 19 Delta1 và 19 Delta2 của MFCC	57
Thử nghiệm 4	*prm60*	MFCC+Delta12 + năng lượng + Delta1 và Delta2 của năng lượng	60
Thử nghiệm 5	*prm79*	prm60 + F0 + cường độ + 4 formant + 4 dải thông + 9 đặc trưng phổ	79
Thử nghiệm 6	*prm87*	prm79 + 8 biến thể F0	87
Thử nghiệm 7	*FeaSpec*	Các đặc trưng phổ	9
Thử nghiệm 8	*MFCC+FeaSpec*	19 MFCC + 9 đặc trưng phổ	28

Thử nghiệm 9	*MFCC+Delta1 +FeaSpec*	19 MFCC + 19 Delta1 + 9 đặc trưng phổ	47
Thử nghiệm 10	*MFCC+Delta12 +FeaSpec*	19 MFCC + 19 Delta1 và 19 Delta2 của MFCC + 9 đặc trưng phổ	66
Thử nghiệm 11	*MFCC+Delta12+ một trong 9 đặc trưng phổ*	19 MFCC + 19 Delta1 và 19 Delta2 của MFCC + một trong 9 đặc trưng phổ	58
Thử nghiệm 12	*prm60+F0+biến thể F0*	prm60 + F0 + 8 biến thể F0	69
Thử nghiệm 13	*prm79 + một trong 8 biến thể F0*		80

3.3.1 Thử nghiệm 1 đến Thử nghiệm 6

3.3.1.1 Nhận dạng đối với từng tập ngữ liệu

+ Với tập ngữ liệu T1: Kết quả cho thấy, nhìn chung tỷ lệ nhận dạng đúng tăng dần khi M tăng lên. Khi sử dụng bộ prm87 để nhận dạng, tỷ lệ nhận dạng đúng trung bình là 98,96% đạt cao nhất so với năm trường hợp còn lại và nằm trong khoảng từ 97,53% - 99,97%.

+ Với tập ngữ liệu T2: Khi sử dụng bộ tham số prm87, tỷ lệ nhận dạng đúng đạt cao nhất so với các bộ tham số còn lại và nằm trong khoảng 93% - 99,11%. Với 5 bộ tham số còn lại, tỷ lệ nhận dạng đúng nằm trong khoảng từ 72,29% - 85,71%.

+ Với tập ngữ liệu T3: Kết quả nhận dạng cho thấy, bộ tham số prm87 vẫn cho tỷ lệ nhận dạng đúng cao nhất và trung bình là 85,44%. Đặc biệt, trong thử nghiệm này, kết quả nhận dạng đạt tỷ lệ cao nhất là 90,14% với $M = 16$ còn thấp nhất là 80,54% với $M = 256$.

+ Với tập ngữ liệu T4: Với thử nghiệm với T4, tỷ lệ nhận dạng đúng cho bộ tham số prm87 cao hơn hẳn so với các bộ tham số còn lại. Khi $M = 1024$, tỷ lệ này đạt cao nhất là 94,22% còn tỷ lệ nhận dạng đúng trung bình là 90,76%. Các bộ tham số còn lại có tỷ lệ nhận dạng đúng thấp hơn và trong khoảng từ 52,69% - 69,40%.

3.3.1.2 Nhận dạng đối với từng cảm xúc

+ Với tập ngữ liệu T1: Cả bốn cảm xúc đều đạt tỷ lệ nhận dạng đúng cao nhất khi sử dụng tập tham số prm87 với tỷ lệ trung bình nhận dạng đúng lần lượt là 99,66%, 98,77%, 97,7%, 90,64% cho các cảm xúc bình thường, tức, vui và buồn. Khi sử dụng tập tham số *prm87* và $M = 4096$, tỷ lệ nhận nhầm giữa các cảm xúc là thấp nhất.

