LỜI CAM ĐOAN

Tôi xin cam đoan tất cả các nội dung trong luận án "Nhận dạng cảm xúc cho tiếng Việt nói" là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả trong luận án là trung thực và chưa từng được tác giả khác công bố. Việc tham khảo các nguồn tài liệu đã được thực hiện trích dẫn và ghi nguồn tài liệu tham khảo đúng quy định.

Hà Nội, ngày tháng năm 2019

TẬP THỂ HƯỚNG DẪN KHOA HỌC

PGS.TS. Trịnh Văn Loan

TS. Nguyễn Hồng Quang

TÁC GIẢ LUẬN ÁN

Đào Thị Lệ Thủy

LỜI CẢM ƠN

Để hoàn thành luận án này không chỉ là sự cố gắng nỗ lực của cá nhân tôi mà còn có sự hỗ trợ và giúp đỡ tận tình của các thầy hướng dẫn, nhà trường, bộ môn và gia đình. Vì vậy, tôi muốn bày tỏ lòng biết ơn của mình đến các thầy cô, đồng nghiệp và gia đình đã giúp đỡ để tôi có được kết quả này.

Trước hết, tôi xin gửi lời cảm ơn sâu sắc tới hai người thầy hướng dẫn của tôi, PGS.TS. Trịnh Văn Loan và TS. Nguyễn Hồng Quang. Hai thầy đã luôn tận tình giúp đỡ tôi trong suốt quá trình nghiên cứu, đưa ra những lời khuyên, những định hướng khoa học và phương pháp thực hiện rất quý báu để tôi có thể triển khai thực hiện và hoàn thành luận án của mình.

Tiếp theo, tôi xin trân trọng cảm ơn Trường Đại học Bách khoa Hà Nội, Viện Công nghệ Thông tin và Truyền thông, Bộ môn Kỹ thuật Máy tính đã tạo điều kiện thuận lợi cho tôi trong quá trình học tập tại Trường. Tôi xin chân thành cảm ơn các thầy cô, đồng nghiệp của Trường Cao đẳng nghề Công nghệ cao Hà Nội, nơi tôi làm việc đã giúp đỡ và động viên tôi trong suốt quá trình nghiên cứu.

Cuối cùng tôi muốn bày tỏ lòng biết ơn sâu sắc tới cha mẹ và gia đình đã luôn bên cạnh ủng hộ, động viên giúp đỡ tôi vượt qua những trở ngại khó khăn để hoàn thành luận án này.

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT 6
DANH MỤC CÁC BẢNG 8
DANH MỤC CÁC HÌNH ẢNH VÀ ĐỒ THỊ 10
MỞ ĐẦU 13
Chương 1. TỔNG QUAN VỀ CẢM XÚC VÀ NHẬN DẠNG CẢM XÚC TIẾNG NÓI 17
1.1 Cảm xúc tiếng nói và phân loại cảm xúc 17
1.2 Nghiên cứu về nhận dạng cảm xúc 21
1.3 Sơ đồ chung cho hệ thống nhận dạng cảm xúc tiếng nói 26
1.4 Một số bộ phân lớp thường dùng cho nhận dạng cảm xúc 26
1.4.1 Bộ phân lớp phân tích phân biệt tuyến tính LDA 26
1.4.2 Bộ phân lớp phân tích khác biệt toàn phương QDA 27
1.4.3 Bộ phân lớp k láng giềng gần nhất k-NN 28
1.4.4 Bộ phân lớp hỗ trợ véctơ SVC 28
1.4.5 Bộ phân lớp máy hỗ trợ véctơ SVM 28
1.4.6 Bộ phân lớp HMM 29
1.4.7 Bộ phân lớp GMM [63] 30
1.4.7.1 Mô hình hỗn hợp Gauss 30
1.4.7.2 Cực đại hóa khả hiện 36
1.4.7.3 EM cho Gauss hỗn hợp 37
1.4.7.4 Thuật toán EM cho mô hình Gauss hỗn hợp 41
1.4.8 Bộ phân lớp ANN 41
1.5 Một số kết quả nhận dạng cảm xúc được thực hiện trong và ngoài nước 42
1.6 Kết chương 1 48
Chương 2. NGỮ LIỆU CẢM XÚC VÀ CÁC THAM SỐ ĐẶC TRƯNG CHO CẢM XÚC TIẾNG VIỆT NÓI 49
2.1 Phương pháp xây dựng ngữ liệu cảm xúc 49
2.2 Một số bộ ngữ liệu cảm xúc hiện có trên thế giới 51
2.3 Ngữ liệu cảm xúc tiếng Việt 53

2.4 Tham số đặc trưng của tín hiệu tiếng nói dùng cho nhận dạng cảm xúc 55
2.4.1 Đặc trưng của nguồn âm và tuyến âm 55
2.4.2 Đặc trưng ngôn điệu 61
2.5 Tham số đặc trưng dùng cho nhận dạng cảm xúc tiếng Việt 64
2.5.1 Các hệ số MFCC 64
2.5.2 Năng lượng tiếng nói 66
2.5.3 Cường độ tiếng nói 66
2.5.4 Tần số cơ bản F0 và các biến thể của F0 66
2.5.5 Các formant và dải thông tương ứng 67
2.5.6 Các đặc trưng phổ 67
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 70
2.6.1 Phân tích phương sai ANOVA và kiểm định T 70
2.6.1.1 Phân tích phương sai one-way ANOVA 70
2.6.1.2 Kiểm định T 71
2.6.2 Ảnh hưởng của tham số đặc trưng đến phân biệt các cảm xúc 71
2.7 Đánh giá sự phân lớp của bộ ngữ liệu cảm xúc tiếng Việt 74
2.7.1 Kết quả phân lớp với LDA 74
2.7.2 Thử nghiệm nhận dạng cảm xúc tiếng Việt dựa trên bộ phân lớp IBk, SMO và Trees J48 75
2.7.2.1 Công cụ, ngữ liệu và tham số sử dụng 75
2.7.2.2 Kết quả thử nghiệm 76
2.8 Kết chương 2 78
Chương 3. NHẬN DẠNG CẢM XÚC TIẾNG VIỆT NÓI VỚI MÔ HÌNH GMM 80
3.1 Mô hình GMM cho nhận dạng cảm xúc 80
3.2 Công cụ, tham số và ngữ liệu sử dụng 83
3.3 Các thử nghiệm nhận dạng 84
3.3.1 Thử nghiệm 1 đến Thử nghiệm 6 85
3.3.1.1 Nhận dạng đối với từng tập ngữ liệu 85
3.3.1.2 Nhận dạng đối với từng cảm xúc 88

3.3.1.3 So sánh kết quả của 6 thử nghiệm 91
3.3.2 Thử nghiệm 7 đến Thử nghiệm 10 92
3.3.3 Thử nghiệm 11 94
3.3.4 Thử nghiệm 12 96
3.3.5 Thử nghiệm 13 99
3.4 Đánh giá sự ảnh hưởng của tần số cơ bản 102
3.5 Quan hệ giữa số thành phần Gauss M và tỷ lệ nhận dạng 104
3.6 Kết chương 3 105
Chương 4. NHẬN DẠNG CẢM XÚC TIẾNG VIỆT NÓI SỬ DỤNG MÔ HÌNH DCNN 106
4.1 Mô hình mạng nơron lấy chập 106
4.1.1 Lấy chập 106
4.1.2 Kích hoạt phi tuyến 110
4.1.3 Lấy gộp 110
4.1.4 Kết nối đầy đủ 111
4.2 Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt 112
4.3 Ngữ liệu, tham số và công cụ dùng cho thử nghiệm 115
4.4 Thử nghiệm nhận dạng cảm xúc tiếng Việt bằng mô hình DCNN 117
4.5 Kết chương 4 121
KẾT LUẬN VÀ ĐỊNH HƯỚNG PHÁT TRIỂN 122
1. Kết luận 122
2. Định hướng phát triển 123
DANH MỤC CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ CỦA LUẬN ÁN 124
TÀI LIỆU THAM KHẢO 125
PHỤ LỤC 144
A. Danh sách các câu được chọn để thể hiện cảm xúc của bộ ngữ liệu thử nghiệm nhận dạng cảm xúc tiếng Việt nói 144
B. Kết quả thử nghiệm nhận dạng cảm xúc với bộ ngữ liệu tiếng Đức dùng công cụ Alize dựa trên mô hình GMM 144

DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

Chữ viết tắt	Chữ viết đầy đủ	Ý nghĩa
ANN	Artificial Neural Network	Mạng nơron nhân tạo
CNN	Convolutional Neural Networks	Mạng nơron lấy chập
DCNN	Deep Convolutional Neural Networks	Mạng nơron lấy chập sâu
ELU	Exponential Linear Unit	Đơn vị kích hoạt phi tuyến mũ
FIR	Finite Impulse Response	Đáp ứng xung hữu hạn
GMM	Gaussian Mixture Model	Mô hình hỗn hợp Gauss
GMVAR	Gaussian Mixture Vector Autoregressive	Mô hình tự hồi qui véctơ hỗn hợp Gauss
HMM	Hidden Markov Model	Mô hình Markov ẩn
IBk	Instance Based k	Tên gọi bộ phân lớp k láng giềng gần nhất trong Weka
IEMOCAP	Interactive Emotional dyadic Motion Capture database	Dữ liệu cảm xúc đa thể thức
Im-SFLA	Improved Shuffled Frog Leaping Algorithm	Thuật toán nhảy vọt trộn cải tiến
k-NN	k- Nearest Neighbor	Bộ phân lớp k- láng giềng gần nhất
LDA	Linear Discriminant Analysis	Phân tích phân biệt tuyến tính
LFPC	Logarit Frequency Power Coefficients	Các hệ số công suất theo logarit tần số
LMT	Logistic Model Tree	Cây mô hình logic
LP	Linear Prediction	Tiên đoán tuyến tính
LPCC	Linear Predictive Cepstral Coefficients	Các hệ số cepstrum tiên đoán tuyến tính
MFCC	Mel Frequency Cepstral Coefficients	Các hệ số cepstrum theo thang đo tần số Mel
OCON	One-Class-in-One Neural Network	Mạng nơron một lớp trong một
PCA	Principal Component Analysis	Phân tích thành phần chính
PLPC	Perceptual Linear Prediction Coefficients	Các hệ số tiên đoán tuyến tính cảm nhận

QDA	Quadratic Discriminant Analysis	Phân tích phân biệt toàn phương
RASTA	Relative Spectral Transform	Biến đổi phổ tương đối
ReLU	Rectified Linear Unit	Đơn vị chỉnh lưu tuyến tính
SFFS	Sequential Floating Forward Search	Thuật toán tìm kiếm chuyển tiếp nổi tuần tự
SFS	Sequential Floating Search	Thuật toán tìm kiếm nổi tuần tự
SMO	Sequential Minimal Optimization	Thuật toán tối ưu hóa tối thiểu tuần tự cho bộ phân lớp véctơ hỗ trợ
STE	Short Time Energy	Năng lượng trong thời gian ngắn
SVC	Support Vector Classifier	Bộ phân lớp véctơ hỗ trợ
SVM	Support Vector Machine	Máy véctơ hỗ trợ
UBM	Universal Background Model	Mô hình nền tổng quát

DANH MỤC CÁC BẢNG

Bảng 1.1 Cảm xúc cơ bản theo Nisimura và cộng sự (nguồn: [20]) 20
Bảng 1.2 Tỷ lệ nhận dạng các cảm xúc dựa trên ANN (nguồn: [87]) 45
Bảng 1.3 Kết quả nhận dạng cảm xúc của một số bộ phân lớp phổ biến (nguồn: [6]) 45
Bảng 2.1 Một số bộ ngữ liệu cảm xúc (nguồn: [6]) 51
Bảng 2.2 Ngữ liệu cảm xúc tiếng Việt dùng cho thử nghiệm 54
Bảng 2.3 Sử dụng thông tin của nguồn kích thích cho các nghiên cứu khác nhau về tiếng nói (nguồn: [133]) 58
Bảng 2.4 Sử dụng thông tin của tuyến âm cho các nghiên cứu khác nhau về xử lý tiếng nói (nguồn: [133]) 60
Bảng 2.5 Sử dụng thông tin về ngôn điệu cho các nghiên cứu khác nhau về tiếng nói (nguồn: [133]) 63
Bảng 2.6 Các tham số đặc trưng được dùng cho nhận dạng cảm xúc tiếng Việt. ... 69
Bảng 2.7 Giá trị thống kê F và P-value của phân tích ANOVA cho các tham số đặc trưng 72
Bảng 2.8 Giá trị $P - value$ của kiểm định T với các tham số đặc trưng cho từng cặp cảm xúc 73
Bảng 2.9 Tỷ lệ (%) nhận dạng cảm xúc với 384 tham số 76
Bảng 2.10 Tỷ lệ (%) nhận dạng cảm xúc chỉ dùng 228 tham số liên quan đến MFCC 77
Bảng 2.11 Tỷ lệ (%) nhận dạng cảm xúc chỉ dùng 48 tham số liên quan đến F0 và năng lượng 77
Bảng 3.1 Các thử nghiệm nhận dạng cảm xúc với GMM 84
Bảng 3.2 Ma trận nhầm lẫn nhận dạng các cảm xúc với T1 88
Bảng 3.3 Ma trận nhầm lẫn nhận dạng các cảm xúc với T2 89
Bảng 3.4 Ma trận nhầm lẫn nhận dạng các cảm xúc với T3 90
Bảng 3.5 Ma trận nhầm lẫn nhận dạng các cảm xúc với T4 91
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 95
Bảng 3.7 Tỷ lệ nhận dạng đúng trung bình đối với 4 tập ngữ liệu khi kết hợp prm60 với $F0$ và biến thể $F0$ 99
Bảng 3.8 Tập tham số prm79 kết hợp với một trong 8 biến thể của F0 99
Bảng 3.9 Tỷ lệ nhận dạng đúng trung bình đối với 4 tập ngữ liệu khi kết hợp prm79 với từng biến thể $F0$ 102

Bảng 4.1 Cấu trúc mạng DCNN cho nhận dạng cảm xúc tiếng Việt trong trường hợp 260 tham số..... 113
Bảng 4.2 Phân chia ngữ liệu T1 (phụ thuộc cả người nói và nội dung) 116
Bảng 4.3 Phân chia ngữ liệu T2 (phụ thuộc người nói và độc lập nội dung) 116
Bảng 4.4 Phân chia ngữ liệu T3 (độc lập người nói và phụ thuộc nội dung) 116
Bảng 4.5 Phân chia ngữ liệu T4 (độc lập cả người nói và nội dung) 116
Bảng 4.6 Năm tập tham số thử nghiệm nhận dạng với DCNN 116
Bảng B.1. Bộ ngữ liệu tiếng Đức với bốn cảm xúc vui, buồn, tức và bình thường 145
Bảng B.2. Kết quả nhận dạng cảm xúc tiếng Đức trong trường hợp 1..... 145
Bảng B.3. Kết quả nhận dạng cảm xúc tiếng Đức trong trường hợp 2..... 145

DANH MỤC CÁC HÌNH ẢNH VÀ ĐỒ THỊ

Hình 1.1 Phân bố 8 cảm xúc trên mặt phẳng cảm xúc 2 chiều Arousal và Valence (nguồn: [11]) 18
Hình 1.2 Sơ đồ chung cho hệ thống nhận dạng cảm xúc tiếng nói 26
Hình 1.3 Phân bố Gauss đơn biến đơn thể hiện với $\mu = 0$ và $\sigma = 1$ 31
Hình 1.4 Hàm khả hiện đối với phân bố Gauss. 32
Hình 1.5 Minh họa hỗn hợp 3 thành phần Gauss trong không gian 2 chiều 33
Hình 1.6 Đồ thị biểu diễn một mô hình hỗn hợp trong đó phân bố kết hợp được biểu diễn dưới dạng $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$ 34
Hình 1.7 Đồ thị biểu diễn một mô hình Gauss hỗn hợp 36
Hình 1.8 Phân bố của 2 tập dữ liệu 2D và PDF tương ứng theo GMM 39
Hình 1.9 Minh họa thuật toán EM, phân bố dữ liệu và đánh giá PDF theo EM 40
Hình 1.10 Phân cấp cảm xúc 2 tầng 3 tầng theo Lugger và Yang (nguồn: [98]) 46
Hình 2.1 Các đoạn tín hiệu của âm vô thanh, hữu thanh và tín hiệu sai số LP tương ứng 56
Hình 2.2 Phân tích trong miền tần số để có phổ tiếng nói 57
Hình 2.3 Các đặc trưng ngôn điệu của tiếng nói 61
Hình 2.4 Sơ đồ tính hệ số MFCC 65
Hình 2.5 Kết quả phân lớp cảm xúc giọng nam và nữ bằng LDA 75
Hình 2.6 Kết quả phân lớp cảm xúc cả giọng nam và nữ bằng LDA 75
Hình 3.1 Sơ đồ mô hình GMM tổng quát cho nhận dạng cảm xúc 81
Hình 3.2 Mô hình Gauss của 4 cảm xúc 82
Hình 3.3 Mô hình Gauss của 6 cặp cảm xúc 82
Hình 3.4 Kết quả nhận dạng cảm xúc đối với T1 86
Hình 3.5 Kết quả nhận dạng cảm xúc đối với T2 86
Hình 3.6 Kết quả nhận dạng cảm xúc đối với T3 87
Hình 3.7 Kết quả nhận dạng cảm xúc đối với T4 87
Hình 3.8 Tỷ lệ nhận dạng đúng trung bình của bốn cảm xúc ứng với 6 tập tham số cho T1 88
Hình 3.9 Tỷ lệ nhận dạng đúng trung bình của bốn cảm xúc ứng với 6 tập tham số cho T2 89
Hình 3.10 Tỷ lệ nhận dạng đúng trung bình của bốn cảm xúc ứng với 6 tập tham số với T3 90
Hình 3.11 Tỷ lệ nhận dạng đúng trung bình của bốn cảm xúc ứng với 6 tập tham số với T4 91

Hình 3.12 Tỷ lệ nhận dạng đúng trung bình cảm xúc của 4 tập ngữ liệu trong 6 thử nghiệm 92
Hình 3.13 Tỷ lệ nhận dạng sử dụng MFCC và các đặc trưng phổ với T1..... 93
Hình 3.14 Tỷ lệ nhận dạng đúng trung bình cho 7 tập tham số đã nêu với T1. 94
Hình 3.15 Tỷ lệ nhận dạng đúng cao nhất và thấp nhất tương ứng với đặc trưng phổ cho các giá trị của M..... 95
Hình 3.16 Tỷ lệ nhận dạng đúng trung bình khi kết hợp prm60+F0+các biến thể của F0 đối với T1 97
Hình 3.17 Tỷ lệ nhận dạng đúng trung bình khi kết hợp prm60+F0+các biến thể của F0 đối với T2 97
Hình 3.18 Tỷ lệ nhận dạng đúng trung bình khi kết hợp prm60+F0+các biến thể của F0 đối với T3 98
Hình 3.19 Tỷ lệ nhận dạng đúng trung bình khi kết hợp prm60+F0+các biến thể của F0 đối với T4 98
Hình 3.20 Tỷ lệ nhận dạng đúng trung bình của các cảm xúc cho từng tập tham số đối với T1 100
Hình 3.21 Tỷ lệ nhận dạng đúng trung bình của các cảm xúc cho từng tập tham số đối với T2 100
Hình 3.22 Tỷ lệ nhận dạng đúng trung bình của các cảm xúc ứng cho từng tập tham số đối với T3..... 101
Hình 3.23 Tỷ lệ nhận dạng đúng trung bình của các cảm xúc cho từng tập tham số đối với T4 101
Hình 3.24 Tỷ lệ nhận dạng trung bình cả 4 cảm xúc theo từng biến thể F0 và prm79 cho các tập ngữ liệu T1 đến T4, với M=512. 103
Hình 3.25 Quan hệ giữa số thành phần Gauss M và tỷ lệ nhận dạng đúng trung bình của Thử nghiệm từ 1 đến 6 với 4 tập ngữ liệu..... 104
Hình 3.26 Quan hệ giữa số thành phần Gauss M và tỷ lệ nhận dạng đúng trung bình các Thử nghiệm từ 1 đến 3 và từ 7 đến 10 với T1. 104
Hình 4.1 Mô tả bước lấy chập dùng bộ lọc kích thước 5×5 107
Hình 4.2 Mô tả chi tiết lấy chập dùng bộ lọc kích thước 5×5 108
Hình 4.3 Mô tả bước lấy chập của mạng nơron dùng bộ lọc kích thước 5×5 108
Hình 4.4 Mô tả bước lấy chập của mạng nơron dùng 3 bộ lọc kích thước 5×5 109
Hình 4.5 Ví dụ sử dụng max-pooling 111
Hình 4.6 Mô tả cách thực hiện max-pooling với zero padding 111
Hình 4.7 Phổ mel của tín hiệu tiếng nói làm ảnh đầu vào cho lớp thứ nhất trong trường hợp mô hình baseline 112
Hình 4.8 Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt với 260 tham số.... 114

Hình 4.9 Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt với 264 tham số.... 114
Hình 4.10 Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt với 267 tham số.. 115
Hình 4.11 Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt với 294 tham số.. 115
Hình 4.12 Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt với 296 tham số.. 115
Hình 4.13 Kết quả nhận dạng với 5 tập tham số cho 4 tập ngữ liệu..... 118
Hình 4.14 Tỷ lệ nhận dạng trung bình của các thử nghiệm với 5 tập tham số..... 119
Hình 4.15 Tỷ lệ nhận dạng đúng cao nhất của từng cảm xúc đối với từng thử nghiệm 119
Hình 4.16 Tỷ lệ nhận dạng đúng trung bình của mỗi cảm xúc đối với từng tập ngữ liệu 120

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 cảm thán, 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. Chính vì vậy, chẳng hạn nếu người nói diễn đạt câu nói theo cảm xúc tức thì hệ thống nhận dạng là cảm xúc tức. Mặc dù người nói đang tức song diễn đạt câu nói lại theo cảm xúc bình thường thì hệ thống nhận dạng là cảm xúc bình thường.

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

Trong những năm gần đây, sự huyền bí của cảm xúc tiếng nói đã làm tăng sự thu hút mối quan tâm nghiên cứu tương tác người - máy. Đây là mối quan tâm mới nhất hiện nay nhằm làm cho mối tương tác giữa con người và máy móc trở nên tự nhiên như tương tác giữa người với người. Đã có các nghiên cứu về cảm xúc cũng như nhận dạng cảm xúc với các ngôn ngữ khác nhau nhằm hỗ trợ các ứng dụng tương tác đó. Chương này sẽ trình bày một số khái niệm cơ bản liên quan đến cảm xúc tiếng nói và tổng quan về nhận dạng cảm xúc tiếng nói trong và ngoài nước.

1.1 Cảm xúc tiếng nói và phân loại cảm xúc

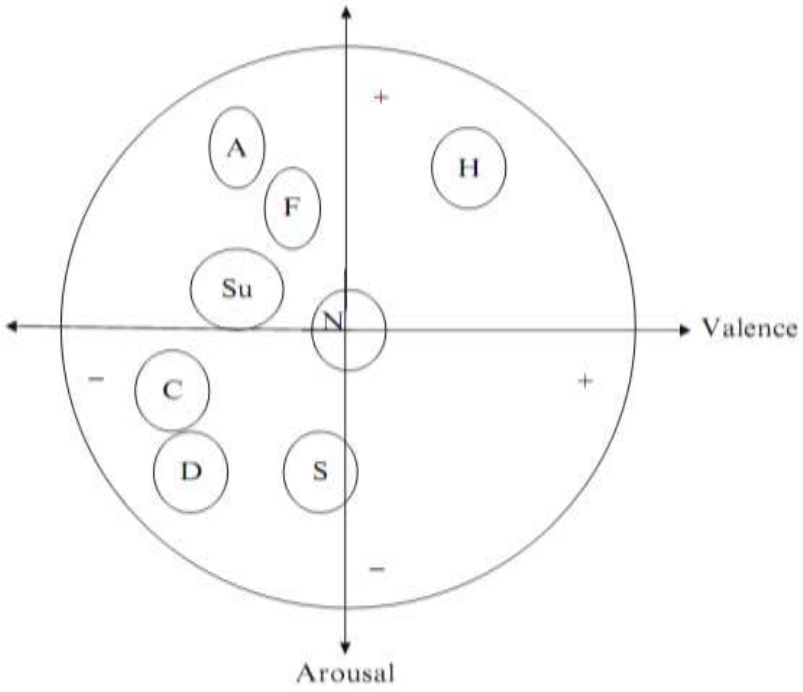
Theo Từ điển Bách khoa Việt Nam [1], "Cảm xúc phản ứng tình cảm chủ quan mạnh của con người và động vật cao cấp phát sinh khi nhận được kích thích từ bên ngoài và bên trong cơ thể. Cảm xúc là một trong những hình thức phản ánh thực tế khách quan trong bộ não và được biểu hiện bằng thái độ của người và động vật với sự vật và các hiện tượng xung quanh. Cảm xúc kèm theo biểu hiện sinh lý (thay đổi sắc mặt, nhịp tim, nhịp thở, hoạt động của các tuyến nội tiết, trạng thái cơ thể) và trạng thái tâm lý. Cảm xúc đơn giản nhất là cảm giác bẩm sinh do tác nhân có ý nghĩa quan trọng đối với tồn tại của cơ thể (thức ăn, nhiệt độ, đau,...). Cảm xúc có ý nghĩa quan trọng đối với sự tích luỹ kinh nghiệm của cá thể, cho phép con người và động vật tập nhiễm những tập tính có ích, tránh được điều bất lợi cho cơ thể".

Hay nói theo một cách khác: Cảm xúc xét về mặt tâm lý có thể được xem như là một trải nghiệm phức hợp của ý thức (tâm lý), cảm giác cơ thể (sinh lý) và hành vi (action-speech). Nói chung cảm xúc là biểu thị tổng hợp trải nghiệm chủ thể, hành vi biểu cảm, và hoạt động của hệ thần kinh [2].

Có nhiều cách khác nhau để phân loại cảm xúc. Đã có các nghiên cứu đưa ra hơn 300 trạng thái cho những cảm xúc khác nhau [3], [4]. Cũng có nghiên cứu khác trong đó các tác giả lại đưa ra 107 loại cảm xúc [5]. Tuy nhiên, nhìn chung, 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].

Trong tâm lý học, biểu hiện của cảm xúc được xem như là đáp ứng đối với các kích thích có liên quan đến sự thay đổi các đặc tính sinh lý [9], [10]. Về mặt sinh lý, một cảm xúc được xác định như là sự chia tách đối với đường cơ sở trung tính (homoeostatic) [9]. Dựa trên những thay đổi này, các tính chất của cảm xúc có thể

được giải thích trong không gian ba chiều. Trục V (Valence) biểu diễn cho cảm xúc mang tính tích cực hoặc tiêu cực. Trục A (Arousal) biểu diễn cho cảm xúc hào hứng hay thờ ơ. Trục P (Power) biểu diễn cho sự điều khiển của các giác quan thông qua cảm xúc [11]. Hình chiếu trong không gian cảm xúc ba chiều, lên mặt phẳng hai chiều với các trục A và V, được thể hiện trên Hình 1.1.



Hình 1.1 *Phân bố 8 cảm xúc trên mặt phẳng cảm xúc 2 chiều Arousal và Valence (nguồn: [11])*
A (tức), C (buồn), D (ghê tởm), F (sợ), H (vui), N (trung tính), S (mỉa mai), Su (ngạc nhiên)

Về mặt sinh lý của cơ chế tạo cảm xúc, người ta đã phát hiện ra rằng hệ thống thần kinh được kích thích bởi sự biểu hiện của cảm xúc hưng phấn cao như giận dữ, vui và sợ hãi. Hiện tượng này 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 là lớn hơn, trung bình tần số cơ bản sẽ cao hơn và phạm vi biến thiên cũng rộng hơn [12]. 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 cao độ, năng lượng, nhịp điệu, chất lượng giọng nói, và tín hiệu tiếng nói có độ tương quan lớn với những cảm xúc chính [13].

Có thể xét cảm xúc theo góc độ tín hiệu tiếng nói như sau. Sự thay đổi tâm lý và sinh lý là do những trải nghiệm về cảm xúc dẫn tới một số phản ứng. Tiếng nói là một trong những kết quả quan trọng của trạng thái cảm xúc của con người. Tín hiệu tiếng nói được tạo ra do tuyến âm được kích thích bởi tín hiệu nguồn [14]. Do đó, thông tin đặc trưng của tiếng nói có thể được trích rút từ đặc tính của tuyến âm và đặc

tính của nguồn âm. Những đặc trưng cảm xúc có trong tiếng nói có thể được xác định từ đặc tính của nguồn âm, sự thay đổi cấu hình của tuyến âm với các cảm xúc khác nhau, siêu đoạn tính (thời hạn, chu kỳ cơ bản, năng lượng) và thông tin ngôn ngữ. Các đặc tính hoạt động của thanh môn và cấu hình tuyến âm cũng đóng một vai trò quan trọng trong việc biến đổi các cảm xúc khác nhau trong quá trình nói.

Do những yếu tố chủ quan ẩn chứa bên trong cảm xúc nên sẽ không có sự phân loại nhất quán cảm xúc tạo cơ sở chung cho nghiên cứu cảm xúc. Vì vậy, các cách tiếp cận khác nhau được sử dụng cho cảm nhận dấu hiệu khác nhau của các cảm xúc và phân biệt cảm xúc từ các tâm trạng khác nhau. Scherer [15] đã phân loại các trạng thái tình cảm như sau:

- Cảm xúc (tức, buồn, vui mừng, sợ hãi, xấu hổ, tự hào, phấn chấn, tuyệt vọng)
- Tâm trạng (vui vẻ, nản lòng, dễ cáu, bơ phờ, chán nản)
- Thái độ giữa các cá nhân với nhau (dè dặt, lạnh lùng, thân thiện, thông cảm, khinh bỉ)
- Sở thích/quan điểm (thích, yêu, ghét, coi trọng, ao ước)
- Khuynh hướng biểu cảm (lo lắng, hồi hộp, hấp tấp, khinh khỉnh, thù địch)

Các trạng thái này phân biệt với nhau theo các đặc điểm chỉ định như cường độ, thời hạn, sự đồng bộ hoá, tiêu điểm sự kiện, đánh giá suy luận, tính thay đổi nhanh chóng, các ảnh hưởng đến hành vi.

Khác với tâm trạng, cảm xúc thường rất cô đọng và kéo dài trong khoảng thời gian ngắn. Để có thể phân biệt các trạng thái cảm xúc khác nhau, nghiên cứu [16] đã phân loại các trạng thái biểu cảm thành biểu cảm tích cực và biểu cảm tiêu cực. Trong mỗi biểu cảm lại phân thành tâm trạng và cảm xúc. Tâm trạng có thời hạn dài hơn, thường kéo dài trong nhiều ngày như tâm trạng phấn khởi, mãn nguyện hay u sầu. Còn cảm xúc thì có thể trong vài phút như vui mừng, buồn, chán ghét, sợ hãi hay tức giận.

Để thiết lập một hệ thống nhận dạng cảm xúc trong tiếng nói, thông thường sẽ dễ dàng và thuận lợi hơn nếu chỉ nhận dạng một số lượng giới hạn các cảm xúc, có nghĩa là tập các cảm xúc cơ bản. Có một số cách tiếp cận để định nghĩa và xác định tập cảm xúc này. Descarté đã đề xuất ý tưởng phân biệt các cảm xúc cơ bản và thứ cấp [17]. Trong nghiên cứu [18], các cảm xúc cơ bản hoặc cơ sở nói chung được giới thiệu là "biểu diễn các mẫu có mối quan hệ sống còn khi đáp ứng với sự kiện, các mẫu đáp ứng này đã được chọn lọc qua lịch sử tiến hoá của loài người trên thế giới này" còn các cảm xúc khác theo một cách nào đó là dẫn xuất từ cảm xúc cơ bản. Cornelius đã đề xuất " Big Six" như là các cảm xúc cơ bản hay sơ cấp bao gồm vui, buồn, sợ, chán, tức và ngạc nhiên. Trong khi đó, Plutchik [19] lại phân biệt 8 loại cảm xúc cơ bản là sợ, tức, vui, buồn, chấp nhận, chán, đề phòng và ngạc nhiên. Nisimura và cộng sự (2006) [20] thậm chí đưa ra 16 cảm xúc cơ bản (gồm cả trạng thái trung tính) có tính đến các cảm xúc đã được Schlosberg [21] và Ekman [22] đề xuất (Bảng 1.1).

Bảng 1.1 *Cảm xúc cơ bản theo Nisimura và cộng sự (nguồn: [20])*

Tức	Coi khinh	Mãn nguyện	Chán nản
Phấn khích	Sợ	Mừng	Vui đùa
Trung tính	Hài lòng	Vội vã	Buồn
Ngạc nhiên	Căng thẳng	Mệt mỏi	Coi khinh

Một cách tiếp cận khác khá đơn giản là nhóm các cảm xúc được phân loại theo cách đánh giá của Fujisawa và Cook [23]. Các cảm xúc được chia thành 3 nhóm:

- Biểu cảm tích cực (vui, thoả mãn, hài lòng)
- Biểu cảm tiêu cực (buồn, tức, khó chịu)
- Biểu cảm pha trộn (bấp bênh, căng thẳng, hồi hộp)

Một mặt, nếu theo cách này thì các cảm xúc có thể được phân loại dễ dàng nhưng mặt khác thì các cảm xúc như buồn và tức cũng được nhóm vào một lớp mặc dù chúng rất khác nhau. Tương tự như vậy, trong [24] định nghĩa 6 cảm xúc và gán chúng vào 4 nhóm cảm xúc chủ yếu như sau:

- Vui (hạnh phúc)
- Buồn (chán, đau buồn)
- Tức (giận dữ, sợ hãi)
- Trung tính (thái độ trung lập)

Như vậy, nhìn chung có 4 cảm xúc cơ bản tức, sợ, vui, buồn và các cảm xúc này xuất hiện phần lớn trong các tài liệu nghiên cứu tiêu biểu cho hành vi cảm xúc [25]. Các cảm xúc như vậy tương ứng với các vấn đề liên quan trong cuộc sống, chẳng hạn tức có thể được xem như phản ứng với tranh đua, sợ là phản ứng với nguy hiểm, vui là phản ứng với sự cộng tác còn buồn là phản ứng với mất mát [26].

Con người hiểu được ý muốn của thông điệp do có những cảm xúc quan trọng được thêm vào thông tin ngữ âm. Vì vậy, cần phải phát triển các hệ thống có thể xử lý các cảm xúc kèm theo nội dung cần truyền tải [27]. Các mục tiêu cơ bản của xử lý tiếng nói có cảm xúc là nhận dạng những cảm xúc thể hiện trong tiếng nói và tổng hợp những cảm xúc mong muốn trong tiếng nói để truyền tải ý định nội dung. Từ góc độ kỹ thuật, sự nhận biết các cảm xúc tiếng nói có thể được xem như là sự phân loại hoặc phân biệt các cảm xúc. Tổng hợp các cảm xúc có thể được xem như là sự lồng ghép các hiểu biết về cảm xúc trong quá trình tổng hợp tiếng nói. Các hiểu biết về cảm xúc được thu thập từ các mô hình cảm xúc đã được thiết kế để trích chọn các đặc trưng về cảm xúc.

Lời nói mà không có cảm xúc sẽ không tự nhiên và đơn điệu. Hầu hết các hệ thống xử lý tiếng nói hiện nay có thể xử lý tiếng nói tự nhiên được ghi âm trong phòng thu. Tuy nhiên, trong các kịch bản giao tiếp trong thế giới thực hiện nay, hệ thống xử lý tiếng nói phải có khả năng xử lý các cảm xúc đã được nhúng vào chính hệ thống đó. Mạch cảm xúc thể hiện trong tiếng nói có thể được phát hiện dựa trên các đặc

điểm khác nhau được trích chọn từ nguồn âm, tuyến âm và các thành phần ngôn điệu của tiếng nói.

Cảm xúc của con người là đa dạng và không thể đo lường một cách chính xác bằng các phương tiện đo đạc thông 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 [28]. Các nhà nghiên cứu cũng đã phân loại cảm xúc thành rất nhiều cảm xúc khác nhau như đã trình bày ở trên. 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 rũ rượi, buồn thiu, buồn tênh, buồn mênh mang, buồn cười, v.v.. [1]. Do đó, trong các nghiên cứu, các tác giả chủ yếu thực hiện nhận dạng các cảm xúc tiêu biểu hay gặp nhất trong cuộc sống hàng ngày. Trong khuôn khổ có hạn, luận án cũng đi theo hướng như vậy bằng cách tập trung vào 4 loại cảm xúc mang tính đại diện là vui, buồn, tức và bình thường.

1.2 Nghiên cứu về nhận dạng cảm xúc

Giao tiếp bằng tiếng nói là phương thức nhanh và tự nhiên nhất trong giao tiếp giữa người với người. Thực tế này đã thúc đẩy các nhà nghiên cứu nghĩ rằng, sử dụng tiếng nói là một phương pháp nhanh và hiệu quả cho sự tương tác giữa con người và máy. Tuy nhiên, điều này đòi hỏi máy phải có đủ thông minh để nhận ra tiếng nói của con người. Trong những năm gần đây, đã có rất nhiều nghiên cứu về nhận dạng tiếng nói, trong đó đề cập đến quá trình chuyển đổi tiếng nói của con người sang dạng chuỗi các từ [6]. Mặc dù đã có những tiến bộ lớn trong nhận dạng tiếng nói song vẫn còn xa so với tương tác tự nhiên giữa con người với nhau vì máy móc hiện tại chưa hiểu được hoàn toàn chính xác trạng thái cảm xúc của người nói. Điều này đã tạo ra một lĩnh vực nghiên cứu mới gần đây, cụ thể là nhận dạng cảm xúc tiếng nói, được định nghĩa là hiểu được các trạng thái cảm xúc của người nói từ trong tiếng nói của họ. Các nghiên cứu thấy rằng, nhận dạng cảm xúc tiếng nói có thể được sử dụng để trích rút những ngữ nghĩa hữu ích từ tiếng nói và do đó cải thiện được hiệu năng của hệ thống nhận dạng tiếng nói [29].

Nhận dạng cảm xúc tiếng nói đặc biệt hữu ích cho các ứng dụng đòi hỏi sự tương tác tự nhiên giữa người - máy như các ứng dụng hướng dẫn bằng máy tính mà đáp ứng của những hệ thống này đối với người sử dụng phụ thuộc vào cảm xúc được phát hiện [30]. Chẳng hạn, nhận dạng cảm xúc sẽ hữu ích cho hệ thống điều khiển trong xe hơi mà thông tin trạng thái tinh thần của người lái xe có thể được cung cấp cho hệ thống để hướng dẫn người lái xe an toàn. Nó cũng có thể được sử dụng như một công cụ chẩn đoán trong chữa bệnh [31]. Nó có thể cũng hữu ích trong hệ thống dịch tự động, trong đó các trạng thái cảm xúc của người nói đóng vai trò quan trọng trong giao tiếp giữa các bên. Ví dụ, trên buồng lái máy bay, người ta đã thấy rằng các hệ thống nhận dạng tiếng nói được huấn luyện đối với tiếng nói có biểu hiện cảm xúc đạt được hiệu năng tốt hơn so với hệ thống được huấn luyện bằng giọng thông thường

[32]. Nhận dạng cảm xúc tiếng nói cũng đã được sử dụng trong các ứng dụng thuộc trung tâm tổng đài và truyền thông di động [33] trong đó mục tiêu chính của việc sử dụng nhận dạng cảm xúc tiếng nói là để thích ứng với yêu cầu của hệ thống, phát hiện sự thất vọng hay bực bội trong giọng của người nói.

Hiện nay, nghiên cứu nhận dạng cảm xúc tiếng nói có nhiều thách thức vì những lẽ sau. Thứ nhất, thường không biết một cách rõ ràng những đặc trưng nào của tiếng nói là mạnh nhất trong việc phân biệt các cảm xúc. Tính đa dạng về mặt âm học do các câu khác nhau, người nói, phong cách nói, và tốc độ nói khác nhau lại làm tăng thêm trở ngại vì những thuộc tính này ảnh hưởng trực tiếp đến phần lớn các đặc trưng tiếng nói được trích rút phổ biến như cao độ, đường bao năng lượng [34]. Vả lại, có thể cùng một câu nói lại có chứa nhiều cảm xúc, mỗi cảm xúc tương ứng với một phần khác nhau của câu nói đó. Thêm vào đó, rất khó xác định ranh giới giữa các phần trong câu nói. Vấn đề thứ hai là một cảm xúc nào đó được biểu hiện còn thường phụ thuộc vào người nói khác nhau, văn hóa và môi trường khác nhau của người nói. Hầu hết các nghiên cứu đã tập trung vào phân lớp cảm xúc trong cùng một ngôn ngữ, và giả thiết rằng không có sự khác biệt văn hóa giữa người nói với nhau. Tuy vậy, các nhiệm vụ phân lớp đa ngôn ngữ cũng đã được nghiên cứu [35]. Một vấn đề khác là người ta có thể trải qua một cảm xúc nhất định như buồn trong nhiều ngày, nhiều tuần, thậm chí hàng tháng. Trong trường hợp như thế, những cảm xúc khác sẽ là thoáng qua và sẽ không kéo dài hơn một vài phút. Kết quả là, bộ nhận dạng cảm xúc tự động sẽ không phát hiện rõ ràng liệu cảm xúc kéo dài hay thoáng qua.

Cảm xúc không có định nghĩa thống nhất chung [36]. Tuy nhiên, con người biết được cảm xúc khi họ cảm nhận được. Vì lẽ đó, các nhà nghiên cứu có thể nghiên cứu và định nghĩa các khía cạnh khác nhau của cảm xúc. Như đã trình bày ở mục 1.1, đa số cho rằng cảm xúc có thể được đặc trưng trong hai chiều: kích hoạt (activation) và hóa trị (valence) [37]. Kích hoạt là tổng năng lượng cần thiết để thể hiện một cảm xúc nhất định.

Tuy nhiên, không thể phân biệt các cảm xúc mà chỉ dùng kích hoạt. Chẳng hạn, cả cảm xúc tức và vui đều tương ứng với kích hoạt cao nhưng chúng lại truyền tải cảm xúc khác nhau. Sự khác biệt này được đặc trưng theo hướng hóa trị. Thật đáng tiếc các nhà nghiên cứu không có sự nhất trí nào hoặc liệu các đặc trưng âm học có tương quan với chiều này không [38]. Vì vậy, trong khi phân lớp giữa cảm xúc kích hoạt cao và cảm xúc kích hoạt thấp có thể đạt được độ chính xác cao thì phân lớp giữa cảm xúc khác nhau vẫn đang là thách thức.

Một vấn đề quan trọng trong việc nhận dạng cảm xúc tiếng nói là sự cần thiết xác định một tập những cảm xúc quan trọng phải được phân lớp theo một hệ nhận dạng cảm xúc tự động. Các nhà ngôn ngữ học đã thống kê rất nhiều các trạng thái của các cảm xúc khác nhau. Tuy nhiên, việc nhận dạng một tập các cảm xúc lớn như vậy là khó khăn. Do đó, các nghiên cứu thường chủ yếu tập trung vào một số cảm xúc cơ bản nhất trong cuộc sống.

Các nghiên cứu lý thuyết và thực nghiệm về các hình thái biểu hiện cảm xúc thông qua tiếng nói và khuôn mặt trong hệ thống giao tiếp đa thể thức đã được nghiên cứu

trên thế giới. Đã có những phương pháp sử dụng các cảm biến sinh học đo lường các đại lượng vật lý liên quan đến cảm xúc, phiên dịch cử chỉ và biểu hiện khuôn mặt sử dụng camera, xử lý ngôn ngữ tự nhiên với các từ khoá biểu hiện cảm xúc và biến thiên cao độ âm thanh để nhận dạng ngôn điệu, phân loại các đặc điểm ngữ điệu được trích rút từ tín hiệu tiếng nói.

Ngày nay, hơn bao giờ hết máy tính được xem như cộng sự. Người dùng máy tính có khuynh hướng áp dụng các chuẩn xã hội cho máy tính của họ. Ví dụ, họ trở nên nổi khùng nếu máy tính phạm lỗi hoặc họ hài lòng nếu máy tính ca tụng họ làm việc thành công (Reeves và Nass 1996) [39]. Hơn nữa, mối quan hệ như vậy sẽ được củng cố khi người dùng có thể cá thể hoá giao diện, chẳng hạn bằng cách áp đặt các chủ đề cho màn hình nền của họ và sẽ cảm thấy tương tác thuận lợi hơn với hệ thống. Khái niệm "quan hệ" giữa máy tính và người dùng sẽ được tăng cường khi máy tính có thể đáp ứng được tình trạng và trạng thái cảm xúc của người dùng [40], [41]. Để có thể làm cho hệ thống đối thoại có tính thông minh như thế, cần phải phân loại, phân tích và nhận dạng cảm xúc.

Đối với hệ thống giao tiếp đơn thể thức chỉ sử dụng tiếng nói, đã có các nghiên cứu nhận dạng cảm xúc từ tín hiệu tiếng nói sử dụng mô hình Markov ẩn HMM. Dựa trên ngữ liệu tiếng nói có cảm xúc, tập các đặc điểm ngữ điệu được lựa chọn và HMM đã được huấn luyện để nhận dạng một số cảm xúc với người nói khác nhau. Do các tham số của mô hình đa dạng, nhiều bộ nhận dạng đã được thiết lập đồng thời. Tuỳ theo kết quả đầu ra của bộ nhận dạng cảm xúc mà thay đổi tiến trình và cách thức đối thoại. Trong trường hợp này, nhờ có mô hình trạng thái người nói và mô hình tình huống, chiến lược đối thoại được thay đổi để thích nghi và lựa chọn phong thái đối thoại thích ứng. Chẳng hạn, nếu người nói diễn đạt với tâm trạng bình thường, phát âm rõ ràng thì hệ thống giao tiếp không cần kèm theo những động thái để xác nhận và đối thoại có thể duy trì trong thời gian ngắn. Tuy nhiên, nếu người nói tỏ ra tức giận và diễn đạt không rõ ràng, hệ thống cần làm cho người nói bình tĩnh và thường cần có những câu hỏi để xác nhận. Điều này cũng có thể lại dẫn tới làm cho người nói tức giận. Chủ yếu có hai phương pháp để mô hình hoá ảnh hưởng của tham số cảm xúc được dùng để điều khiển: một là cách tiếp cận dựa trên quy tắc trong đó mỗi tình huống của hành vi người nói được bao hàm bằng một quy tắc chứa đáp ứng thích hợp, hai là cách tiếp cận có tính phỏng đoán ngẫu nhiên trong đó cần mô hình hoá xác suất của các đáp ứng thích hợp phụ thuộc vào ngôn điệu của người nói trước đó và các tham số điều khiển tương ứng.

Do không thể đo lường các cảm xúc bằng các phương tiện một cách khách quan và khó phân biệt các cảm xúc một cách rõ ràng nên dẫn tới tính nhập nhằng trong các giai đoạn phát triển hệ thống nhận dạng cảm xúc trong đó cùng một ngôn điệu của ngữ liệu huấn luyện song có thể xảy ra tình trạng các cảm xúc khác nhau sẽ được gán nhãn mà nguyên nhân là sự khác nhau về cảm nhận của những người gán nhãn. Từ đó, cũng có thể thấy, với cùng một ngữ liệu huấn luyện, trong trường hợp này hệ thống không thể thực hiện nhận dạng tốt hơn người gán nhãn.

Holzapfel và cộng sự (2002) [42] đã đề xuất việc tích hợp cảm xúc vào cấu trúc đặc trưng kiểu đa chiều. Cấu trúc này không chỉ chứa thông tin về ngữ nghĩa mà còn chứa thông tin bổ sung mô tả người nói và tình trạng. Theo đó, trạng thái đối thoại của họ được đặc trưng bằng 7 biến bao gồm kiểu cảm xúc, kiểu hành vi tiếng nói, ý định của người dùng và các phép đo tin cậy. Để tương tác với robot có tính đến cảm xúc, các tác giả đã đề xuất chiến lược thao tác trong không gian giá trị của các biến trạng thái 7 chiều. Chiến lược này cũng quyết định cách phiên dịch như thế nào về cảm xúc, chẳng hạn xem tức giận như là phản ứng đối với hệ thống bị hỏng.

Brown và Levinson (1987) [43] đã thảo luận về ảnh hưởng của biểu cảm và sự tế nhị đối với phong cách ngôn ngữ và kết quả này đã được Walker và cộng sự (1997) [44] đưa vào các tác tử nhân tạo có cá tính. Các tác giả đã đề xuất sự ứng biến phong cách ngôn ngữ để làm cho các tác tử này hướng đến quan hệ người - người và như vậy tương tác trở nên đáng tin hơn. Lý thuyết của các tác giả đã dựa trên hành vi tiếng nói để biểu diễn trừu tượng ngôn điệu và đặt kế hoạch cho ứng biến. Có thể có sự thay đổi trong nội dung ngữ nghĩa, dạng cú pháp và thể hiện về mặt âm học. Chiến lược để thực hiện một ý định nào đó được lựa chọn dựa trên 2 tham số: khoảng cách xã hội giữa các người dùng và hệ thống đối thoại, thứ hạng áp đặt cho hành vi tiếng nói hiện tại (thấp cho tin tốt như chấp nhận, cao cho tin xấu như loại bỏ).

Ngoài vấn đề xem xét và kiểm tra giải pháp do người dùng đề nghị, hệ thống hướng dẫn thông minh được sử dụng cho các lệnh có trợ giúp máy tính. Mô hình cảm xúc có kết hợp gợi ý đối với các hệ thống như vậy đã được [45] nghiên cứu. Cấu trúc cảm xúc của họ phân biệt các hành vi theo các cấp: cấp cơ bản, cấp thứ hai và cấp thứ ba.

Với các ngôn ngữ có thanh điệu như tiếng Trung [46] hoặc tiếng Thái, cao độ được dùng để phân biệt nghĩa của từ. Hơn nữa, với ngôn ngữ có thanh điệu, ngữ điệu cũng được sử dụng. Nghiên cứu trong [47] đã thêm vào mỉa mai và ngạc nhiên để biểu thị trạng thái cảm xúc của người nói. Trong tiếng nói tổng hợp, sử dụng yếu tố ngữ điệu sẽ làm cho tiếng nói tự nhiên hơn [48], [49], đồng thời phát hiện trạng thái cảm xúc của người nói [23], [50], [51].

Đối với các nghiên cứu hiện tại, có một số cách tiếp cận để phân loại và nhận dạng cảm xúc, từ việc phiên dịch biểu cảm khuôn mặt và cử chỉ trong hệ thống đa thể thức [52] tới đo lường vật lý [53], [54], [55], phân tích ngữ nghĩa hoặc kết hợp các thể thức này. Đối với nhận dạng cảm xúc dựa trên tiếng nói, bộ nhận dạng có thể bao gồm từ điển đã được đơn giản hoá, mô hình ngôn ngữ và mô hình âm học, việc huấn luyện và nhận dạng được thực hiện theo cùng cách. Có một số cách tiếp cận để gán nhãn ngữ liệu tiếng nói cảm xúc. Nếu chỉ có một véctơ đặc trưng được trích rút từ dạng sóng, tương ứng chỉ cần gán nhãn mỗi phát ngôn với một cảm xúc mà không xét đến khoảng lặng hoặc những thay đổi khác trong dạng sóng.

Cùng với phương pháp nhận dạng cảm xúc dựa trên tín hiệu tiếng nói, trạng thái cảm xúc của một lời nói có thể được xác định bằng cách xem xét nội dung văn bản (text) dùng cho phát ngôn. Một mặt, bởi vì thao tác được thực hiện trên văn bản, phương pháp này tự nó không cần đến phân tích tín hiệu phức tạp và phương pháp

phân loại nhưng mặt khác, phải giả thiết văn bản là kết quả của nhận dạng đúng, có nghĩa là bộ nhận dạng tiếng nói trước đó đã thực hiện tin cậy.

Giả thiết ta có câu với cảm xúc trung tính "Tôi muốn về vào thứ Hai". Câu này có thể được mở rộng để biểu thị cảm xúc vui thành câu " Ồi tuyệt quá, tôi thích về vào ngày thứ Hai" hoặc biểu thị cảm xúc tức giận "Chết tiệt, tôi phải về vào thứ Hai". Như phần text nhấn mạnh cho thấy, phần lớn thông tin về cảm xúc có liên quan chặt chẽ với các từ khoá cần được nhận ra. Danh mục các từ khoá mang dấu hiệu cảm xúc như vậy đã được nghiên cứu đối với tiếng Anh [8], [56].

Hiện nay, những kết quả nghiên cứu về nhận dạng cảm xúc đã được công bố hầu như mới chỉ tập trung vào một số ngôn ngữ thông dụng trên thế giới. Đối với tiếng Việt, các nghiên cứu được thực hiện còn rất ít. Hiện tại, nghiên cứu về cảm xúc tiếng Việt đã được thực hiện ở cấp độ ngôn ngữ nhưng còn ít nghiên cứu đã được thực hiện ở phương diện xử lý tín hiệu. Có thể nói, bộ ngữ liệu đầu tiên về cảm xúc tiếng Việt là bộ ngữ liệu đã được Lê Thị Xuyến xây dựng trong luận án tiến sĩ của mình [57]. Bộ ngữ liệu có nội dung gồm 5 câu và 2 người nói (một nam, một nữ). Các câu này cũng được hai người Pháp nói tương ứng bằng tiếng Pháp. Người nói tự tập luyện thể hiện cảm xúc của mình theo các câu và cuối cùng mới được ghi âm. Trong số 5 câu, có 4 câu được thể hiện với 12 cảm xúc khác nhau: bình thường*, lừa dối, bất ngờ*, vui*, tức giận*, hài lòng (thỏa mãn), xác nhận, chán nản*, khuyên bảo, nghi ngờ*, mỉa mai* và hối hận. Câu còn lại được thể hiện bằng 7 cảm xúc (bảy cảm xúc được đánh dấu *). Dựa trên bộ ngữ liệu này, Lê Thị Xuyến đã nghiên cứu các tín hiệu tiếng nói đại diện cho thái độ tâm lý và biểu cảm, mối quan hệ giữa các sự kiện âm thanh và kết quả của các thử nghiệm nhận thức, trải nghiệm chéo trong cả hai ngôn ngữ.

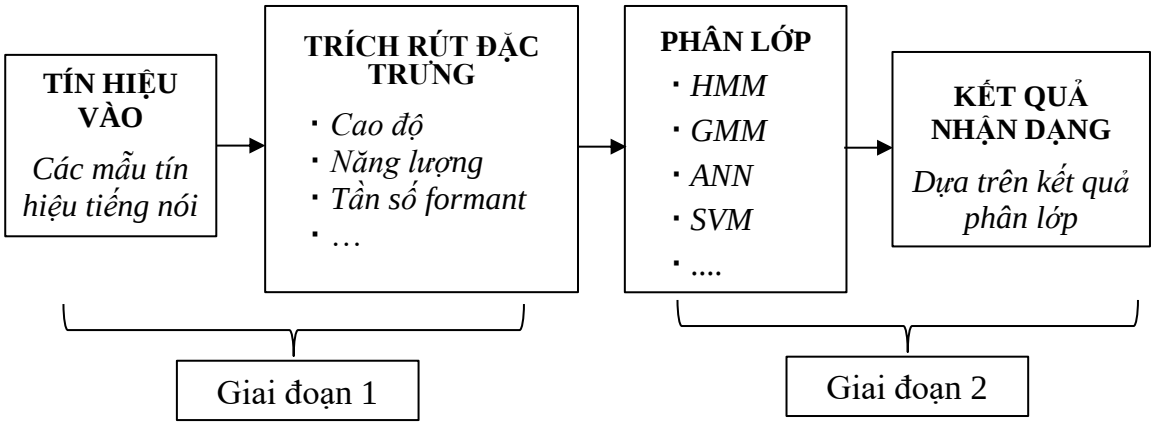
Về mặt ngôn ngữ, có thể kể đến công trình "Ngữ điệu tiếng Việt sơ khảo" của Đỗ Tiến Thắng công bố năm 2009 [58]. Trong [58], tác giả đã xét đến các ngữ điệu với chức năng ngữ pháp như ngữ điệu cấu tạo và ngữ điệu mục đích, ngữ điệu tình thái với chức năng biểu cảm, ngữ điệu hàm ý với chức năng lôgic, ngữ điệu hành vi và ngữ điệu hội thoại với chức năng dụng học.

Có thể nói, các nghiên cứu về tiếng nói tiếng Việt với giọng trần thuật (bình thường) đã có nhiều kết quả rất tốt. Trong khi đó, chưa có nhiều nghiên cứu về phương diện cảm xúc trong tổng hợp hay nhận dạng tiếng Việt. 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 [59], [60], [61] 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 [62] 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 nam nghệ sĩ và một nữ nghệ sĩ phát âm 19 câu ở năm trạng thái cơ bản: bình thường, vui, buồn, tức giận và rất tức giận.

Phần trên của luận án đã trình bày tình hình chung trong và ngoài nước về nhận dạng cảm xúc tiếng nói. Nội dung tiếp theo sau đây của luận án sẽ khái quát hóa một số bộ phân lớp thường sử dụng cho nhận dạng cảm xúc.

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

Nhìn chung, hệ thống nhận dạng cảm xúc tiếng nói xét theo phương diện xử lý tín hiệu của các ngôn ngữ khác nhau thường được thực hiện theo sơ đồ khối trên Hình 1.2.



Hình 1.2 *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 là giai đoạn tiền xử lý. Từ ngữ liệu sẵn có, giai đoạn này trích rút các đặc trưng thích hợp như tần số $F0$, năng lượng, formant và dải thông tương ứng, ...
- Giai đoạn 2 là phân loại cảm xúc dựa trên các bộ phân lớp, bộ phân lớp sẽ quyết định giọng nói có cảm xúc nào.

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. Hiện nay, các bộ phân lớp truyền thống đã được sử dụng hầu như trong tất cả các hệ thống nhận dạng cảm xúc tiếng nói. Có nhiều kiểu bộ phân lớp khác nhau để nhận dạng cảm xúc tiếng nói như HMM (Hidden Markov Model), GMM (Gaussian Mixture Model), SVM (Support Vector Machines), ANN (Artificial Neural Network), k-NN (k-Nearest Neighbor), … Nói chung, không có một sự thỏa thuận về bộ phân lớp nào là thích hợp nhất cho phân lớp cảm xúc. Dường như mỗi bộ phân lớp có ưu thế và hạn chế riêng của nó. 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

Phân tích dữ liệu là bước then chốt trong bất kỳ quá trình nhận dạng mẫu và liên quan chặt chẽ với hiệu năng và tính phức tạp của bộ phân lớp. Trên thực tế, nếu như các đặc trưng được trích rút từ tín hiệu vẫn còn mang ý nghĩa vật lý (biên độ, tần số,

đường bao…) thì các đặc tính có được sau bước phân tích dữ liệu sẽ mất đi ý nghĩa vật lý trong không gian biểu diễn mới. Có nhiều kỹ thuật được dùng để phân loại dữ liệu, trong đó kỹ thuật PCA (Principal Component Analysis) và LDA (Linear Discriminant Analysis) là hai kỹ thuật thường được sử dụng để phân loại dữ liệu và giảm chiều. Mục tiêu của LDA là tối thiểu hóa khoảng cách của các véctơ trong cùng một lớp và cực đại hóa khoảng cách giữa các tâm lớp.

Giả sử các đối tượng thuộc vào N lớp, π_n là xác suất tiên nghiệm để một đối tượng đến từ lớp thứ n, $f_n(x) = P(X = x|Y = n)$ là hàm mật độ xác suất để đối tượng X lấy giá trị x khi đang ở lớp thứ n, giả định $f_n(x)$ là hàm chuẩn Gauss đa thể hiện (phương trình (1.1)) [63].

$$f_n(x) = N(x; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \quad (1.1)$$

Trong đó: $\boldsymbol{\mu}$ là kỳ vọng, Σ là ma trận hiệp phương sai, D là số chiều của không gian vào.

Định lý Bayes [64] được mô tả ở phương trình (1.2) cho phép tính xác suất hậu nghiệm để đối tượng có giá trị bằng x khi thuộc vào lớp n.

$$P(Y = n|X = x) = \frac{\pi_n f_n(x)}{\sum_{l=1}^{N} \pi_l f_l(x)} \quad (1.2)$$

Đối tượng sẽ được nhận vào lớp nào có giá trị xác suất hậu nghiệm tương ứng lớn nhất (phương trình (1.2)). Với phương pháp phân tích phân biệt tuyến tính LDA, giả sử mỗi lớp có riêng giá trị kỳ vọng μ_n song tất cả các lớp đều có chung ma trận hiệp phương sai Σ. Phương trình (1.3) sau đây được gọi là các hàm phân biệt tuyến tính (Linear Discriminant Functions).

$$\delta_n(x) = x^T\Sigma^{-1}\mu_n - \frac{1}{2}\mu_n^T\Sigma^{-1}\mu_n + \log\pi_n \quad (1.3)$$

Vì $\delta_n(x)$ là hàm tuyến tính của x nên phương pháp này được gọi là phương pháp phân biệt tuyến tính.

1.4.2 Bộ phân lớp phân tích khác biệt toàn phương QDA

Với bộ phân lớp khác biệt toàn phương QDA (Quadratic Discriminant Analysis), giả sử mỗi lớp có ma trận hiệp phương sai riêng Σ_n, khi đó hàm phân biệt sẽ được biểu diễn bằng phương trình (1.4) [65].

$$\delta_n(x) = -\frac{1}{2}\log|\Sigma_n| - \frac{1}{2}(x - \mu_n)^T\Sigma_n^{-1}(x - \mu_n) + \log\pi_n \quad (1.4)$$

Các tham số μ_n và Σ_n trong các phương trình (1.3) và (1.4) sẽ được xác định trong quá trình huấn luyện dựa vào các dữ liệu huấn luyện.

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

Với mỗi đối tượng x trong tập thử nghiệm, cần tính giá trị $Y_n(x)$ theo phương trình (1.5).

$$Y_n(x) = \frac{1}{k}\Sigma_{x_i \in N_k(x)} y_i \quad (1.5)$$

Trong phương trình (1.5), $N_k(x)$ là láng giềng của x, bao gồm k điểm gần x nhất trong tập huấn luyện, y_i là trọng số của điểm trong tập huấn luyện. Đối tượng x được nhận dạng vào lớp n nếu $Y_n(x)$ đạt giá trị lớn nhất khi so sánh với các giá trị $Y_n(x)$.

1.4.4 Bộ phân lớp hỗ trợ véctơ SVC

Bộ phân lớp SVC (Support Vector Classifier) là sự mở rộng của bộ phân lớp phân biệt tuyến tính với lề cực đại (maximal margin classifier), cho phép phân lớp với các lớp không thể phân tách bằng một biên giới tuyến tính [65]. Lề cực đại được xác định như sau: với mỗi mẫu trong tập huấn luyện, tính khoảng cách trực giao đến biên giới phân lớp; lề là khoảng cách trực giao tối thiểu tìm được. Bộ phân lớp này chọn biên giới phân lớp có lề đạt giá trị lớn nhất, nghĩa là biên giới phân lớp phân biệt tốt nhất các mẫu trong tập huấn luyện. Các véctơ nằm trên lề được gọi là các véctơ hỗ trợ. Phân lớp SVC sẽ tìm biên giới phân lớp phù hợp nhất với đa số các mẫu và chấp nhận một số mẫu huấn luyện bị phân lớp sai (được điều chỉnh bằng tham số *C* như sẽ trình bày trong phương trình (1.7) dưới đây). Phiên bản mở rộng của phương pháp này là máy hỗ trợ véctơ SVM.

1.4.5 Bộ phân lớp máy hỗ trợ véctơ SVM

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

Xét bài toán sử dụng SVM để phân chia các mẫu thành 2 lớp. Giả sử tập huấn luyện bao gồm N mẫu $x_i, i = 1,2, \dots, N$. Các mẫu này được phân vào lớp $y_i, i = 1,2, \dots, N$; y_i chỉ lấy các giá trị -1 hoặc 1. Biên giới phân lớp được biểu diễn bằng vế trái của phương trình (1.6).

$$f(x) = \beta_0 + \sum_{i=1}^{N} \alpha_i k(x, x_i) \quad (1.6)$$

Thực chất đa phần các giá trị α_i đều bằng 0, chỉ trừ những giá trị α_i của các véctơ hỗ trợ. Các giá trị này bị giới hạn theo phương trình (1.7).

$$0 \le \alpha_i \le C, i = 1, 2, \dots, N \quad (1.7)$$

C là giá trị cho phép các mẫu bị vi phạm. Khi C càng nhỏ, lề sẽ càng rộng, và ngược lại khi C càng lớn, lề sẽ càng hẹp, k là hàm kernel của hệ thống, u và v là hai véctơ của tập huấn luyện. Với bộ phân lớp hỗ trợ véctơ SVC, k được tính theo phương trình (1.8).

$$k(u, v) = u^T v \tag{1.8}$$

Với SVM, hàm k được sử dụng để biến đổi không gian tham số, và được tính theo phương trình (1.9) với γ là hệ số biến đổi của hàm k.

$$k(u, v) = \exp\{-\gamma|u - v|^2\} \tag{1.9}$$

Khi đó, thuật toán thực hiện tìm các giá trị β_0 và α_i theo phương trình (1.10) với k là ma trận $N \times N$ tính trên tất cả các cặp mẫu sử dụng trong quá trình huấn luyện.

$$\min_{\beta_0, \alpha} \sum_{i=1}^{N} \big(1 - y_i f(x_i)\big) + \frac{1}{2}\alpha^T k \alpha \tag{1.10}$$

Quá trình phân lớp được thực hiện bằng cách tính hàm f (phương trình (1.6)) trên mẫu cần thử nghiệm. Tùy vào dấu của hàm f mà mẫu thử nghiệm sẽ được phân vào một trong hai lớp.

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

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

So với bộ phân lớp SVM, các bộ phân lớp trên đã sử dụng toàn bộ dữ liệu huấn luyện để xây dựng biên giới phân lớp. Trong khi đó, phân lớp SVM chỉ sử dụng các véctơ hỗ trợ để quyết định biên giới phân lớp. Bộ phân lớp hỗ trợ véctơ SVC chỉ sử dụng biên giới phân lớp tuyến tính, còn bộ phân lớp SVM lại cho phép xây dựng biên giới phi tuyến với sự mở rộng số lượng tham số lớn. Về mặt thực chất, phân lớp SVC có thể coi là phân lớp SVM với hàm nhân tuyến tính (được tính theo phương trình 1.8).

1.4.6 Bộ phân lớp HMM

Bộ phân lớp HMM đã được dùng rộng rãi trong các ứng dụng như nhận dạng tiếng nói rời rạc và tiếng nói liên tục [6] vì HMM liên quan về mặt vật lý với cơ chế tạo tín hiệu tiếng nói. HMM là quá trình ngẫu nhiên kép chứa chuỗi Markov bậc nhất mà các trạng thái của nó bị ẩn đối với người quan sát. Gắn với mỗi trạng thái là một quá

trình ngẫu nhiên tạo nên chuỗi quan sát. Như vậy, mỗi trạng thái ẩn của mô hình nắm bắt được cấu trúc thời gian của dữ liệu. Về mặt toán, để mô hình hóa một chuỗi véctơ dữ liệu quan sát được, $x_1 \dots x_T$ bằng một HMM, giả thiết tồn tại chuỗi Markov ẩn để tạo ra dãy quan sát này. Ký hiệu K là số trạng thái, $\pi_i, i = 1, \dots K$ là các xác suất trạng thái khởi đầu đối với chuỗi Markvo ẩn còn $a_{ij}, j = 1, \dots K$ là xác suất chuyển đổi từ trạng thái i sang trạng thái j. Thông thường, các tham số HMM được ước lượng dựa trên nguyên lý cực đại khả hiện. Bằng cách giả thiết các dãy trạng thái thực sự là $s_1, \dots, s_T$, khả hiện của dữ liệu quan sát được cho bởi công thức (1.11).

$$\begin{aligned} p(\mathbf{x}_1, s_1 \dots, \mathbf{x}_\mathrm{T}, s_T) &= \pi_{s_1} b_{s_1}(\mathbf{x}_1) a_{s_1,s_2} b_{s_2}(\mathbf{x}_2) \dots a_{s_{T-1},s_T} b_{s_T}(\mathbf{x}_T) \\ &= \pi_{s_1} b_{s_1}(\mathbf{x}_1) \prod_{t=2}^{T} a_{s_{t-1},s_t} b_{s_t}(\mathbf{x}_t) \end{aligned} \tag{1.11}$$

trong đó: $b_i(\mathbf{x}_t) \equiv P(\mathbf{x}|s_\mathrm{t} = i)$ là mật độ quan sát của trạng thái thứ i. Mật độ này có thể là rời rạc với HMM rời rạc hoặc là mật độ hỗn hợp Gauss đối với HMM liên tục. Bởi vì dãy trạng thái thực chủ yếu là chưa biết nên phải lấy tổng cho tất cả các dãy trạng thái có thể có để tìm ra khả hiện của dãy dữ liệu đã cho, nghĩa là:

$$p(\mathbf{x}_1, \dots, \mathbf{x}_\mathrm{T}) = \sum_{s_1,\dots,s_T} \left(\pi_{s_1}, b_{s_1}(\mathbf{x}_1) \prod_{t=2}^{T} a_{s_{t-1}s_t} b_{s_t}(\mathbf{x}_t) \right) \tag{1.12}$$

Để tính hàm khả hiện với độ phức tạp $O(KT)$ có thể sử dụng các thuật toán rất hiệu quả như các thuật toán tiến và lùi [66] [67]. Trong giai đoạn huấn luyện, các tham số HMM được xác định như là các tham số cực đại hóa khả hiện của (1.12). Điều này đạt được bằng cách sử dụng thuật toán cực đại hóa kỳ vọng EM trong [68].

1.4.7 Bộ phân lớp GMM [63]

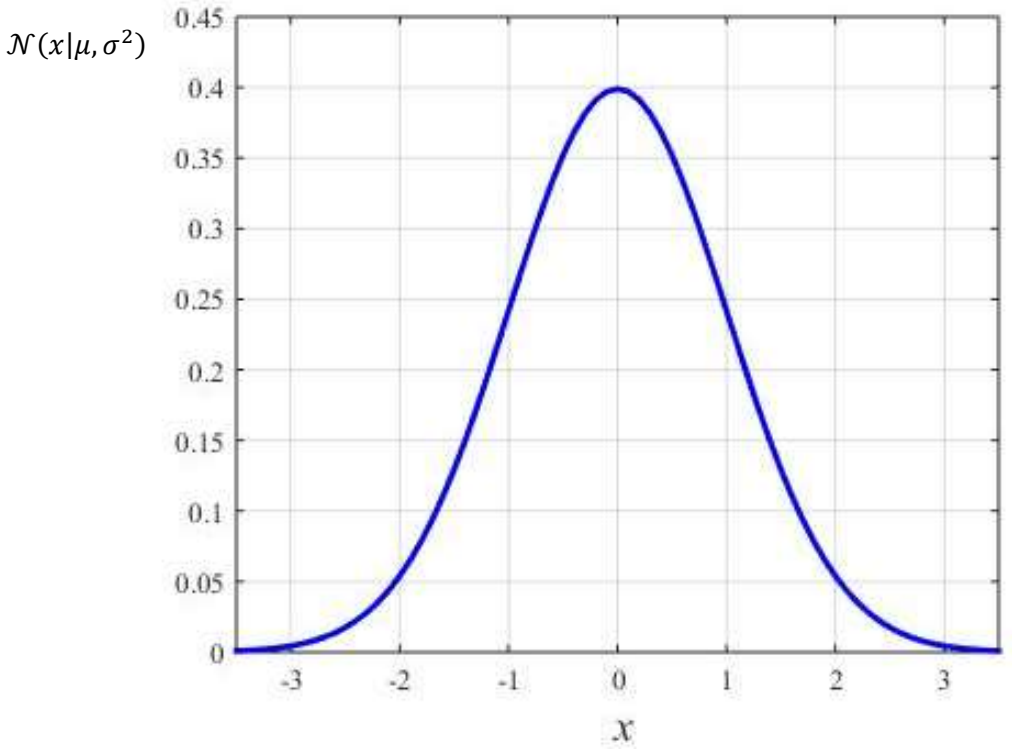
Mô hình GMM là mô hình xác suất để đánh giá mật độ bằng cách sử dụng tổ hợp lồi của các mật độ chuẩn đa thể hiện. GMM có thể được xem như HMM liên tục đặc biệt chứa chỉ một trạng thái [69]. GMM rất hiệu quả khi mô hình hóa các phân bố đa thể thức và các yêu cầu về việc huấn luyện ít hơn nhiều so với yêu cầu của HMM liên tục tổng quát. Do vậy, GMM thích hợp hơn 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 dữ liệu huấn luyện bởi vì tất cả các phương trình huấn luyện và nhận dạng đều dựa trên giả thiết rằng tất cả các vectơ là độc lập.

1.4.7.1 Mô hình hỗn hợp Gauss

Phân bố Gauss còn được gọi là phân bố chuẩn được sử dụng rộng rãi như là mô hình cho phân bố của các biến liên tục. Trong trường hợp biến đơn x, phân bố Gauss có thể được viết dưới dạng:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1.13}$$

Phân bố Gauss như trên được chi phối bởi 2 tham số: kỳ vọng μ và phương sai σ^2. Hình 1.3 là ví dụ cho phân bố Gauss đơn biến đơn thể hiện với $\mu = 0, \sigma = 1$.



Hình 1.3 *Phân bố Gauss đơn biến đơn thể hiện với* $\mu = 0$ *và* $\sigma = 1$

Từ (1.13) có thể thấy phân bố Gauss thỏa mãn:

$$\mathcal{N}(x|\mu,\sigma^2) \geq 0 \tag{1.14}$$

Phân bố Gauss được chuẩn hóa:

$$\int_{-\infty}^{\infty}\mathcal{N}(x|\mu,\sigma^2)dx = 1 \tag{1.15}$$

Kỳ vọng của x theo phân bố Gauss:

$$\mathbb{E}[x] = \int_{-\infty}^{\infty}\mathcal{N}(x|\mu,\sigma^2)xdx = \mu \tag{1.16}$$

Mô ment bậc 2:

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty}\mathcal{N}(x|\mu,\sigma^2)x^2dx = \mu^2 + \sigma^2 \tag{1.17}$$

Từ (1.16) và (1.17), phương sai của x sẽ là:

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \tag{1.18}$$

Trường hợp véc tơ **x** có *D* chiều, phân bố Gauss đa thể hiện có dạng:

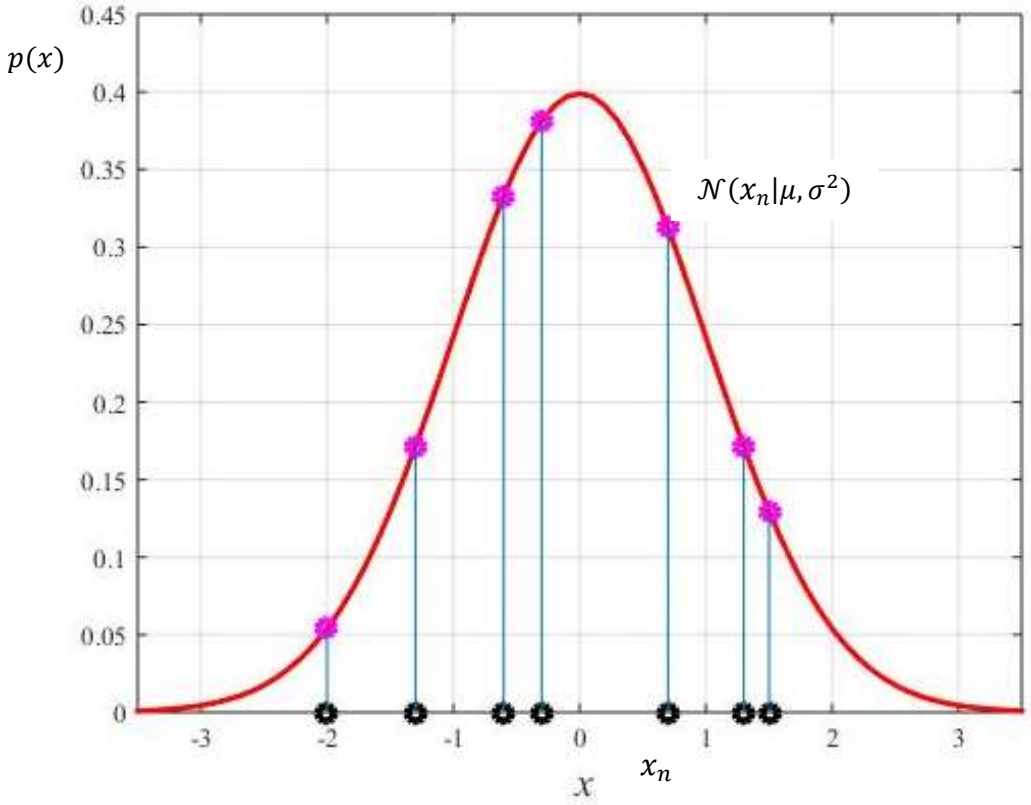
$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{1.19}$$

Trong biểu thức trên, $\boldsymbol{\mu}$ là véc tơ kỳ vọng có D chiều, $\boldsymbol{\Sigma}$ là ma trận hiệp phương sai có kích thước $D \times D$ còn $|\boldsymbol{\Sigma}|$ là định thức của $\boldsymbol{\Sigma}$.

Giả thiết tập dữ liệu của các quan sát là $\mathbf{x}$ với $\mathbf{x} = (x_1, \dots, x_N)^{\mathrm{T}}$ biểu diễn cho N quan sát của biến vô hướng x. Cũng giả thiết các quan sát được sinh ra một cách độc lập từ phân bố Gauss có kỳ vọng μ và phương sai σ^2 chưa biết và ta muốn xác định các tham số này từ tập dữ liệu. Các điểm dữ liệu được sinh ra một cách độc lập từ cùng một phân bố sẽ được gọi là có phân bố giống nhau và độc lập (*independent and identically distributed* – i.i.d). Bởi vì tập dữ liệu $\mathbf{x}$ là i.i.d, nên có thể viết như sau cho xác suất của tập dữ liệu với μ và σ^2:

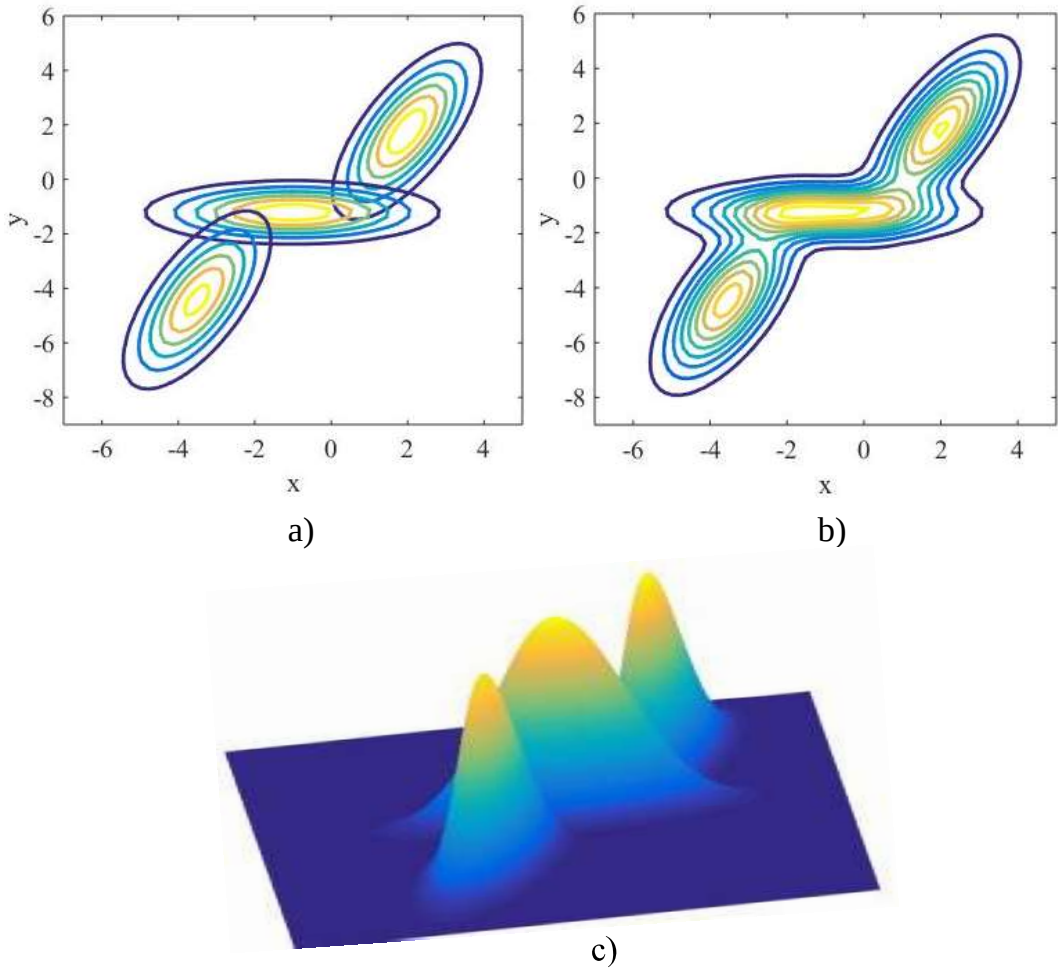
$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) \tag{1.20}$$

Một khi được xem như là hàm của μ và σ^2, đây là hàm khả hiện Gauss và có thể được diễn dịch như Hình 1.4.



Hình 1.4 *Hàm khả hiện đối với phân bố Gauss.*

Phân bố Gauss có những thuộc tính giải tích quan trọng song để mô hình hóa các tập dữ liệu thực lại có hạn chế. Vì vậy, việc xếp chồng tuyến tính các phân bố Gauss sẽ đặc trưng tốt hơn cho đặc tính phức tạp của tập dữ liệu thực. Bằng cách sử dụng số lượng đủ lớn các thành phần Gauss, điều chỉnh kỳ vọng và phương sai của chúng cũng như điều chỉnh các hệ số của tổ hợp tuyến tính, có thể xấp xỉ phần lớn các phân bố liên tục bất kỳ với độ chính xác tùy ý.



Hình 1.5 *Minh họa hỗn hợp 3 thành phần Gauss trong không gian 2 chiều*
a) Đường bao mật độ không đổi cho 3 thành phần hỗn hợp; b) Đường bao của mật độ xác suất biên $p(\mathbf{x})$ *của phân bố hỗn hợp, trọng số lần lượt là 0,5, 0,3 và 0,2;*
c) Biểu diễn phân bố $p(\mathbf{x})$ *theo bề mặt.*

Xét trường hợp xếp chồng của *K* phân bố Gauss như sau:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{1.21}$$

Đây là trường hợp phân bố Gauss hỗn hợp. Mỗi một phân bố $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ được gọi là một thành phần của hỗn hợp có kỳ vọng và phương sai riêng $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ tương ứng.

Hình 1.5 cho thấy phân bố Gauss có 3 thành phần. Tham số π_k là các hệ số hỗn hợp. Tích phân cả hai vế của (1.21) đối với $\mathbf{x}$ và lưu ý cả $p(\mathbf{x})$ và các thành phần Gauss riêng rẽ đều được chuẩn hóa, ta có:

$$\sum_{k=1}^{K} \pi_k = 1 \tag{1.22}$$

Yêu cầu $p(\mathbf{x}) \geq 0$, $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \geq 0$ dẫn tới $\pi_k \geq 0$ đối với mọi *k*. Kết hợp với (1.22) sẽ có:

$$0 \le \pi_k \le 1 \tag{1.23}$$

Các hệ số hỗn hợp cũng thỏa mãn điều kiện như là xác suất. Tương đương với biểu thức (1.21), có thể viết:

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k) \tag{1.24}$$

Trong đó, $\pi_k = p(k)$ là xác suất tiên nghiệm của thành phần thứ *k*.

$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = p(\mathbf{x}|k)$ là xác suất có điều kiện của **x** đối với *k*. Một đại lượng quan trọng nữa là xác suất hậu nghiệm $p(k|\mathbf{x})$. Từ định lý Bayes, ta có:

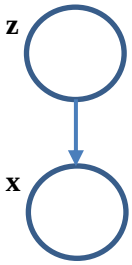
$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k) \quad = \frac{p(k)p(\mathbf{x}|k)}{\sum_l p(l)p(\mathbf{x}|l)} \quad = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_l \pi_l \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \tag{1.25}$$

Dạng phân bố hỗn hợp Gauss được chi phối bởi các tham số $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ và $\boldsymbol{\Sigma}$, trong đó $\boldsymbol{\pi} \equiv \{\pi_1, \dots, \pi_K\}$, $\boldsymbol{\mu} \equiv \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}$ và $\boldsymbol{\Sigma} \equiv \{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K\}$. Để thiết lập giá trị của các tham số này có thể dùng cực đại khả hiện (likelihood). Từ (1.21), logarit của hàm khả hiện cho bởi:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{1.26}$$

Trong đó $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$. Có thể thấy rằng đây là trường hợp phức tạp hơn nhiều so với phân bố Gauss đơn do có tổng theo *k* bên trong logarit. Vì vậy lời giải của các tham số không còn dưới dạng giải tích nữa. Trong trường hợp này có thể sử dụng cực đại kỳ vọng để nhận được lời giải.

Giả thiết biến **z** nhị phân ngẫu nhiên *K* chiều có một trong *K* cách biểu diễn trong đó phần tử đặc biệt $z_k = 1$ còn các phần tử khác bằng 0. Vì thế giá trị z_k thỏa mãn $z_k \in \{0,1\}$ còn $\sum_k z_k = 1$. Có thể thấy có *K* trạng thái đối với véc tơ **z** tương ứng với nó có phần tử khác 0. Định nghĩa phân bố kết hợp $p(\mathbf{x}, \mathbf{z})$ theo phân bố biên $p(\mathbf{z})$ và phân bố có điều kiện $p(\mathbf{x}|\mathbf{z})$ tương ứng với mô hình trên Hình 1.6.



Hình 1.6 *Đồ thị biểu diễn một mô hình hỗn hợp trong đó phân bố kết hợp được biểu diễn dưới dạng* $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$

Phân bố biên đối với **z** được quy định tùy thuộc vào các hệ số hỗn hợp π_k sao cho $p(z_k = 1) = \pi_k$.