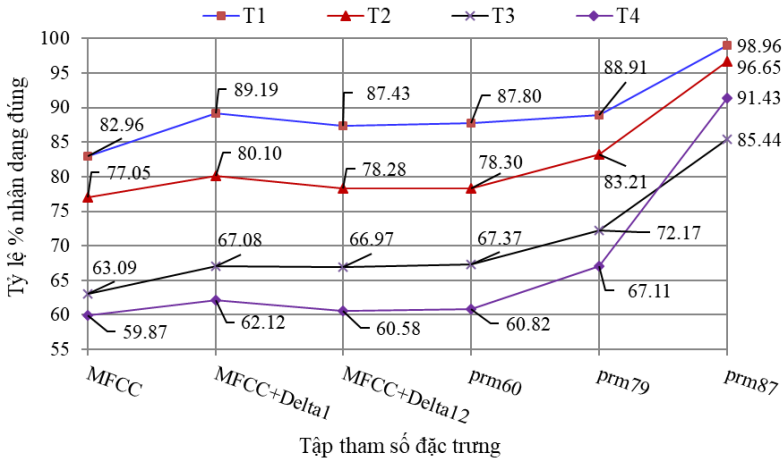
+ Với tập ngữ liệu T2: Tỷ lệ nhận dạng đúng nhận được khi sử dụng

prm87 lần lần lượt là 98,82% (vui), 97,24% (bình thường), 94,97% (tức) và 86,88% (buồn). Nếu dùng bộ tham số *prm87* và $M = 128$ thì tỷ lệ nhận dạng nhầm lẫn giữa các cảm xúc sẽ thấp nhất. Tính trung bình, tỷ lệ nhận dạng đúng của 4 cảm xúc là 93% còn tỷ lệ nhận nhầm là 0,42%.

+ Với tập ngữ liệu T3: Tỷ lệ nhận dạng cao nhất khi sử dụng tập tham số prm87 đối với cảm xúc vui là 91,15%, tức là 91,98%, bình thường là 95,52% và buồn là 68,13%. Tỷ lệ nhận dạng nhầm lẫn từ cảm xúc bình thường sang cảm xúc buồn là 23,42% và là tỷ lệ cao nhất. Tỷ lệ nhận dạng đúng trung bình của 4 cảm xúc đối với T3 là 80,54% còn trung bình tỷ lệ nhận dạng nhầm lẫn là 2,7%.

+ Với tập ngữ liệu T4: , khi sử dụng tập tham số *prm87*, tỷ lệ nhận dạng đúng các cảm xúc đều tăng cao: vui (97,17%), tức (98,15%), bình thường (97,08%), trừ cảm xúc buồn giảm xuống (64,33%) so với ba cảm xúc còn lại. Tỷ lệ nhận nhầm từ cảm xúc bình thường sang buồn là cao nhất và bằng 25,43% còn tỷ lệ nhận nhầm từ cảm xúc tức sang vui chỉ bằng 1,14%. Các cặp cảm xúc khác có tỷ lệ nhận nhầm bằng 0%. Tỷ lệ nhận dạng đúng trung bình của 4 cảm xúc là 84,42%, tỷ lệ nhận nhầm trung bình là 2,21%.

3.3.1.3 So sánh kết quả của 6 thử nghiệm



Hình 3.12 *Tỷ lệ nhận dạng đúng trung bình cảm xúc của 4 thử nghiệm*

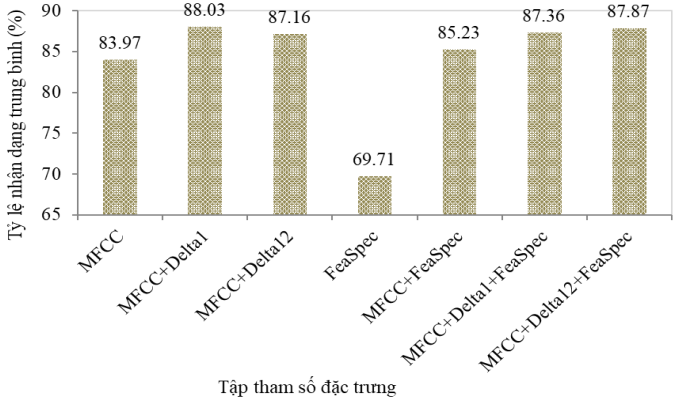
Tỷ lệ nhận dạng đúng trung bình của các cảm xúc đối với T1 cao nhất và bằng 89,21%, tiếp đến là tập ngữ liệu T2 bằng 82,27%, với tập ngữ liệu T3 là 70,35% còn tập ngữ liệu T4 là 66,99%. Điều này là phù hợp vì trong thử nghiệm với T1, giai đoạn huấn luyện và nhận dạng đều có chung

người nói, nội dung nói giống nhau chỉ khác nhau ở thời điểm phát âm. Vì vậy, tỷ lệ nhận dạng sẽ đạt cao nhất.

Qua thử nghiệm 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*.

3.3.2 Thử nghiệm 7 đến Thử nghiệm 10

Các thử nghiệm trong phần này được thực hiện với tập ngữ liệu T1. Các tập tham số bao gồm: *FeaSpec*, *MFCC+FeaSpec*, *MFCC+Delta1+FeaSpec*, *MFCC+Delta12+FeaSpec*. Kết quả của các thử nghiệm này được so sánh với kết quả thử nghiệm với 3 tập tham số liên quan đến MFCC.



Hình 3.14 *Tỷ lệ nhận dạng đúng trung bình cho 7 tập tham số đã nêu với T1*

Hình 3.14 thống kê tỷ lệ nhận dạng đúng trung bình cho 7 thử nghiệm. Tỷ lệ nhận dạng đúng trung bình là thấp nhất khi chỉ dùng đặc trưng phổ và bằng 69,71%. Tỷ lệ nhận dạng đúng trung bình đạt cao nhất bằng 88,03% khi dùng *MFCC+Delta1*. Tỷ lệ nhận dạng đúng trung bình đạt cao nhất bằng 88,03% khi dùng *MFCC+Delta1*. Nếu dùng *MFCC+Delta12* 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+Delta12+FeaSpec*.

3.3.3 Thử nghiệm 11

Kết quả đánh giá ảnh hưởng của từng đặc trưng phổ khi được kết hợp với *MFCC+Delta1* trên tập ngữ liệu T1 được trình bày ở Bảng 3.6.

Bảng 3.6 *Tỷ lệ nhận dạng trung bình của M khi kết hợp MFCC+Delta1 với mỗi đặc trưng phổ cho các cảm xúc đối với T1*

Thứ tự	Tham số	Tỷ lệ (%) nhận dạng đúng cho từng cảm xúc			
		Vui	Buồn	Tức	Bình thường
1	*Harmonicity*	88,41	90,43	89,41	85,20
2	*Center of gravity*	88,78	90,76	89,31	85,09
3	*Standard deviation*	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	*Central spectral moment*	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	*Standard deviation of LTAS*	88,48	90,46	90,13	85,65

3.3.4 Thử nghiệm 12

Trong phần này, luận án đã nghiên cứu và đánh giá việc nhận dạng cảm xúc sử dụng tập tham số *prm60* kết hợp với tần số cơ bản và các biến thể của nó. Có 3 trường hợp đã được tiến hành bao gồm: chỉ dùng *prm60*, *prm60+F0* và *prm60+F0+biến thể F0*. Các trường hợp này được thực hiện với cả bốn tập ngữ liệu T1, T2, T3 và T4.

+ Kết quả thử nghiệm đối với T1: Kết quả nhận dạng khi sử dụng bộ tham số *pm60+F0+8 biến thể* của $F0$ cho tỷ lệ nhận dạng cao hơn hẳn so với chỉ dùng *prm60* hoặc *prm60+F0*, độ chính xác của thử nghiệm dùng bộ tham số này đã đạt trung bình từ 96,49% đến 99,93%. Nếu chỉ dùng *prm60+F0* thì tỷ lệ này tăng ít và gần như xấp xỉ bằng tỷ lệ của *prm60*.

+ Kết quả thử nghiệm đối với T2: Tỷ lệ nhận dạng khi sử dụng $F0$ và các biến thể của $F0$ cao hơn hẳn so với chỉ dùng *prm60*, độ chính xác trung bình từ 91,83% - 98,82%. Khi sử dụng *prm60*, tỷ lệ này là 72,86% - 81,36%.

+ Kết quả thử nghiệm đối với T3: Kết quả nhận dạng đối với tập ngữ liệu T3 cũng cho thấy, khi thêm $F0$ và biến thể $F0$, tỷ lệ nhận dạng cũng tăng lên đáng kể. Tỷ lệ nhận dạng cao nhất đạt được là 94,39% khi sử dụng *prm60+F0* và $M = 16$.

+ Kết quả thử nghiệm đối với T4: Thử nghiệm kết hợp *prm60* với $F0$ và biến thể của $F0$ cũng cho thấy, kết quả nhận dạng cao hơn hẳn so với chỉ sử dụng *prm60*. Tỷ lệ nhận dạng cao nhất đạt 94,95% đối với *prm60+F0+biến thể F0*. Nếu chỉ sử dụng *prm60*, tỷ lệ nhận dạng đạt được chỉ từ 52,69% - 64,99%.