Tham số $\{\pi_k\}$ phải thỏa mãn: $0 \leq \pi_k \leq 1$ (1.27)

Cùng với: $$\sum_{k=1}^{K} \pi_k = 1 \quad (1.28)$$

Vì **z** dùng một trong *K* cách biểu diễn nên có thể viết phân bố này dưới dạng:

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \quad (1.29)$$

Tương tự như vậy, phân bố có điều kiện của **x** với một giá trị đặc biệt của **z** cũng là phân bố Gauss:

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (1.30)$$

Công thức này cũng có thể được viết dưới dạng:

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \quad (1.31)$$

Phân bố kết hợp cho bởi $p(\mathbf{z})p(\mathbf{x}|\boldsymbol{z})$ còn phân bố biên của **x** là tổng của các phân bố kết hợp lấy cho tất cả các trạng thái có thể có của **x**:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (1.32)$$

Ở đây đã sử dụng công thức (1.30) và (1.31). Như vậy phân bố biên của **x** là phân bố Gauss hỗn hợp có dạng (1.21). Nếu có các quan sát $\mathbf{x}_1, \ldots, \mathbf{x}_N$ và phân bố biên có dạng $p(\mathbf{x}) = \sum_z p(\mathbf{x}, \mathbf{z})$ nên với mỗi điểm dữ liệu quan sát được $\mathbf{x}_n$ sẽ có biến tiềm ẩn $\mathbf{z}_n$.

Từ đó sẽ có công thức tương đương của phân bố Gauss hỗn hợp tương ứng với một biến tiềm ẩn được biểu diễn tường minh. Như vậy, có thể làm việc với phân bố kết hợp $p(\mathbf{x}, \mathbf{z})$ thay cho làm việc với phân bố biên $p(\mathbf{x})$ và điều này dẫn tới đơn giản hóa rất quan trọng thông qua thuật toán cực đại hóa kỳ vọng (EM – Expectation Maximization).

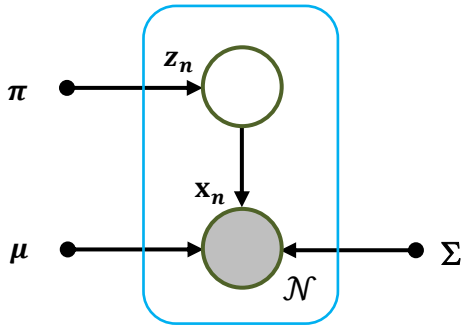
Một đại lượng khác đóng vai trò quan trọng là xác suất có điều kiện của **z** với **x** đã cho. Sử dụng ký hiệu $\gamma(z_k)$ cho $p(z_k = 1|\mathbf{x})$ và $\gamma(z_k)$ được xác định theo định lý Bayes như sau:

$$\begin{aligned} \gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \end{aligned} \quad (1.33)$$

π_k là xác suất tiên nghiệm để $z_k = 1$ còn $\gamma(z_k)$ là xác suất hậu nghiệm tương ứng khi đã có quan sát **x**. $\gamma(z_k)$ có thể xem như là đại lượng đóng vai trò trách nhiệm dẫn tới phần tử *k* sẽ lấy quan sát **x**.

1.4.7.2 Cực đại hóa khả hiện

Giả thiết có tập dữ liệu quan sát $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ và ta muốn mô hình hóa dữ liệu này bằng phân bố Gauss hỗn hợp. Có thể biểu diễn tập dữ liệu này như là ma trận $\mathbf{X}$ có kích thước $N \times D$ trong đó hàng n là $\mathbf{x}_n^{\mathrm{T}}$. Tương tự như vậy, các biến ẩn được biểu diễn bằng ma trận $\mathbf{Z}$ kích thước $N \times K$ với các hàng là $\mathbf{z}_n^{\mathrm{T}}$. Giả thiết rằng các điểm dữ liệu có phân bố độc lập nên có thể biểu diễn mô hình Gauss hỗn hợp đối với tập dữ liệu này bằng cách biểu diễn đồ họa như trên Hình 1.7.



Hình 1.7 *Đồ thị biểu diễn một mô hình Gauss hỗn hợp*

Hình 1.7 biểu diễn cho một tập N điểm ngẫu nhiên độc lập có phân bố giống nhau $\{\mathbf{x}_n\}$, với các điểm tiềm ẩn $\{\mathbf{z}_n\}$, trong đó $n = 1,\ldots, N$.

Từ (1.21), log của hàm khả hiện cho bởi:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{1.34}$$

Ở đây có vấn đề quan trọng liên quan tới việc cực đại hóa khả hiện cho mô hình Gauss hỗn hợp do có tính đơn điệu. Để đơn giản, xét phân bố Gauss hỗn hợp trong đó các thành phần của nó có các ma trận hiệp phương sai cho bởi $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbf{I}$ với $\mathbf{I}$ là ma trận đơn vị. Kết luận được rút ra cũng sẽ đúng với trường hợp ma trận hiệp phương sai tổng quát. Giả thiết một trong các thành phần của mô hình hỗn hợp chẳng hạn thành phần thứ j có trung bình là $\boldsymbol{\mu}_j$ chính xác bằng một trong những điểm dữ liệu sao cho $\boldsymbol{\mu}_j = \mathbf{x}_n$ đối với một giá trị nào đó của n. Điểm dữ liệu này sẽ tham gia vào số hạng trong hàm khả hiện dưới dạng:

$$\mathcal{N}(\mathbf{x}_n|\mathbf{x}_n, \sigma_j^2 \mathbf{I}) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j} \tag{1.35}$$

Nếu xét giới hạn khi $\sigma_j \to 0$, số hạng này sẽ tiến tới vô hạn. Vì thế log của hàm khả hiện cũng tiến tới vô hạn. Như vậy, việc cực đại hóa của hàm log khả hiện là bài toán được đặt ra không thích hợp bởi vì tính đơn điệu như thế luôn luôn có mặt và xuất hiện bất cứ khi nào một trong những thành phần của phân bố Gauss chạm tới một điểm dữ liệu cụ thể. Vấn đề này không xảy ra với phân bố Gauss đơn. Như vậy lưu ý rằng khi áp

dụng cực đại khả hiện đối với mô hình hỗn hợp Gauss phải theo các bước để tránh tìm ra lời giải vô lý và tránh đi tìm cực đại địa phương của hàm khả hiện.

Vấn đề khác liên quan tới lời giải cực đại khả hiện là với bất kỳ nghiệm cực đại khả hiện nào thì hỗn hợp K phần tử sẽ có $K!$ nghiệm tương đương ứng với $K!$ cách gán K tập các tham số cho K thành phần. Nói cách khác, đối với điểm đã cho bất kỳ (không suy biến) trong không gian các giá trị tham số sẽ có $K! - 1$ điểm nữa có cùng phân bố.

1.4.7.3 EM cho Gauss hỗn hợp

Xét các điều kiện cần phải được thỏa mãn tại cực đại của hàm khả hiện.

Đạo hàm của $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ trong (1.34) đối với trung bình $\boldsymbol{\mu}_k$ của các thành phần Gauss và gán bằng 0, ta có:

$$0 = -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \boldsymbol{\Sigma}_k (x_n - \mu_k) \tag{1.36}$$

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad *$$

Ở đây đã dùng công thức (*) đối với phân bố Gauss. Lưu ý là các xác suất hậu nghiệm $\gamma(z_{nk})$ trong (1.33) xuất hiện rất tự nhiên bên vế phải. Nhân cả 2 vế với $\boldsymbol{\Sigma}_k^{-1}$ và sắp xếp lại ta có:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{1.37}$$

Trong đó định nghĩa:

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{1.38}$$

Có thể xem N_k như là số lượng thực tế của các điểm đã được gán cho cụm k. Lưu ý tới dạng lời giải này. Có thể thấy rằng nhận được trung bình $\boldsymbol{\mu}_k$ đối với thành phần Gauss thứ k bằng cách lấy trung bình có trọng số của tất cả các điểm trong tập dữ liệu trong đó trọng số đối với dữ liệu $\mathbf{x}_n$ là xác suất hậu nghiệm $\gamma(z_{nk})$ mà thành phần k tạo nên $\mathbf{x}_n$.

Nếu cho đạo hàm của $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ đối với $\boldsymbol{\Sigma}_k$ bằng 0 và lý luận tương tự, bằng cách sử dụng nghiệm cực đại khả hiện đối với ma trận hiệp phương sai của Gauss đơn biến, ta có:

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \tag{1.39}$$

Công thức này có cùng một dạng với kết quả tương ứng cho hàm Gauss đơn biến khớp cho tập dữ liệu trong đó mỗi điểm dữ liệu được gán trọng số bằng xác suất hậu

nghiệm tương ứng. Còn mẫu số cho bởi số lượng thực tế của các điểm đã gán cho thành phần tương ứng.

Cuối cùng, cực đại hóa $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ với các hệ số hỗn hợp π_k có tính đến ràng buộc (1.22) tức là tổng các hệ số hỗn hợp phải bằng 1. Điều này đạt được bằng cách sử dụng nhân tử Lagrange và cực đại hóa đại lượng sau đây:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right) \tag{1.40}$$

Từ đó có:

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}\left(\mathbf{x}_n\middle|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} + \lambda \tag{1.41}$$

Ở đây lại thấy xuất hiện của xác suất hậu nghiệm. Nếu nhân cả 2 vế với π_k và lấy tổng theo k với ràng buộc (1.22) sẽ có $\lambda = -N$. Dùng kết quả này để loại λ và sắp xếp lại ta có:

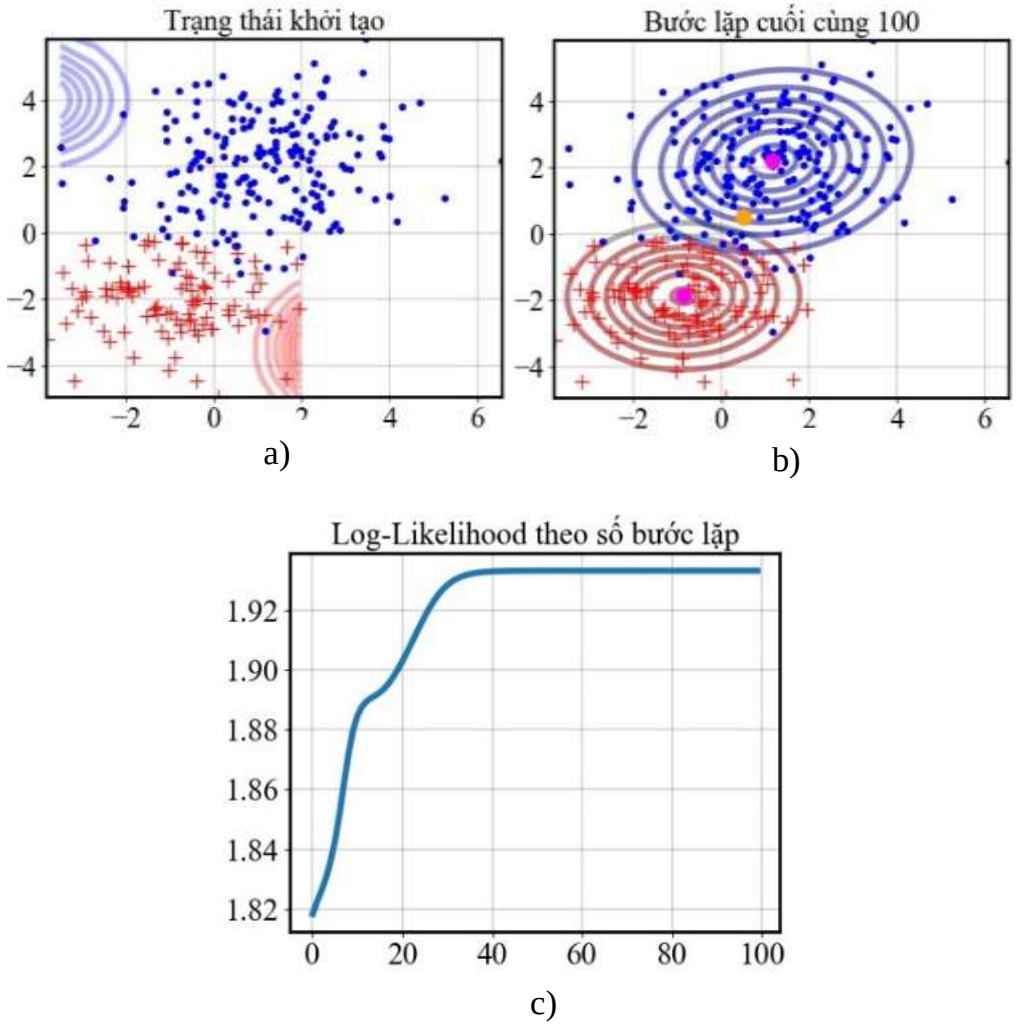
$$\pi_k = \frac{N_k}{N} \tag{1.42}$$

Như vậy hệ số hỗn hợp đối với thành phần thứ k là trung bình của xác suất hậu nghiệm mà thành phần đó lấy để có các điểm dữ liệu.

Cần nhấn mạnh rằng các kết quả (1.33), (1.35) và (1.38) không phải là nghiệm dưới dạng giải tích đối với các tham số của mô hình hỗn hợp vì xác suất hậu nghiệm $\gamma(z_{nk})$ phụ thuộc vào các tham số này theo một cách phức tạp thông qua (1.29). Tuy nhiên các kết quả này gợi ý cách thức truy hồi đơn giản để tìm ra nghiệm của bài toán cực đại khả hiện, chính là một trường hợp của thuật toán EM cho riêng mô hình Gauss hỗn hợp. Đầu tiên chọn các giá trị khởi tạo cho trung bình, hiệp phương sai và các hệ số hỗn hợp sau đó sẽ luân phiên giữa hai cập nhật: bước E và bước M.

Trong bước kỳ vọng (bước E) sẽ dùng các giá trị hiện tại của các tham số để đánh giá các xác suất hậu nghiệm cho bởi (1.29). Sau đó dùng các xác suất này ở bước cực đại hóa (bước M) để đánh giá lại trung bình phương sai và các hệ số hỗn hợp bằng cách dùng các kết quả (1.33), (1.35) và (1.38). Lưu ý rằng bằng cách làm như thế, đầu tiên sẽ đánh giá các giá trị trung bình mới bằng cách dùng (1.33) sau đó dùng các giá trị mới này để tìm ra các giá trị hiệp phương sai bằng cách dùng (1.35) trong khi giữ nguyên các kết quả tương ứng đối với phân bố Gauss đơn biến.

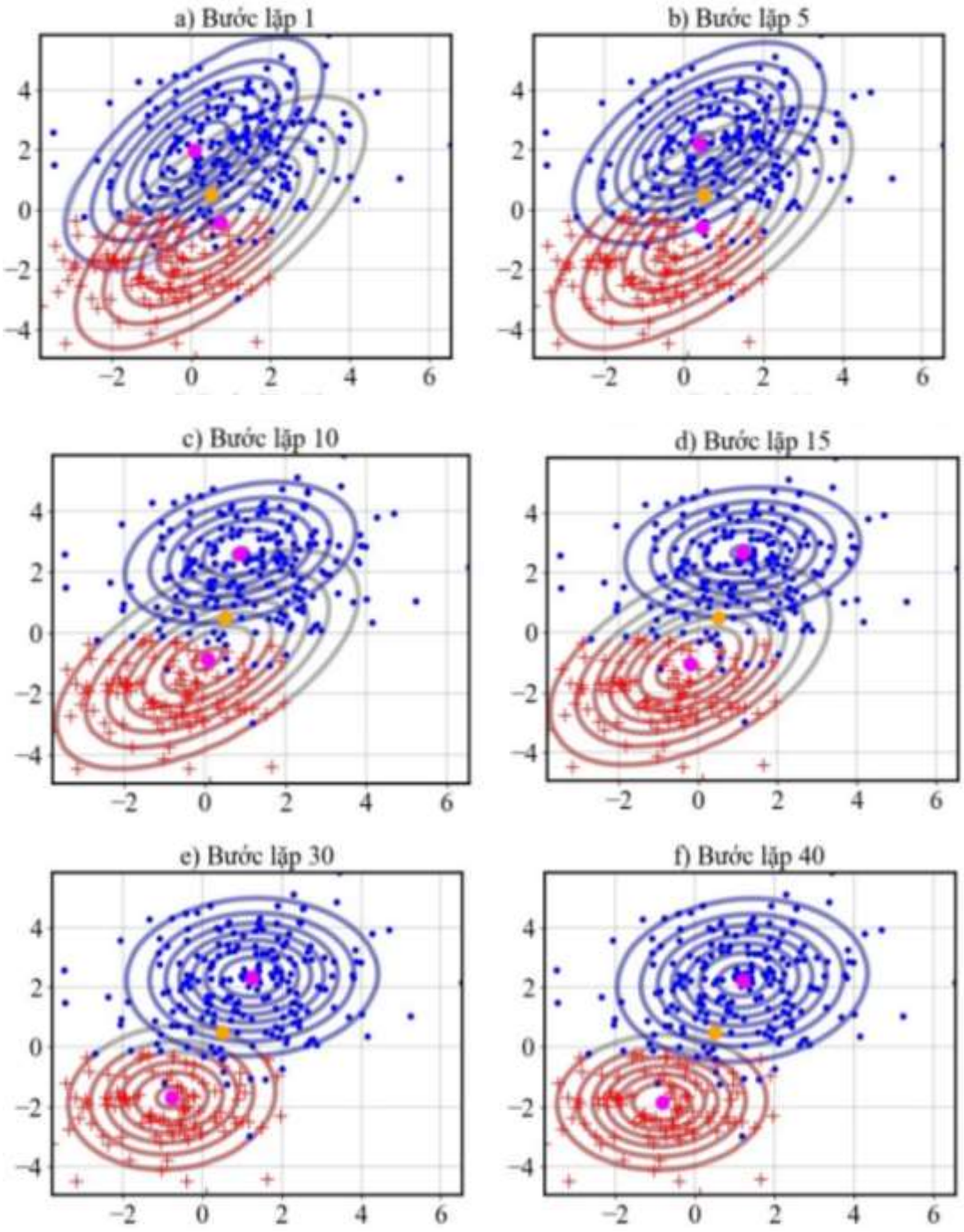
Có thể chỉ ra rằng mỗi cập nhật đối với các tham số là kết quả từ bước E sau đó là bước M sẽ đảm bảo làm tăng hàm log khả hiện. Trên thực tế, thuật toán được coi là hội tụ khi có sự thay đổi của hàm log khả hiện hoặc là tương đương như vậy có sự thay đổi các tham số rơi xuống thấp hơn một ngưỡng nào đó.



Hình 1.8 *Phân bố của 2 tập dữ liệu 2D và PDF tương ứng theo GMM*
a) Phân bố của 2 tập dữ liệu 2D và khởi tạo ban đầu của EM; b) PDF của 2 tập dữ liệu sau bước lặp 100 của EM; c) Log-Likelihood theo số bước lặp

Có thể minh họa thuật toán EM cho hỗn hợp của hai phân bố Gauss trên Hình 1.9 với tập dữ liệu trên Hình 1.8. Ở đây có hỗn hợp của hai phân bố Gauss được sử dụng. Các tâm được khởi tạo ban đầu như Hình 1.8, ma trận hiệp phương sai được khởi tạo với giá trị như nhau còn tỷ lệ của 2 thành phần được khởi tạo là 0,5 và 0,5. Hình 1.9 cho thấy các điểm dữ liệu mầu xanh lam và mầu đỏ cùng với cấu hình khởi tạo của mô hình hỗn hợp trong đó đường bao PDF cho hai thành phần Gauss được biểu diễn bằng các đường tròn màu xanh lam và màu đỏ. Hình 1.9 a) cho thấy kết quả của bước lặp đầu tiên.

Tình trạng sau bước M đầu tiên được biểu diễn trên Hình 1.9 c) trong đó trung bình của phân bố Gauss xanh đã được chuyển thành trung bình của tập dữ liệu được lấy trọng số bằng xác suất của mỗi điểm dữ liệu thuộc về cụm màu xanh, nói cách khác, là điểm chuyển tới trọng tâm của màu xanh. Cũng như thế, hiệp phương sai của phân bố Gauss màu xanh được đặt bằng hiệp phương sai màu xanh. Kết quả cũng tương tự như vậy đối với thành phần đỏ. Các Hình 1.9 d), e), f) cho thấy kết quả sau các bước của thuật toán EM thứ 15, 30 và 40 được hoàn thành tương ứng.



Hình 1.9 *Minh họa thuật toán EM, phân bố dữ liệu và đánh giá PDF theo EM a) Sau bước lặp 1/100; b) Sau bước lặp 5/100; c) Sau bước lặp 10/100; d) Sau bước lặp 15/100; e)Sau bước lặp 30/100; f) Sau bước lặp 40/100*

Hình 1.9 f) là thuật toán tiến tới hội tụ. Lưu ý là thuật toán EM cần nhiều bước lặp hơn để tiến tới hội tụ so với thuật toán *K-mean* và mỗi bước lặp cần số lượng tính toán lớn hơn nhiều. Do đó thường chạy thuật toán *K-mean* để tìm ra khởi tạo thích hợp cho mô hình Gauss hỗn hợp sau đó mới áp dụng thuật toán EM. Các ma trận hiệp phương sai có thể khởi tạo một cách thuận tiện như là các ma trận hiệp phương sai của các cụm mà thuật toán *K-mean* đã tìm ra, còn các hệ số hỗn hợp có thể đặt bằng tỷ lệ các điểm đã được gán cho các cụm tương ứng.

1.4.7.4 Thuật toán EM cho mô hình Gauss hỗn hợp

Cho mô hình Gauss hỗn hợp, mục tiêu là cực đại hóa hàm khả hiện đối với các tham số (bao gồm trung bình, hiệp phương sai của các thành phần và các hệ số hỗn hợp)

- Khởi tạo trung bình $\boldsymbol{\mu}_k$, hiệp phương sai $\boldsymbol{\Sigma}_k$ và các hệ số hỗn hợp π_k đồng thời đánh giá giá trị khởi tạo của logarit khả hiện.
- Bước E: đánh giá các xác suất hậu nghiệm sử dụng giá trị các tham số hiện tại

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \Sigma_j)} \tag{1.43}$$

- Bước M: đánh giá lại các tham số bằng cách sử dụng các xác suất hậu nghiệm hiện tại

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{1.44}$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k^{new})(\mathbf{x}_n - \boldsymbol{\mu}_k^{new})^{\mathrm{T}} \tag{1.45}$$

$$\pi_k^{new} = \frac{N_k}{N} \tag{1.46}$$

Trong đó:
$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{1.47}$$

Đánh giá logarit khả hiện :

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{1.48}$$

Và kiểm tra hội tụ của các tham số hoặc của logarit khả hiện. Nếu tiêu chí hội tụ không thỏa mãn, trở lại bước 2.

Như vậy mô hình hỗn hợp Gauss - GMM là một dạng mô hình thống kê được xây dựng từ việc huấn luyện các tham số thông qua dữ liệu học. Về cơ bản, mô hình GMM xấp xỉ một hàm mật độ xác suất bằng cách kết hợp các hàm mật độ Gauss.

Đối với hướng tiếp cận mô hình GMM để giải quyết bài toán nhận dạng cảm xúc của người nói, 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.

1.4.8 Bộ phân lớp ANN

Bộ phân lớp ANN có một số ưu thế so với GMM và HMM. ANN được biết có hiệu quả hơn khi mô hình hóa các ánh xạ phi tuyến. Cũng như vậy, hiệu năng phân

lớp của ANN thường tốt hơn GMM và HMM khi số mẫu huấn luyện tương đối ít. Phần lớn các ANN có thể được phân thành 3 dạng chính: mạng truyền thẳng MLP (Multi Linear Perceptrol), mạng hồi quy RNN (Recurrent Neural Networks) và mạng hướng hàm cơ bản RBF (Radial Basis Functions) [6]. Trong đó, dạng RBF ít dùng trong nhận dạng cảm xúc tiếng nói. MLP thường dùng phổ biến trong nhận dạng cảm xúc tiếng nói vì dễ cài đặt và thuật toán huấn luyện được định nghĩa rõ ràng khi cấu trúc ANN được hoàn toàn qui định. Tuy nhiên, các bộ phân lớp ANN có nhiều tham số cần phải lựa chọn thiết kế như dạng hàm kích hoạt nơron, số lượng lớp ẩn và số nơron trong mỗi lớp. Trên thực tế, hiệu năng của ANN phụ thuộc rất nhiều vào các tham số này. Như vậy, trong một số hệ thống nhận dạng cảm xúc thường sử dụng từ 2 bộ phân lớp ANN trở lên [29]. Sơ đồ kết hợp thích hợp thường dùng để tổ hợp các đầu ra của các bộ phân lớp ANN riêng rẽ.

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

Hiện nay, các mô hình nhận dạng cảm xúc tiếng nói khá đa dạng. Vì vậy, đã có nhiều nghiên cứu phối hợp thử nghiệm các phương pháp thực hiện mô hình khác nhau. Trong [70], các tác giả đã thực hiện nhận dạng 5 cảm xúc với 25 mẫu giọng nói cho tiếng Mandarin (Trung Quốc). Có 60 tham số đặc trưng liên quan đến MFCC, tần số cơ bản, biến thiên trung bình qua trục không trong thời gian ngắn và năng lượng trong thời gian ngắn được sử dụng để nhận dạng. Kết quả nhận dạng dựa vào bộ phân lớp GMM đã đạt được tỷ lệ nhận dạng chính xác trung bình khoảng 80%. Trong [71], tác giả đã sử dụng HMM và GMM để nhận dạng 5 cảm xúc vui, buồn, tức, ngạc nhiên và bình thường với kết quả đạt được về độ chính xác từ 67,39% - 82,49% khi dùng HMM và từ 68,39% - 78,27% khi dùng GMM.

Nghiên cứu [72] dùng bộ phân lớp GMM với bộ dữ liệu tiếng nói KISMET chứa 726 giọng nói. Các cảm xúc được thể hiện gồm tán thành, chú ý, ngăn chặn, dịu dàng và trung tính. Tiêu chí lựa chọn mô hình dựa trên kurtosis đã được sử dụng để xác định số thành phần Gauss tối ưu cho mỗi mô hình [73]. Do số lượng phát ngôn sẵn có hạn chế nên mỗi lần lấy 100 phát ngôn để nhận dạng. Kỹ thuật lựa chọn đặc trưng SFS được dùng để lựa chọn các đặc trưng tốt nhất từ tập chứa các đặc trưng cao độ và năng lượng hình [6]. Độ chính xác cực đại đạt được là 78,7% với 5 đặc trưng tốt nhất. Bằng cách sử dụng sơ đồ phân lớp tuần tự có phân cấp, độ chính xác nhận dạng tăng lên 81,94%. GMM cũng được dùng với một số dữ liệu khác như BabyEars [74]. Ngữ liệu này gồm 509 giọng nói, trong đó có 212 giọng cảm xúc tán thành, 149 giọng cảm xúc lôi cuốn, 148 giọng cảm xúc ngăn chặn. Lỗi đánh giá chéo được đo với số thành phần của GMM từ 1-100. Hiệu năng trung bình tốt nhất khoảng 75% tương ứng với số thành phần mô hình bằng 10.

Kết quả tương tự với dữ liệu FERMUS III [75] chứa 5250 mẫu gồm các cảm xúc cơ bản và cảm xúc trung tính. Có 16 thành phần GMM được dùng để mô hình hóa cảm xúc. Độ chính xác trung bình đạt 74,83% đối với nhận dạng độc lập người nói và 98,7% đối với nhận dạng phụ thuộc người nói. Các kết quả này dựa trên đánh giá chéo.

Để mô hình hóa cấu trúc thời gian của dữ liệu, GMM thích hợp với việc xử lý tự hồi qui vectơ trong mô hình GMVAR được dùng cho nhận dạng cảm xúc trong bộ ngữ liệu Berlin [76]. Các cảm xúc được nhận dạng là vui, buồn, chán, tức, ghê tởm và bình thường. Cảm xúc ghê tởm được bớt đi vì số lượng phát ngôn ít. GMVAR có độ phân lớp chính xác là 76% so với 71% của HMM, 67% của k-NN và 55% của mạng nơron. Thêm vào đó, GMVAR cho độ chính xác 90% khi chia các cảm xúc thành 3 lớp còn HMM chỉ đạt 86%.

Để nhận dạng cảm xúc cho tiếng nói được ghi lại trong 10 giờ từ một trung tâm trả lời các cuộc gọi cấp cứu y tế bằng tiếng Pháp, các tác giả trong [77] sử dụng máy hỗ trợ véctơ SVM và mô hình cây logic LMT để phân loại hai cảm xúc tiêu cực và tích cực. Tỷ lệ nhận dạng đạt khoảng 82%. Với nghiên cứu [78], tác giả đã sử dụng mô hình GMM thực hiện nhận dạng bốn cảm xúc vui, buồn, tức và bình thường. Ngữ liệu gồm 30 giọng nam và 25 giọng nữ, kết quả nhận dạng đúng trung bình đạt 60%. Trong [79], các tác giả đã sử dụng mô hình nhận dạng SVM cho 6 cảm xúc: vui, buồn, chán, ghê tởm, sợ hãi, bình thường với ngữ liệu thử nghiệm sử dụng bộ ngữ liệu tiếng Đức. Có 182 tham số đặc trưng được sử dụng để nhận dạng bao gồm:

- Giá trị trung bình, phương sai, trung vị, giá trị min, max và phạm vi của biên độ tiếng nói
- Giá trị trung bình và phương sai của năng lượng tiếng nói
- Giá trị trung bình, phương sai, trung vị, giá trị min, max và dải cao độ tiếng nói
- Giá trị trung bình, phương sai, giá trị min, max và phạm vi của 4 formant đầu tiên
- Năng lượng của 22 băng tần con đầu tiên theo thang Bark
- Giá trị trung bình, phương sai, giá trị min, max và phạm vi của 12 hệ số MFCC
- Các đặc trưng dạng phổ: center of gravity, độ lệch chuẩn, skewness và kurtosis
- Trung bình và độ lệch chuẩn của chu kỳ xung thanh môn, giá trị tuyệt đối của biến động cục bộ, trung bình tương đối của nhiễu loạn, vi sai của chênh lệch chu kỳ và hệ số nhiễu loạn theo chu kỳ năm điểm.

Trong trường hợp này, tỷ lệ nhận dạng trung bình là 77,4% - 81,5% cho bốn cảm xúc vui, buồn, bình thường và sợ hãi. Bộ ngữ liệu tiếng Đức cũng được dùng để nhận dạng với nghiên cứu của [80], trong đó giới tính được nhận dạng trước khi nhận dạng cảm xúc. Khi có thêm thông tin về giới tính, hiệu năng nhận dạng của hệ thống đã được cải thiện, tỷ lệ nhận dạng tổng thể tăng 2% - 4% so với khi không có thông tin về giới tính.

Với mô hình HMM, cũng đã có rất nhiều nghiên cứu về nhận dạng cảm xúc sử dụng mô hình này. Nói chung, HMM cũng cho độ chính xác phân lớp tương thích với các bộ phân lớp đã biết khác. Mô hình HMM được dùng trong [81] để phân lớp 7 cảm xúc: vui, buồn, tức, ghê tởm, sợ hãi, ngạc nhiên và bình thường. Các đặc trưng LFPC, LPCC, MFCC được trích chọn từ tín hiệu tiếng nói. Ngữ liệu thử nghiệm là tiếng Trung và tiếng Myanma. Trong 720 file tiếng nói có 432 file được dùng để huấn luyện còn lại để nhận dạng. Tỷ lệ trung bình tốt nhất là 78,5% và 75,5% cho tiếng Trung và tiếng Myanma trong khi dùng người nghe để phân lớp chỉ đạt 65,8%. Như

vậy, hệ thống nhận dạng cảm xúc tiếng nói tốt hơn đối với riêng ngữ liệu này. Tuy nhiên, kết quả này không thể được tổng quát hóa trừ khi có một nghiên cứu toàn diện hơn và có nhiều ngữ liệu hơn được thực hiện.

HMM cũng được dùng trong nhiều nghiên cứu khác như trong [82], [83]. Trong [82], độ chính xác đạt 70,1% đối với việc phân lớp các giọng nói thành 4 lớp với ngữ liệu SUSAS độc lập nội dung. Nghiên cứu [83] có 2 hệ thống được đề xuất, hệ thống thứ nhất là hệ thống thông thường trong đó mỗi cảm xúc được mô hình hóa bằng HMM liên tục có 12 thành phần Gauss cho mỗi trạng thái. Hệ thống thứ 2 có HMM liên tục 3 trạng thái được xây dựng cho mỗi lớp âm vị. Có 46 âm vị được nhóm thành 5 lớp: nguyên âm, âm lướt (glide), âm mũi, âm tắc và âm xát. Mỗi trạng thái được mô hình hóa bằng 16 thành phần Gauss. Dữ liệu tiếng nói TMIT được dùng để huấn luyện HMM cho mỗi lớp âm vị. Để đánh giá kết quả, các file tiếng nói của ngữ liệu tiếng nói có cảm xúc được ghi âm tại chỗ chứa các cảm xúc tức, vui, buồn. Mỗi file tiếng nói được phân lớp theo mức âm vị và nhận được chuỗi âm vị. Đối với mỗi phát ngôn tiếng nói cần nhận dạng, mô hình tổng quát HMM được xây dựng cho phát ngôn đó bao gồm các lớp HMM theo âm vị được kết nối theo cùng trình tự của dãy âm vị tương ứng với số lượng khung bắt đầu và kết thúc của mỗi đoạn được xác định theo thuật toán Viterbi. Thủ tục này được lặp lại cho mỗi cảm xúc và tiêu chí cực đại khả hiện được dùng để xác định cảm xúc được biểu thị. Ngữ liệu tiếng nói chứa 704 file cho huấn luyện và 176 cho nhận dạng. Độ chính xác tổng thể nhận được đối với HMM phụ thuộc lớp âm vị là 76,12% so với 55,68% khi dùng SVM với các đặc trưng ngôn điệu và 64,77% đối với HMM nhận dạng cảm xúc nói chung. Dựa trên kết quả nhận được này, các tác giả cho rằng mô hình hóa dựa trên âm vị cho kết quả nhận dạng cảm xúc tốt hơn.

Mô hình ANN cũng được sử dụng trong nhiều nghiên cứu. Mục đích chính của [29] là sử dụng mô hình ANN để phân lớp 8 cảm xúc: vui, trêu chọc, sợ, buồn, ghê tởm, tức, ngạc nhiên và bình thường với ngữ liệu cảm xúc ghi âm tại chỗ. Bộ phân lớp cơ bản là OCON bao gồm 8 mạng nơron con MLP và một bộ điều khiển logic quyết định. Mỗi mạng nơron con chứa 2 lớp ẩn cùng với các lớp vào và lớp ra. Lớp ra chỉ chứa một nơron, còn đầu ra là giá trị tương tự từ 0-1. Mỗi mạng nơron con được huấn luyện để nhận dạng một trong tám cảm xúc. Trong giai đoạn nhận dạng, đầu ra của mỗi ANN qui định véctơ tiếng nói đầu vào là do cảm xúc nào tạo là phù hợp hơn cả. Bộ điều khiển logic quyết định tạo ra một giả thiết nào đó dựa trên các đầu ra của 8 mạng nơron con. Hệ thống này được áp dụng cho ngữ liệu tiếng nói ghi âm tại chỗ gồm 100 người nói, mỗi người nói phát âm 100 từ, mỗi từ 8 lần, mỗi lần cho một cảm xúc đã nói ở trên. Độ chính xác phân lớp nhận dạng tốt nhất chỉ đạt 52,87% khi huấn luyện 30 giọng và nhận dạng cho các giọng còn lại, tức là phân lớp độc lập người nói.

Độ phân lớp chính xác tương tự cũng đạt được trong [84] với cấu trúc mạng nơron AllClass-in-One. Bốn cấu hình mạng đã được thử nghiệm trong nghiên cứu này trong đó mạng nơron chỉ có một lớp ẩn chứa 56 nơron, lớp vào có 7 hoặc 8 nơron còn lớp ra có 14 hoặc 26 nơron. Độ chính xác phân lớp đạt được tốt nhất trong trường hợp này là 51,19%. Các mô hình phân lớp là phụ thuộc người nói.

Nghiên cứu trong [85] đạt kết quả tốt hơn. Với nghiên cứu này có 3 cấu hình ANN được sử dụng. Cấu hình thứ nhất là bộ phân lớp MLP 2 lớp thông thường, ngữ liệu cũng được ghi âm tại chỗ gồm 700 phát ngôn cho 5 cảm xúc: vui, tức, buồn, sợ hãi và bình thường. Tập con của ngữ liệu chứa 369 phát ngôn được lựa chọn dựa trên các quyết định chủ quan và được phân chia ngẫu nhiên thành tập con huấn luyện chứa 70% phát ngôn còn lại tập con 30% để nhận dạng. Độ chính xác trung bình vào khoảng 65%. Với cấu hình thứ 2, độ chính xác trung bình là 70% trong đó cấu hình sơ đồ kết hợp tăng cường (boostrap aggregation (bagging)) đã sử dụng. Sơ đồ này là phương pháp tạo ra nhiều phiên bản của bộ phân lớp và dùng các bộ phiên bản đó để có bộ phân lớp được kết hợp với độ phân lớp chính xác cao hơn [86]. Cuối cùng, cấu hình thứ 3 cho độ phân lớp chính xác trung bình là 63%.

Trong nghiên cứu [87], các tác giả đã sử dụng mạng nơron ANN và kỹ thuật bộ lọc thông cao để phân loại 5 cảm xúc vui, buồn, tức, bình thường và ghê tởm với các tham số đặc trưng MFCC. Ngữ liệu thử nghiệm là 10 mẫu tiếng nói của 5 người nói với 5 cảm xúc, tỷ lệ nhận dạng đúng tương đối tốt cho từng cảm xúc được thống kê trong Bảng 1.2.

Bảng 1.2 *Tỷ lệ nhận dạng các cảm xúc dựa trên ANN (nguồn: [87])*

Cảm xúc	Vui	Buồn	Tức	Bình thường	Ghê tởm
Tỷ lệ nhận dạng	95%	94,16%	92,74%	91,58%	93,42%

Kết quả nhận dạng cảm xúc đã được các nhà nghiên cứu công bố khá nhiều đồng thời việc tìm kiếm các tham số đặc trưng và mô hình nhận dạng để cải thiện hiệu năng, nâng cao tỷ lệ nhận dạng vẫn tiếp tục được thực hiện với nhiều phương pháp cho các ngôn ngữ khác nhau. Mới đây, các tác giả trong nghiên cứu [86] đã sử dụng kỹ thuật học sâu để thực hiện nhận dạng 5 cảm xúc: tức, vui, buồn, bình thường và thất vọng. Ngữ liệu được sử dụng là IEMOCAP gồm 12 giờ ghi âm được chia thành 5 buổi thu âm. Mỗi buổi gồm hai diễn viên (một nam và một nữ) thực hiện các kịch bản cảm xúc cũng như các kịch bản ngẫu hứng. Tổng cộng, ngữ liệu này gồm 10039 file tiếng nói với thời lượng trung bình mỗi file là 4,5 giây. Mô hình mạng nơron lấy chập CNN cho kết quả nhận dạng đúng trung bình đạt 64,78% với ngữ liệu độc lập người nói. Ngoài ra, có nhiều bộ phân lớp khác được dùng cho các nhận dạng cảm xúc như bộ phân lớp k-NN [30], phân lớp mờ [88], cây quyết định [89]. Tuy nhiên các bộ phân lớp đã nói ở phần trên đặc biệt là GMM và HMM được sử dụng nhiều để nhận dạng cảm xúc. Bảng 1.3 so sánh hiệu năng của các bộ phân lớp thông dụng nhận dạng cảm xúc tiếng nói.

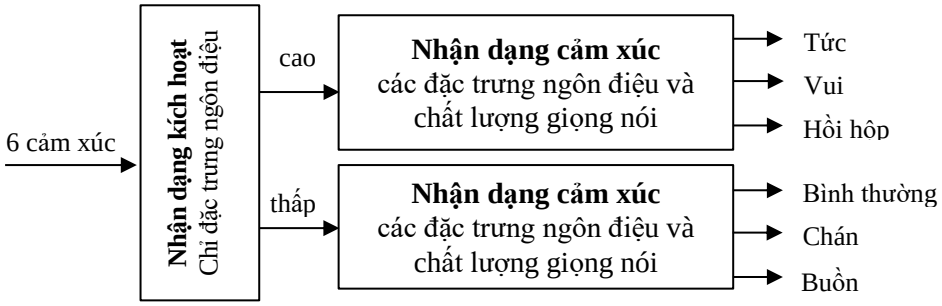
Bảng 1.3 *Kết quả nhận dạng cảm xúc của một số bộ phân lớp phổ biến (nguồn: [6])*

Bộ phân lớp	**HMM**	**GMM**	**ANN**	**SVM**
Độ chính xác phân lớp trung bình	75,5-78,5% ([81], [90])	74,83-81,94% ([72], [74])	51,19-52,82% ([29], [35]), 63-70% ([91])	75,45-81,29% ([30])

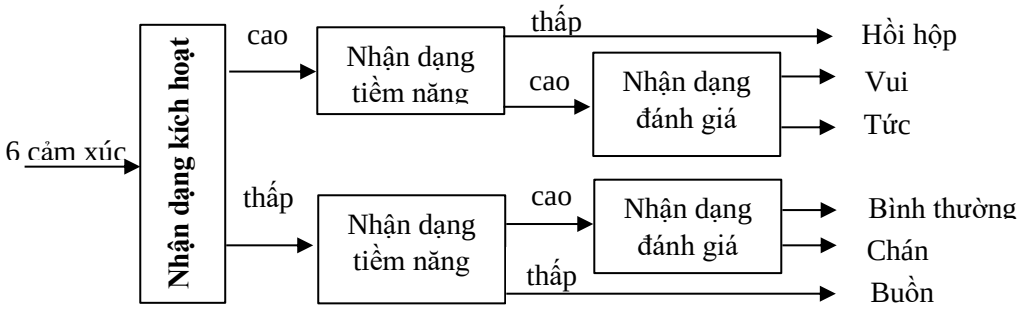
Bộ phân lớp	HMM	GMM	ANN	SVM
Thời gian huấn luyện trung bình	Nhỏ	Nhỏ nhất	Lớn	Lớn
Độ nhạy đối với khởi tạo mô hình	Nhạy	Nhạy	Nhạy	Không nhạy

Một cách lựa chọn khác là kết hợp sử dụng nhiều bộ phân lớp để nhận dạng cảm xúc tiếng nói [92], [93]. Có ba phương pháp kết hợp các bộ phân lớp [93], [94]: phân cấp, nối tiếp, và song song. Trong phương pháp phân cấp, các bộ phân lớp được sắp xếp theo một cấu trúc hình cây, trong đó tập các lớp ứng viên sẽ càng nhỏ nếu đi theo chiều sâu của cây. Tại các bộ phân lớp ở nút lá, chỉ còn một lớp sau khi quyết định. Với phương pháp nối tiếp, các bộ phân lớp được đặt trong một hàng đợi, mỗi bộ phân lớp làm giảm số lượng các lớp ứng viên cho bộ phân lớp kế tiếp [43], [95] [96]. Đối với phương pháp song song, tất cả các bộ phân lớp làm việc độc lập và một thuật toán trộn quyết định được áp dụng cho các kết quả đầu ra của chúng [97].