3.3.5 Thử nghiệm 13

Thử nghiệm 13 sử dụng tập tham số gồm *prm79* kết hợp với một trong 8 biến thể $F0$ nhằm xem xét ảnh hưởng của mỗi biến thể này với từng cảm xúc. Có 8 tập tham số được đánh số từ S1 đến S8 với số lượng tương ứng các tham số được trình bày trong Bảng 3.8

Bảng 3.8 *Tập tham số prm79 kết hợp với một trong 8 biến thể của F0*

Bộ tham số	**Tên bộ tham số**	**Các tham số đặc trưng ứng với các chỉ số ở Bảng 2.6**	**Số lượng**
S1	*Prm79+dF0*	*prm79* + đạo hàm của *F0*	80
S2	*prm79+F0NormAver*	*prm79* + chuẩn hóa F0 theo giá trị trung bình của *F0*	80
S3	*prm79+F0NormMinMax*	*prm79* + chuẩn hóa *F0* theo giá trị max *F0* và min *F0*	80
S4	*prm79+F0NormAverStd*	*prm79* + chuẩn hóa *F0* theo giá trị trung bình và độ lệch chuẩn của *F0*	80
S5	*prm79+dLogF0*	*prm79* + đạo hàm của log*F0*	80
S6	*prm79+LogF0NormMinMax*	*prm79* + chuẩn hóa log*F0* theo giá trị min của log*F0* và max của log*F0*	80
S7	*prm79+LogF0NormAver*	*prm79* + chuẩn hóa log*F0* theo giá trị trung bình của log*F0*	80
S8	*prm79+LogF0NormAverStd*	*prm79* + chuẩn hóa log*F0* theo trung bình và độ lệch chuẩn của log*F0*	80

Bảng 3.9 *Tỷ lệ (%) nhận dạng trung bình các cảm xúc đối với 4 tập ngữ khi sử dụng kết hợp prm79 với biến thể F0*

Tập ngữ liệu	**Tập tham số**								
	prm79	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
T1	86,80	96,73	96,66	96,70	96,66	96,75	96,73	96,73	96,73
T2	81,18	94,50	93,92	94,21	94,46	94,41	94,45	94,07	94,42
T3	70,38	83,52	81,92	77,51	83,20	83,30	81,94	82,55	82,95
T4	65,39	88,25	88,37	88,11	88,20	88,07	88,79	88,31	88,22

Kết quả thử nghiệm đối với từng cảm xúc cho các tập ngữ liệu đều cho tỷ lệ nhận dạng cao hơn khi thêm một trong 8 biến thể của F0 vào tập prm79 so với chỉ dùng prm79.

3.4 Đánh giá sự ảnh hưởng của tần số cơ bản

Các nghiên cứu thử nghiệm đã trình bày ở mục 3.3 cho thấy tần số cơ bản có tầm ảnh hưởng rất lớn đến kết quả nhận dạng các cảm xúc tiếng Việt. Khi các tham số liên quan trực tiếp đến $F0$ được thêm vào, tỷ lệ nhận dạng tăng đáng kể so với việc bổ sung các tham số liên quan trực tiếp đến phổ. Khi thêm 8 biến thể của $F0$ (từ prm79 lên prm87), tỷ lệ nhận dạng trung bình tăng mạnh nhất đối với T4 là 24,32%.

Kết quả trong Thử nghiệm 12 cũng cho thấy, tỷ lệ nhận dạng tăng lên rất nhiều đối với cả 4 tập ngữ liệu khi sử dụng tập tham số *prm60+F0+biến thể F0* so với chỉ sử dụng tập tham số prm60.

Các kết quả nhận dạng đối với từng cảm xúc được trình bày trong Thử nghiệm 13 cho kết quả nhận dạng tốt khi kết hợp tập tham số *prm79* với một trong 8 biến thể của $F0$. Luận án đã thử nghiệm nhận dạng sử dụng các biến thể $F0$ và 79 tham số khác cho các tập ngữ liệu từ T1 đến T4, với M=512. Với T1, các biến thể $F0$ (18), (19), (20), (22) và (23) đã cho tỷ lệ nhận dạng tăng lên tối đa và đạt 100%. Khi thêm biến thể $F0$ (23) thì T1, T3 và T4 có tỷ lệ nhận dạng cao nhất và tỷ lệ này lần lượt là 100%, 87,42% và 93,46%.

3.5 Quan hệ giữa số thành phần Gauss và tỷ lệ nhận dạng

Các thử nghiệm nhận dạng cảm xúc với mô hình GMM cho thấy, tỷ lệ nhận dạng thay đổi theo số thành phần Gauss được sử dụng trong mô hình. 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 [2]. 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 theo thời gian tính toán cần thiết và độ chính xác nhận dạng theo yêu cầu.

3.6 Kết chương 3

Chương 3 của luận án đã trình bày các kết quả nghiên cứu về nhận dạng cảm xúc tiếng Việt nói dựa trên mô hình nhận dạng GMM cùng với các bộ tham số đặc trưng khác nhau.

GMM là một mô hình khá thích hợp cho nhận dạng cảm xúc tiếng Việt. Tỷ lệ nhận dạng với tập ngữ liệu cảm xúc tiếng Việt phụ thuộc cả người nói và nội dung đạt tới 99,97% khi sử dụng bộ tham số *prm87*, với

ngữ liệu độc lập cả người nói và nội dung đạt 97,58% khi sử dụng bộ tham số *prm79* kết hợp với biến thể *LogF0NormMinMax* của $F0$.

Với những kết quả nhận dạng đã phân tích và đánh giá trong chương này, luận án đề xuất một mô hình tốt để nhận dạng cảm xúc tiếng Việt với GMM là cần phải kết hợp *MFCC*, các đặc trưng phổ và đặc biệt là tần số cơ bản $F0$ và biến thể của $F0$.

Chương 4. NHẬN DẠNG CẢM XÚC TIẾNG VIỆT SỬ DỤNG MÔ HÌNH DCNN SÂU

4.1 Mô hình mạng nơron lấy chập

Mạng nơron lấy chập CNN là một trong những giải thuật học sâu cho kết quả tốt nhất hiện nay trong hầu hết các bài toán về thị giác máy như phân lớp, nhận dạng. Về cơ bản CNN là một kiểu mạng ANN truyền thẳng, trong đó kiến trúc chính gồm nhiều thành phần được ghép nối với nhau theo cấu trúc nhiều tầng bao gồm: lấy chập (Convolution), lấy gộp (Pooling), kích hoạt phi tuyến (Non-linear activation) và kết nối đầy đủ (Fully-connected).

4.1.1 Lấy chập

Lấy chập là thao tác đầu tiên quan trọng nhất trong cấu trúc của mạng học sâu CNN. Đầu vào của phép lấy chập là một mảng các giá trị của dữ liệu. Để thực hiện lấy chập, một bộ lọc (filter) còn gọi là kernel được di chuyển qua các vị trí trên toàn bộ ma trận ảnh. Thao tác lấy chập được thực hiện tại các vị trí mà bộ lọc đi qua. Ý nghĩa của thao tác lấy chập là xác định khả năng xuất hiện các mẫu tại các vị trí nhất định trong ảnh. Mỗi mẫu được biểu diễn bằng trọng số của cửa sổ tương ứng với một bộ lọc. Mỗi vị trí của bộ lọc sẽ tính được một giá trị theo công thức:

$$y = \sum_i w_i x_i + b \tag{4.1}$$

Trong công thức (4.1), x_i bao gồm các điểm ảnh phổ nằm trong phạm vi cửa sổ đang quét, b là hệ số độ lệch.

4.1.2 Kích hoạt phi tuyến

Sau mỗi lớp lấy chập, đầu ra của ánh xạ lấy chập thường được cho qua hàm kích hoạt phi tuyến để tăng tính phi tuyến của mô hình và toàn mạng. Một số hàm kích hoạt phi tuyến thường dùng như ReLU (Rectified Linear Unit), ELU (Exponential Linear Unit).

4.1.3 Lấy gộp

Tầng Pool (hay còn gọi subsampling hoặc downsampling) là một trong những thành phần tính toán chính trong cấu trúc CNN. Xét về mặt toán

học, pooling thực chất là quá trình tính toán trên ma trận đầu vào trong đó mục tiêu đạt được sau khi tính toán là giảm kích thước ma trận nhưng vẫn làm nổi bật lên được đặc trưng có trong ma trận đầu vào. Có nhiều toán tử pooling như sum-pooling, max-pooling, L2-pooling song max-pooling thường được sử dụng.

4.1.4 Kết nối đầy đủ

Kết nối đầy đủ là cách kết nối các nơron ở hai tầng với nhau trong đó tầng sau kết nối đầy đủ với các nơron ở tầng trước nó. Trong CNN, tầng này thường được sử dụng ở các tầng phía cuối của kiến trúc mạng kết nối với đầu ra của mạng. Lớp này cơ bản là lấy thông tin đầu vào (có thể là đầu ra của lớp lấy chập hoặc kích hoạt phi tuyến hoặc lớp gộp) còn đầu ra là véctơ N chiều với N là số lớp cần phân lớp.