Phương pháp phân cấp đã được áp dụng trong [98] cho việc phân lớp các phát ngôn của ngữ liệu cảm xúc Berlin [76], trong đó mục tiêu chính là để cải thiện phân lớp cảm xúc độc lập người nói. Những cảm xúc được lựa chọn để phân loại gồm: tức, vui, buồn, chán, lo lắng và bình thường [99]. Hệ thống phân lớp có phân cấp xuất phát từ việc nghiên cứu tâm trạng cảm xúc trong đó các cảm xúc được thể hiện trong ba chiều: kích hoạt (activation) (được xem như arousal), tiềm năng (potency) (được xem như power) và đánh giá (evaluation) (được xem như pleasure). Hệ thống phân lớp phân cấp 2 tầng và 3 tầng đã được đề xuất trong [98], bộ phân lớp Naïve Bayesian đã được sử dụng cho tất cả các phân loại. Cả hai hệ thống được thể hiện trong Hình 1.10.



a) *Hệ thống phân cấp 2 tầng*



b) Hệ thống phân cấp 3 tầng

Hình 1.10 *Phân cấp cảm xúc 2 tầng 3 tầng theo Lugger và Yang (nguồn: [98])*

Trên Hình 1.10, các đặc trưng ngôn điệu bao gồm số liệu thống kê của cao độ, năng lượng, thời hạn, phương thức phát âm (articulation), tỷ lệ biến thiên qua trục không. Các đặc trưng chất lượng tiếng nói được tính toán là các tham số của phổ nguồn kích thích và được gọi là các gradient phổ [100]. Hệ thống phân cấp 2 tầng cho độ chính xác phân lớp là 83,5%, cao hơn khoảng 9% so với kết quả mà các tác giả đã thu được trong một nghiên cứu trước đó với cùng các đặc trưng chất lượng tiếng nói [101]. Đối với phân lớp 3 tầng, độ chính xác phân lớp tăng lên tới 88,8%. Trong hai nghiên cứu này, độ chính xác phân lớp dựa trên đánh giá chéo tỷ lệ 1:10 song véctơ dữ liệu đánh giá đã được sử dụng cho cả lựa chọn đặc trưng (trong đó dùng thuật toán SFFS) và nhận dạng.

Cả 3 phương pháp tiếp cận kết hợp các bộ phân lớp được áp dụng để nhận dạng cảm xúc tiếng nói trong [93]. Các tác giả đã dùng cùng một thiết lập thử nghiệm như trong nghiên cứu trước đó. Khi các véctơ đánh giá được dùng cho cả lựa chọn đặc trưng và nhận dạng, độ chính xác phân lớp với cách tiếp cận phân cấp, nối tiếp và song song được tăng lên lần lượt là 88,6%, 96,5% và 92,6%. Khi dữ liệu đánh giá và nhận dạng khác nhau, độ chính xác phân lớp giảm xuống đáng kể còn 58,6%, 59,7%, 61,8% và 70,1% tương ứng với bộ phân lớp đơn, phân lớp phân cấp, nối tiếp và song song.

Với nghiên cứu [102], các tác giả đã thực hiện nhận dạng cảm xúc cho ngôn ngữ Malayalam – một trong những ngôn ngữ phía nam của Ấn Độ. Hệ thống nhận dạng đã sử dụng các đặc trưng MFCC, năng lượng trong thời gian ngắn (STE) và cao độ. Hai mô hình ANN và SVM được sử dụng để nhận dạng các cảm xúc. Kết quả cho thấy trường hợp sử dụng phương pháp nhận dạng ANN có độ chính xác cao 88,4% còn trường hợp sử dụng phương pháp SVM là 78,2%.

Trên đây là các kết quả nhận dạng cảm xúc đã được thực hiện cho các ngôn ngữ không phải tiếng Việt. Ngoài các nghiên cứu trong [58], [57], [59], [60], [61], [62] đã được trình bày trong mục 1.2, có thể nêu lại các công trình nghiên cứu sau đây liên quan đến cảm xúc tiếng Việt. Các tác giả trong nghiên cứu [103] đã sử dụng SVM 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 [104], [105] có kết hợp với sinh viên Việt Nam xây dựng ngữ liệu cảm xúc tiếng Việt theo cách đóng kịch biểu lộ cảm xúc. Trong nghiên cứu [104], ngữ liệu bao gồm phát ngôn của 2 giọng nam và 2 giọng nữ là sinh viên Việt Nam với 6 cảm xúc vui, bình thường, buồn, ngạc nhiên, tức, sợ hãi. Các tham số đặc trưng được sử dụng là MFCC, năng lượng tín hiệu, cao độ và các formant F1, F2, F3. Mô hình nhận dạng được dùng là GMM cho thấy, tỷ lệ nhận dạng cao nhất là 96,5% đối với cảm xúc bình thường và thấp nhất là 76,5% đối với cảm xúc buồn. Còn trong nghiên cứu [105], ngữ liệu bao gồm các phát ngôn của 6 người nói với 20 câu và những cảm xúc như trong [104]. Với mô hình nhận dạng sử dụng SVM dựa trên tối ưu hóa của thuật toán Im-SFLA (Improved Shuffled Frog Leaping Algorithm), tỷ lệ nhận dạng tiếng Việt đạt 96,5% đối với cảm xúc bình thường và giảm xuống còn 84,1% cho cảm xúc bất ngờ.

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. Nhiều bộ phân lớp đã được thử nghiệm để nhận dạng cảm xúc như HMM, k-NN, GMM, ANN và SVM. Tuy nhiên, rất khó để quyết định bộ phân lớp nào thực hiện tốt nhất cho phân loại cảm xúc vì việc nhận dạng được thực hiện trên các ngữ liệu cảm xúc khác nhau với các thiết lập thử nghiệm khác nhau. 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.

Hiệu năng của các bộ nhận dạng hiện nay vẫn cần phải được cải thiện một cách đáng kể. Độ chính xác phân lớp trung bình của các hệ thống nhận dạng cảm xúc không phụ thuộc người nói đạt khoảng 80% trong hầu hết các kỹ thuật được đề xuất [6]. Trong một số trường hợp, chẳng hạn như [29], tỷ lệ đạt vào khoảng 50%. Đối với phân lớp phụ thuộc người nói, độ chính xác nhận dạng vượt trên 90% trong một vài nghiên cứu [30], [89], [106].

Hầu hết các nhóm nghiên cứu hiện nay tập trung nhiều vào nghiên cứu các đặc trưng tiếng nói và mối quan hệ của chúng với nội dung cảm xúc. Các đặc trưng mới cũng đã được phát triển như các đặc trưng dựa trên TEO. Ngoài ra, cũng có sử dụng các kỹ thuật lựa chọn đặc trưng khác nhau để tìm ra các đặc trưng tốt nhất cho nhận dạng cảm xúc. Tuy nhiên, kết quả thu được không nhất quán từ các nghiên cứu khác nhau. Lý do chính có thể là do thực tế chỉ từng ngữ liệu cảm xúc riêng biệt được xem xét trong mỗi nghiên cứu.

Cũng cần lưu ý rằng, phần lớn các kỹ thuật phân lớp hiện nay chưa xét đến mô hình cấu trúc thời gian của dữ liệu huấn luyện và cũng chỉ có một vài nghiên cứu xem xét áp dụng hệ thống phân lớp dùng nhiều bộ phân lớp (MCS) để nhận dạng cảm xúc [92], [93].

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.

Dựa trên các nghiên cứu đã trình bày ở trên, nội dung tiếp theo của luận án sẽ nghiên cứu về ngữ liệu cảm xúc tiếng Việt, các tham số tín hiệu tiếng nói có ảnh hưởng tới chất lượng nhận dạng cảm xúc tiếng Việt và sử dụng một số mô hình để thực hiện nhận dạng thử nghiệm cảm xúc tiếng Việt.

Các nội dung nghiên cứu chính của chương 1 đã được công bố trong các bài báo số 2 và 4 trong danh mục các công trình nghiên cứu của luận án:

2. *"Emotion recognition and corpus for Vietnamese emotion recognition*", Tạp chí Khoa học và Công nghệ, ĐHSPKT Hưng Yên, số 7, ISSN 2354-0575, trang 51-56.
4. "*So sánh hiệu năng một số phương pháp nhận dạng cảm xúc tiếng Việt nói*", Kỷ yếu Hội nghị khoa học công nghệ quốc gia lần thứ IX, Nghiên cứu cơ bản và Ứng dụng Công nghệ Thông tin, Cần Thơ, trang 656-662.

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

Xây dựng ngữ liệu và phân tích đánh giá bộ ngữ liệu là khâu quan trọng trong quá trình thử nghiệm nhận dạng cảm xúc. Chất lượng của bộ ngữ liệu có ảnh hưởng không nhỏ đến kết quả nhận dạng các cảm xúc. Một bộ ngữ liệu tốt được dùng cho nhận dạng với thuật toán đơn giản sẽ tốt hơn so với việc sử dụng thuật toán nhận dạng phức tạp song lại được thử nghiệm trên bộ ngữ liệu không phản ánh rõ các cảm xúc. Vì vậy, phần này sẽ trình bày các phương pháp để xây dựng một bộ ngữ liệu có cảm xúc đã được các nghiên cứu thảo luận. Từ đó, bộ ngữ liệu cảm xúc tiếng Việt cũng được xây dựng dùng cho thử nghiệm nghiên cứu của luận án. Các tham số đặc trưng của tín hiệu tiếng nói cũng là yếu tố quan trọng ảnh hưởng tới kết quả nhận dạng cảm xúc. Phần này cũng trình bày các kết quả đánh giá bộ ngữ liệu cảm xúc tiếng Việt và các tham số đặc trưng của tín hiệu tiếng nói được sử dụng.

2.1 Phương pháp xây dựng ngữ liệu cảm xúc

Một vấn đề quan trọng cần được xem xét trong việc đánh giá hệ thống tiếng nói có cảm xúc là chất lượng của ngữ liệu được sử dụng để phát triển và đánh giá hiệu năng của hệ thống [6]. Mục tiêu và phương pháp thu thập ngữ liệu tiếng nói thay đổi rất nhiều tùy theo các mục đích phát triển hệ thống tiếng nói sau đó.

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 (mô phỏng)
- 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

Ngữ liệu tiếng nói có cảm xúc dựa trên đóng kịch thường được thu thập từ các nghệ sĩ nhà hát hoặc đài phát thanh có kinh nghiệm. Các nghệ sĩ sẽ được thử thể hiện các cảm xúc khác nhau theo nội dung của câu nói trung tính về mặt ngôn ngữ sau đó thể hiện các cảm xúc này trong phòng ghi âm. Việc ghi âm được thực hiện trong các phiên khác nhau để tính đến sự biến đổi theo thời gian của cả mức độ biểu cảm lẫn cơ chế tạo tiếng nói của con người. Đây là phương pháp dễ dàng hơn và đáng tin cậy trong việc thu thập ngữ liệu tiếng nói có cảm xúc trong một phạm vi rộng. Hơn 60% các ngữ liệu thu thập cho nghiên cứu về tiếng nói là thuộc loại ngữ liệu này. Các cảm xúc thu thập theo cách mô phỏng có rất nhiều trong tự nhiên và thường được biểu hiện mạnh mẽ, kết hợp phần lớn các khía cạnh nổi bật đối với cảm xúc [107]. Nói chung, các cảm xúc đóng kịch có xu hướng được biểu thị mạnh mẽ hơn so với cảm xúc thực [6], [108]. Một ví dụ về ngữ liệu được xây dựng theo phương pháp này là ngữ liệu Berlin về tiếng nói có cảm xúc [109].

Ngữ liệu tiếng nói có cảm xúc suy diễn được thu thập bằng cách mô phỏng các tình huống cảm xúc nhân tạo mà người nói (chủ thể) không biết trước các cảm xúc cần biểu thị. Người nói bị lôi kéo vào những cảm xúc gắn với tình huống, trong đó các bối cảnh tình huống được tạo ra cố định thông qua đối thoại để suy diễn các cảm xúc khác nhau của chủ thể mà các chủ thể không biết trước các loại cảm xúc cần diễn đạt. Ngữ liệu suy diễn có thể tự nhiên hơn ngữ liệu đóng kịch, nhưng các chủ thể sẽ có thể không diễn đạt được cảm xúc một cách thích hợp nếu như họ biết trước đang được ghi âm. Đôi khi, ngữ liệu này được ghi âm bằng cách yêu cầu chủ thể trao đổi bằng lời với máy tính và máy tính do người điều khiển. Người điều khiển không biết gì về chủ thể đối thoại [110].

Không giống như cảm xúc đóng kịch, cảm xúc tự nhiên được biểu thị một cách thích hợp hơn. Nhưng đôi khi rất khó để nhận biết một cách rõ ràng các cảm xúc này và chúng được xem như là cảm xúc không rõ ràng. Ngữ liệu trong trường hợp này có thể được ghi âm từ các cuộc đối thoại qua tổng đài điện thoại, ghi âm từ buồng lái máy bay trong các tình huống bất thường, đối thoại giữa bệnh nhân và bác sĩ,…Tuy vây, rất khó để có được dải cảm xúc rộng trong các trường hợp đó. Việc phân loại cảm xúc sẽ mang tính chủ quan cao và thường không thống nhất và đây là loại ngữ liệu liên quan đến tính pháp lý như tính riêng tư và bản quyền.

Các tác giả trong [111] cũng phân biệt hai loại ngữ liệu đối với việc xây dựng ngữ liệu tiếng nói có cảm xúc. Đó là ngữ liệu nhân tạo (ngữ liệu đóng kịch) và ngữ liệu tự nhiên đời thực. Ngữ liệu có cảm xúc tự nhiên có thể được xây dựng hoặc là bằng cách ghi âm tình huống thực như đối thoại ở tổng đài điện thoại hoặc phỏng vấn truyền hình [112]. Hoặc một phương pháp rất phổ biến để xây dựng ngữ liệu là thu thập các ngữ liệu mà không cần có sự điều khiển nào, tức là ghi âm tiếng nói hàng ngày hoặc các chương trình mạn đàm. Ngay cả trong các tình huống thực như vậy, đặc tính tiếng nói có thể được điều khiển nhờ sự trợ giúp của cộng sự hoặc các diễn viên không chuyên tham gia vào tương tác. Hơn nữa, ngữ liệu có thể được xây dựng trong phòng thí nghiệm hoặc trong đời thực. Thách thức trong việc ghi âm đời thực như ghi âm chương trình mạn đàm hay tiếng nói hàng ngày là làm thế nào để có được ngữ liệu với chất lượng kỹ thuật tốt, đặc biệt không có nhiễu nền bên ngoài. Trong các ứng dụng đặc biệt, khi cảm xúc cần được nhận dạng ở môi trường nhiễu thì ngữ liệu có một số dạng nhiễu nền lại là hữu dụng.

Như vậy, 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. 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,

dễ đạt được số lượng 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, ngữ liệu sử dụng trong nghiên cứu của luận án cũng được xây dựng theo phương pháp này.

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

Theo thống kê trong [113], đã có nhiều ngữ liệu cảm xúc được xây dựng cho các ngôn ngữ khác nhau trên thế giới với số lượng bộ ngữ 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 ngữ liệu này, có một số ngữ liệu được xây dựng đồng thời cho 2, 3 hoặc 4 ngôn ngữ khác nhau. Bảng 2.1 dưới đây thống kê một số bộ ngữ liệu dùng cho các ngôn ngữ khác nhau.

Bảng 2.1 *Một số bộ ngữ liệu cảm xúc (nguồn: [6])*

STT	Ngữ liệu	Các thông số chung về bộ ngữ liệu
1	LDC Emotional Prosody Speech and Transcripts [114]	- Ngôn ngữ: tiếng Anh - Có 7 nghệ sĩ gồm các diễn viên chuyên nghiệp, 10 câu nói - Có 15 cảm xúc: Bình thường, hoảng loạn, lo lắng, nóng giận, tức giận, lạnh lùng, tuyệt vọng, buồn, hân hoan, vui, quan tâm, chán, xấu hổ, kiêu ngạo, khinh thường
2	Berlin emotional database [115]	- Được công bố và dùng miễn phí - Ngôn ngữ: tiếng Đức - Có 800 phát ngôn, 10 nghệ sĩ chuyên nghiệp - Có 7 cảm xúc: tức, vui, buồn, sợ hãi, ghê tởm, chán nản, bình thường
3	Danish emotional database [116]	- Được công bố song cần phí bản quyền - Ngôn ngữ: tiếng Đan Mạch - Có 4 nghệ sĩ không chuyên - Có 5 cảm xúc: tức, vui, buồn, ngạc nhiên, bình thường
4	Natural [117]	- Không công bố - Ngôn ngữ: tiếng Mandarin (Trung Quốc) lấy từ cuộc gọi trung tâm - Gồm 388 phát ngôn, 11 người nói - Có 2 cảm xúc: tức, bình thường
5	ESMBS [118]	- Không công bố - Ngôn ngữ: tiếng Mandarin (Trung Quốc) - Gồm 720 phát ngôn, 12 người nói là nghệ sĩ không chuyên - Có 6 cảm xúc: tức, buồn, ghê tởm, sợ hãi, ngạc nhiên
6	INTERFACE [119]	- Tiếng Anh (186 phát ngôn), tiếng Slovenian (190 phát ngôn), tiếng Tây Ban Nha (184 phát ngôn), tiếng Pháp (175 phát ngôn)

STT	Ngữ liệu	Các thông số chung về bộ ngữ liệu
		- Có 8 cảm xúc: tức, ghê tởm, sợ hãi, vui, ngạc nhiên, buồn, bình thường nói chậm, bình thường nói nhanh - Người nói là các nghệ sĩ
7	KISMET [120]	- Không công bố - Ngôn ngữ: tiếng Anh Mỹ - Có 1002 phát ngôn - Có 7 cảm xúc: bằng lòng (tán thành), niềm nở, ngăn cấm, dễ chịu, tán thành, thu hút, bình thường - Có 3 người nói là nghệ sĩ không chuyên (nữ)
8	BabyEars [121]	- Không công bố - Ngôn ngữ: tiếng Anh - Có 509 phát ngôn - Có 12 nghệ sĩ (6 nam + 6 nữ) - Có 3 cảm xúc: tán thành, thu hút, ngăn cấm
9	MPEG-4 [122]	- Không công bố - Ngôn ngữ: tiếng Anh - Có 2440 phát ngôn, 35 người nói U.S. American movies - Có 7 cảm xúc: vui, tức, ghê tởm, sợ hãi, buồn, ngạc nhiên, bình thường
10	FERMUS III [123]	- Công bố nhưng cần phí bản quyền - Ngôn ngữ: tiếng Đức, tiếng Anh - Có 2829 phát ngôn, 13 nghệ sĩ - Có 7 cảm xúc: vui, tức, ghê tởm, sợ hãi, buồn, ngạc nhiên, bình thường
11	KES [124]	- Không công bố - Ngôn ngữ: tiếng Hàn - Có 5400 phát ngôn, 10 nghệ sĩ không chuyên - Có 4 cảm xúc: bình thường, vui, buồn, tức
12	CLDC [125]	- Không công bố - Ngôn ngữ: tiếng Trung - Có 1200 phát ngôn, 4 nghệ sĩ không chuyên - Có 6 cảm xúc: vui, tức, ngạc nhiên, sợ hãi, buồn, bình thường
13	Amir et al. [126]	- Không công bố - Các nghệ sĩ không chuyên gồm 60 nghệ sĩ Hebrew và 1 nghệ sĩ Russian - Có 6 cảm xúc: tức, ghê tởm, sợ hãi, vui, bình thường, buồn
14	Pereira [127]	- Không công bố - Ngôn ngữ: tiếng Anh - Có 5 cảm xúc: nóng giận, giận lạnh lùng, vui, buồn, bình thường - Có 8 phát ngôn, 2 nghệ sĩ không chuyên

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. BKEmo được xây dựng tại Đại học Bách khoa Hà Nội [128] và có thể coi đây là bộ ngữ liệu có số lượng người nói, câu nói, số lượng phát ngôn khá đồng đều và có thể dùng cho các nghiên cứu về nhận dạng cảm xúc tiếng Việt.

Bộ ngữ liệu BKEmo đã được thu âm từ giọng của 56 người gồm 28 nam và 28 nữ 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, tránh tình trạng ngữ liệu gồm rất nhiều cách biểu lộ khác nhau cho cùng một cảm xúc nhưng mỗi loại lại chỉ có vài câu gây khó khăn trong việc tìm quy luật. Tiếng nói được ghi âm với tần số lấy mẫu 16kHz, 16bit/mẫu. Mỗi cảm xúc được phát âm 4 lần. Việc ghi âm được tiến hành trong studio dùng cho lồng tiếng phim. Dung lượng ghi âm khoảng 4GB.

Kịch bản thu âm gồm 55 câu và cả 55 câu được phát âm cho 4 cảm xúc: vui, buồn, tức và bình thường. Các câu có chứa từ cảm thán hoặc không chứa từ cảm thán sẽ cần được biểu lộ cùng một kiểu cảm xúc trong số bốn cảm xúc.

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.

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. Để tránh sự không đồng đều về số lượng người nói, câu nói, giới tính cho mỗi cảm xúc như đã trình bày trong phần 2.2, luận án đã lựa chọn các file ngữ liệu tiếng nói tốt nhất của 8 giọng nam và 8 giọng nữ với bốn cảm xúc vui, buồn, tức giận, bình thường. Các file cảm xúc được lựa chọn với 22 câu có nội dung khác nhau, mỗi câu được nói 4 lần ở các thời điểm 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. Tổng số file là $16 \times 4 \times 22 \times 4 = 5632$ file. Các file này đã được nghe lại để loại bỏ những file chưa thể hiện đúng cảm xúc hoặc kiểu diễn đạt cho một cảm xúc không đồng đều nên số file còn lại là 5584 file. Trong đó, số lượng file cảm xúc cho mỗi giọng nam và nữ là 2792 file. Mỗi cảm xúc có 1396 file.

Nói chung để thực hiện các thử nghiệm, bộ ngữ liệu tiếng nói có thể được phân chia theo độc lập hoặc phụ thuộc người nói, độc lập hoặc phụ thuộc nội dung. Như vậy chỉ có bốn khả năng để lựa chọn ngữ liệu cho thử nghiệm: ngữ liệu phụ thuộc cả người nói lẫn nội dung, ngữ liệu phụ thuộc người nói và độc lập về nội dung, ngữ liệu độc lập người nói và phụ thuộc nội dung, ngữ liệu độc lập cả người nói và nội dung. Vì vậy luận án đã chia ngữ liệu theo cách này để đánh giá chi tiết hơn về kết quả nhận dạng cảm xúc đối với người nói và đối với nội dung.

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). Mỗi tập ngữ liệu lại được chia thành 2 phần: một nửa dùng cho huấn luyện và nửa còn lại dùng cho nhận dạng.

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

Sau đây, bốn tập ngữ liệu nêu 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.

Với ngữ liệu phụ thuộc cả người nói và nội dung T1, tập ngữ liệu huấn luyện và nhận dạng đều có nội dung câu nói và giọng người nói như nhau tuy nhiên thời điểm phát ngôn là khác nhau cho cùng một câu nói. Vì vậy, T1 vẫn hoàn toàn có ý nghĩa thực tế do tính đa dạng của tín hiệu tiếng nói. Tính đa dạng của tín hiệu tiếng nói thể hiện ở chỗ cùng một người nói cùng một âm song các tín hiệu tiếng nói cho âm này không hoàn toàn như nhau khi phát âm ở các thời điểm khác nhau [129]. Như đã trình bày ở trên, đối với ngữ liệu BKEmo, mỗi câu được người nói phát ngôn 4 lần ở 4 thời điểm khác nhau. Ngữ liệu phụ thuộc người nói nhưng độc lập nội dung có giọng người nói như nhau nhưng số lượng câu nói được chia đôi, nội dung của 11 câu dùng để huấn luyện khác với nội dung 11 câu còn lại dùng để nhận dạng. Đối với ngữ liệu độc lập người nói, phụ thuộc nội dung, số lượng người nói được chia đôi và người nói dùng cho huấn luyện khác với người nói dùng cho nhận dạng. Còn đối với ngữ liệu độc lập người nói và độc lập nội dung, cả nội dung câu nói lẫn người nói dùng cho huấn luyện đều khác so với nội dung câu nói và người nói dùng cho nhận dạng.

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

Trong nghiên cứu [25], các tác giả đã đưa ra các đặc tính sử dụng để chú giải và nhận dạng cảm xúc bao gồm các đặc tính âm học và các đặc tính ngôn ngữ. Các đặc tính âm học là các đặc trưng của nguồn âm, tuyến âm và ngôn điệu. Các đặc tính ngôn ngữ là các đặc trưng về các dấu hiệu ngôn ngữ nhận diện cảm xúc, ví dụ từ vựng, ngữ nghĩa. Các đặc trưng này đã được xác định để tham gia vào nhận dạng cảm xúc trong các hệ thống nhận dạng cảm xúc tiếng nói và đã được thực hiện trong các công trình nghiên cứu cho nhiều ngôn ngữ như tiếng Anh, tiếng Đức, tiếng Trung… Nội dung sau đây của luận án sẽ trình bày về các đặc trưng của nguồn âm, tuyến âm và các đặc trưng ngôn điệu xét theo phương diện tín hiệu tiếng nói có liên quan đến biểu thị cảm xúc.

2.4.1 Đặc trưng của nguồn âm và tuyến âm

Trong quá trình phát âm, dao động của dây thanh tạo ra dãy xung hầu như tuần hoàn kích thích cho tuyến âm. Bộ lọc đảo đối với tín hiệu tiếng nói có thể loại bỏ đi thông tin về tuyến âm và cho ra sai số tiên đoán tuyến tính LP. Nói chung, sai số tiên đoán tuyến tính được coi là xấp xỉ của tín hiệu nguồn âm. Phân tích đảo LP sẽ loại bỏ các mối tương quan bậc thấp đối với các mẫu tiếng nói và giữ lại các mối tương quan có bậc cao hơn [130].

Các mối tương quan bậc thấp là liên quan giữa các mẫu tiếng nói kề nhau, còn các mối tương quan bậc cao liên quan đến các mẫu cách xa nhau. Do sự hiện diện của các mối tương quan bậc cao, tín hiệu sai số LP được coi như là tín hiệu ngẫu nhiên và không chứa thông tin nào ngoài tần số cơ bản của tiếng nói. Các đặc tính tìm thấy từ sai số LP có thể chứa những thông tin hữu ích, và nó có thể được sử dụng để phát triển các hệ thống xử lý tiếng nói khác nhau.

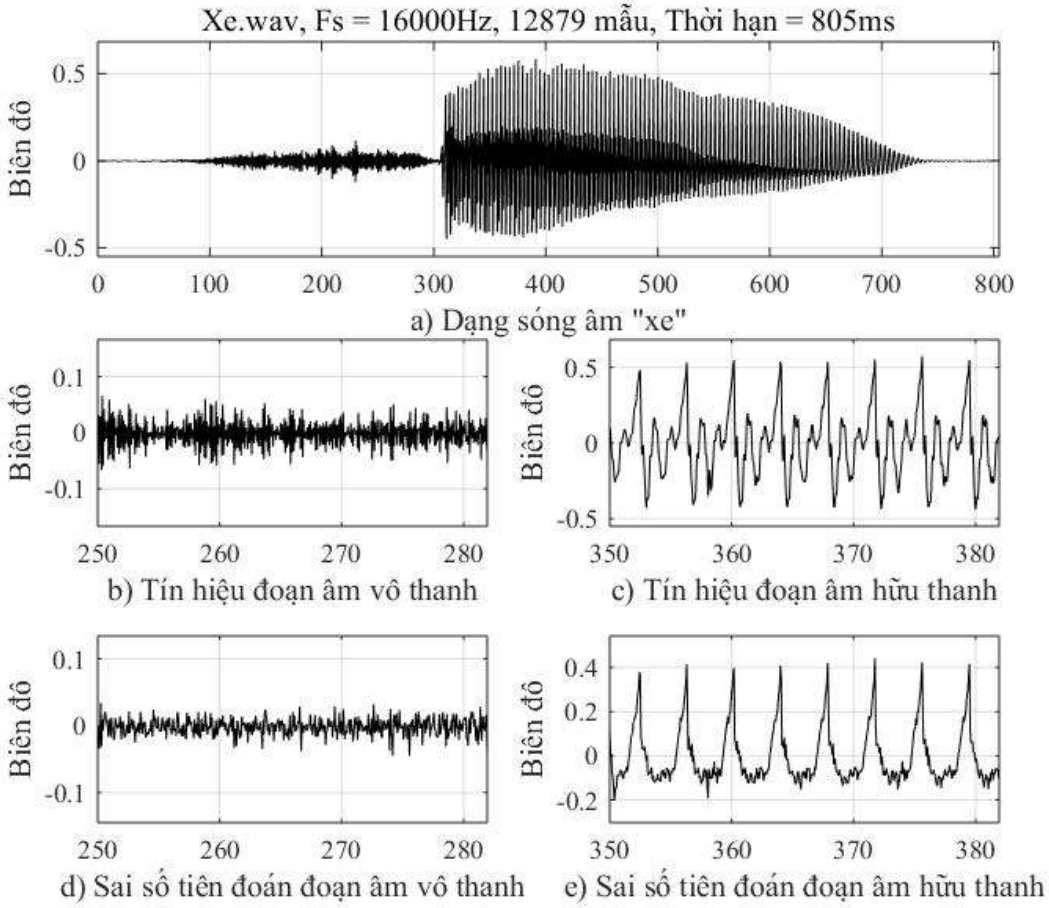
Ngoài những đặc tính của tuyến âm, những đặc tính của nguồn âm sẽ bổ sung thêm thông tin vì đặc tính của tuyến âm chỉ biểu diễn cho các mối tương quan bậc thấp giữa các mẫu tiếng nói. Các đặc tính được sử dụng để biểu diễn cho hoạt động của thanh môn, mà chủ yếu là dao động của dây thanh được xem là các đặc tính của nguồn âm hoặc nguồn kích thích. Trên Hình 2.1 là ví dụ các đoạn tín hiệu của âm vô thanh, hữu thanh và tín hiệu sai số LP tương ứng của âm "*xe*" trong tiếng Việt khi phân tích LPC. Các sai số tiên đoán LP của đoạn hữu thanh là một chuỗi các dao động ứng với khoảng đóng của thanh môn, nơi mà sai số tiên đoán là lớn nhất.

Các tương quan bậc cao có trong sai số LP có thể chứa những thông tin cảm xúc cùng với các tính năng khác của tiếng nói. Chuỗi các dao động gây ra do dây thanh rung đóng vai trò kích thích cho tuyến âm. Tuyến âm có thể được coi là một chuỗi các ống có tiết diện thay đổi. Chuỗi các cấu hình tuyến âm trong quá trình tạo các âm khác nhau được xem như đặc tính của tuyến âm đối với các đơn vị âm.

Trong quá trình phát âm, tuyến âm hoạt động như bộ cộng hưởng và tăng cường một số thành phần tần số nào đó tùy thuộc vào cấu hình của khoang miệng. Các formant là cộng hưởng của tuyến âm tại một thời điểm nhất định. Formant được đặc

trưng bởi dải thông và biên độ cộng hưởng, các tham số này là duy nhất cho mỗi đơn vị âm. Chuỗi các cấu hình tuyến âm cũng vì vậy mà mang thông tin về cảm xúc, cùng với các thông tin khác liên quan đến các đơn vị âm.

Những đặc tính này được thấy rõ trong Hình 2.1: a) là dạng sóng tín hiệu tiếng nói của âm "*xe*", b) là tín hiệu phần âm vô thanh của tín hiệu tiếng nói, c) là tín hiệu phần âm hữu thanh, d) là tín hiệu sai số tiên đoán của phần âm vô thanh còn e) là tín hiệu sai số tiên đoán của phần âm hữu thanh của tín hiệu tiếng nói tương ứng. Đối với các phân tích trong miền tần số, tín hiệu tiếng nói thường được chia thành các khung với độ rộng 20-30ms, độ dịch của mỗi khung là 10ms.



Hình 2.1 *Các đoạn tín hiệu của âm vô thanh, hữu thanh và tín hiệu sai số LP tương ứng*

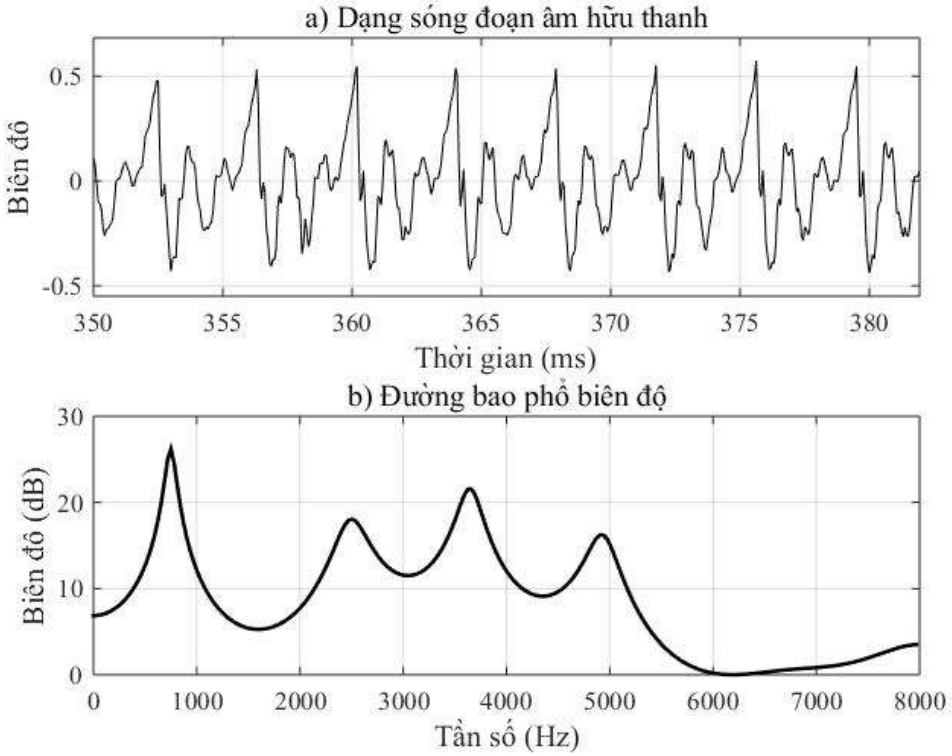
Từ biến đổi Fourier sẽ được phổ của khung tiếng nói. Xuất phát từ phổ biên độ, các đặc tính như: các hệ số cepstrum tiên đoán tuyến tính (LPCC), các hệ số cepstrum theo thang tần số mel (MFCC), các hệ số tiên đoán tuyến tính cảm nhận (PLPC) và các dẫn xuất của chúng được tính toán để biểu diễn cho các đặc tính của tuyến âm [131], [132]. Nhìn chung, các đặc tính được trích chọn từ tuyến âm bao gồm các đặc trưng phổ, các đặc tính mức đoạn hoặc mức hệ thống.

MFCC, LPCC, PLPC là những đặc tính được biết rộng rãi trong miền tần số và được dùng trong các nghiên cứu khác nhau về tiếng nói. Từ các đặc tính này, có thể nhận được thông tin về cấu hình của tuyến âm trong quá trình tạo ra các âm khác

nhau. Các cực đại trong phổ được gọi là các formant [131], [132]. Các formant cho biết các tần số cộng hưởng của tuyến âm. Hình 2.2 là một đoạn tiếng nói và phổ LP tương ứng của âm hữu thanh.

Sau khi loại bỏ các đặc tính của tuyến âm từ tín hiệu tiếng nói sẽ nhận được tín hiệu nguồn âm [133]. Để đạt được điều đó, đầu tiên phải tiên đoán các thông tin của tuyến âm bằng cách dùng các hệ số của bộ lọc (các hệ số tiên đoán tuyến tính LPC) từ tín hiệu tiếng nói, sau đó tách lấy nguồn âm bằng bộ lọc đảo.

Tín hiệu nhận được sau bộ lọc đảo là sai số tiên đoán và chứa phần lớn thông tin về nguồn âm. Việc phân tích các khung tín hiệu tiếng nói có mục đích nghiên cứu các đặc tính của tín hiệu thanh môn, các thời điểm đóng-mở thanh môn, cường độ kích thích…



Hình 2.2 *Phân tích trong miền tần số để có phổ tiếng nói*

a) Đoạn âm hữu thanh, b) Đường bao phổ biên độ tương ứng

Các đặc tính hoạt động của thanh môn riêng cho cảm xúc có thể được đánh giá bằng cách dùng các đặc điểm của nguồn âm. Tín hiệu sai số tiên đoán và thông lượng không khí qua thanh môn được xem như có tương quan với nhau. Trong các nghiên cứu về tiếng nói, có rất ít thử nghiệm khai thác thông tin nguồn âm cho các hướng nghiên cứu khác [133]. Lý do có thể là:

- Đặc trưng phổ có tính phổ biến
- Tín hiệu nguồn âm có được từ phân tích LP phần lớn được xem như là sai số tiên đoán hoặc là thành phần không thể đoán trước được của tín hiệu đã được tiên đoán [134].

- Sai số tiên đoán chủ yếu chứa các tương quan bậc cao hơn và việc thu thập các tương quan bậc cao hơn này thường chưa được biết rõ [135].

Việc tham số hóa tín hiệu sai số tiên đoán là khó khăn, tuy nhiên tín hiệu này chứa các thông tin giá trị vì nó là kích thích ban đầu của tuyến âm trong quá trình tạo tiếng nói. Tín hiệu sai số LP chủ yếu chứa các tương quan bậc cao trong số các mẫu của nó, vì các tương quan bậc nhất và bậc hai đã được loại bỏ trong quá trình phân tích LP [136]. Các tương quan bậc cao này có thể được thu thập ở mức độ nào đó bằng cách sử dụng các đặc điểm như cường độ kích thích, đặc tính của thông lượng không khí qua thanh môn, dạng xung thanh môn, các đặc tính đóng-mở của thanh môn,… Bảng 2.3 thống kê một số nghiên cứu sử dụng thông tin nguồn kích thích trong các ứng dụng xử lý tiếng nói khác nhau [133].

Bảng 2.3 *Sử dụng thông tin của nguồn kích thích cho các nghiên cứu khác nhau về tiếng nói (nguồn: [133])*

STT	Các đặc trưng	Mục đích và ứng dụng	Tài liệu tham khảo
1	Năng lượng sai số tiên đoán	Nhận dạng nguyên âm và người nói	H. Wakita (1976) [138]
2	Sai số tiên đoán	Các thể hiện của kích thích quan trọng được xác định	Rao K. S và cộng sự (2007) [141]
3	Các quan hệ bậc cao của các mẫu sai số tiên đoán	Phân loại dữ liệu âm thanh	Bajpai A. Và Yegnanarayana B. (2004) [140]
4	Sai số tiên đoán	Tăng cường chất lượng tiếng nói trong môi trường có nhiều người nói	Yegnanarayana B và cộng sự (2009) [142]
5	Sai số tiên đoán	Đặc trưng cho độ to, hiệu ứng Lombard, tốc độ nói, đoạn có tiếng cười	G. Bapineedu và cộng sự (2009) [143]
6	Tín hiệu kích thích thanh môn	Phân tích quan hệ giữa các trạng thái cảm xúc của người nói và hoạt động của thanh môn Phân tích các rối loạn liên quan đến cảm xúc	Cummings K. E. And Clements M. A. (1995) [144]
7	Tín hiệu nguồn kích thích	Phân biệt các cảm xúc trong tiếng nói liên tục	Zhen-Hua Ling, Yu Hu, Ren-Hua Wang (2005) [145]

Các nghiên cứu hiện tại dựa trên các đặc trưng nguồn âm đã chỉ rõ rằng, thông tin nguồn âm chứa toàn bộ các đặc trưng của tiếng nói như: nội dung truyền tải, người nói, ngôn ngữ và các thông tin riêng về cảm xúc [133]. Thông tin về cao độ được trích rút từ sai số tiên đoán đã được sử dụng thành công để nhận dạng người nói [137], năng lượng sai số tiên đoán đã được dùng để nhận dạng người nói và nguyên âm trong [138]. Các đặc trưng cepstrum được dẫn xuất từ sai số tiên đoán đã được dùng

để thu thập thông tin riêng về người nói [139]. Kết hợp các đặc điểm dẫn xuất từ các sai số tiên đoán và cepstrum đã được dùng để tối thiểu hóa tỷ lệ lỗi trong trường hợp nhận dạng người nói. Các tương quan bậc cao trong số các mẫu của sai số tiên đoán cũng được dùng để phân loại các thể loại âm thanh liên quan đến: thể thao, tin tức, phim hoạt hình, âm nhạc trong môi trường sạch và môi trường có nhiễu [140].

Các thời điểm kích thích quan trọng nhận được từ tín hiệu sai số tiên đoán trong quá trình tạo âm hữu thanh được dùng để xác định độ trễ tương đối giữa các đoạn tiếng nói của những người nói khác nhau trong một môi trường nhiều người nói, và sau đó các đại lượng này được dùng để cải thiện tiếng nói của từng người nói riêng rẽ [142]. Các thuộc tính liên quan đến thời điểm dây thanh đóng được khai thác trong [146] để cải thiện tiếng nói có tiếng vọng. Các tham số được trích rút tại thời điểm dây thanh đóng được khai thác để phân tích cường độ, hiệu ứng Lombard, tốc độ nói và phát hiện đoạn có tiếng cười trong tiếng nói.

Ngoài các đặc trưng về nguồn âm được sử dụng trong xử lý tiếng nói, các đặc trưng về tuyến âm cũng được sử dụng trong nhiều nghiên cứu. Nói chung, khung tín hiệu tiếng nói có độ rộng 20-30ms thường được dùng để trích rút các đặc trưng của tuyến âm. Các đặc tính tuyến âm được phản ánh rất rõ khi phân tích tín hiệu tiếng nói trong miền tần số. Biến đổi Fourier ngược của khung tín hiệu tiếng nói sẽ cho phổ trong thời gian ngắn. Các đặc tính như formant, dải thông, phổ năng lượng và đường bao phổ có thể nhận được từ phổ. Biến đổi Fourier của logarit của biên độ phổ sẽ cho cepstrum của khung tiếng nói

Các hệ số MFCC và LPCC là các đặc trưng phổ biến được dẫn xuất từ miền cepstrum đặc trưng cho thông tin tuyến âm. Những đặc điểm này của tuyến âm cũng được biết đến như là các đặc tính phổ, đặc tính phân đoạn hoặc đặc tính hệ thống. Nói chung, các đặc tính 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.

Các 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. Một số công trình nghiên cứu quan trọng về nhận dạng cảm xúc sử dụng các đặc tính phổ sẽ được trích dẫn sau đây. Các đặc tính MFCC được dùng để phân biệt các thông tin về tiếng nói và âm nhạc [147]. Nghiên cứu cho thấy các đặc tính MFCC 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, RASTA PLPC 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, chán, trung tính, vui, buồn trong tiếng phổ thông Trung Quốc [106], [148].

Các hệ số công suất theo logarit tần số (LFPC) được dùng để biểu diễn các thông tin riêng về cảm xúc nhằm phân loại 6 cảm xúc trong [12]. Với nghiên cứu [12], mô hình Markov ẩn có 4 tầng được dùng như bộ phân lớp để thực hiện phân loại cảm xúc. Hiệu năng của các tham số LFPC có thể tương đương với các đặc tính MFCC,

LPCC thông thường và LFPC có thể có hiệu năng tốt hơn [12], [149]. Các đặc tính MFCC được trích rút từ các thành phần tần số thấp (20-300Hz) của tín hiệu tiếng nói được sử dụng để mô hình hóa biến thiên $F0$. Đó chính là các đặc tính MFCC tần thấp và được dùng để nhận dạng cảm xúc trong ngữ liệu cảm xúc tiếng Anh và tiếng Thụy Điển. Bảng 2.4 thống kê các nghiên cứu sử dụng thông tin của tuyến âm cho các xử lý về tiếng nói [133].