4.2 Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt

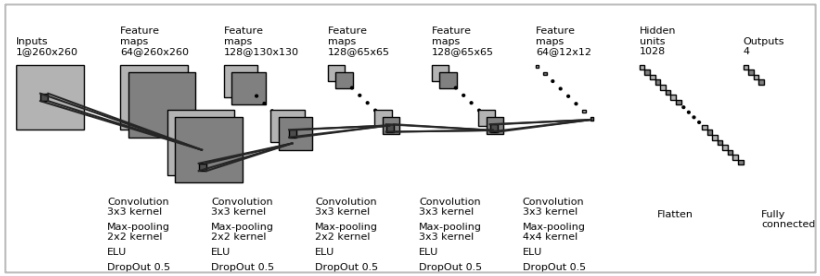
Tín hiệu tiếng nói đều có thể được biểu diễn bằng hình ảnh phổ mel để làm ảnh đầu vào cho CNN. Vì vậy, có thể sử dụng mô hình CNN để nhận dạng cảm xúc tiếng nói nói riêng và cho các xử lý tín hiệu tiếng nói nói chung. Cấu hình đầy đủ của mạng nơron DCNN sâu để huấn luyện được mô tả như Bảng 4.1 trong trường hợp mô hình baseline với 260 tham số.

Bảng 4.1 *Cấu trúc mạng DCNN cho nhận dạng cảm xúc tiếng Việt trong trường hợp 260 tham số*

Layer Index	Layer (type)	Output Shape	Param #
1	BatchNormalization	(260, 260, 1)	1040
	Convolution2D (3×3)	(260, 260, 64)	640
	BatchNormalization	(260, 260, 64)	256
	ELU	(260, 260, 64)	0
	MaxPooling2D (2×2)	(130, 130, 64)	0
	Dropout (0,5)	(130, 130, 64)	0
2	Convolution2D(3×3)	(130, 130, 128)	73856
	BatchNormalization	(130, 130, 128)	512
	ELU	(130, 130, 128)	0
	MaxPooling2D(2×2)	(65, 65, 128)	0
	Dropout (0,5)	(65, 65, 128)	0
3	Convolution2D(3×3)	(65, 65, 128)	147584
	BatchNormalization	(65, 65, 128)	512
	ELU	(65, 65, 128)	0
	MaxPooling2D(2×2)	(32, 32, 128)	0
	Dropout (0,5)	(32, 32, 128)	0

Layer Index	Layer (type)	Output Shape	Param #
4	Convolution2D(3×3)	(32, 32, 128)	147584
	BatchNormalization	(32, 32, 128)	512
	ELU	(32, 32, 128)	0
	MaxPooling2D(2×2)	(10, 10, 128)	0
	Dropout (0,5)	(10, 10, 128)	0
5	Convolution2D	(10, 10, 64)	73792
	BatchNormalization	(10, 10, 64)	256
	ELU	(10, 10, 64)	0
	MaxPooling2D(2×2)	(2, 2, 64)	0
	Dropout (0.5)	(2, 2, 64)	0
	Flatten	(256)	0
	Dense	(4)	1028

Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt với trường hợp sử dụng tập 260 tham số được trình bày trên Hình từ 4.8.



Hình 4.8 *Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt với 260 tham số*

Đối với mô hình có số lượng tham số lớn hơn 260, cấu hình mạng có thể dễ dàng được suy diễn theo cách tương tự.

4.3 Ngữ liệu và tham số dùng cho thử nghiệm

Để thực hiện các thử nghiệm với DCNN, bốn tập ngữ liệu T1, T2, T3 và T4 trong Bảng 2.2 của Chương 2 được phân chia theo tỷ lệ số file tiếng nói là 2-1-1 tương ứng với huấn luyện - đánh giá - thử nghiệm.

Các tham số sử dụng nhận dạng cảm xúc với mô hình DCNN được thống kê trong Bảng 4.6. Trong đó, các thử nghiệm được thực hiện đối với năm tập tham số và bốn tập ngữ liệu.