Các đặc tính MFCC tần thấp cho thấy có kết quả tốt hơn so với các đặc tính cao độ trong trường hợp nhận dạng cảm xúc [150]. Các hệ số MFCC được tính thông qua 3 lớp âm vị: nguyên âm có trọng âm, nguyên âm không có trọng âm và phụ âm được dùng để nhận dạng cảm xúc độc lập với người nói. Các đặc tính này được xem như là lớp các đặc tính phổ. Độ chính xác phân loại sẽ cao hơn khi dùng lớp các đặc tính phổ so với dùng lớp các đặc trưng ngôn điệu hoặc đặc trưng phổ phát âm. Việc kết hợp lớp các đặc trưng phổ với các đặc trưng ngôn điệu sẽ cải thiện hiệu năng nhận dạng cảm xúc.

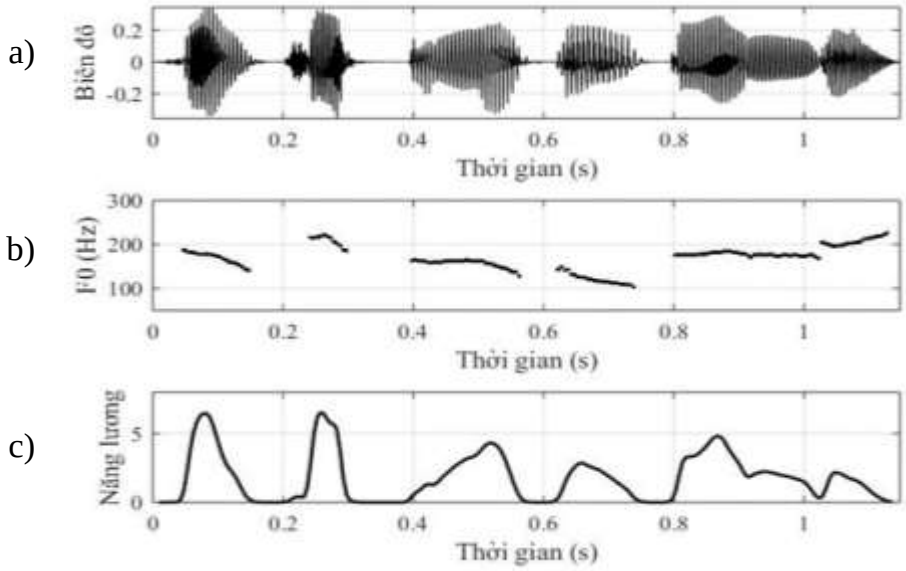
Bảng 2.4 *Sử dụng thông tin của tuyến âm cho các nghiên cứu khác nhau về xử lý tiếng nói (nguồn: [133])*

STT	Các đặc trưng	Mục đích và ứng dụng	Tài liệu tham khảo
1	Các đặc trưng MFCC	Phân biệt tiếng nói và âm nhạc: Các MFCC bậc cao chứa nhiều hơn thông tin liên quan đến âm nhạc và MFCC bậc thấp chứa nhiều hơn thông tin liên quan đến tiếng nói	Mubarak O. M. Và cộng sự (2005) [147]
2	MFCC, LPCC, RASTA, các hệ số PLPC, các hệ số công suất trong miền logarit tần số	Phân loại 4 cảm xúc trong tiếng Trung Quốc: tức, vui, trung tính, buồn	Pao T. L. Và cộng sự (2005, 2007) [106], [148]
3	Tổ hợp các MFCC và các đặc tính MFCC bậc thấp	Phân loại cảm xúc cho tiếng Thụy Điển và tiếng Anh	Neiberg D. Và cộng sự (2006) [150]
4	Các đặc tính MFCC của phụ âm, nguyên âm có và không có trọng âm	Phân loại cảm xúc cho tiếng Anh,	Bitouk D., Verma R. And Nenkova A. (2010) [151]
5	Các đặc tính phổ dùng biến đổi Fourier và Chirp	Mô hình hóa trạng thái cảm xúc khi căng thẳng	Sigmund M. (2007) [152]

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. Chúng biểu diễn cho chất lượng tổng thể tiếng nói như nhịp điệu, trọng âm, ngữ điệu, âm sắc, cảm xúc… Chu kỳ cơ bản, thời hạn, năng lượng và các dẫn xuất tương ứng đã được sử dụng rộng rãi để biểu diễn cho các đặc trưng của ngôn điệu [153], [154].

Các đặc trưng ngôn điệu liên quan đến cảm xúc là các đặc trưng có liên quan đến các tham số siêu đoạn tính (supra-segmental) hoặc thời gian dài. Con người cảm nhận được các cảm xúc có trong tiếng nói bằng cách khai thác các đặc trưng ngôn điệu và ở đây các đặc trưng này được khai thác để phân loại cảm xúc. Hình 2.3 là đoạn tín hiệu tiếng nói và các đặc trưng ngôn điệu tương ứng. Từ các thảo luận ở trên, có thể thấy tầm quan trọng của việc khai thác nguồn âm, tuyến âm, các đặc trưng ngôn điệu để nhận được thông tin riêng về cảm xúc.



Hình 2.3 *Các đặc trưng ngôn điệu của tiếng nói*
a) Tín hiệu tiếng nói có phân đoạn âm tiết, b) Biến thiên F0, c) Biến thiên năng lượng

Nội dung sau đây sẽ trình bày khái quát về việc khai thác các đặc trưng ngôn điệu đã được thực hiện trong các hệ thống xử lý tiếng nói cho một số ngôn ngữ khác nhau trên thế giới [2].

Trong khi nói, có những đại lượng đặc trưng như: thời hạn, ngữ điệu, cường độ cho các chuỗi âm khác nhau. Sự hợp thành các ràng buộc về ngôn điệu như vậy (thời hạn, thanh điệu và cường độ) làm cho tiếng nói tự nhiên. Có thể phát hiện dễ dàng việc thiếu ngôn điệu trong tiếng nói. Ngôn điệu có thể được xem như các đặc điểm của tiếng nói đi cùng với các đơn vị âm lớn hơn như âm tiết, từ, đoạn câu và câu. Từ đó, ngôn điệu thường được xem như thông tin siêu đoạn tính.

Ngôn điệu biểu lộ cho cấu trúc của luồng tiếng nói. Ngôn điệu được biểu diễn về mặt âm học bởi thời hạn, ngữ điệu (đường bao *F0*) và năng lượng. Các đại lượng này thường biểu thị các thuộc tính của tiếng nói cảm thụ như: ngữ điệu, năng lượng. Ngữ

điệu và năng lượng thường được con người sử dụng để nghiên cứu xử lý tiếng nói trong đó có nhận dạng cảm xúc [155], [156]. Tính biểu cảm của con người có thể được thu thập thông qua các đặc trưng ngôn điệu. Ngôn điệu có thể được phân biệt theo 4 mức biểu hiện chính [156]. Đó là mức ý định về mặt ngôn ngữ, mức cấu âm, mức thể hiện về mặt âm học, mức cảm thụ. Đối với mức ngôn ngữ, ngôn điệu có quan hệ với các yếu tố ngôn ngữ khác nhau của phát ngôn để thể hiện tính tự nhiên cần có.

Chẳng hạn, sự phân biệt về mặt ngôn ngữ có thể được truyền thông qua sự phân biệt giữa câu hỏi và diễn đạt thông thường, hoặc sự nhấn mạnh ngữ nghĩa đối với một phần tử nào đó. Ở mức cấu âm, ngôn điệu được biểu hiện về mặt vật lý như là một chuỗi các chuyển động của bộ phận cấu âm. Như vậy, các biểu hiện của ngôn điệu bao gồm chủ yếu các biến thiên về mặt biên độ của các chuyển động cấu âm cũng như biến thiên áp suất không khí. Hoạt động của cơ trong hệ thống hô hấp cũng như dọc theo tuyến âm dẫn tới bức xạ sóng âm.

Thể hiện về mặt âm học của ngôn điệu có thể được nhận thấy và lượng tử hóa bằng cách phân tích các tham số âm học như tần số cơ bản $F0$, cường độ và thời hạn. Chẳng hạn, các âm tiết có trọng âm sẽ 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, từ ngôn điệu và thông qua 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.

Trong quá trình cảm thụ, ngôn điệu có thể được biểu thị tùy thuộc vào trải nghiệm chủ quan của người nghe, như khoảng dừng, độ dài, âm điệu và độ to của tiếng nói cảm thụ được. Rất khó để xử lý hoặc phân tích ngôn điệu thông qua cơ chế tạo tiếng nói và cảm thụ tiếng nói.

Các đặc tính như giá trị cực tiểu, cực đại, trung bình, phương sai, phạm vi và độ lệch chuẩn của năng lượng cũng như các đặc tính tương tự của tần số cơ bản đã được dùng như là nguồn thông tin quan trọng về ngôn điệu để phân biệt các cảm xúc [107], [157]. Một số nghiên cứu của [13], [157] cũng đã thử nghiệm đo lường độ dốc của đường bao $F0$ khi lên xuống, tốc độ cấu âm, số lượng và thời hạn của khoảng dừng để đặc trưng cho cảm xúc.

Các đặc trưng ngôn điệu được trích rút từ các đơn vị ngôn ngữ nhỏ hơn như các âm tiết và ở mức phụ âm và nguyên âm cũng được dùng để phân tích cảm xúc [157]. Tầm quan trọng của đường bao ngôn điệu dẫn tới các ngữ cảnh có cảm xúc khác nhau đã được nghiên cứu trong [158], [159]. Nghiên cứu [160] cho thấy, 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, buồn và vui. Hiệu năng nhận dạng cảm xúc trung bình đạt được khoảng 55% khi sử dụng các phân tích các yếu tố phân biệt.

Trong nghiên cứu [161], dãy các đặc trưng 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 trưng $F0$ đã nêu cùng với bộ phân lớp láng giềng k gần nhất.

Các đặc trưng siêu đoạn tính trong thời gian ngắn như tần số cơ bản, năng lượng, vị trí formant và dải tần tương ứng, dải động của $F0$, năng lượng và đường bao formant, tốc độ nói đã được sử dụng để phân tích các cảm xúc trong [162]. Quan hệ phức hợp giữa tần số cơ bản, thời hạn và các tham số năng lượng đã được khai thác để phát hiện cảm xúc [163]. Bảng 2.5 cho thấy một số các công trình nghiên cứu quan trọng về nhận dạng cảm xúc có sử dụng các đặc trưng ngôn điệu [133].

Bảng 2.5 *Sử dụng thông tin về ngôn điệu cho các nghiên cứu khác nhau về tiếng nói (nguồn: [133])*

STT	Các đặc trưng	Mục đích và ứng dụng	Tài liệu tham khảo
1	Khởi đầu với việc sử dụng 86 đặc trưng ngôn điệu, sau đó 6 đặc trưng tốt nhất đã được chọn	Định danh các cảm xúc cho tiếng Basque. Dùng GMM và hiệu năng đạt được là 92%	Luengo I., Navas E., Hernáez I., and Sánchez J. (2005) [167]
2	Véctơ đặc trưng ngôn điệu có 3 chiều bao gồm *F0*, năng lượng và thời hạn	Phân loại 7 cảm xúc cho tiếng Đức. Kết quả nhận dạng cảm xúc đạt khoảng 51% cho trường hợp không phụ thuộc người nói, dùng mạng nơron	Iliou T. And Anagnostopoulos C.-N. (2009) [168]
3	Đặc trưng *F0* và công suất được trích rút theo từng khung, âm tiết và từ	Nhận dạng cảm xúc cho tiếng Trung Quốc. Tổ hợp các đặc trưng của khung, âm tiết và từ cho kết quả nhận dạng cảm xúc 90%	Kao Y. Hao and Lee L. Shan [169]
4	Các đặc trưng dựa trên thời hạn, năng lượng và *F0*	Nhận dạng cảm xúc cho tiếng Trung Quốc. Sử dụng mạng nơron. Dữ liệu cho nhiều người nói và đa ngôn ngữ	Zhu A. And Luo Q. (2007) [170]
5	8 đặc trưng ngôn điệu và đặc trưng chất lượng tiếng nói	Phân loại 6 cảm xúc (giận, hồi hộp, chán, vui, buồn) cho tiếng Đức. Phân lớp cảm xúc dùng bộ phân lớp Bayes	Lugger M. And Yang B (2007) [101] [138]
6	Các đặc trưng dựa trên *F0*, năng lượng và thời hạn	Phân loại 6 cảm xúc cho tiếng Trung Quốc sử dụng SVM và thuật toán sinh. Kết quả nhận dạng đạt 88%	Wang Y., Du S., and Zhan S. (2008) [171]
7	Các đặc trưng dựa trên ngôn điệu và chất lượng tiếng nói	Phân loại 4 cảm xúc cho tiếng Trung Quốc (giận, phấn khởi, trung tính, buồn) sử dụng SVM đạt kết quả nhận dạng 76%	Zhang S. (2008) [172]

Có thể thấy rằng phần lớn các nghiên cứu về nhận dạng cảm xúc đều thực hiện bằng cách sử dụng đặc trưng ngôn điệu thống kê ở mức phát ngôn (tổng thể) [49], [50], [161], [163], [164], [165]. Rất ít nghiên cứu về hành vi động của các mẫu ngôn điệu (chi tiết) để phân tích cảm xúc [160], [166]. Việc phân tích ngôn điệu cơ bản của tiếng nói được thực hiện trong [141] ở mức câu, từ, và mức âm tiết chỉ sử dụng các thống kê bậc nhất của các tham số ngôn điệu cơ bản.

Thời hạn cũng là một trong những tham số ảnh hưởng nhiều nhất đến cảm xúc theo Cahn [13] 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 [173]. Nghiên cứu trong [174] 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.

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

Trong giao tiếp thông thường giữa người với người, ngoài nội dung của thông điệp cần trao đổi, 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 Việt nói nói riêng sau đó đưa ra được các mô hình tổng hợp, nhận dạng tiếng nói có cảm xúc.

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. Ở mục 2.4, luận án đã trình bày các tham số đặc trưng của tín hiệu tiếng nói dùng cho nhận dạng cảm xúc và một số các nghiên cứu đã sử dụng các đặc trưng này để phân loại các cảm xúc. Dựa theo các nghiên cứu về các đặc trưng của tín hiệu tiếng nói và những ứng dụng của nó trong các nghiên cứu thử nghiệm về nhận dạng cảm xúc, trong mục này luận án sẽ đề xuất các tham số đặc trưng được trình bày dưới đây dùng cho thử nghiệm nhận dạng cảm xúc tiếng Việt nói.

2.5.1 Các hệ số MFCC

Như đã trình bày trong mục 2.4.1, MFCC là các hệ số ceptrum theo tần số Mel được sử dụng rộng rãi trong nhận dạng tiếng nói và được Davis và Mermelstein giới thiệu trong những năm 1980. Đã có nhiều nghiên cứu cho thấy các đặc tính MFCC được xem như là tập các đặc điểm của hệ thống để phân loại các cảm xúc. MFCC đã được sử dụng để phân loại các cảm xúc cho nhiều ngôn ngữ khác nhau như tiếng Trung Quốc [175], [176], tiếng Đức [177], [178] hay nhận dạng cảm xúc trên ngữ liệu đa thể thức IEMOCAP [179]… Do vậy, MFCC được coi là các đặc trưng cơ bản của tín hiệu tiếng

nói và được sử dụng trong các hệ thống nhận dạng về cảm xúc tiếng nói. Quá trình tính toán các hệ số MFCC này thường được thực hiện theo sơ đồ Hình 2.4.

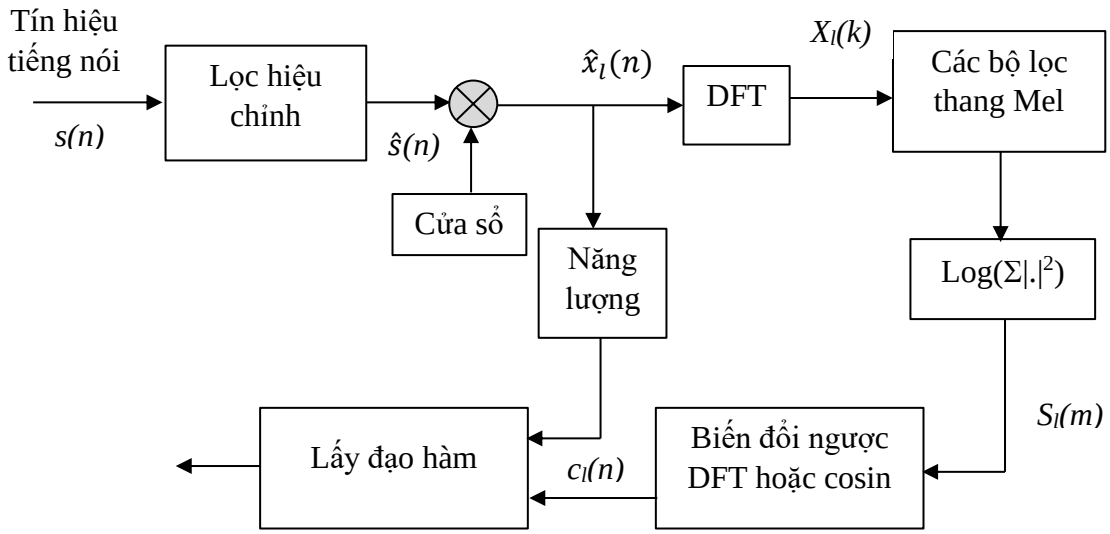
Các bước được thực hiện trên sơ đồ như sau:

(1). Tiền xử lý tín hiệu tiếng nói: Tín hiệu tiếng nói được đưa qua bộ lọc hiệu chỉnh có tác dụng bù lại phổ tín hiệu nguồn âm hữu thanh chủ yếu tập trung ở miền tần thấp và hiệu ứng bức xạ tại môi tương đương với bộ lọc thông cao. Bởi vì tiếng nói có sự suy giảm khoảng 20dB/decade khi lên tần số cao do đặc điểm sinh lý của hệ thống phát âm của con người nên bước xử lý này sẽ tăng cường tín hiệu lên một giá trị gần 20dB/decade để khắc phục sự suy giảm đó.

Để thực hiện lọc hiệu chỉnh thường dùng bộ lọc đáp ứng xung hữu hạn FIR (Finite Impulse Response) có hàm truyền đạt:

$$H_{hc}(z) = 1 - a_{hc}z^{-1} \quad (2.1)$$

Trong đó, a_{hc} là hệ số hiệu chỉnh, thường có giá trị là 0,95.



Hình 2.4 *Sơ đồ tính hệ số MFCC*

(2). Chia tín hiệu tiếng nói thành chuỗi các khung với kích thước khung là 20ms và độ dời khung là 10ms. Sau khi chia khung, tín hiệu được đưa qua cửa sổ Hamming

(3). Biến đổi tín hiệu về miền tần số: Tại bước này, với mỗi khung tín hiệu, sử dụng bộ biến đổi Fourier rời rạc DFT để chuyển tín hiệu về miền tần số. Công việc tính toán được thực hiện bằng thuật toán FFT.

(4). Phổ Mel được tính bằng cách cho tín hiệu DFT đi qua qua băng bộ lọc Mel: Phổ của mỗi khung tín hiệu sau khi thu được qua DFT được xử lý qua các bộ lọc số được áp dụng để lọc các tín hiệu theo các dải tần số khác nhau. Phản ứng của tai người với các thành phần của tần số là không tuyến tính. Sự khác nhau về tần số ở vùng tần số thấp (<1KHz) dễ được nhận biết bởi con người hơn là ở vùng tần số cao. Lọc theo thang tần số Mel mô phỏng tính chất này bằng cách dùng các bộ lọc được

phân bố theo một hàm phi tuyến trong khoảng không gian tần số, thông thường là hàm Mel:

$$Mel(f) = 2595log_{10}(1 + \frac{f}{700}) \quad (2.2)$$

Đầu ra của các bộ lọc là tổng các thành phần phổ được lọc. Các bộ lọc này mô phỏng quá trình xử lý của hệ thống thính giác.

(5). Lấy log của đầu ra các bộ lọc và dùng biến đổi cosin rời rạc DCT trên các véctơ log của phổ sẽ được các hệ số MFCC.

Các hệ số MFCC có thể được lấy đạo hàm để có thông tin biến thiên theo thời gian của các véctơ đặc trưng như đạo hàm bậc nhất, đạo hàm bậc 2 của MFCC.

2.5.2 Năng lượng tiếng nói

Năng lượng tiếng nói cũng là một tham số có liên quan đến cảm xúc của tiếng nói. Năng lượng được tính bằng tổng của căn bậc hai độ lớn các thành phần FFT rời rạc của tín hiệu. Tổng này sau đó được chuẩn hóa theo độ dài cửa sổ.

2.5.3 Cường độ tiếng nói

Cường độ tiếng nói là đặc trưng của ngôn điệu tiếng nói. Các cảm xúc hưng phấn thường có cường độ lớn hơn như khi vui thì người ta nói to hơn khi buồn, hay khi tức giận thường nói to hơn bình thường. Đây cũng là tham số quan trọng ảnh hưởng tới phân biệt các cảm xúc vui, buồn, tức hay bình thường.

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

Tiếng Việt là ngôn ngữ có thanh điệu nên ngoài các tham số đặc trưng về nguồn âm, tuyến âm thì đặc trưng về ngôn điệu có tầm quan trọng ảnh hưởng đến nhận dạng cảm xúc. Tiếng Việt có sáu thanh điệu: thanh ngang, thanh huyền, thanh sắc, thanh hỏi, thanh ngã và thanh nặng [180]. 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. Kết luận này cũng phù hợp với việc sử dụng tham số $F0$ để tổng hợp tiếng Việt có cảm xúc trong nghiên cứu đã được công bố ở bài báo "Tổng hợp tiếng Việt có cảm xúc" tại Chuyên san các công trình nghiên cứu phát triển Công nghệ Thông tin và Truyền thông của tạp chí Bưu chính Viễn thông, tập V-2, số 18 (38), trang 67-77.

Luận án sử dụng các biến thể của $F0$ được tính theo các công thức sau đây:

- Đạo hàm $F0$:

$$dF0(t) = dF0(t)/dt \quad (2.3)$$

- Chuẩn hóa $F0$ theo giá trị trung bình của $F0$ cho mỗi file:

$$F0NormAver(t) = F0(t)/\overline{F0(t)} \quad (2.4)$$

- Chuẩn hóa $F0$ theo giá trị min $F0$ và max $F0$ cho mỗi file:

$$F0NormMinMax(t) = \frac{F0(t) - \min F0(t)}{\max F0(t) - \min F0(t)} \quad (2.5)$$

- Chuẩn hóa $F0$ theo trung bình và độ lệch chuẩn của $F0$:

$$F0NormAverStd(t) = \frac{F0(t) - \overline{F0(t)}}{\sigma F0(t)} \quad (2.6)$$

- Đạo hàm $LogF0$:

$$dLogF0(t) = dLogF0(t)/dt \quad (2.7)$$

- Chuẩn hóa $LogF0$ theo giá trị min $LogF0$ và max $LogF0$ cho mỗi file:

$$LogF0NormMinMax(t) = \frac{LogF0(t) - \min LogF0(t)}{\max\ LogF0(t) - \min\ LogF0(t)} \quad (2.8)$$

- Chuẩn hóa $LogF0$ theo trung bình $LogF0$ cho mỗi file:

$$LogF0NormAver(t) = LogF0(t)/\overline{LogF0(t)} \quad (2.9)$$

- Chuẩn hóa $LogF0$ theo trung bình và độ lệch chuẩn của $LogF0$ cho mỗi file:

$$LogF0NormAverStd(t) = \frac{LogF0(t) - \overline{LogF0(t)}}{\sigma LogF0(t)} \quad (2.10)$$

2.5.5 Các formant và dải thông tương ứng

Nói chung, formant đại diện cho chuỗi các cấu hình tuyến âm, do đó phân tích formant sử dụng giá trị, vị trí và băng thông của chúng có thể giúp trích xuất được những thông tin cụ thể liên quan đến cảm xúc từ tín hiệu tiếng nói. Tham số formant rất quan trọng trong nhận dạng giọng nói. Vì vậy, một sự thay đổi nhỏ trong các tham số này gây ra sự khác biệt về cảm nhận, có thể dẫn đến sự biểu hiện của những cảm xúc khác nhau.

Dải thông tương ứng với formant không ảnh hưởng đến thông tin ngữ âm mà đại diện cho một số thông tin cụ thể của người nói. Khi có sự thay đổi trong dải thông, formant cũng gây ra những trường hợp thay đổi về cảm xúc. Dải thông formant là dải tần số đo được ở khoảng 3dB tính từ đỉnh cực đại tương ứng trở xuống.

2.5.6 Các đặc trưng phổ

Các đặc trưng phổ như các thành phần hài (*harmonicity*), trọng tâm phổ (c*enter of gravity),* mômen trung tâm *(central spectral moment),* độ lệch chuẩn tần số *(standard deviation),* giá trị trung bình của phổ (*mean*), độ lệch *(skewness),* độ nhọn (*kurtosis),* độ dốc (*slope*) và độ lệch chuẩn của phổ trung bình dài hạn *(standard deviation of LTAS-Long Term Average Spectrum)* cũng được xem là các tham số đặc trưng có liên quan đến cảm xúc tiếng nói. Theo Praat [181], các thành phần hài đại diện cho mức độ tuần hoàn và còn được gọi là tỷ lệ sóng hài-nhiễu HNR (Harmonics-to-Noise

Ratio). *Harmonicity* được biểu diễn theo 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 0 dB có nghĩa là năng lượng trong sóng hài và trong nhiễu bằng nhau [181]. Giả sử $S(f)$ là phổ phức, trong đó f là tần số, trọng tâm phổ được cho bởi công thức (2.11).

$$\frac{\int_0^\infty f|S(f)|^p df}{\int_0^\infty |S(f)|^p df} \tag{2.11}$$

Ở đây $\int_0^\infty |S(f)|^p df$ 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, trọng tâm phổ là 377 Hz. Đối với nhiễu trắng ở tần số 22050 Hz, 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 (2.12) với f_c là trọng tâm phổ.

$$\frac{\int_0^\infty (f - f_c)^n |S(f)|^p df}{\int_0^\infty |S(f)|^p df} \tag{2.12}$$

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 (2.12). 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:

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j \quad (2.13)$$

Để tiến hành các thử nghiệm nhận dạng, các tham số đặc trưng cho tiếng nói có cảm xúc trong bộ ngữ liệu cảm xúc tiếng Việt đã được trích chọn bằng bộ công cụ Praat [181] và Alize [182]. Các tham số này được đề xuất trong Bảng 2.6. Phạm vi xác định $F0$ phụ thuộc vào giới tính. Đối với giọng nữ, giá trị $F0$ tối đa là 350 Hz, và giá trị này là 200 Hz đối với giọng nam.

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

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

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

Luận án sẽ sử dụng phân tích phương sai ANOVA và kiểm định *T* (Tukey's test) để đánh giá ảnh hưởng của một số tham số cơ bản như tần số cơ bản $F0$ trung bình, năng lượng trung bình, các đặc trưng phổ của bộ ngữ liệu cảm xúc tiếng Việt đã được trình bày cụ thể trong mục 2.3. Mục 2.6.1 sau đây sẽ trình bày khái quát về phương pháp phân tích phương sai ANOVA và kiểm định T.

2.6.1 Phân tích phương sai ANOVA và kiểm định T

2.6.1.1 Phân tích phương sai one-way ANOVA

Các phân tích ANOVA [64] 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 dùng từ hai phương pháp xử lý trở lên. Các đặc tính phân biệt các phương pháp xử lý hoặc các quần thể với nhau được gọi là các nhân tố (factor) và 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 $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_I$ là gần với nhau. Các thủ tục kiểm tra dựa trên việc so sánh phép đo độ chênh lệch giữa các $\bar{x}_i$ 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 (2.14) và (2.15).

$$MSTr = \frac{J}{I-1}[(\bar{X}_{1.} - \bar{X}_{..})^2 + (\bar{X}_{2.} - \bar{X}_{..})^2 + \cdots + (\bar{X}_{I.} - \bar{X}_{..})^2]$$
$$= \frac{J}{I-1}\sum_{i}(\bar{X}_{i.} - \bar{X}_{..})^2 \quad (2.14)$$

$$MSE = \frac{S_1^2 + S_2^2 + \cdots + S_I^2}{I} \quad (2.15)$$

Trong công thức (2.14), 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 (2.15), S_i^2 là phương sai mẫu thứ i. Thử nghiệm thống kê cho one-way ANOVA là $F = MSTr/MSE$.

2.6.1.2 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 [64]). 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 (2.16) với $Q_{(\alpha,I,I(J-1))}$ là giá trị của phân bố *Student* tại mức ý nghĩa α.

$$\bar{X}_i - \bar{X}_j - Q_{(\alpha,I,I(J-1))}\sqrt{MSE/J} \leq \mu_i - \mu_j \leq \bar{X}_i - \bar{X}_j + Q_{(\alpha,I,I(J-1))}\sqrt{MSE/J} \quad (2.16)$$

với mọi i, j và ($i = 1, \ldots, I$ và $j = 1, \ldots, J$) với ($i < j$)

Giả sử $I = 4$ thì 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 ($P > F$).

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

Luận án đã tiến hành các thử nghiệm để xem xét yếu tố đặc trưng của tín hiệu tiếng nói đã trình bày trong Bảng 2.6 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. Trong luận án, phương pháp one-way ANOVA đã được sử dụng để phân tích sự ảnh hưởng của các đặc trưng đến sự phân biệt các cảm xúc với nhau. Ngữ liệu cảm xúc đã giới thiệu ở mục 2.3 được đưa vào đánh giá. Giả thuyết cần kiểm định là các giá trị trung bình 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. Bảng 2.7 thống kê giá trị F và $P - value$ của phân tích ANOVA cho các đặc trưng đã đề xuất.

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 được trình bày trong Bảng 2.7.

Bảng 2.7 *Giá trị thống kê F và P-value của phân tích ANOVA cho các tham số đặc trưng*

STT	Tham số đặc trưng	Giá trị F	Giá trị $P - value$
1	*Harmonicity*	218,4038	5,7727E-134
2	*Center of gravity*	1317,4812	0,0
3	*Standard deviation*	1397,2756	0,0
4	*Skewness*	1114,7564	0,0
5	*Kurtosis*	594,1112	0,0
6	*Central spectral moment*	1664,5398	0,0
7	*Mean*	783,2338	0,0
8	*Slope*	1218,3402	0,0
9	*Standard deviation of LTAS*	751,0346	0,0
10	*Intensity*	877,2506	0,0
11	*Formant1*	414,7949	3,008715E-243
12	*Band1*	423,7496	5,269112E-248
13	*Formant2*	1066,9129	0,0
14	*Band2*	417,9563	6,265008E-245
15	*Formant3*	418,7876	2,265692E-245
16	*Band3*	342,0899	6,349E-204
17	*Formant4*	198,3841	2,954376E-122
18	*Band4*	321,2435	2,071804E-192
19	*F0*	453,9051	6,902E-264
20	*dF0*	25,6372	1,8217E-16
21	*F0NormAver*	3241,4807	0,0
22	*F0NormMinMax*	257,7462	1,1788E-156
23	*F0NormAverStd*	3241,481	0,0
24	*dLogF0*	91,85127	5,104602E-58
25	*LogF0NormMinMax*	186,8734	1,792956E-115
26	*LogF0NormAver*	1168,8149	0,0
27	*LogF0NormAverStd*	3001,683	0,0

Có thể thấy rằng, trong phần lớn các trường hợp, $P - value = 0{,}05$ được sử dụng làm ngưỡng cắt cho quyết định khác biệt [64]. Nếu giá trị $P - value$ nhỏ hơn 0,05, giả thuyết cho rằng không có sự khác biệt giữa các cặp cảm xúc sẽ bị loại bỏ và khẳng

định có sự khác biệt đáng kể. Bảng 2.7 cho thấy, các giá trị $P - value \cong 0$, có nghĩa xác suất để các giá trị trung bình $\mu_1 = \mu_2 = \mu_3 = \mu_4$ của các cảm xúc là rất thấp. Như vậy, giả thuyết không có sự phân biệt giữa các cảm xúc bị bác bỏ và có thể khẳng định rằng có thể phân biệt được các cảm xúc với nhau dựa vào các tham số trên.

Bảng 2.8 *Giá trị $P - value$ của kiểm định T với các tham số đặc trưng cho từng cặp cảm xúc*

Thứ tự	Tham số đặc trưng	$P - value$					
		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	*Center of gravity*	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	1,3024E-06
3	*Standard deviation*	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 spectral moment*	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	*Standard deviation of LTAS*	3,7683E-09	0,61003	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
10	*Intensity*	3.7683E-09	0,4524	3,7683E-09	3,76 83E-09	3,7683E-09	3,7683E-09
11	*Formant1*	3,7683E-09	0,0032784	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
12	*Band1*	3,7683E-09	3,7683E-09	3,7694E-09	4,0303E-09	3,7683E-09	3,7683E-09
13	*Formant2*	3,7683E-09	3,9206E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
14	*Band2*	3,7683E-09	3,7683E-09	0,0001134	3,7683E-09	3,7683E-09	3,7683E-09
15	*Formant3*	3,7683E-09	3,7683E-09	7,2108E-05	3,7683E-09	3,7683E-09	3,7683E-09
16	*Band3*	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,8514E-09	3,7683E-09
17	*Formant4*	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	0,015535	3,7683E-09
18	*Band4*	3,7683E-09	0,0013824	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
19	*F0*	3,7683E-09	0,018149	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
20	*dF0*	2,8295E-07	0,0092189	0,99973	3,7683E-09	1,6984E-07	1,6984E-07
21	*F0NormAver*	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
22	*F0NormMinMax*	3,7683E-09	0,014222	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
23	*F0NormAverStd*	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
24	*dLogF0*	3,7683E-09	0,18898	0,42845	3,7683E-09	3,7683E-09	0,96305
25	*LogF0NormMinMax*	4,4826E-07	3,84E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
26	*LogF0NormAver*	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09
27	*LogF0NormAverStd*	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09	3,7683E-09

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 khả năng phân biệt các cặp cảm xúc này, luận án đã tiến hành thử nghiệm dùng kiểm định *T* và kết quả được thống kê trong Bảng 2.8.

Bảng 2.8 cho thấy, các giá trị $P - value$ là rất nhỏ khi đánh giá cho từng tham số đối với từng cặp cảm xúc. Các cặp cảm xúc được phân biệt tốt nhất với hầu hết 27 tham số đặc trưng là buồn-vui, buồn-tức. Điều này là phù hợp vì trong thực tế hai cặp cảm xúc này cũng dễ được cảm nhận phân biệt một cách rõ ràng. Các cặp cảm xúc còn lại cũng được phân biệt rõ, tuy nhiên các tham số *mean, standard deviation of LTA, intensity* có ảnh hưởng ít hơn với cặp cảm xúc bình thường-tức. Các tham số *dF0, dLogF0* ảnh hưởng ít đến cặp cảm xúc bình thường-vui. Cũng khó phân biệt đối với cặp cảm xúc tức-vui khi sử dụng tham số *dLogF0*.

Cũng cần lưu ý, việc xét một tham số nào đó để thấy rằng có độ phân biệt cao chỉ có ý nghĩa riêng biệt đối với tham số đó chứ không mang tính tổng thể cho cả mô hình nhận dạng. Trên thực tế, việc xây dựng một mô hình không chỉ sử dụng mỗi một tham số nào đó mà sử dụng một tập tham số khác nhau. Trong luận án không đánh giá sự ảnh hưởng của các tham số MFCC bởi vì MFCC là tham số đặc trưng cơ bản cần có của tín hiệu tiếng nói và được sử dụng rộng rãi trong các nghiên cứu về nhận dạng tiếng nói nói chung và nhận dạng cảm xúc nói riêng.

Các giá trị $P - value$ trong Bảng 2.7 và 2.8 được tính toán bởi MatLab với độ chính xác double-precision và được giữ nguyên theo định dạng hiển thị của MatLab. Trên thực tế, các giá trị rất nhỏ như xxxE-243, xxxE-248... có thể được coi là giá trị 0.

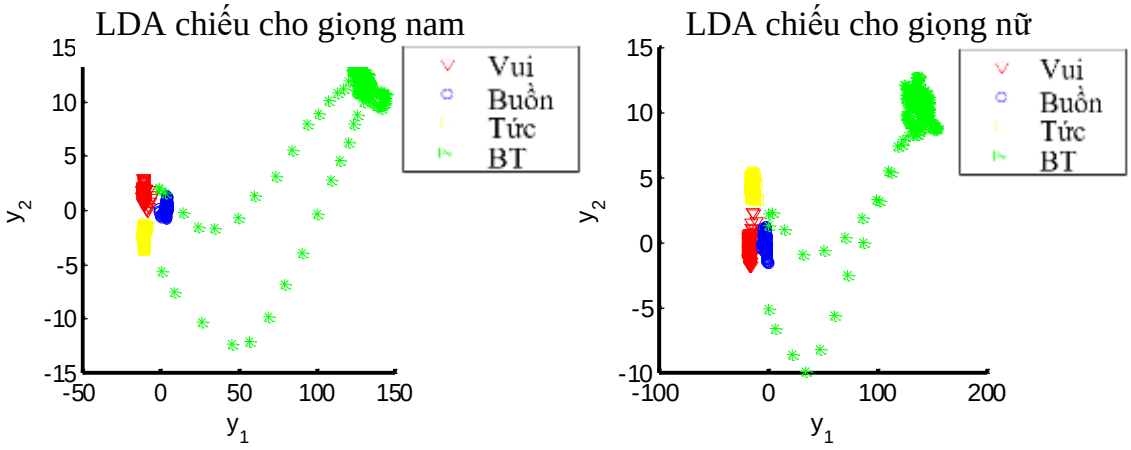
2.7 Đánh giá sự phân lớp của bộ ngữ liệu cảm xúc tiếng Việt

2.7.1 Kết quả phân lớp với LDA

Để đánh giá sự phân lớp của bộ ngữ liệu cảm xúc tiếng Việt dùng cho các thử nghiệm trong luận án, phương pháp phân lớp LDA đã được sử dụng. Tổng số file tiếng nói đưa vào phân lớp là 5584 file, trong đó có 2792 file tiếng nói tương ứng với mỗi giới tính.

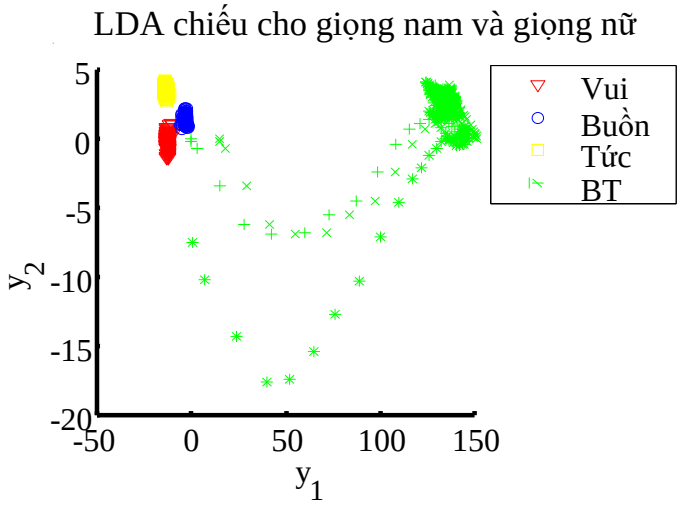
Kết quả phân lớp bằng phương pháp LDA trên Hình 2.5 cho thấy, 4 cảm xúc vui, buồn, tức, bình thường được phân lớp tương đối rõ ràng cho cả giọng nam và giọng nữ. Trong 4 cảm xúc, cảm xúc bình thường được phân biệt rõ nhất so với 3 cảm xúc còn lại.

Từ Hình 2.5 cho giọng nữ, cảm xúc bình thường được quan sát thấy rất rõ rệt. Ba cảm xúc vui, buồn và tức tuy phân bố gần nhau hơn nhưng chúng vẫn phân bố theo từng vùng của từng cảm xúc và cũng dễ quan sát và cũng dễ thấy có độ phân biệt.



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 biệt 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

Trước khi nghiên cứu thử nghiệm nhận dạng với mô hình GMM và DCNN, luận án đã tiến hành thử nghiệm nhận dạng cảm xúc trên các bộ phân lớp IBk, Trees J48, SMO để góp phân đánh giá tính phân lớp của bộ ngữ liệu cảm xúc tiếng Việt.

2.7.2.1 Công cụ, ngữ liệu và tham số sử dụng

Các bộ phân lớp IBk, Trees J48, SMO thuộc bộ công cụ Weka, gồm tập hợp các thuật toán học máy dùng cho khai phá ngữ liệu do Đại học Waikato, NewZealand phát triển [183]. SMO [184] là thuật toán tối ưu hóa cực tiểu lần lượt để huấn luyện bộ phân lớp hỗ trợ véctơ dùng kernel đa thức hoặc Gauss. IBk là bộ phân lớp *k* láng giềng gần nhất sử dụng độ đo khoảng cách Ơclit [183]. Bộ phân lớp Trees J48 [185] 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 toán C4.5 và thuật toán này được dùng để tạo cây quyết định do Ross Quinlan phát triển.

Ngữ liệu dùng cho các thử nghiệm này là tập ngữ liệu T1 đã được trình bày trong Chương 2 gồm 5584 file tương ứng với 4 cảm xúc và được 16 nghệ sĩ (8 giọng nam và 8 giọng nữ) thể hiện. Số file trong bộ ngữ liệu 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 với 384 tham số [186]. Các tham số này được tính toán như sau:

Với mỗi file tín hiệu tiếng nói sẽ được phân tách thành một tập các khung tín hiệu với độ dài khung 25 ms và độ dịch khung 10 ms. Sau đó, 16 giá trị đặc trưng được tính toán cho mỗi khung bao gồm:

- 12 hệ số MFCC, tỷ lệ biến thiên qua trục không (Zero-Crossing Rate)
- Xác suất âm hữu thanh, tần số cơ bản, năng lượng

Mỗi giá trị đặc trưng này lại được tính đạo hàm bậc nhất theo thời gian và thu được 32 tham số. Tất cả 32 tham số lại được tính toán với 12 giá trị thống kê sau đây:

- Giá trị cực đại, cực tiểu, vị trí xuất hiện cực đại, vị trí xuất hiện cực tiểu, giá trị trung bình, dải giá trị (độ chênh lệch giữa giá trị lớn nhất và giá trị nhỏ nhất)
- Độ dốc, độ lệch và lỗi trung bình bình phương của xấp xỉ tuyến tính
- 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*).

Như vậy, số tham số đặc trưng được tính sẽ là 32 × 12 = 384 tham số cho mỗi file tiếng nói.

2.7.2.2 Kết quả thử nghiệm

Kết quả nhận dạng cảm xúc tiếng Việt sử dụng các bộ phân lớp IBk, Trees J48, SMO được thống kê trong Bảng 2.9, Bảng 2.10 và Bảng 2.11.