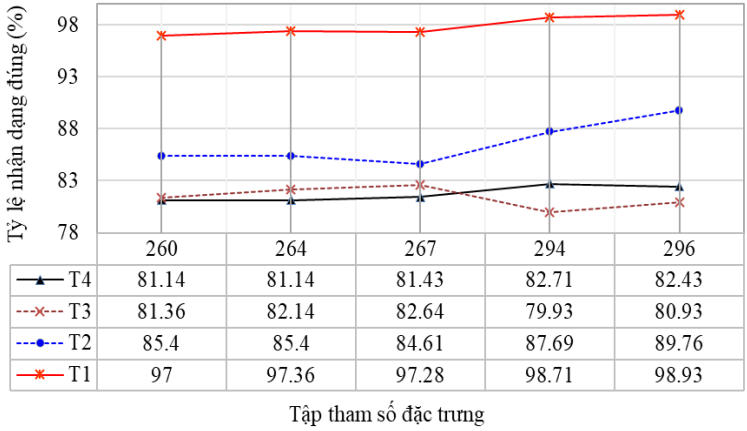
Bảng 4.6 *Năm tập tham số thử nghiệm nhận dạng với DCNN*

Tập tham số	Các tham số sử dụng
260	260 hệ số *MFCC*

Tập tham số	Các tham số sử dụng
264	- 260 hệ số *MFCC* - Tần số cơ bản *F*0 - 3 biến thể của *F*0: *F0NormMinMax, logF0NormAver, logF0NormMinMax*
267	- 264 tham số - 3 biến thể *F*0: *F0NormAver, F0NormAverStd, logF0NormAverStd*
294	- 260 hệ số *MFCC* - *Intensity*, *F*0 - 5 biến thể *F*0: *F0NormAver, F0NormMinMax, F0NormAverStd, logF0NormMinMax, logF0NormAverStd* - 4 formant và dải thông tương ứng - 5 đặc trưng phổ: *harmonicity, centre of gravity, central moment, skewness, kurtosis* - 14 hệ số đáp ứng xung của bộ lọc đảo của tuyến âm
296	- 294 tham số - 2 tham số liên quan đến *F*0: *dF0, logF0NormAver*

4.4 Thử nghiệm nhận dạng cảm xúc tiếng Việt bằng mô hình DCNN

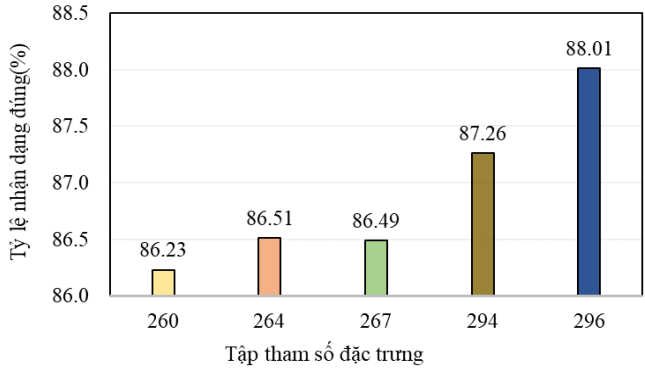
Trong các thử nghiệm với 5 bộ tham số khác nhau, tỷ lệ nhận dạng đạt cao nhất ứng với tập ngữ liệu T1, T2 khi sử dụng 296 tham số. Đối với T3, tỷ lệ nhận dạng là cao nhất khi sử dụng 267 tham số, còn đối với T4 cao nhất khi sử dụng 294 tham số. Tỷ lệ nhận dạng trung bình của tất cả các thử nghiệm đối với từng bộ tham số được trình bày trên Hình 4.13.



	260	264	267	294	296
T4	81.14	81.14	81.43	82.71	82.43
T3	81.36	82.14	82.64	79.93	80.93
T2	85.4	85.4	84.61	87.69	89.76
T1	97	97.36	97.28	98.71	98.93

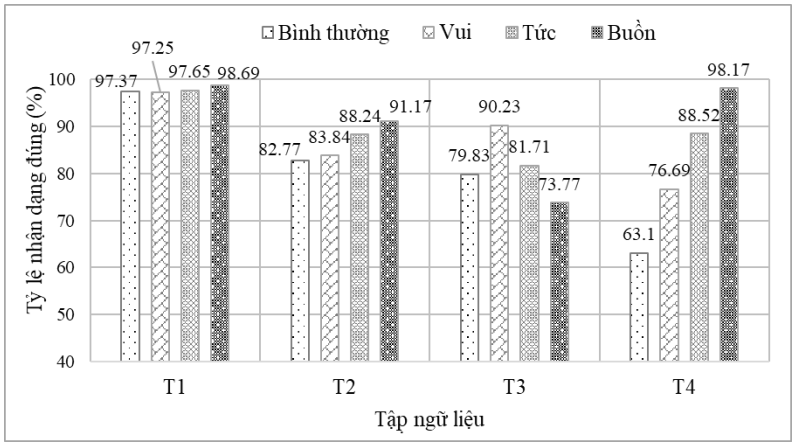
Hình 4.13 *Kết quả nhận dạng với 5 tập tham số cho 4 tập ngữ liệu*