Bảng 2.9 là tỷ lệ nhận dạng cảm xúc dùng tất cả 384 tham số. Bảng 2.9 cho thấy, tỷ lệ nhận dạng đúng trung bình cao nhất cho cả 4 cảm xúc đạt 98,17% khi sử dụng bộ phân lớp IBk còn tỷ lệ nhận dạng đúng trung bình thấp nhất là 80,64% khi sử dụng bộ phân lớp Trees J48.

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	Tỷ lệ 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 là tỷ lệ nhận dạng cảm xúc đối với trường hợp chỉ sử dụng 288 tham số liên quan đến MFCC. Tỷ lệ nhận dạng trung bình đối với bộ phân lớp IBk là 98,4%, SMO là 92.8%, Trees J48 là 79,3%.

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	Tỷ lệ trung bình
IBk	Tức	**98,28**	1,29	0,29	0,14	**98,4**
	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	**92,8**
	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	**79,3**
	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**	

Đối với Bảng 2.11, khi số tham số giảm xuống còn 48 tham số liên quan đến $F0$ và năng lượng thì tỷ lệ nhận dạng đúng đều giảm so với trường hợp dùng cả 384 tham số song vẫn giữ quy luật tỷ lệ nhận dạng đúng trung bình cao nhất cho bộ phân lớp IBk và thấp nhất cho bộ phân lớp Trees J48. Trường hợp chỉ sử dụng các tham số liên quan đến $F0$ và năng lượng, tỷ lệ nhận dạng đúng trung bình cao nhất giảm xuống còn 82,59% và tỷ lệ nhận dạng đúng trung bình thấp nhất giảm xuống còn 75,25%.

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	Tỷ lệ 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**	

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 trong các tài liệu [104], [105] hoặc kết quả nhận dạng cảm xúc của một số ngôn ngữ khác [187] [188], [189], [190].

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 IBk cho kết quả nhận dạng tốt nhất đối với cả bốn cảm xúc. Với việc thử nghiệm nhận dạng trên công cụ Weka dùng 3 bộ phân lớp trên, có thể thấy bộ ngữ liệu cảm xúc tiếng Việt đã đề xuất có chất lượng đảm bảo để tiến hành các thử nghiệm nhận dạng cảm xúc trong luận án. Chương sau của luận án sẽ đi sâu vào nghiên cứu thử nghiệm nhận dạng cảm xúc cho tiếng Việt nói với mô hình nhận dạng GMM dựa trên các tập ngữ liệu và tham số sử dụng khác nhau.

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 cảm xúc tiếng Việt được đề xuất cho các thử nghiệm trong đề tài bao gồm 5584 file tiếng nói có cảm xúc của các nghệ sĩ được thu âm với bốn cảm xúc vui, buồn, tức và bình thường. 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.

Các nghiên cứu tính toán, phân tích và đánh giá các tham số đặc trưng của tiếng nói có ảnh hưởng đến cảm xúc tiếng Việt cũng đã được trình bày. Các tham số đặc trưng của tín hiệu tiếng nói trong bộ ngữ liệu tiếng Việt đã được trích chọn bằng bộ công cụ Alize và Praat. 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.

Các nội dung nghiên cứu chính của chương 2 đã được công bố trong các bài báo số 3, 4, 5, 6 trong danh mục các công trình nghiên cứu của luận án:

3. *Cảm xúc trong tiếng nói và phân tích thống kê ngữ liệu cảm xúc tiếng Việt*, Chuyên san Các công trình Nghiên cứu, Phát triển và Ứng dụng Công nghệ Thông tin, Tạp chí Bưu chính Viễn thông, tập V-1, số 15 (35), trang 86-98.
4. *So sánh hiệu năng một số phương pháp nhận dạng cảm xúc tiếng Việt nói*, Kỷ yếu Hội nghị khoa học công nghệ quốc gia lần thứ IX, Nghiên cứu cơ bản và ứng dụng công nghệ thông tin, Cần Thơ, trang 656-662.
5. *Tổng hợp tiếng Việt có cảm xúc*, Chuyên san Các công trình nghiên cứu phát triển Công nghệ Thông tin và Truyền thông, Tạp chí Bưu chính Viễn thông, Tập V-2, Số 18 (38), trang 67-77.
6. *Ảnh hưởng của đặc trưng phổ tín hiệu tiếng nói đến nhận dạng cảm xúc tiếng Việt*, Kỷ yếu Hội nghị khoa học công nghệ quốc gia lần thứ X, Nghiên cứu cơ bản và Ứng dụng Công nghệ Thông tin, Đà Nẵng, trang 36-43.

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

Để nhận dạng cảm xúc tiếng nói, đã có nhiều bộ phân lớp được dùng như HMM, GMM, SVM, ANN, k-NN và nhiều bộ phân lớp khác. Như đã trình bày, trên thực tế không có sự thỏa thuận về một bộ phân lớp nào đó là thích hợp nhất cho nhận dạng cảm xúc. Bởi lẽ mỗi bộ phân lớp có ưu thế và hạn chế riêng của nó. Chương này của luận án sẽ trình bày kết quả thử nghiệm nhận dạng cảm xúc với mô hình nhận dạng GMM sử dụng bộ ngữ liệu cảm xúc tiếng Việt đã được trình bày trong Chương 2.

3.1 Mô hình GMM cho nhận dạng cảm xúc

Mục 1.4.7 đã trình bày khá chi tiết về mô hình GMM và phạm vi ứng dụng của mô hình này. Trên thực tế, GMM đã được dùng phổ biến cho các trường hợp định danh người nói [182], định danh ngôn ngữ [191] định danh phương ngữ [192] hoặc phân lớp thể loại âm nhạc [193]. Phần này sẽ trình bày cụ thể việc sử dụng mô hình GMM cho nhận dạng cảm xúc tiếng Việt trong khuôn khổ của luận án.

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, 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:

$$p(\boldsymbol{X}|\lambda_j) = \sum_{m=1}^{M} \pi_m N(\boldsymbol{X}; \boldsymbol{\mu}_m, \Sigma_m) \tag{3.1}$$

Trong (3.1), π_m là các trọng số của hỗn hợp thỏa mãn điều kiện $\sum_{m=1}^{M} \pi_m = 1$, $N(\boldsymbol{X}; \boldsymbol{\mu}_m, \Sigma_m)$ là các hàm mật độ thành phần với phân bố Gauss D chiều có dạng:

$$N(\boldsymbol{X}; \boldsymbol{\mu}_m, \Sigma_m) = \frac{1}{(2\pi)^{D/2}|\Sigma_m|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_m)^T \Sigma_m{}^{-1} (\boldsymbol{X}-\boldsymbol{\mu}_m)} \tag{3.2}$$

Trong (3.2), $\boldsymbol{\mu}_m$ là véctơ kỳ vọng $\boldsymbol{\mu}_m \in \mathbb{R}^D$ 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 gồm các véctơ kỳ vọng, các ma trận hiệp phương sai và các trọng số cho M thành phần: $\lambda_j = \{\boldsymbol{\mu}_m, \Sigma_m, \pi_m\}_j, m = 1,2, \dots, M$.

Trên thực tế, việc xác định mô hình GMM λ_j của cảm xúc j sẽ được thực hiện theo thuật toán cực đại hóa kỳ vọng (EM: Expectation-Maximization). Thuật toán này sẽ xác định cực đại khả hiện (ML: Maximum-Likelihood) của log khả hiện log $(p(\boldsymbol{X}|\lambda_j)$ [194]. Trong quá trình huấn luyện, thường dùng thuật toán EM để tìm ra các đại lượng

đặc trưng của mô hình. Hình 3.1 là sơ đồ mô hình GMM cho nhận dạng cảm xúc tiếng nói. Trong đó:

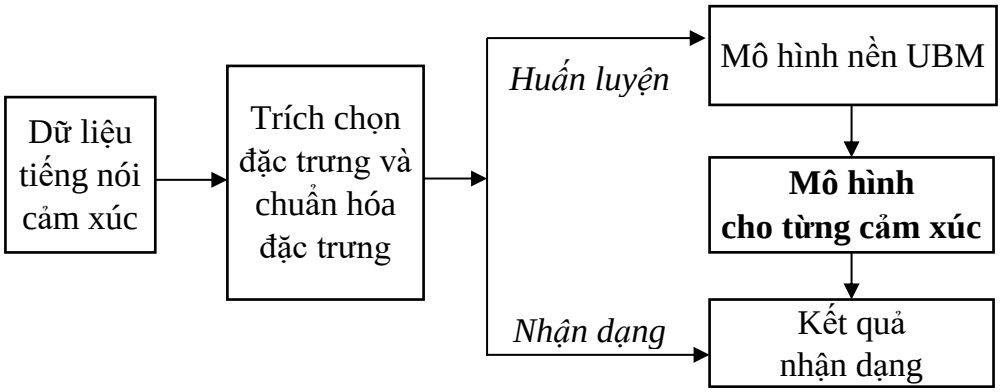
- Phần huấn luyện mô hình: Từ các véctơ tham số của tập ngữ liệu, huấn luyện được thực hiện qua các bước:

+ Khởi tạo mô hình

+ Huấn luyện mô hình nền UBM: Tạo ra một mô hình chung từ tập ngữ liệu chung với các đặc trưng chung cho cảm xúc ($\lambda = \{\mu_i, \pi_i, \Sigma_i\}$)

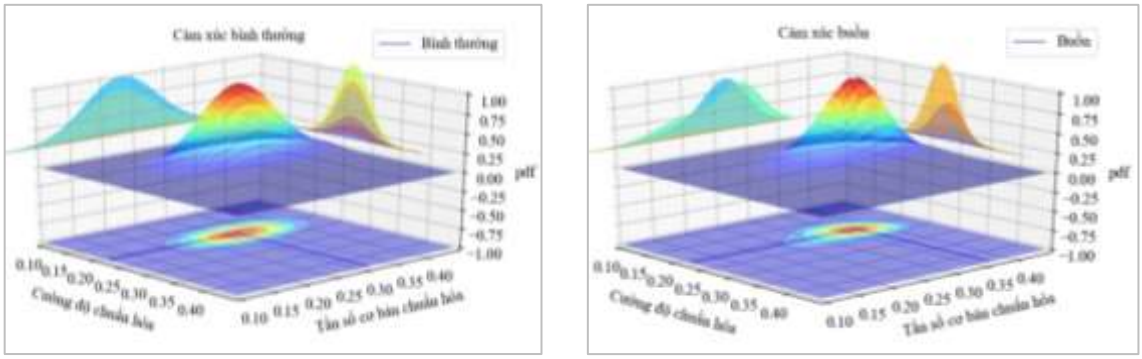
+ Dựa trên mô hình nền đã được huấn luyện, tập ngữ liệu của bốn cảm xúc sẽ được huấn luyện cho từng mô hình cảm xúc. Sẽ có 4 mô hình cảm xúc tương ứng với 4 cảm xúc như sau: $\lambda_{vui} = \{\mu_v, \pi_v, \Sigma_v\}$, $\lambda_{buồn} = \{\mu_b, \pi_b, \Sigma_b\}$, $\lambda_{tức} = \{\mu_t, \pi_t, \Sigma_t\}$, $\lambda_{bình\ thường} = \{\mu_{bt}, \pi_{bt}, \Sigma_{bt}\}$)

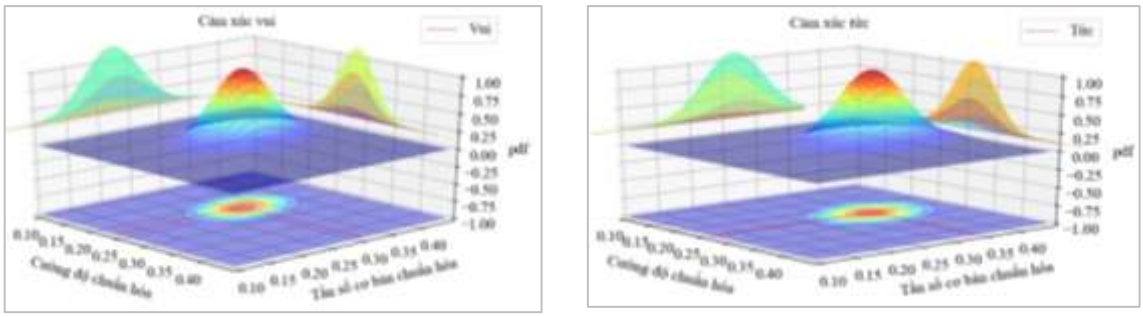
Mỗi mô hình cảm xúc có các đặc trưng kỳ vọng, phương sai, ma trận hiệp phương sai và một giá trị của số thành phần Gauss *M* tương ứng.



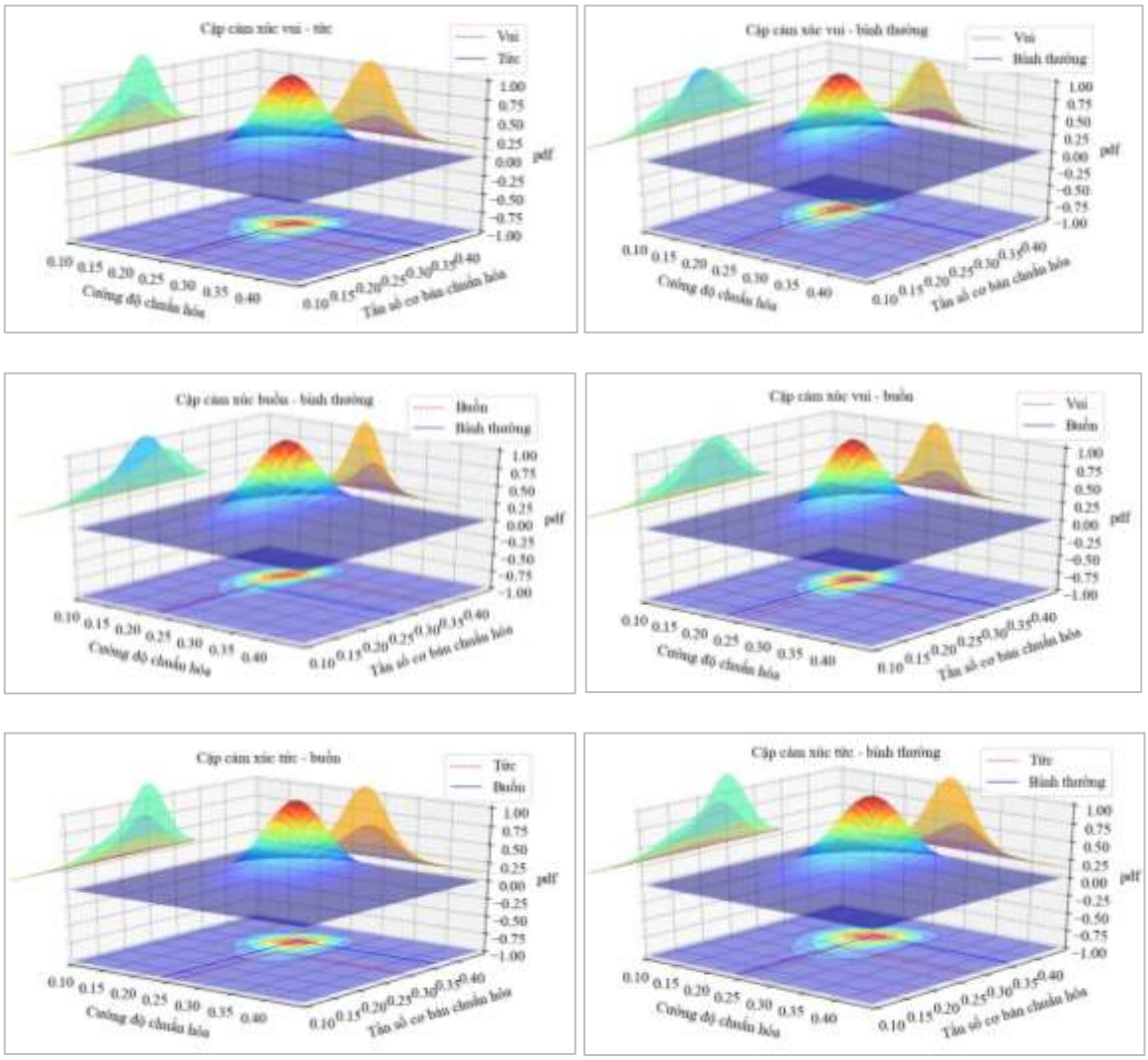
Hình 3.1 *Sơ đồ mô hình GMM tổng quát cho nhận dạng cảm xúc*

Hình 3.2 là các ví dụ minh họa cho mô hình GMM tương ứng với 4 cảm xúc: vui, buồn, tức và bình thường của bộ ngữ liệu BKEmo. Hình 3.3 là ví dụ minh họa cho mô hình GMM của 6 cặp cảm xúc tương ứng với 4 cảm xúc này. Véctơ tham số đặc trưng của các ví dụ ở đây chỉ là 2 chiều bao gồm tần số cơ bản và cường độ.





Hình 3.2 *Mô hình Gauss của 4 cảm xúc*



Hình 3.3 *Mô hình Gauss của 6 cặp cảm xúc*

- Phần thử nghiệm nhận dạng: Từ ngữ liệu đầu vào dùng để nhận dạng cảm xúc, các véctơ tham số đặc trưng của tập ngữ liệu được đưa vào mô hình, mô hình sẽ thực hiện tính toán kết quả và đưa ra điểm số.

3.2 Công cụ, tham số và ngữ liệu sử dụng

Trong luận án, bộ công cụ Alize đã được sử dụng để đánh giá mô hình GMM và thực hiện nhận dạng cảm xúc còn 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.

Alize là một thư viện mã nguồn mở hỗ trợ trong lĩnh vực nhận dạng tự động người nói và có thể mở rộng cho các ứng dụng nhận dạng khác như nhận dạng phương ngữ, nhận dạng cảm xúc [182]. Alize được phát triển từ một hệ thống có sẵn tại KTH (Kungliga Tekniska högskolan – Royal Institute of Technology), gọi là GIVES (General Identity Verification System) trong chương trình nghiên cứu của Pháp. Alize cung cấp các công cụ giúp thực hiện các xử lý như: trích chọn đặc trưng, chuẩn hóa đặc trưng, huấn luyện mô hình và nhận dạng. Kết quả sử dụng Alize để nhận dạng phương ngữ tiếng Việt đã được công bố ở bài báo "Nghiên cứu và thử nghiệm nhận dạng phương ngữ tiếng Việt" trên Tạp chí Khoa học và Công nghệ, ĐHSPKT Hưng Yên, số 4, ISSN 2354-0575, trang 96-101.

Luận án cũng đã thực hiện thử nghiệm đánh giá bộ công cụ Alize với mô hình GMM nhận dạng cảm xúc trên bộ ngữ liệu tiếng Đức [115]. Tuy nhiên, bộ ngữ liệu này có số lượng người nói, số lượng giọng nói cho các cảm xúc rất ít và phân bố không đồng đều như thống kê trong Phụ lục 2 cho thấy nên rất khó để so sánh kết quả nhận dạng cảm xúc với bộ ngữ liệu tiếng Việt. Kết quả thử nghiệm dùng Alize để nhận dạng cảm xúc theo mô hình GMM đối với bộ ngữ liệu cảm xúc tiếng Đức đã được cho ở Phụ lục 2.

Bộ tham số sử dụng trong phần thử nghiệm để nhận dạng bốn cảm xúc thuộc Chương 3 gồm MFCC (19 hệ số), đạo hàm bậc nhất, đạo hàm bậc hai của MFCC, $F0$ và biến thể của $F0$, năng lượng, cường độ, formant và dải thông tương ứng, các tham số đặc trưng phổ. Các tham số này đã được trình bày chi tiết trong mục 2.5 của Chương 2. Để đánh giá và xác định giá trị của các tham số hoặc của từng tham số đối với mô hình nhận dạng, luận án đã thử nghiệm với nhiều tổ hợp tham số khác nhau để đánh giá kết quả nhận dạng với các tổ hợp tham số này, từ đó xác nhận những tham số nào có ảnh hưởng tích cực đến nhận dạng cảm xúc tiếng Việt nói.

Mỗi thử nghiệm nhận dạng được tiến hành với số thành phần Gauss M tăng từ 16 đến 8192 theo lũy thừa 2 ($M = 2^n, n = 4,5,\dots,13$). Qua thử nghiệm thấy rằng, khi M tăng cao, thời gian tính toán cũng tăng lên nhưng tỷ lệ nhận dạng tăng không đáng kể. Do vậy, luận án đã sử dụng M tăng từ 2^4 đến 2^{13} mà không tăng thêm nữa.

Ngữ liệu dùng cho các thử nghiệm trong mục 3.3 sau đây gồm 4 tập ngữ liệu T1, T2, T3 và T4. Chi tiết về các tập ngữ liệu này đã được trình bày trong Bảng 2.2 của Chương 2.

3.3 Các thử nghiệm nhận dạng

Luận án đã tiến hành 13 thử nghiệm nhận dạng cảm xúc với mô hình GMM. Bộ tham số và số lượng các tham số của 13 thử nghiệm này được trình bày trong Bảng 3.1.

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

Các thử nghiệm	Tập tham số	Ghi chú	Số lượng
Thử nghiệm 1	*MFCC*	19 MFCC	19
Thử nghiệm 2	*MFCC+Delta1*	19 MFCC + 19 Delta1 của 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

Các tham số trong Bảng 3.1 gồm các đặc trưng phổ (*FeaSpec*) đã được trình bày chi tiết ở mục 2.5.6, tám biến thể *F0* đã được trình bày chi tiết ở mục 2.5.4. Ký hiệu *Delta1* của *MFCC* để chỉ đạo hàm bậc nhất của *MFCC*, ký hiệu *Delta2* của *MFCC* để chỉ đạo hàm bậc hai của *MFCC*, ký hiệu *Delta12* của *MFCC* để chỉ đạo hàm bậc nhất và đạo hàm bậc hai của *MFCC*. Ký hiệu *prm60* để chỉ tập gồm 60 tham số, *prm79* để chỉ tập gồm 79 tham số, *prm87* để chỉ tập gồm 87 tham số.

Tiếng Việt là ngôn ngữ có thanh điệu nên các đặc trưng liên quan đến tần số cơ bản có ảnh hưởng đáng kể đến nhận dạng cảm xúc. Điều này đã được phân tích và đánh giá trong Chương 2 của luận án. Mục này sẽ trình bày các kết quả thử nghiệm nhận dạng cảm xúc không chỉ sử dụng riêng đặc trưng *MFCC* mà còn kết hợp với đặc trưng phổ, cường độ, năng lượng, formant và dải thông tương ứng cùng với $F0$ và biến thể của $F0$.

Lý do sử dụng các tham số như trong Bảng 2.6 và việc chia các bộ tham số như trong Bảng 3.1 để tiến hành thử nghiệm có thể được giải thích như sau. *MFCC* đã được sử dụng phổ biến trong các hệ thống xử lý tiếng nói như nhận dạng người nói, nhận dạng tiếng nói, nhận dạng cảm xúc… nên *MFCC* được xem như các tham số đặc trưng cơ bản của các hệ thống này. Có thể nói rằng *MFCC* là các tham số cơ bản liên quan đến phổ tín hiệu tiếng nói đã được hội tụ và dựa trên độ nhạy của hệ thống thính giác. Ngoài ra, các tham số đặc trưng từ (8) đến (15) trong Bảng 2.6 cũng là các tham số liên quan đến phổ tín hiệu tiếng nói đã được xác định thống kê. Đặc biệt, các tham số đặc trưng *skewness* (11) và *kurtosis* (12) liên quan chặt chẽ đến phân bố chuẩn mà GMM đã sử dụng. Quy luật biến thiên khác nhau của $F0$ sẽ xác định sáu thanh điệu khác nhau của tiếng Việt. Mặt khác, các quy luật biến thiên $F0$ của một từ hoặc một câu cũng tham gia biểu hiện cảm xúc [8]. Do đó, $F0$ và các tham số biến thể của $F0$ từ (16) đến (23) có liên quan chặt chẽ với tiếng Việt và cảm xúc của tiếng nói.

Sau đây là nội dung chi tiết của các thử nghiệm đã được tiến hành.

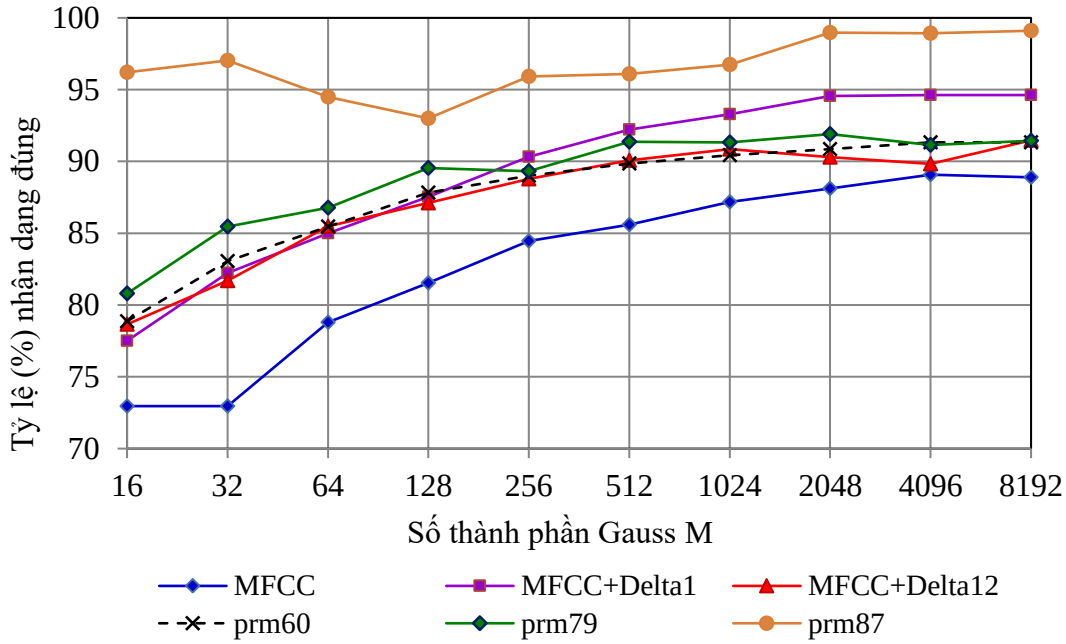
3.3.1 Thử nghiệm 1 đến Thử nghiệm 6

Các Thử nghiệm từ 1 đến 6 thực hiện nhận dạng cảm xúc với sáu tập tham số bao gồm *MFCC, MFCC+Delta1, MFCC+Delta12, prm60, prm79* và *prm87*. Sau đây là kết quả thử nghiệm nhận dạng đối với từng tập ngữ liệu và với từng cảm xúc.

3.3.1.1 Nhận dạng đối với từng tập ngữ liệu

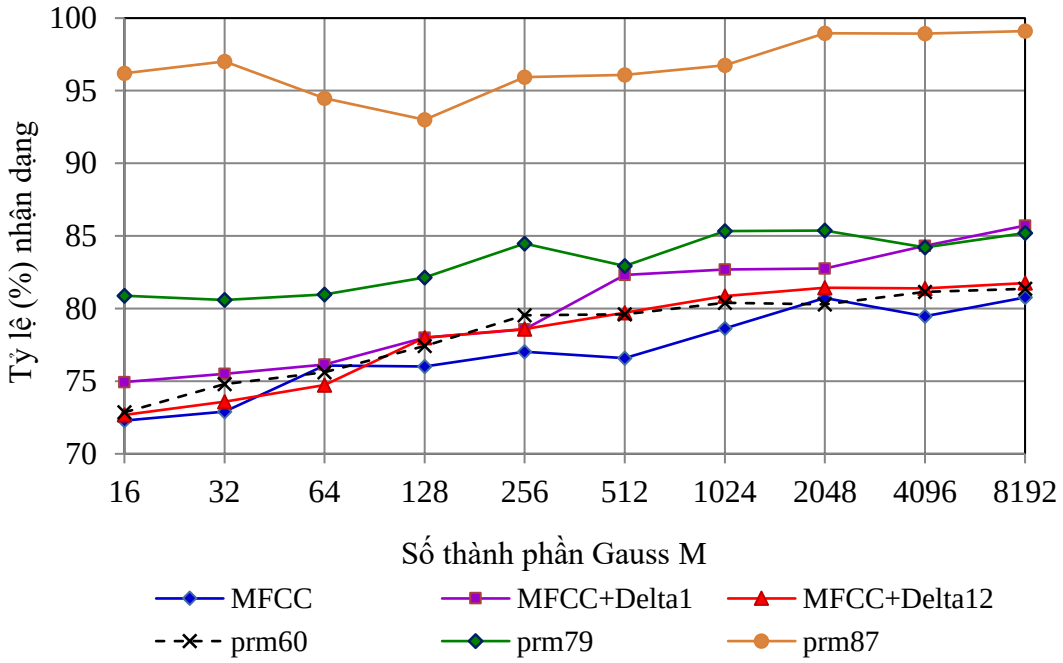
Hình 3.4 là kết quả nhận dạng cảm xúc với lần lượt các tập tham số tương ứng với các Thử nghiệm từ 1 đến 6 cho tập ngữ liệu T1. Có thể thấy, nhìn chung tỷ lệ nhận dạng đúng tăng dần khi M tăng lên. Khi sử dụng tập *prm87* để nhận dạng, tỷ lệ nhận dạng đúng trung bình là 98,96% đạt cao nhất so với năm trường hợp còn lại và nằm trong khoảng từ 97,53% - 99,97%. Trường hợp chỉ dùng *MFCC*, tỷ lệ nhận dạng đúng là thấp nhất và nằm trong khoảng từ 72,96% - 88,90%, tỷ lệ trung bình đạt 82,96%. Bốn tập tham số còn lại (*MFCC+Delta1, MFCC+Delta12, prm60, prm79*) có tỷ lệ nhận dạng xấp xỉ nhau trong khoảng từ 87,43% - 89,19%. Khi $M > 128$, tập

tham số *MFCC+Delta1* cho tỷ lệ nhận dạng cao hơn so với tập tham số *MFCC+Delta12, prm60, prm79*.



Hình 3.4 *Kết quả nhận dạng cảm xúc đối với T1*

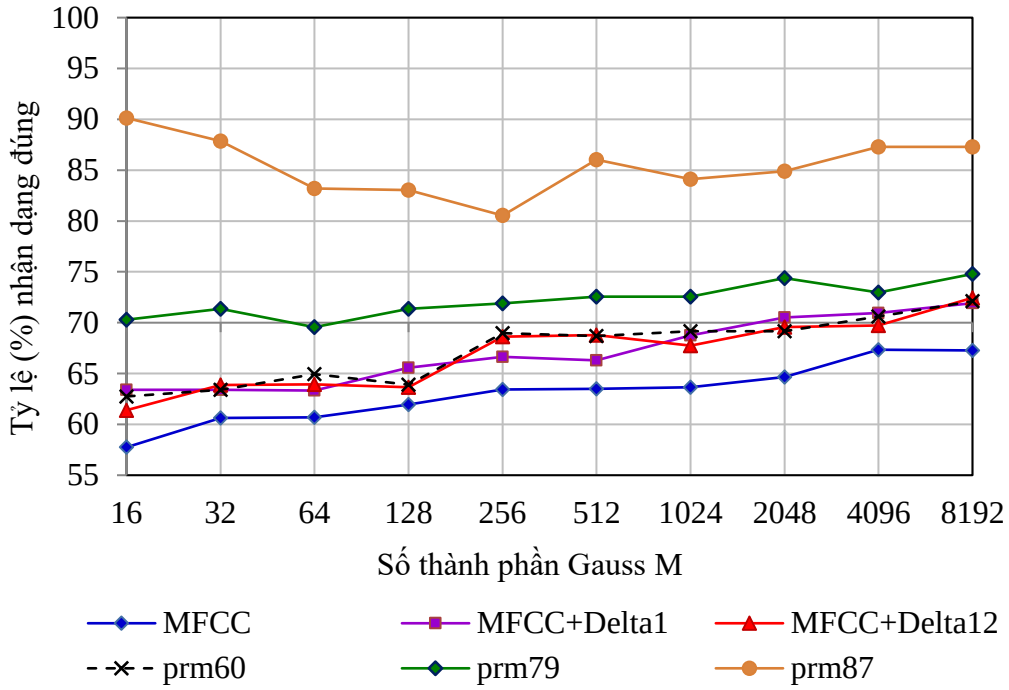
Kết quả nhận dạng đúng trung bình của bốn cảm xúc cho từng tập tham số với tập ngữ liệu T2 được trình bày trên Hình 3.5.



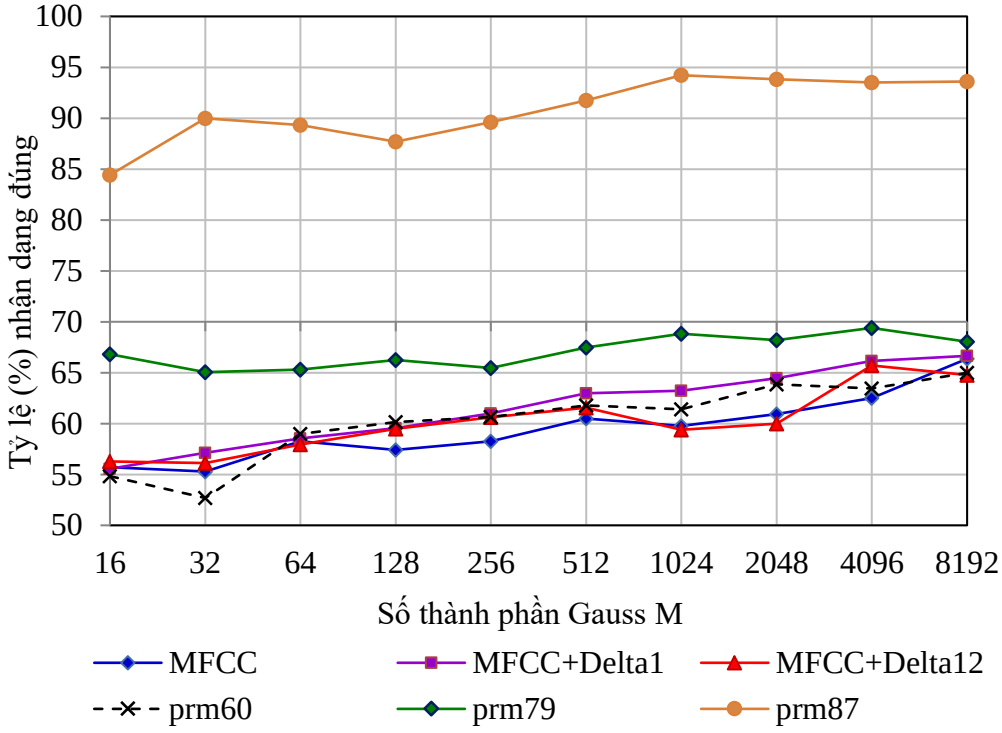
Hình 3.5 *Kết quả nhận dạng cảm xúc đối với T2*

Hình 3.5 cho thấy, khi sử dụng tập tham số *prm87*, tỷ lệ nhận dạng đúng vẫn đạt cao nhất so với các tập tham số còn lại và nằm trong khoảng 93% - 99,11%. Với 5 tập tham số còn lại, tỷ lệ nhận dạng đúng nằm trong khoảng từ 72,29% - 85,71%. Trường hợp tập tham số chỉ có *MFCC*, tỷ lệ nhận dạng đúng vẫn là thấp nhất. Hai trường hợp dùng *MFCC+Delta12* và *prm60* có tỷ lệ nhận dạng đúng xấp xỉ nhau.

Hình 3.6 là kết quả nhận dạng với tập ngữ liệu T3. Kết quả nhận dạng cho thấy, tập 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 $M = 256$. Các trường hợp còn lại, tỷ lệ nhận dạng đúng trong khoảng từ 57,75% - 74,79%. Khi M tăng, tỷ lệ nhận dạng của các tập tham số này cũng tăng nhưng không nhiều chỉ từ 66,97% đến 67,37%.



Hình 3.6 *Kết quả nhận dạng cảm xúc đối với T3*

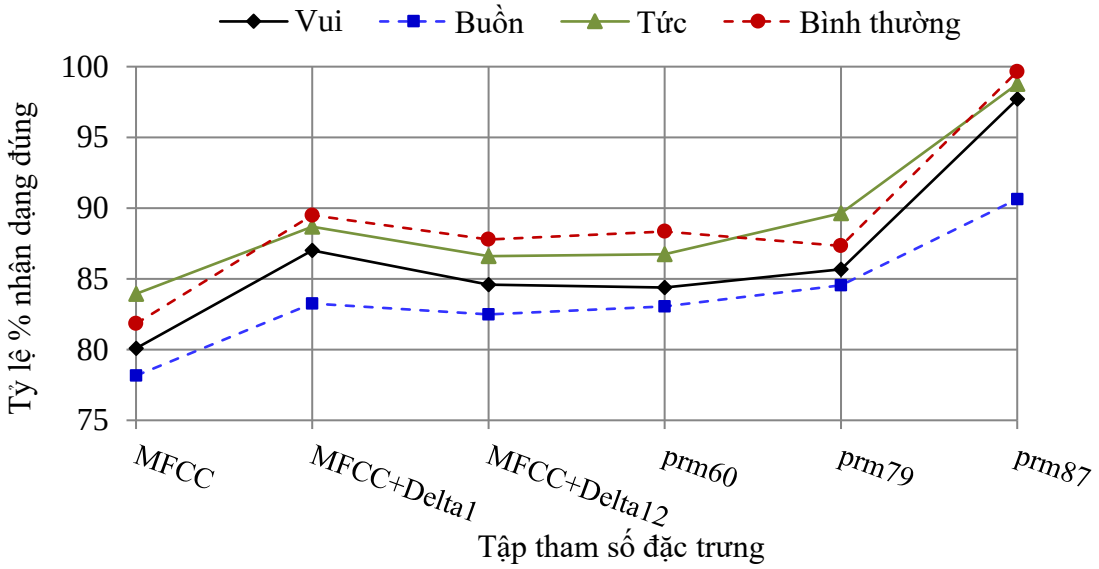


Hình 3.7 *Kết quả nhận dạng cảm xúc đối với T4*

Với tập ngữ liệu T4, Hình 3.7 là kết quả nhận dạng với các tập tham số. Tỷ lệ nhận dạng đúng cho tập tham số *prm87* cao hơn hẳn so với các tập 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 tập 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

Hình 3.8 là kết quả nhận dạng cho từng cảm xúc ứng với từng tập tham số cho T1. Kết quả thống kê cho thấy, tỷ lệ nhận dạng đúng trung bình của cảm xúc buồn là thấp nhất (83,69%). Cảm xúc vui đạt tỷ lệ nhận dạng đúng (86,57%) cao hơn cảm xúc buồn. Hai cảm xúc còn lại có tỷ lệ nhận dạng đúng trung bình cao hơn và xấp xỉ bằng nhau, trong đó cảm xúc tức có tỷ lệ nhận dạng đúng là 89,06% còn cảm xúc bình thường là 89,08%. 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.



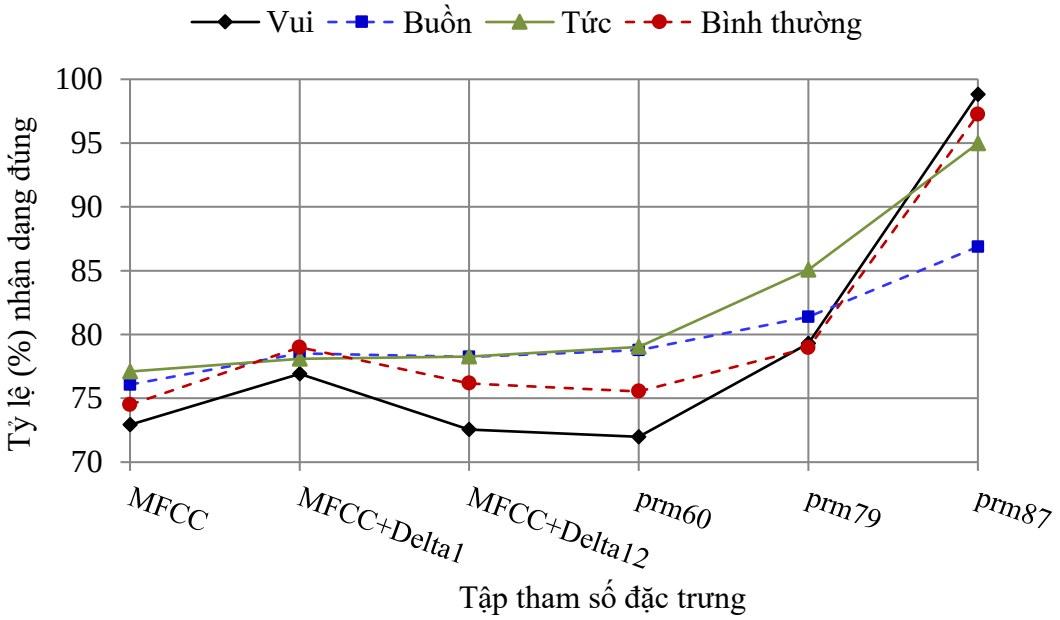
Hình 3.8 *Tỷ lệ nhận dạng đúng trung bình của bốn cảm xúc ứng với 6 tập tham số cho T1*

Khi sử dụng tập tham số *prm87* và $M = 4096$, tỷ lệ nhận nhầm giữa các cảm xúc là thấp nhất. Ma trận nhầm lẫn nhận dạng đối với các cảm xúc được thống kê trong Bảng 3.2.

Bảng 3.2 *Ma trận nhầm lẫn nhận dạng các cảm xúc với T1*

***M*=4096**	**Vui**	**Buồn**	**Tức**	**Bình thường**
Vui	**100**	0	1	0
Buồn	0	**100**	0	0,72
Tức	3,15	0,57	**99,86**	0
Bình thường	0	0,43	0	**100**

Bảng 3.2 cho thấy, tỷ lệ nhận dạng đúng trung bình của bốn cảm xúc là 99,97%, trong đó các cảm xúc vui, buồn, bình thường đều đạt 100% còn cảm xúc tức đạt 99,86%. Tỷ lệ nhận dạng nhầm lẫn từ cảm xúc tức sang vui chỉ là 3,15%. Còn lại, giữa các cặp cảm xúc khác đều có tỷ lệ nhận nhầm nhỏ hơn hoặc bằng 1%.



Hình 3.9 *Tỷ lệ nhận dạng đúng trung bình của bốn cảm xúc ứng với 6 tập tham số cho T2*

Hình 3.9 thống kê tỷ lệ nhận dạng đúng trung bình của bốn cảm xúc ứng với các tập tham số cho tập ngữ liệu T2. Kết quả cho thấy, cảm xúc vui cho tỷ lệ nhận dạng đúng thấp hơn so với ba cảm xúc còn lại khi dùng các tập tham số *MFCC, MFCC+Delta1, MFCC+Delta12, prm60, prm79*. Tuy nhiên, tỷ lệ nhận dạng đúng trung bình với cảm xúc này lại biến thiên tăng mạnh hơn so với 3 cảm xúc còn lại khi dùng tập tham số *prm87*. Với tập tham số *prm87*, cảm xúc buồn có tỷ lệ nhận dạng đúng thấp nhất so với ba cảm xúc còn lại. Tỷ lệ nhận dạng đúng cho các cảm xúc tức và bình thường tăng cao khi dùng tập tham số *prm79* và *prm87*. 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).

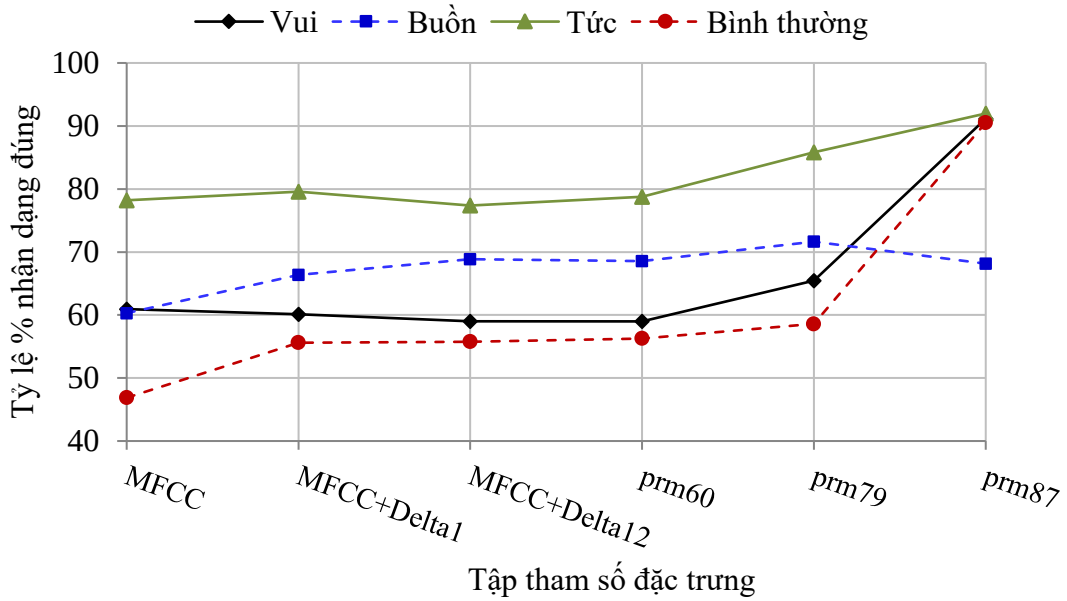
Bảng 3.3 *Ma trận nhầm lẫn nhận dạng các cảm xúc với T2*

***M*=128**	**Vui**	**Buồn**	**Tức**	**Bình thường**
Vui	**97,98**	0	0,29	0
Buồn	0	**93,83**	0	3,01
Tức	0	0,85	**85,09**	0
Bình thường	0	0,86	0	**95,11**

Cũng trong thử nghiệm này, nếu dùng tập tham số *prm87* và $M = 128$ thì tỷ lệ nhận dạng nhầm lẫn giữa các cảm xúc sẽ thấp nhất. Ma trận nhầm lẫn nhận dạng đối với cảm xúc được thống kê trong Bảng 3.3. Tỷ lệ nhận dạng đúng cao nhất là 97,98%

cho cảm xúc vui, còn thấp nhất là 85,09% cho cảm xúc tức. Tỷ lệ nhận dạng nhầm từ cảm xúc buồn sang bình thường là cao nhất và bằng 3,01%. Các trường hợp nhận nhầm còn lại đều có tỷ lệ nhầm nhỏ hơn 1%. Tính trung bình, tỷ lệ nhận dạng đúng của 4 cảm xúc là 93% còn tỷ lệ nhận nhầm là 0,42%.

Hình 3.10 là tỷ lệ nhận dạng đúng cho từng cảm xúc đối với 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%



Hình 3.10 *Tỷ lệ nhận dạng đúng trung bình của bốn cảm xúc ứng với 6 tập tham số với T3*

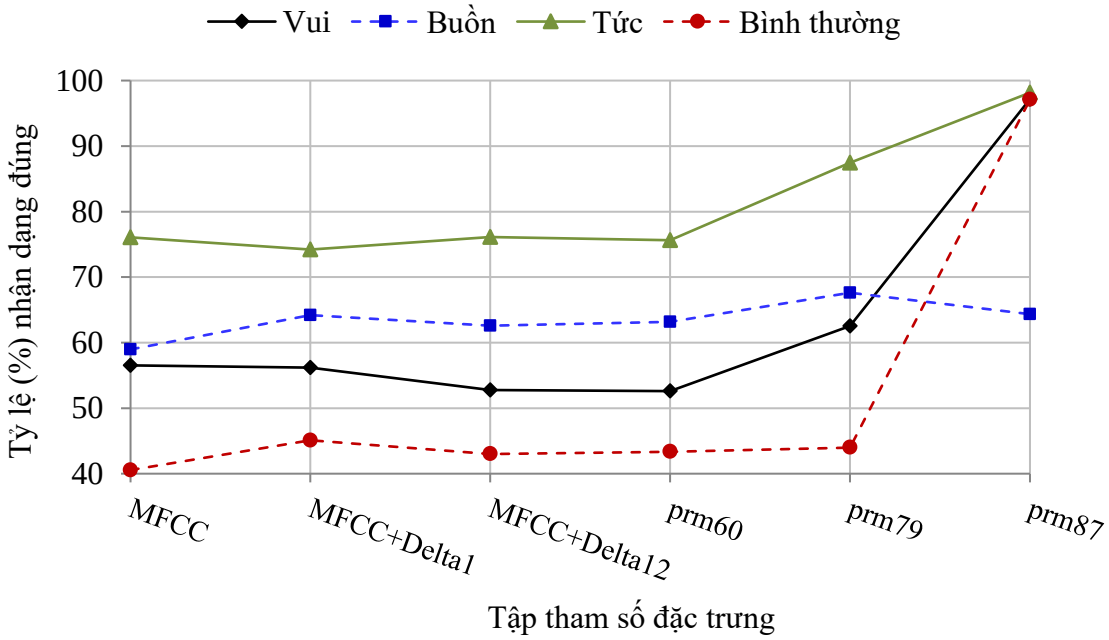
Tỷ lệ nhận dạng nhầm từ cảm xúc bình thường sang cảm xúc buồn là 23,42% và là tỷ lệ cao nhất. Tỷ lệ nhận dạng đúng trung bình của 4 cảm xúc đối với T3 là 80,54% còn trung bình tỷ lệ nhận dạng nhầm là 2,7%. Ma trận nhầm lẫn nhận dạng đối với các cảm xúc được thống kê trong Bảng 3.4 khi sử dụng tập tham số *prm87* ứng với $M = 256$.

Bảng 3.4 *Ma trận nhầm lẫn nhận dạng các cảm xúc với T3*

***M*=256**	**Vui**	**Buồn**	**Tức**	**Bình thường**
Vui	**89,6**	0	0,28	0
Buồn	0	**63,71**	3	0,29
Tức	5,49	0	**79,19**	0
Bình thường	0	23,42	0	**89,66**

Hình 3.11 là tỷ lệ nhận dạng đúng cho từng cảm xúc. Tỷ lệ nhận dạng đúng của cảm xúc tức là tốt nhất cho tất cả các bộ tham số. Tiếp theo, tỷ lệ nhận dạng đúng giảm xuống theo các cảm xúc lần lượt là buồn, vui và bình thường. Nhìn chung, tỷ lệ nhận dạng đúng các cảm xúc biến thiên ít khi dùng các tập tham số *MFCC*, *MFCC+Delta1*, *MFCC+Delta12*, *prm60* và *prm79*: vui (52,60% - 62,52%), buồn

(58,96% - 67,61%), tức (74,21% - 87,44%), bình thường (40,55% - 45,09%). Tuy nhiên, 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.



Hình 3.11 *Tỷ lệ nhận dạng đúng trung bình của bốn cảm xúc ứng với 6 tập tham số với T4*

Tỷ lệ nhận dạng nhầm các cảm xúc là thấp nhất khi dùng tập tham số *prm87* với $M = 16$. Ma trận nhầm lẫn nhận dạng đối với các cảm xúc được được thống kê trong Bảng 3.5.

Bảng 3.5 *Ma trận nhầm lẫn nhận dạng các cảm xúc với T4*

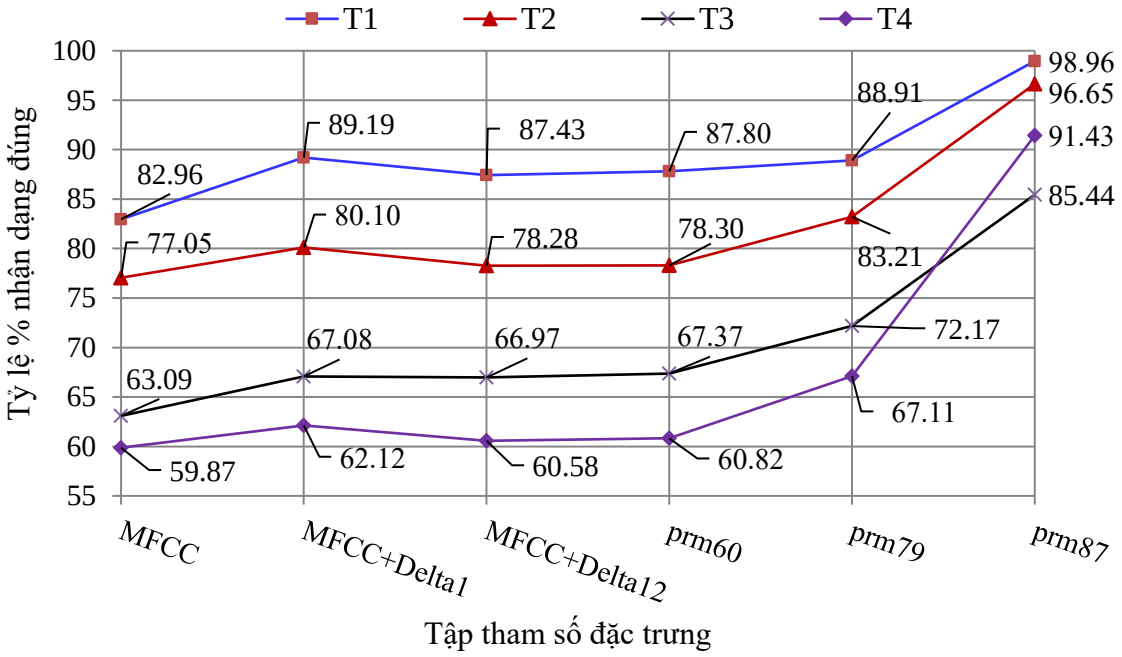
***M*=16**	**Vui**	**Buồn**	**Tức**	**Bình thường**
Vui	**97,43**	0	0	0
Buồn	0	**48,01**	0	0
Tức	1,14	0	**97,44**	0
Bình thường	0	25,43	0	**94,8**

Từ Bảng 3.5 có thể thấy, tỷ lệ nhận dạng đúng cao nhất đạt 97,44% cho cảm xúc tức, thấp nhất đạt 48,01% cho cảm xúc buồn. 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

Kết quả nhận dạng đúng trung bình của 4 tập ngữ liệu tương ứng với 6 tập tham số của 6 thử nghiệm được trình bày trên Hình 3.12.

Hình 3.12 cho thấy, tỷ lệ nhận dạng đúng trung bình của các cảm xúc đối với T1 có kết quả 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 song khác nhau ở thời điểm phát âm. Vì vậy, thông thường tỷ lệ nhận dạng sẽ đạt cao nhất. Đối với T4 là tập ngữ liệu độc lập cả người nói và nội dung, giai đoạn huấn luyện và nhận dạng có người nói, nội dung nói hoàn toàn khác nhau nên tỷ lệ nhận dạng trung bình cho T4 là thấp nhất.



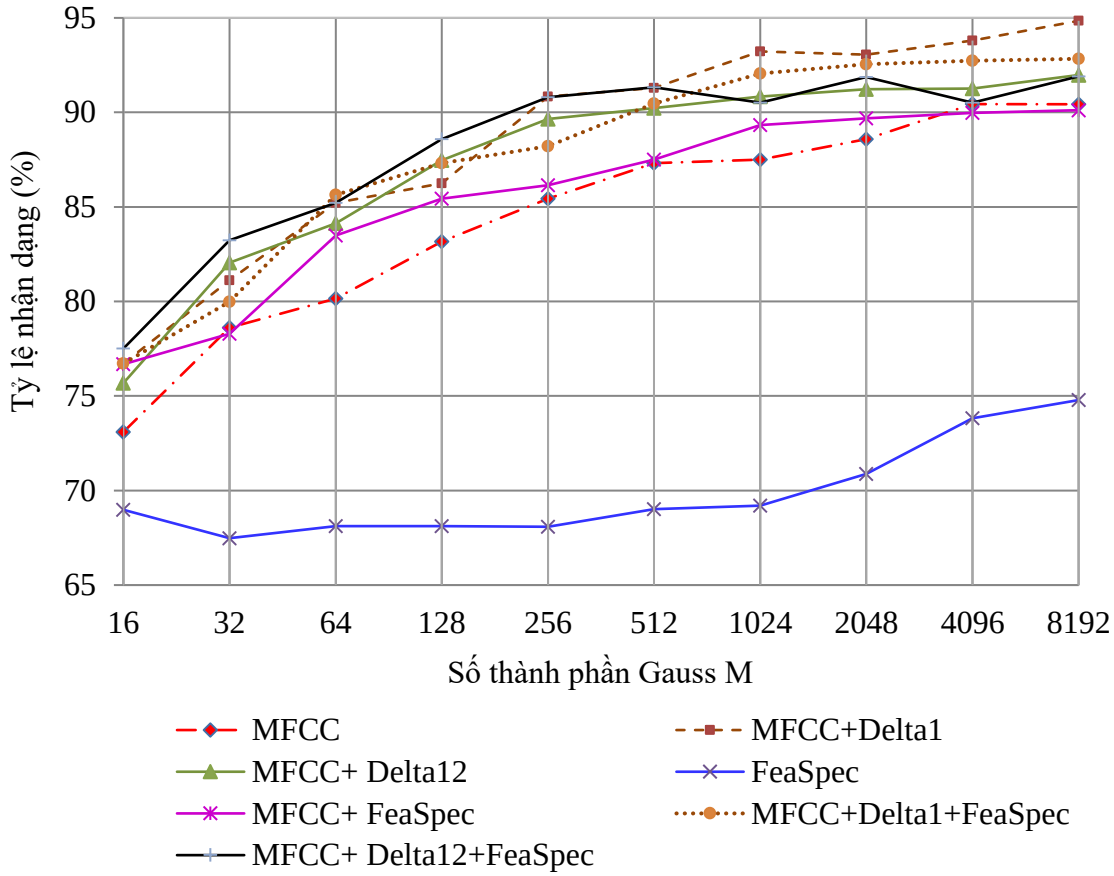
Hình 3.12 *Tỷ lệ nhận dạng đúng trung bình cảm xúc của 4 tập ngữ liệu trong 6 thử nghiệm*

3.3.2 Thử nghiệm 7 đến Thử nghiệm 10

Các thử nghiệm này thực hiện nhận dạng cảm xúc sử dụng *MFCC* và các đặc trưng liên quan đến *MFCC* kết hợp với các đặc trưng phổ để xem xét ảnh hưởng của đặc trưng phổ đến nhận dạng cảm xúc. 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*. Để so sánh kết quả nhận dạng của các tập tham số vừa nêu này với các tập tham số chỉ sử dụng *MFCC* và các đặc trưng liên quan đến *MFCC*, ở đây cũng lấy lại kết quả nhận dạng sử dụng *MFCC* và các tham số liên quan đến *MFCC* đã được trình bày ở mục 3.3.1.

Hình 3.13 đã trình bày kết quả nhận dạng của 7 tập tham số khác nhau. Bẩy tập tham số này bao gồm: *MFCC, MFCC+Delta1, MFCC+Delta12, FeaSpec, MFCC+FeaSpec, MFCC+Delta1+FeaSpec, MFCC+Delta12+FeaSpec* và có thể chia thành 2 nhóm (Nhóm 1: chỉ sử dụng các tham số liên quan đến *MFCC* hoặc chỉ đặc trưng phổ, Nhóm 2: Kết hợp các tham số liên quan đến *MFCC* và các đặc trưng phổ).

Kết quả đạt được của Nhóm 1 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+Delta1*, tỷ lệ nhận dạng chính xác tăng lên và trong khoảng 76,72% - 93,8%. Với trường hợp dùng *MFCC+Delta12*, tỷ lệ nhận dạng chính xác trong khoảng 75,68% - 91,26%. Nếu chỉ dùng các tham số đặc trưng phổ (*FeaSpec*), tỷ lệ nhận dạng sẽ thấp hơn và có giá trị từ 67,48% - 73,82%.



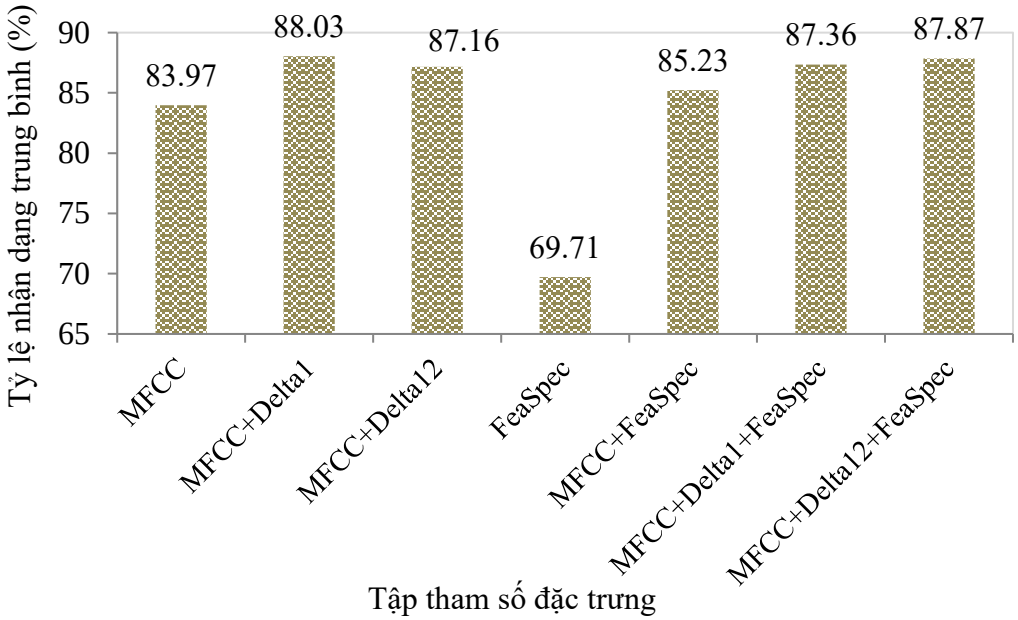
Hình 3.13 *Tỷ lệ nhận dạng sử dụng MFCC và các đặc trưng phổ với T1*

Đối với Nhóm 2, tỷ lệ nhận dạng chính xác đạt từ 76,68% - 89,97% khi sử dụng *MFCC+FeaSpec*. Tỷ lệ này là từ 76,72% - 92,73% khi sử dụng *MFCC+Delta1+FeaSpec* và từ 77,51% - 91,87% khi sử dụng *MFCC+Delta12+FeaSpec*.

Hình 3.14 cũng cho thấy, nhìn chung khi số thành phần Gauss *M* tăng, 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+Delta12+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+Delta1* cao hơn so với các trường hợp còn lại.

Hình 3.14 thống kê tỷ lệ nhận dạng đúng trung bình cho 7 tập tham số đã nói ở trên. 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*.

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*.



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

3.3.3 Thử nghiệm 11

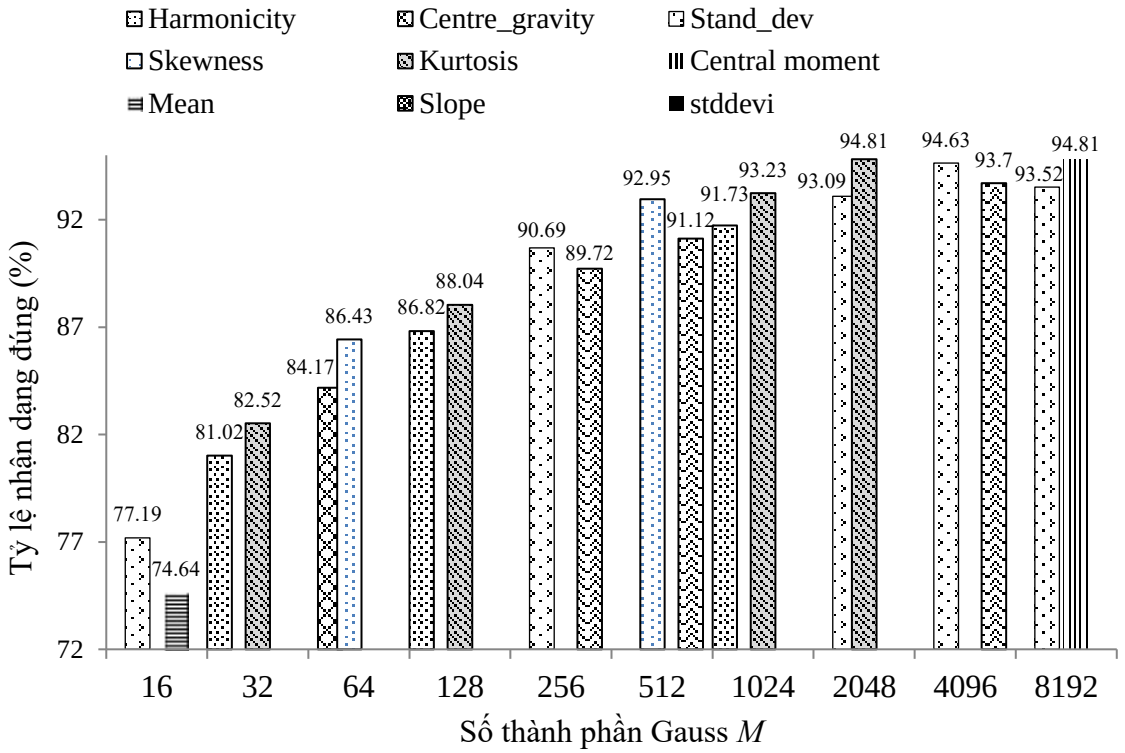
Thử nghiệm 11 cũng được thực hiện nhận dạng với tập ngữ liệu T1 nhằm xem xét ảnh hưởng của từng đặc trưng phổ đối với nhận dạng cảm xúc tiếng Việt. Kết quả nhận dạng khi kết hợp các đặc trưng MFCC với đặc trưng phổ trên Hình 3.14 ở mục 3.3.2 cho thấy kết quả nhận dạng cao nhất khi sử dụng *MFCC+Delta1*. Do vậy, Thử nghiệm 11 sẽ chỉ sử dụng *MFCC+Delta1* kết hợp với từng đặc trưng phổ để xét riêng ảnh hưởng của mỗi đặc trưng phổ. Kết quả nhận dạng trong thử nghiệm này 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 3.15 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 chỉ biểu diễn tỷ lệ nhận dạng đúng cao nhất và thấp nhất tương ứng với đặc trưng phổ.

Từ Hình 3.15 có thể nhận xét:

- 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 (*standard deviation*) 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.

- Đặc trưng mômen trung tâm của phổ chỉ xuất hiện một lần có tỷ lệ nhận dạng cao nhất ứng với $M = 8192$.
- Đặc trưng độ lệch chuẩn của LTAS (*stddevi*) 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.



Hình 3.15 *Tỷ lệ nhận dạng đúng cao nhất và thấp nhất tương ứng với đặc trưng phổ cho các giá trị của M*

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.

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

Kết quả đánh giá ảnh hưởng của từng đặc trưng phổ khi kết hợp mỗi đặc trưng này với *MFCC+Delta1* được trình bày ở Bảng 3.6. Bảng 3.6 cho thấy, khi đặc trưng *kurtosis* của phổ được kết hợp với *MFCC+Delta1*, 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 *MFCC+Delta1* 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%.

Như đã trình bày trong Chương 2, phương pháp thống kê ANOVA và kiểm định *T* đã được sử dụng để đánh giá và kết quả 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. Tất cả các tham số đặc trưng phổ khi kết hợp với *MFCC+Delta1* đều cho tỷ lệ nhận dạng tương đối ngang nhau với từng cảm xúc. Tỷ lệ nhận dạng đúng trên 90% đối với cảm xúc buồn và trên 84% với các cảm xúc còn lại.

Trong số các đặc trưng phổ *harmonicity, centre of gravity, standard deviation, skewness, kurtosis, mean, slope* và s*tandard deviation of LTA*, 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, 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.

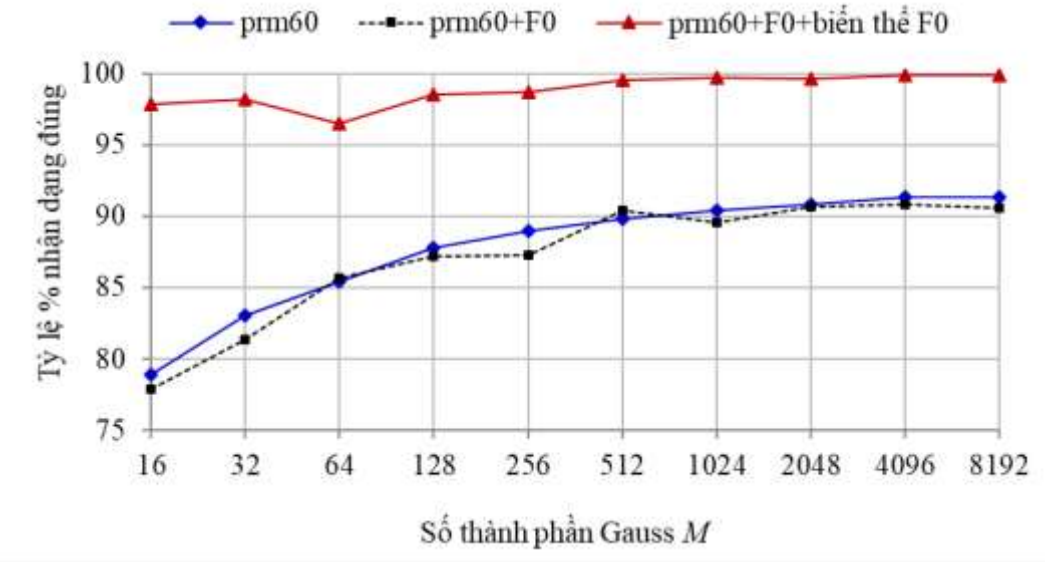
3.3.4 Thử nghiệm 12

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.

Đối với tập ngữ liệu T1, kết quả nhận dạng được trình bày trên Hình 3.16. Khi sử dụng tập tham số *pm60+F0+8* biến thể của *F*0, tỷ lệ nhận dạng cao hơn hẳn so với chỉ dùng *prm60* hoặc *prm60+F0*.

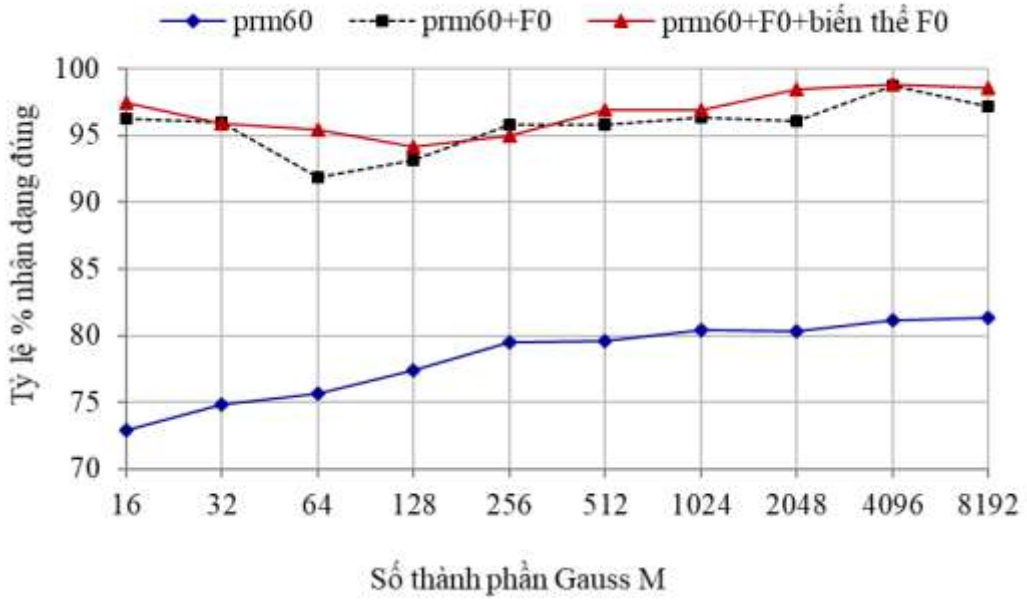
Độ chính xác nhận dạng sử dụng tập tham số *pm60+F0+8* biến thể của *F*0 đã đạt trung bình từ 96,49% đến 99,93%. Nếu chỉ dùng *prm60+F0*, tỷ lệ này tăng ít và gần

như xấp xỉ bằng tỷ lệ của *prm60*. Điều này cho thấy, các biến thể của *F*0 có ảnh hưởng rất lớn đến tỷ lệ nhận dạng cảm xúc tiếng Việt và đã cải thiện đáng kể tỷ lệ nhận dạng đối với trường hợp ngữ liệu phụ thuộc cả người nói và nội dung.



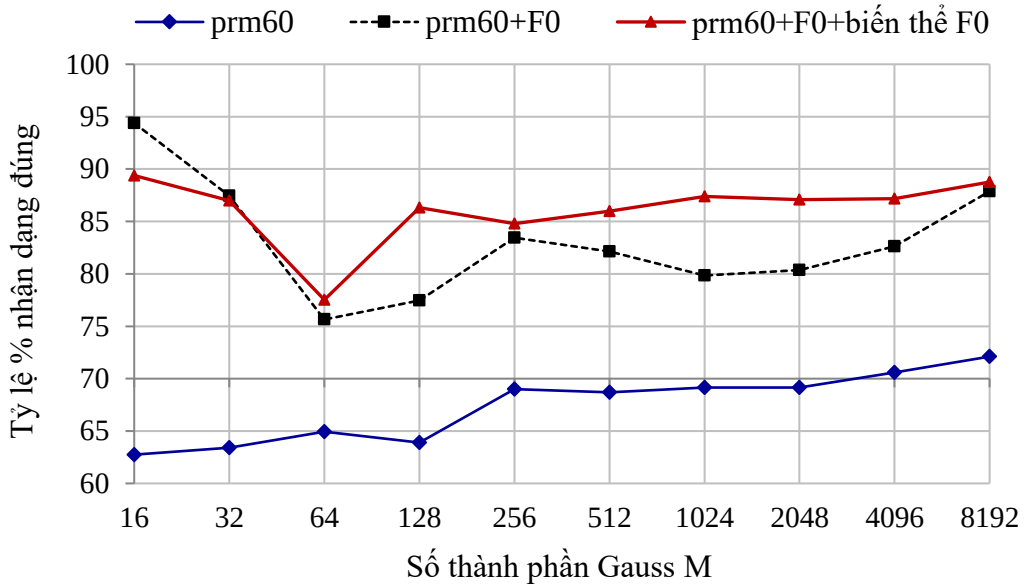
Hình 3.16 *Tỷ lệ nhận dạng đúng trung bình khi kết hợp prm60+F0+các biến thể của F0 đối với T1*

Đối với tập ngữ liệu T2, Hình 3.17 thấy rằng tỷ lệ nhận dạng khi sử dụng *F*0 và các biến thể của *F*0 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 với T2 cũng khẳng định *F*0 và các biến thể của *F*0 có ảnh hưởng rất lớn, cải thiện đáng kể tỷ lệ nhận dạng cảm xúc tiếng Việt.



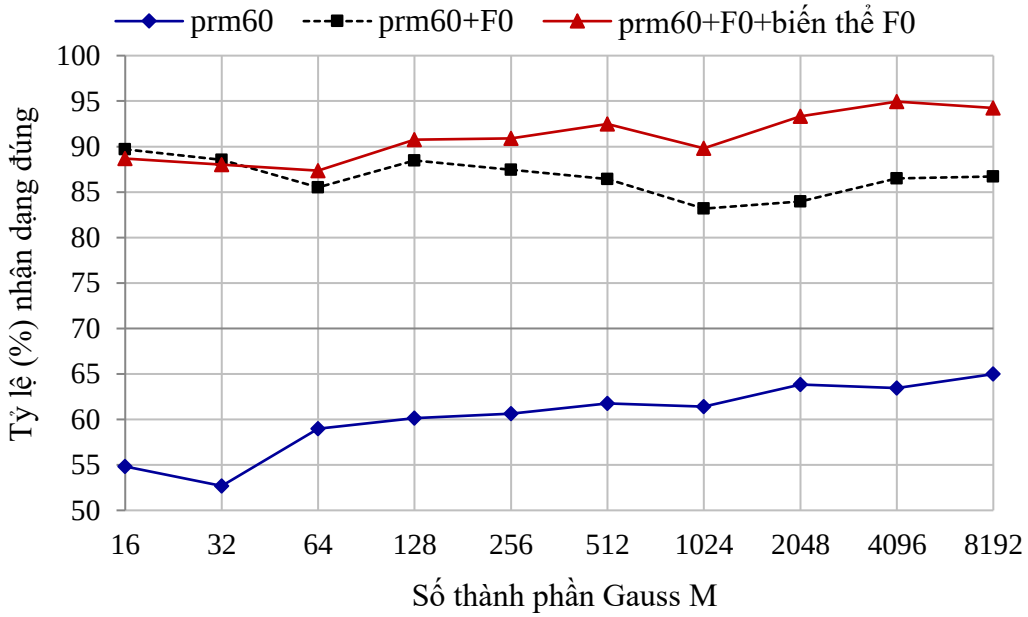
Hình 3.17 *Tỷ lệ nhận dạng đúng trung bình khi kết hợp prm60+F0+các biến thể của F0 đối với T2*

Hình 3.18 là kết quả nhận dạng đối với tập ngữ liệu T3, 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 $M = 16$.



Hình 3.18 *Tỷ lệ nhận dạng đúng trung bình khi kết hợp prm60+F0+các biến thể của F0 đối với T3*

Hình 3.19 là kết quả thử nghiệm nhận dạng đối với tập ngữ liệu T4. Thử nghiệm kết hợp *prm60* với $F0$ và biến thể của $F0$ cũng cho tỷ lệ 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%. Điều này cho thấy $F0$ và biến thể của $F0$ có vai trò quan trọng đối với phân biệt các cảm xúc tiếng Việt.



Hình 3.19 *Tỷ lệ nhận dạng đúng trung bình khi kết hợp prm60+F0+các biến thể của F0 đối với T4*

Kết quả nhận dạng đúng trung bình đối với 4 tập ngữ liệu khi sử dụng kết hợp tập tham số *prm60* với $F0$ và biến thể $F0$ được thống kê trong Bảng 3.7.

Bảng 3.7 *Tỷ lệ nhận dạng đúng trung bình đối với 4 tập ngữ liệu khi kết hợp prm60 với $F0$ và biến thể $F0$*

Tập ngữ liệu	Tỷ lệ nhận dạng (%)		
	prm60	*prm60+F0*	*prm60+F0+biến thể F0*
T1	87,80	87,15	98,86
T2	78,30	95,73	96,76
T3	67,37	83,12	86,14
T4	60,27	86,64	91,05

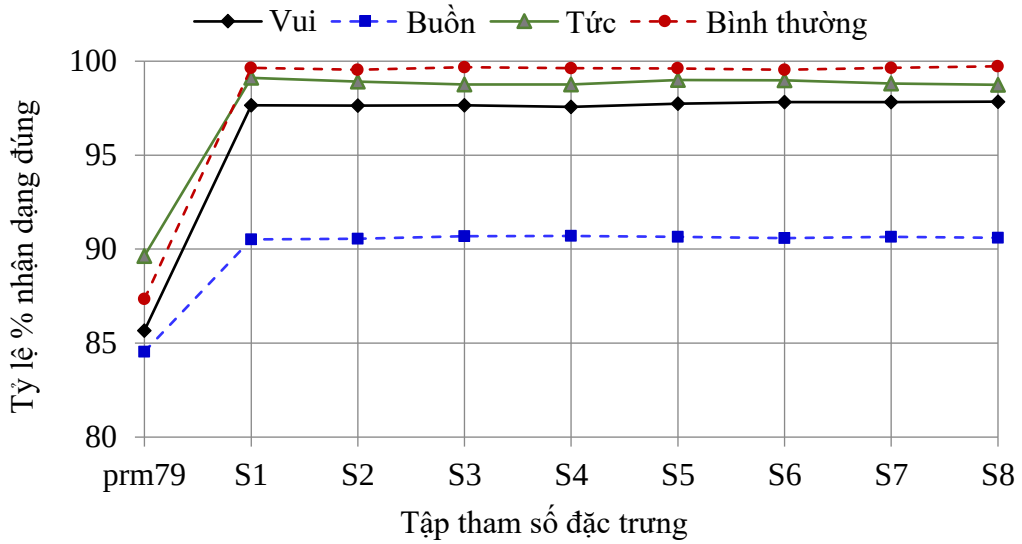
3.3.5 Thử nghiệm 13

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. Các kết quả nhận dạng đối với với 8 tập tham số này được so sánh với kết quả nhận dạng chỉ sử dụng *prm79* trong Thử nghiệm 5 đã trình bày ở mục 3.3.1.

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

Tập tham số	Các tham số	Ghi chú	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

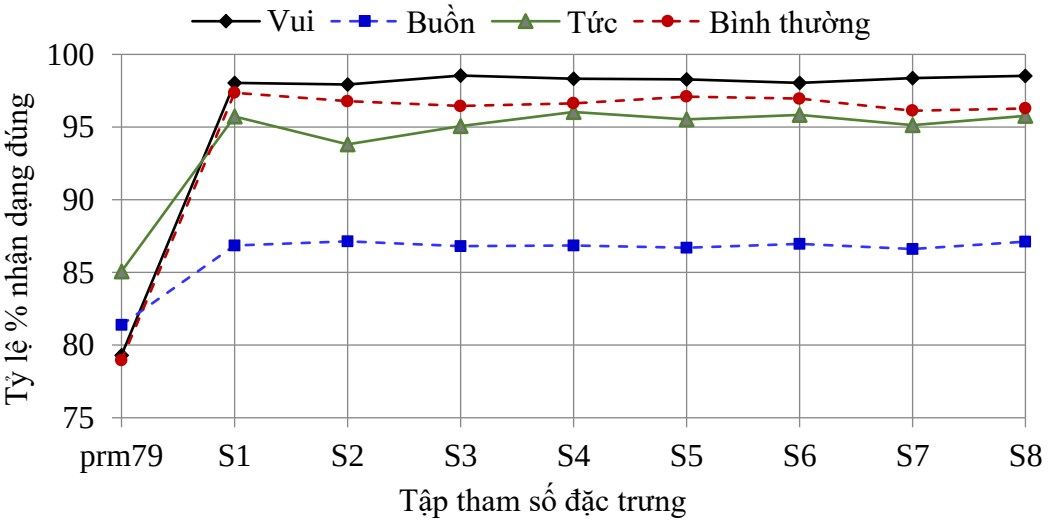
Hình 3.20 là kết quả xét ảnh hưởng của các tập tham số từ S1 đến S8 đối với các cảm xúc trong đó thống kê tỷ lệ nhận dạng đúng trung bình khi kết hợp tập tham số *prm79* với từng biến thể của *F*0 cho T1. Kết quả cho thấy, khi một trong 8 biến thể *F*0 được thêm vào tập tham số prm79 thì tỷ lệ nhận dạng cho các cảm xúc được tăng lên đáng kể (trên 90%), đặc biệt là 3 cảm xúc: bình thường, tức và vui.



Hình 3.20 *Tỷ lệ nhận dạng đúng trung bình của các cảm xúc cho từng tập tham số đối với T1*

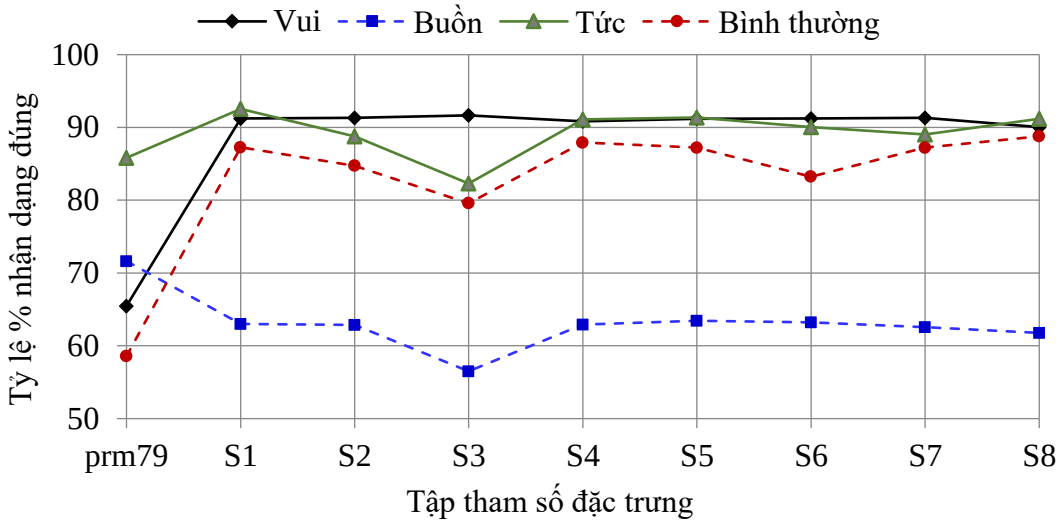
Cảm xúc bình thường luôn cho tỷ lệ cao hơn ba cảm xúc còn lại và đạt xấp xỉ 100% với cùng tập tham số từ S1 đến S8. Nếu xét tỷ lệ nhận dạng đối với mỗi cảm xúc thì cả 8 biến thể *F*0 đều cho tỷ lệ nhận dạng ổn định trên tập T1 và cao hơn tỷ lệ nhận dạng chỉ dùng *prm79*.

Từ Hình 3.21 có thể thấy, đối với T2 cảm xúc vui có tỷ lệ nhận dạng đúng cao nhất, cảm xúc buồn có tỷ lệ nhận dạng thấp nhất. Ảnh hưởng của các tập tham số xét trên từng cảm xúc cũng tương đối ổn định và cao hơn hẳn so với chỉ dùng *prm79*.



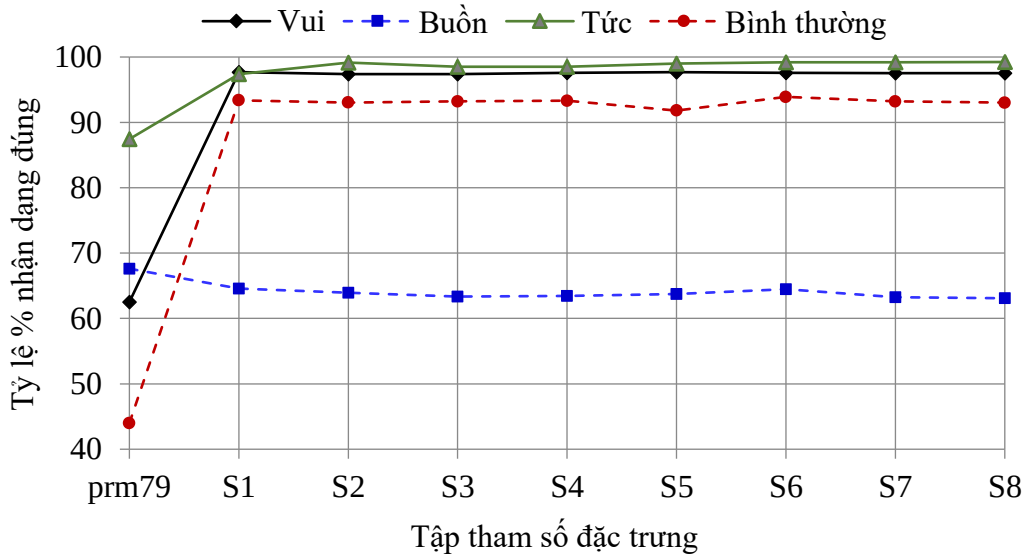
Hình 3.21 *Tỷ lệ nhận dạng đúng trung bình của các cảm xúc cho từng tập tham số đối với T2*

Hình 3.22 là kết quả thống kê tỷ lệ nhận dạng cảm xúc với từng cảm xúc cho tập ngữ liệu T3. Kết quả nhận dạng đối với cảm xúc buồn khi sử dụng tập tham số từ S1 đến S8 thấp hơn so với kết quả nhận dạng sử dụng *prm79*. Cảm xúc vui có tỷ lệ nhận dạng ổn định hơn khi sử dụng các biến thể khác nhau của $F0$. Với ba cảm xúc buồn, tức và bình thường, tập tham số S3 ($prm79{+}F0NormMinMax$) có ảnh hưởng ít hơn so với các tập tham số còn lại.