Hình 4.14 cho thấy tỷ lệ nhận dạng trung bình của các tập ngữ liệu đạt cao nhất khi sử dụng 296 tham số và nhỏ nhất khi sử dụng 260 tham số. Như vậy, việc bổ sung các đặc trưng về năng lượng, phổ, tần số cơ bản $F0$ và biến thể của $F0$, formant và dải thông tương ứng đã tăng tỷ lệ nhận dạng. Đặc biệt, ảnh hưởng của hai tham số liên quan đến tần số cơ bản $F0$ là $dF0$ và $logF0NormAver$ khi được sử dụng trong bộ tham số 296 đã nâng tỷ lệ nhận dạng lên tốt hơn (từ 87,26% lên 88,01%).



Hình 4.14 *Tỷ lệ nhận dạng trung bình của các thử nghiệm với 5 tập tham số*

Tỷ lệ nhận dạng trung bình của từng bộ tham số ứng với từng cảm xúc được thống kê trên Hình 4.16



Hình 4.16 *Tỷ lệ nhận dạng đúng trung bình của mỗi cảm xúc đối với từng tập ngữ liệu*

4.5 Kết chương 4

Chương 4 đã trình bày kết quả nhận dạng bốn cảm xúc sử dụng mô hình DCNN. Tính trung bình, độ chính xác nhận dạng tối đa đạt được là

97,86% đối với phụ thuộc vào nội dung và phụ thuộc vào người nói. Kết quả của các thử nghiệm cũng cho thấy $F0$ và các biến thể của nó góp phần đáng kể vào sự gia tăng độ chính xác của nhận dạng cảm xúc tiếng Việt. Đối với thử nghiệm sử dụng mô hình DCNN, cảm xúc buồn cho tỷ lệ cao hơn các cảm xúc còn lại.

KẾT LUẬN VÀ ĐỊNH HƯỚNG PHÁT TRIỂN

1. Kết luận

Luận án đã thực hiện nghiên cứu về cảm xúc cũng như khái quát các nghiên cứu nhận dạng cảm xúc hiện nay trên thế giới và trong nước từ đó nghiên cứu đánh giá ngữ liệu, tham số đặc trưng, thử nghiệm với các mô hình nhận dạng và đưa ra mô hình chung cho nhận dạng cảm xúc tiếng Việt. Với những mục tiêu đã đề ra ban đầu, luận án đã hoàn thành được các mục tiêu đó. ***Đóng góp khoa học của luận án:***

(1) Sử dụng các phương pháp thích hợp để đánh giá bộ ngữ liệu cảm xúc tiếng Việt từ đó đề xuất được bộ ngữ liệu cảm xúc tiếng Việt dùng cho thử nghiệm nhận dạng cảm xúc tiếng Việt nói.

(2) Nghiên cứu, khai thác và đề xuất được các mô hình GMM, DCNN và các tham số đặc trưng phù hợp cho nhận dạng cảm xúc tiếng Việt nói đồng thời đánh giá được ảnh hưởng của các tham số đặc trưng đến kết quả nhận dạng cảm xúc tiếng Việt với bốn cảm xúc vui, buồn, tức và bình thường.

2. Định hướng phát triển

Từ các kết quả nghiên cứu đã được thực hiện, luận án đề xuất các kiến nghị sau nhằm mở rộng hướng nghiên cứu hiện có:

- Mở rộng nghiên cứu nhận dạng cho các hình thái cảm xúc khác đối với tiếng Việt nói.
- Mở rộng nghiên cứu thử nghiệm nhận dạng với mô hình mạng nơron như điều chỉnh cấu hình mạng, các tham số đầu vào, số lượng tham số.
- Nghiên cứu thử nghiệm với các mô hình nhận dạng khác.
- Tiếp cận hướng nghiên cứu nhằm đảm bảo độ chính xác nhận dạng khi ngữ liệu trong môi trường thực không hoàn toàn như ngữ liệu đã được huấn luyện.
- Kết hợp việc nhận dạng cảm xúc tiếng Việt nói với nhận dạng tiếng Việt nói để góp phần hướng tới xây dựng các hệ thống tương tác người-máy hoạt động hoàn thiện và hiệu quả.