Hình 3.22 *Tỷ lệ nhận dạng đúng trung bình của các cảm xúc ứng cho từng tập tham số đối với T3*

Đối với T4, khi thêm một trong 8 biến thể $F0$, kết quả nhận dạng của ba cảm xúc: vui, tức và bình thường đạt tỷ lệ cao hơn so với *prm79* (Hình 3.23). Đặc biệt đối với cảm xúc bình thường, tỷ lệ nhận dạng đã tăng vọt từ (nhỏ hơn 50%) lên (trên 90%). Đối với cảm xúc vui và tức tỷ lệ nhận dạng đã đạt xấp xỉ 100%. Tỷ lệ nhận dạng của các cảm xúc cũng tương đối ổn định khi sử dụng các tập tham số từ S1 đến S8.



Hình 3.23 *Tỷ lệ nhận dạng đúng trung bình của các cảm xúc cho từng tập tham số đối với T4*

Như vậy, khi thêm một trong 8 biến thể của $F0$ vào tập tham số *prm79*, tỷ lệ nhận dạng các cảm xúc hầu như được nâng lên đáng kể nhất là đối với cảm xúc vui, tức và bình thường. Riêng cảm xúc buồn giảm nhẹ đối với tập ngữ liệu T3 và T4.

Kết quả nhận dạng đúng trung bình đối với 4 tập ngữ liệu khi sử dụng kết hợp *prm79* với từng biến thể $F0$ được thống kê trong Bảng 3.9.

Bảng 3.9 *Tỷ lệ nhận dạng đúng trung bình đối với 4 tập ngữ liệu khi kết hợp prm79 với từng 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

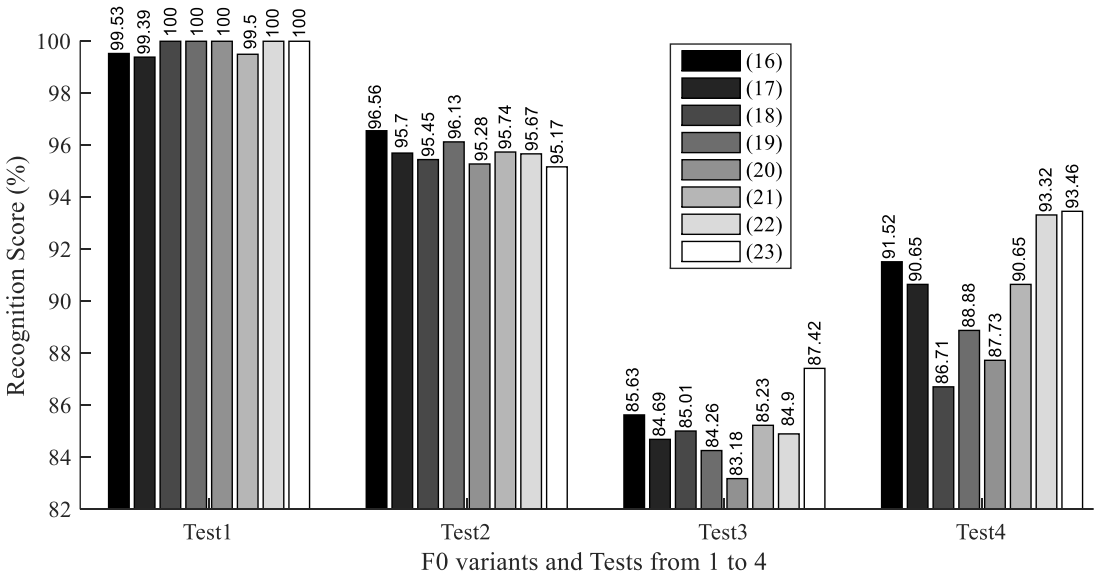
3.4 Đánh giá sự ảnh hưởng của tần số cơ bản

Các nghiên cứu thử nghiệm đã trình bày ở mục 3.3 trên đây cho thấy, tần số cơ bản có tầm ảnh hưởng rất lớn đến kết quả nhận dạng các cảm xúc tiếng Việt. Việc bổ sung tham số *F0* và biến thể của *F0* đã cho tỷ lệ nhận dạng tăng lên đáng kể. Cụ thể, Thử nghiệm 6 đã thực hiện nhận dạng cảm xúc với tập tham số *prm87* và cho thấy, khi các tham số liên quan trực tiếp đến $F0$ được thêm vào, tỷ lệ nhận dạng đối với các tập ngữ liệu đều tăng cao so với việc bổ sung các tham số liên quan trực tiếp đến phổ. Kết quả của Thử nghiệm 6 đã được trình bày trên Hình 3.12 trong mục 3.3.1.3. Tỷ lệ nhận dạng trung bình tăng mạnh nhất đối với T4 là 24,32%. Mức tăng nhỏ nhất là 10,05% đối với T1. Tuy nhiên, sự gia tăng nhỏ nhất trong trường hợp này vẫn lớn hơn mức tăng tối đa trong trường hợp bổ sung đặc trưng phổ (6.29% đối với T4).

Kết quả trong Thử nghiệm 12 cũng chỉ ra, 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*. Đặc biệt với T4, nếu chỉ dùng *prm60* thì tỷ lệ nhận dạng đạt 60,27% còn khi dùng *prm60+F0+biến thể F0* tỷ lệ nhận dạng đạt 91,05%.

Các kết quả nhận dạng đối với từng cảm xúc cho 4 tập ngữ liệu được trình bày trong Thử nghiệm 13 cũng cho kết quả nhận dạng tốt khi kết hợp tập tham số *prm79* với một trong 8 biến thể của $F0$.

Nếu chỉ xét riêng trường hợp nhận dạng với số thành phần Gauss $M = 512$ và sử dụng từng biến thể $F0$ kết hợp với tập tham số *prm79* như trình bày trên Hình 3.26 thì cũng có kết quả tương tự. Việc lấy giá trị $M = 512$ được dựa trên các kết quả thử nghiệm đã trình bày ở trên. Giá trị này của *M* có thể được coi là điểm nằm trong dải tỷ lệ nhận dạng chuyển từ tăng nhanh sang tỷ lệ nhận dạng tăng chậm hơn khi tăng *M*.



Hình 3.24 *Tỷ lệ nhận dạng trung bình cả 4 cảm xúc theo từng biến thể F0 và prm79 cho các tập ngữ liệu T1 đến T4, với M=512.*

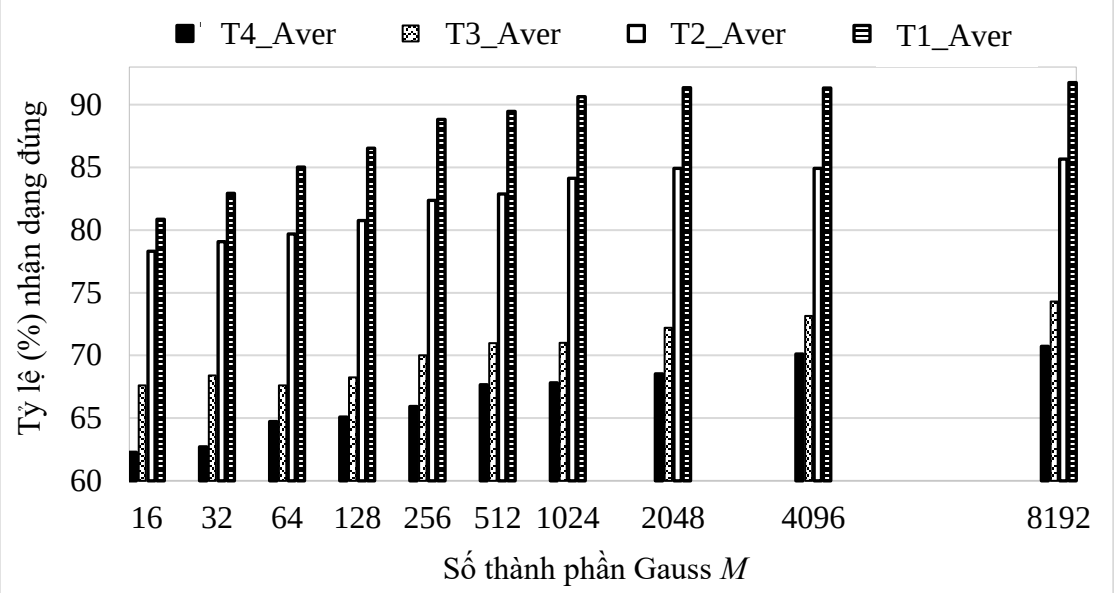
Hình 3.24 cũng cho thấy, ảnh hưởng của các biến thể $F0$ không giống nhau đối với mỗi tập ngữ liệu. Với T1, các biến thể $F0$ (18), (19), (20), (22) và (23) đã cho tỷ lệ nhận dạng tăng lên tối đa và đạt 100%. Tương tự như vậy, T1, T3 và T4 có tỷ lệ nhận dạng cao nhất khi thêm biến thể $F0$ (23) và tỷ lệ này lần lượt là 100%, 87,42% và 93,46%. Trong khi đó, biến thể $F0$ (23) có ảnh hưởng ít nhất đến T2 so với các biến thể còn lại. T2 có tỷ lệ nhận dạng cao nhất và bằng 96,56% đối với trường hợp thêm biến thể $F0$ (16). Tương ứng với T1, T2, T3, và T4, các biến thể $F0$ (17), (23), (20), và (18) có hiệu suất thấp nhất.

Sự gia tăng đáng kể tỷ lệ nhận dạng cảm xúc tiếng Việt khi bổ sung $F0$ và các biến thể $F0$ là hoàn toàn hợp lý vì tần số cơ bản đóng một vai trò rất quan trọng trong ngôn ngữ có thanh điệu như tiếng Việt và tần số cơ bản cũng tham gia tích cực vào biểu thị cảm xúc.

Trên đây luận án đã trình bày các kết quả nhận dạng bốn cảm xúc cơ bản vui, buồn, tức và bình thường với bốn trường hợp tùy theo sự độc lập hay phụ thuộc của người nói và nội dung. Việc nhận dạng dựa trên mô hình GMM cũng được tiến hành với các tập tham số khác nhau. Kết quả cho thấy, tỷ lệ nhận dạng đúng cao nhất khi ngữ liệu phụ thuộc người nói và phụ thuộc nội dung. Tỷ lệ nhận dạng đúng là thấp nhất trong trường hợp ngữ liệu độc lập cả nội dung và người nói. Với ngữ liệu phụ thuộc người nói nhưng độc lập về nội dung và ngữ liệu phụ thuộc nội dung nhưng độc lập người nói, tỷ lệ nhận dạng đúng nằm ở mức trung gian giữa hai trường hợp có tỷ lệ nhận dạng đúng cao nhất và thấp nhất này. Nhận dạng phụ thuộc người nói, độc lập nội dung có tỷ lệ cao hơn nhận dạng độc lập người nói, phụ thuộc nội dung. Với tất cả bốn tập ngữ liệu T1, T2, T3 và T4, tập tham số *prm87* luôn cho tỷ lệ nhận dạng cao hơn cả. Điều này cho thấy tần số cơ bản và các biến thể của nó là những đặc trưng của tín hiệu tiếng Việt nói có ảnh hưởng rất lớn, làm tăng độ chính xác nhận dạng cảm xúc tiếng Việt.

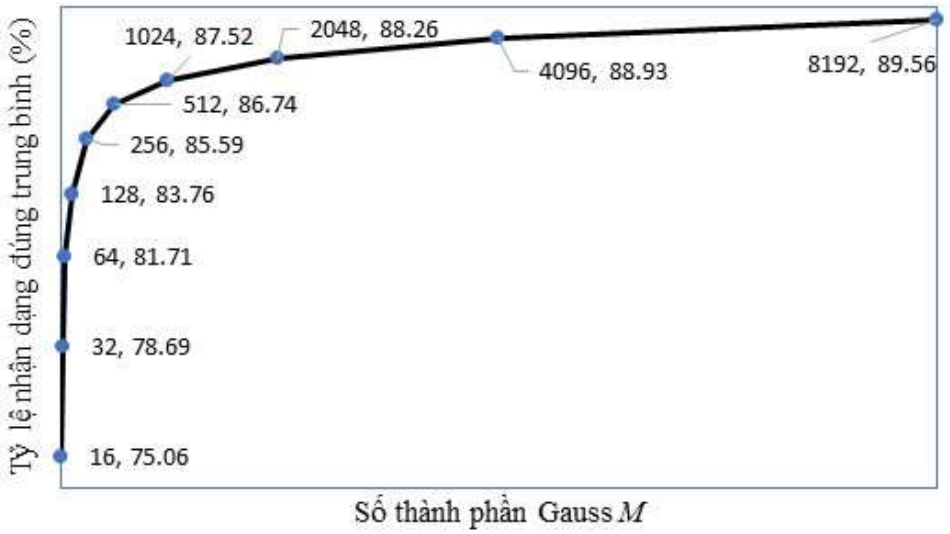
3.5 Quan hệ giữa số thành phần Gauss M và tỷ lệ nhận dạng

Trong mục này, luận án trình bày mối quan hệ giữa số thành phần Gauss *M* và tỷ lệ nhận dạng cảm xúc. 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.



Hình 3.25 *Quan hệ giữa số thành phần Gauss M và tỷ lệ nhận dạng đúng trung bình của Thử nghiệm từ 1 đến 6 với 4 tập ngữ liệu*

Hình 3.25 là mối quan hệ giữa số thành phần Gauss *M* và tỷ lệ nhận dạng trung bình của bốn tập ngữ liệu T1, T2, T3, T4 trong các Thử nghiệm từ 1 đến 6. Hình 3.25 cho thấy, với giá trị *M* thấp (giữa 16 và 512), tỷ lệ nhận dạng tăng đáng kể, khi *M* tăng từ 512 lên 8192, tỷ lệ nhận dạng trung bình tăng rất ít.



Hình 3.26 *Quan hệ giữa số thành phần Gauss M và tỷ lệ nhận dạng đúng trung bình các Thử nghiệm từ 1 đến 3 và từ 7 đến 10 với T1.*

Hình 3.26 là mối quan hệ giữa số thành phần Gauss M và tỷ lệ nhận dạng đúng trung bình các Thử nghiệm từ 1 đến 3 và từ 7 đến 10 với T1cũng cho kết quả tương

tự. Với giá trị M thay đổi từ 16 đến 256, tỷ lệ nhận dạng đúng trung bình tăng 10,53% nhưng khi M tăng từ 512 đến 8192, tỷ lệ nhận dạng đúng trung bình tăng chỉ 2,82%.

Như vậy, 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 [6]. 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 các mô hình nhận dạng GMM cùng với các tập tham số đặc trưng khác nhau.

Từ kết quả nhận dạng có thể thấy rằng, 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 tập tham số *prm87*. Đối với ngữ liệu độc lập cả người nói và nội dung, tỷ lệ nhận dạng đạt 97,58% khi sử dụng tập tham số *prm79* kết hợp với biến thể *LogF0NormMinMax* của $F0$. Kết quả nhận dạng với cảm xúc buồn luôn thấp hơn ba cảm xúc còn lại. Cặp cảm xúc hay nhầm lẫn với nhau là buồn-bình thường và vui-tức. Với những kết quả nhận dạng đã được 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 $F0$.

Tiếp theo trong chương 4 sẽ trình bày về nhận dạng cảm xúc tiếng Việt nói sử dụng mô hình DCNN sâu. Đây là một kỹ thuật nhận dạng mới được sử dụng trong những năm gần đây và đã mang lại hiệu quả tốt khi khai thác mạng nơron học sâu.

Các kết quả nghiên cứu chính của chương 3 đã được công bố trong các bài báo số 1, 3, 6, 7 trong danh mục các công trình nghiên cứu của luận án:

1. *Nghiên cứu và thử nghiệm nhận dạng phương ngữ tiếng Việt*, Tạp chí Khoa học và Công nghệ, ĐHSPKT Hưng Yên, số 4, ISSN 2354-0575, trang 96-101.
3. *Cảm xúc trong tiếng nói và phân tích thống kê ngữ liệu cảm xúc tiếng Việt*, Chuyên san Các công trình Nghiên cứu, Phát triển và Ứng dụng Công nghệ Thông tin, Tạp chí Bưu chính Viễn thông, tập V-1, số 15 (35), trang 86-98.
6. *Ảnh hưởng của đặc trưng phổ tín hiệu tiếng nói đến nhận dạng cảm xúc tiếng Việt*, Kỷ yếu Hội nghị khoa học công nghệ quốc gia lần thứ X, Nghiên cứu cơ bản và ứng dụng công nghệ thông tin, Đà Nẵng, trang 36-43.
7. *GMM for emotion recognition of Vietnamese*, Journal of Computer Science and Cybernetics, V.33, N.3, pp.229-246.

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

Chương 3 đã trình bày các nghiên cứu nhận dạng cảm xúc tiếng Việt theo mô hình GMM. Với GMM, đây là mô hình mang tính truyền thống không tốn nhiều thời gian để huấn luyện và nhận dạng và chỉ cần xác định mô hình theo bộ ba tham số (các véctơ kỳ vọng, các ma trận hiệp phương sai và các trọng số cho *M* thành phần). Trong khi đó, với mạng nơron nói chung, kiến trúc mạng nơron rất phong phú nên khả năng khai thác mô hình mạng nơron cho các ứng dụng là rất lớn.

Trong số các mạng nơron, Convolutional Neural Network (CNN – Mạng nơron lấy chập) là một trong những mô hình Deep Learning tiên tiến giúp xây dựng được những hệ thống thông minh với độ chính xác cao hiện nay như hệ thống lớn xử lý ảnh Facebook, Google hay Amazon đã đưa vào sản phẩm của mình những chức năng thông minh nhận diện khuôn mặt người dùng, phát triển xe hơi tự lái hay máy giao hàng tự động... CNN được sử dụng nhiều trong các bài toán nhận dạng các đối tượng trong ảnh. Chương này trình bày các kết quả thử nghiệm nhận dạng cảm xúc tiếng Việt sử dụng mô hình DCNN.

4.1 Mô hình mạng nơron lấy chập

Mạng nơron lấy chập CNN là một trong những thuật toán học sâu cho kết quả tốt nhất hiện nay trong hầu hết các bài toán về thị giác máy như phân lớp, nhận dạng. Đã có nhiều công trình nghiên cứu ứng dụng mô hình CNN trong nhiều lĩnh vực khác nhau như nhận dạng hình ảnh [195], xử lý tín hiệu tiếng nói [196], nhận dạng cảm xúc theo gương mặt [197], nhận dạng người nói [198], nhận dạng cảm xúc tiếng nói [199], [200], [201], [202], [203], [204], [205], [206] cũng như trong nhiều nhiệm vụ phân tích dữ liệu lớn [207], [208].

Trong [209], các tác giả đã sử dụng DCNN 3 lớp để nhận dạng 7 cảm xúc của ngữ liệu tiếng Đức: vui, buồn, tức, sợ hãi, ghê tởm, chán nản, bình thường. Kết quả nhận dạng đúng trung bình các cảm xúc đạt 56,38%.

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. Phép lấy chập dựa trên lý thuyết xử lý tín hiệu số, thực hiện các xử lý về mặt toán học tính lấy chập để giúp trích xuất được những thông tin quan trọng từ dữ liệu. Đầu vào của phép lấy chập là một mảng các giá trị của dữ liệu. Chẳng hạn, trong phân

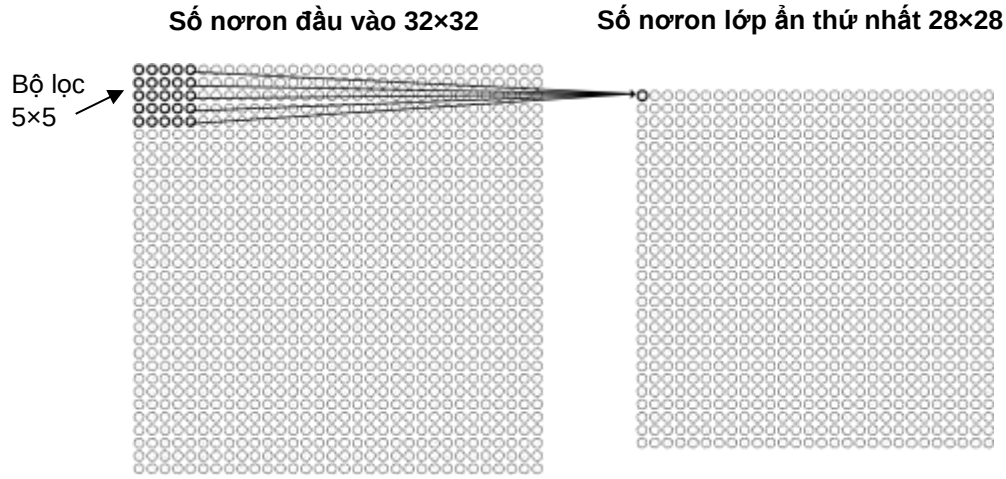
loại ảnh, đầu vào là một ảnh được biểu diễn bằng mảng 32×32×3 các giá trị pixel (mỗi phần tử của mảng có giá trị trong khoảng từ 0 đến 255 biểu diễn cường độ sáng của pixel tại một điểm).

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. Bộ lọc này thực chất là một cửa sổ có kích thước $n \times n$ (kí hiệu $F = n$) chứa các số (các số này chính là trọng số hay tham số). Kích thước của bộ lọc thường là nhỏ (chẳng hạn 3×3 hoặc lớn nhất là 5×5). Bộ lọc sẽ di chuyển từ trái qua phải, từ trên xuống dưới với bước dịch chuyển $S = 1$ cho cả hai chiều, vị trí đầu tiên của bộ lọc là góc trên bên trái. 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:

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

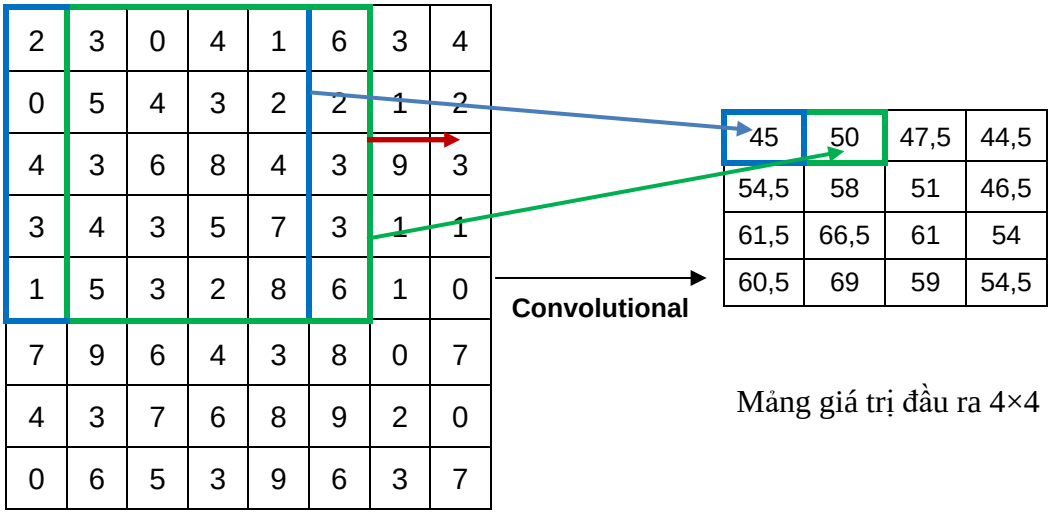
Trong công thức (4.1), x_i bao gồm các điểm ảnh phổ nằm trong phạm vi cửa sổ đang quét, giả sử kích thước bộ lọc 5×5 sẽ có 25 điểm ảnh phổ được quét. Số giá trị phải quét nằm tại cùng vị trí cho tất cả các thành phần, do vậy sẽ có K×5×5 giá trị x_i và w_i tương ứng. Ngoài ra, còn thêm một hệ số độ lệch b trong công thức này. Do vậy, số tham số cần thiết cho thao tác lấy chập bao gồm w_i và b là K×5×5+1 tham số. Giả thiết dùng M bộ lọc, số lượng tham số sẽ là M×(K×5×5+1) tham số.

Sau khi trượt bộ lọc qua tất cả các vị trí và lấy chập sẽ được một mảng với mỗi giá trị là các số được tính bằng cách lấy chập theo công thức trên. Đối với ví dụ mảng đầu vào của ảnh ở trên ta được mảng 28×28×1 các giá trị. Mảng này được gọi là ánh xạ kích hoạt (activation map) hay ánh xạ đặc trưng (feature map). Lý do có mảng 28×28 vì có 784 vị trí khác nhau để bộ lọc 5×5 có thể khớp trên ảnh 32×32, 784 giá trị này được ánh xạ thành mảng 28×28. Hình 4.1 mô tả bước lấy chập cho ví dụ ảnh đầu vào có kích thước 32×32, đầu ra là ánh xạ đặc trưng có kích thước 28×28.

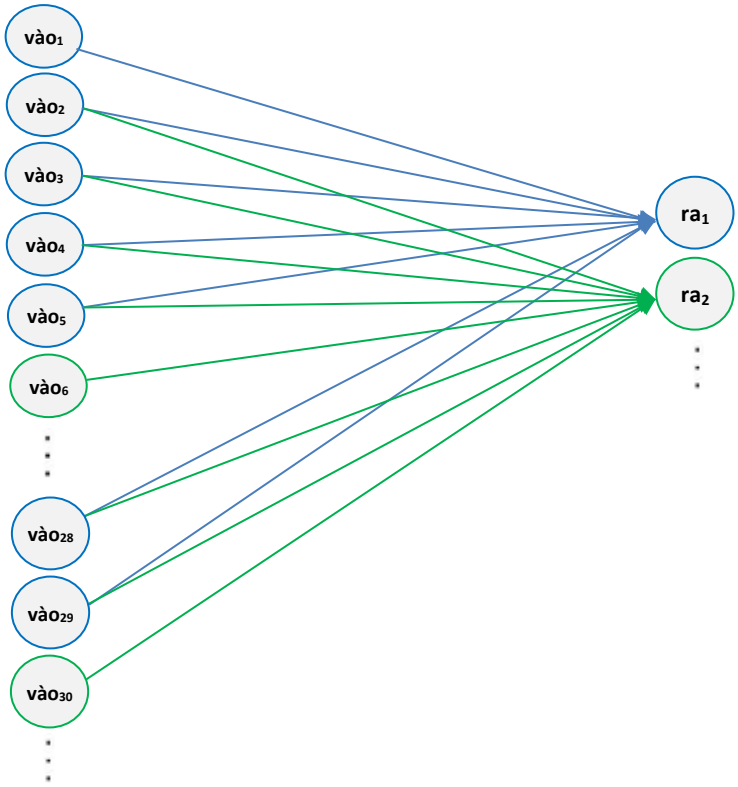


Hình 4.1 *Mô tả bước lấy chập dùng bộ lọc kích thước 5×5*

Hình 4.2 mô tả lấy chập trên một phần ảnh đầu vào với mảng giá trị có kích thước 8×8. Bộ lọc dịch chuyển 5×5 sẽ di chuyển qua toàn bộ phần ảnh và thao tác lấy chập được áp dụng cho 25 nút trên cửa sổ bộ lọc. Giả thiết ma trận trọng số của bộ lọc có các giá trị đều là 0,5. Kết quả thu được ở đầu ra là một mảng ánh xạ đặc trưng có kích thước 4×4. Với mỗi ô vuông màu lam và màu lục tượng trưng cho cửa sổ bộ lọc sẽ được tính toán (lấy chập) cho ra một giá trị tương ứng ở đầu ra. Bước trượt sẽ kiểm soát bộ lọc lấy chập khắp ảnh vào. Mặc định, mỗi lần bộ lọc sẽ dịch chuyển đi một đơn vị. Số lượng đơn vị mà bộ lọc dịch đi được gọi là bước trượt (stride).



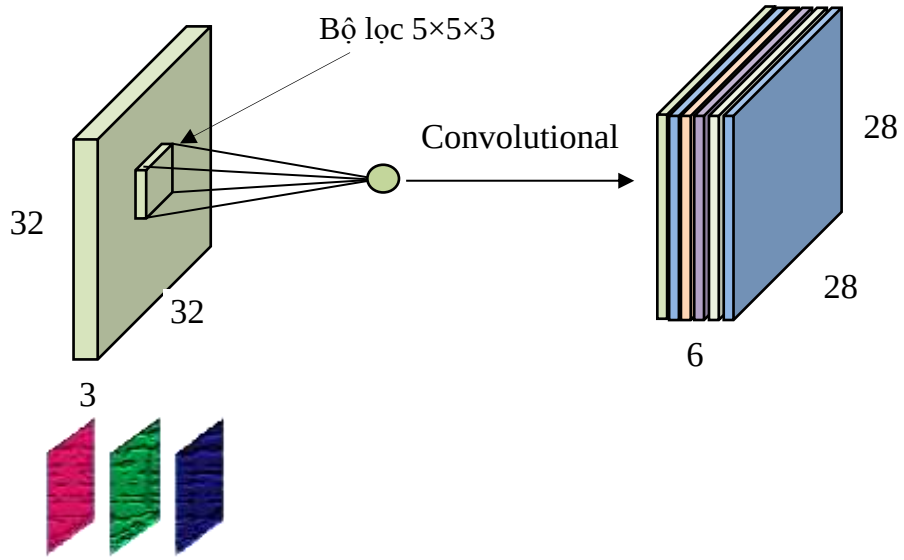
Hình 4.2 *Mô tả chi tiết lấy chập dùng bộ lọc kích thước 5×5*



Hình 4.3 *Mô tả bước lấy chập của mạng nơron dùng bộ lọc kích thước 5×5*

Thao tác lấy chập trên Hình 4.2 có thể được biểu diễn minh họa bằng biểu đồ mạng nơron như Hình 4.3. Các giá trị đầu vào và đầu ra của mạng được biểu diễn bằng một nút với trọng số kết nối là 0,5. Đối với mảng giá trị ảnh đầu vào kích thước 8×8 sẽ có 64 nút vào (vào$_1$…vào$_{64}$) và 16 nút ra (ra$_1$…ra$_{16}$). Vị trí thứ nhất của bộ lọc ứng với màu lam, vị trí thứ hai ứng với màu lục.

Chú ý rằng, các tham số của bộ lọc (trọng số) được giữ nguyên khi bộ lọc di chuyển qua ảnh đầu vào. Điều này cho phép bộ lọc cần được huấn luyện nhận ra một số đặc trưng trong dữ liệu vào. Đối với nhận dạng ảnh, có thể học để nhận biết các hình dạng như đoạn thẳng, cạnh và các hình dạng phân biệt khác. Vì thế bước lấy chập còn được gọi là ánh xạ đặc trưng. Tuy nhiên, để phân loại tốt, ở mỗi tầng lấy chập thường cần nhiều bộ lọc. Ví dụ, Hình 4.4 sử dụng 3 bộ lọc có kích thước 5×5 với đầu vào là ảnh màu kích thước 32×32, đầu ra của lấy chập là ánh xạ đặc trưng xếp chồng có kích thước 28×28×6.



Hình 4.4 *Mô tả bước lấy chập của mạng nơron dùng 3 bộ lọc kích thước 5×5*

Sử dụng lấy chập có những ưu điểm sau:

+ Giảm số lượng tham số: Ở ANN truyền thống, các nơron ở lớp trước sẽ kết nối tới tất cả các nơron ở lớp sau (fully connected) gây nên tình trạng quá nhiều tham số cần học. Đây là nguyên nhân chính gây nên tình trạng quá khớp (overfiting) cũng như làm tăng thời gian huấn luyện. Việc sử dụng lấy chập trong đó cho phép chia sẻ trọng số liên kết (shared weights), cũng như thay vì sử dụng kết nối đầy đủ (fully connected) sẽ sử dụng trường tiếp nhận cục bộ (local receptive fields) giúp giảm tham số.
+ Các tham số trong quá trình sử dụng lấy chập hay giá trị của các filter - kernel sẽ được học trong quá trình huấn luyện.

4.1.2 Kích hoạt phi tuyến

Về cơ bản, lấy chập là một phép biến đổi tuyến tính. Nếu tất cả các nơron được tổng hợp bởi các phép biến đổi tuyến tính thì một mạng nơron đều có thể đưa về dưới dạng một hàm tuyến tính. Khi đó mạng ANN sẽ đưa các bài toán về hồi qui logistic (logistic regression). Do đó, sau mỗi lớp lấy chập, đầu ra của ánh xạ lấy chập được cho qua hàm kích hoạt phi tuyến.

Một số hàm kích hoạt phi tuyến thường dùng như ReLU (Rectified Linear Unit), ELU (Exponential Linear Unit) [210]. ReLU có hàm kích hoạt dạng $f(x) = max(0, x)$ cho các giá trị vào. Về cơ bản, hàm này sẽ thay đổi tất cả các giá trị kích hoạt âm thành 0 và tăng tính phi tuyến của mô hình và toàn mạng mà không ảnh hưởng tới lớp lấy chập. Hàm ELU có dạng như (4.2) với $\propto > 0$:

$$f(x) = \begin{cases} x & nếu\ x \geq 0 \\ \propto (\exp(x) - 1) & nếu\ x < 0 \end{cases} \quad (4.2)$$

Hàm ReLU thực hiện tính toán đơn giản hơn ELU. Tuy nhiên, các nghiên cứu gần đây cho thấy việc sử dụng hàm ELU cho hiệu suất tốt hơn.

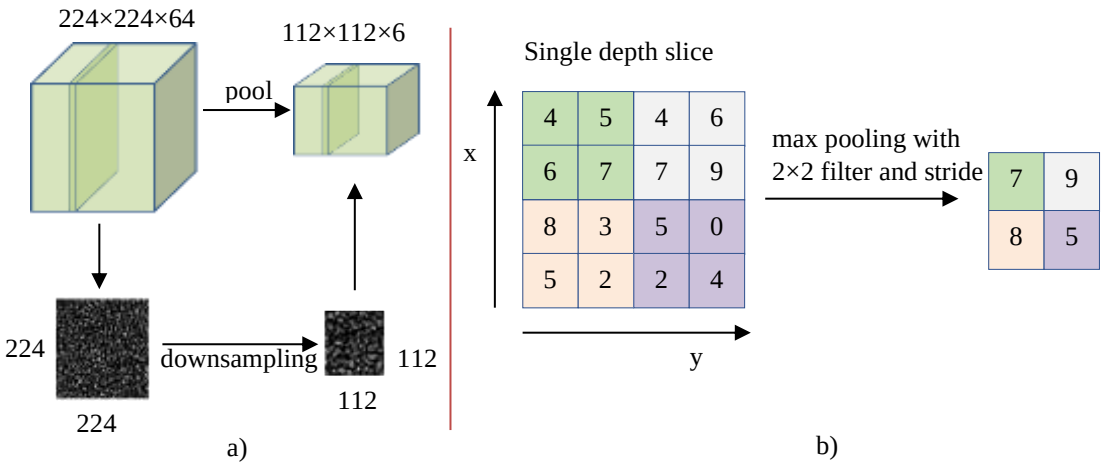
4.1.3 Lấy gộp

Lấy gộp (pooling) hay còn gọi subsampling hoặc downsampling là một trong những thành phần tính toán chính trong cấu trúc CNN. Xét về mặt toán học, lấy gộp 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.

Trong CNN, toán tử pooling được thực hiện độc lập trên mỗi kênh màu của ma trận ảnh đầu vào. Có nhiều toán tử pooling như sum-pooling, max-pooling, L2-pooling song max-pooling thường được sử dụng. Về mặt ý nghĩa, max-pooling xác định vị trí cho tín hiệu mạnh nhất khi áp dụng một loại bộ lọc. Điều này cũng tương tự như là bộ lọc phát hiện vị trí đối tượng trong bài toán phát hiện đối tượng trong ảnh. Nhìn chung, bộ lọc di chuyển thường có kích thước $F = 2$ và bước trượt $S = 2$ được dùng phổ biến. Có ít các thiết lập sử dụng bộ lọc di chuyển có kích thước $F = 3$ và $S = 2$. Và hiếm có thiết lập kích thước lớn hơn 3. Do kích thước bộ lọc di chuyển quá lớn có thể sẽ dẫn đến mất mát một số thông tin hữu ích, điều này làm cho hiệu suất nhận dạng kém đi.

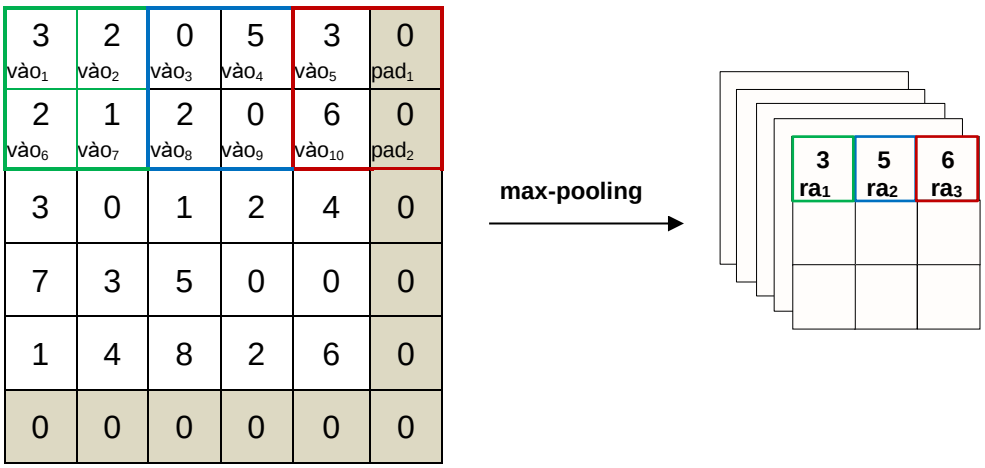
Hình 4.5 là ví dụ về mạng nơron lấy gộp sử dụng toán tử max-pooling (lấy gộp cực đại).

Hình 4.5 a) là cách thức pooling xử lý đối với một đầu vào là kết quả của nhiều bộ lọc ($k = 64$), kích thước của đầu vào là 224×224×64 được thực hiện với $F = 2$, $S = 2$, đầu ra sẽ có kích thước 112×112×64 . Hình 4.5 b) mô tả chi tiết cách thức hoạt động của max-pooling trong đó $F = 2, S = 2$ và kết quả đầu ra ma trận tương ứng.



Hình 4.5 *Ví dụ sử dụng max-pooling*

Padding (đệm) là kỹ thuật thêm các pixel bên ngoài hình ảnh. Với CNN, thường dùng zero padding nghĩa là các giá trị pixel thêm vào đều bằng 0. Nên trong quá trình lấy gộp, để bảo toàn thông tin cho ảnh đầu vào thường sử dụng padding với giá trị bằng 0 để không ảnh hưởng đến quá trình lấy gộp trên ảnh đầu vào. Hình 4.6 mô tả cách thực hiện max-pooling với zero padding và bộ lọc $F = 2$ cho toàn bộ ảnh.



Hình 4.6 *Mô tả cách thực hiện max-pooling với zero padding*

4.1.4 Kết nối đầy đủ

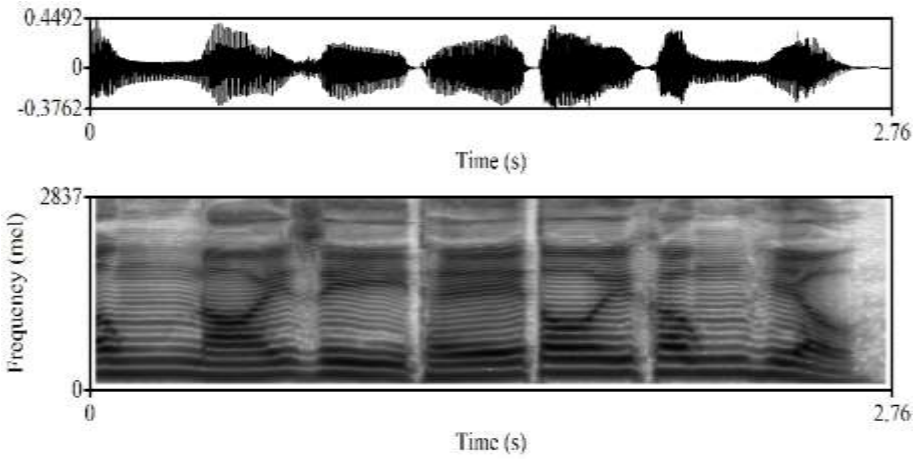
Kết nối đầy đủ là cách kết nối các nơron ở hai tầng với nhau trong đó tầng sau kết nối đầy đủ với các nơron ở tầng trước nó. Đây cũng là dạng kết nối thường thấy ở ANN. Trong CNN, tầng này thường được sử dụng ở các tầng phía cuối của kiến trúc mạng kết nối với đầu ra của mạng. Lớp này cơ bản là lấy thông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

Như vậy, CNN là thuật toán có kiến trúc bao gồm nhiều tầng có chức năng khác nhau trong đó tầng chính hoạt động thông qua cơ chế lấy chập. Trong suốt quá trình huấn luyện, CNN sẽ tự động học được các thông số cho các bộ lọc tương ứng là các đặc trưng theo từng cấp độ khác nhau. Ví dụ trong bài toán phân lớp ảnh, CNN sẽ cố gắng tìm ra các thông số tối ưu cho các bộ lọc tương ứng theo thứ tự điểm ảnh → các cạnh → khuôn hình → bộ mặt → các đặc trưng mức cao (pixel → edges → shapes → facial → high-level features). Đây chính là lý do mà CNN có được kết quả vượt trội so với các thuật toán trước đây.

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.

Hình 4.7 là ví dụ về phổ mel của tín hiệu tiếng nói làm ảnh đầu vào cho lớp thứ nhất trong trường hợp mô hình baseline.



Hình 4.7 *Phổ mel của tín hiệu tiếng nói làm ảnh đầu vào cho lớp thứ nhất trong trường hợp mô hình baseline*

Đối với nhận dạng cảm xúc của tiếng Việt được thực hiện trong luận án, cấu hình đầy đủ của mạng nơron DCNN để huấn luyện được mô tả như Bảng 4.1 trong trường hợp mô hình baseline 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.

Đối với lớp thứ nhất, ảnh đầu vào là ma trận 260×260 (260 hệ số phổ Mel×260 khung). Số khung được chọn là 260 đối với tất cả các file được sử dụng cho các thử nghiệm với độ rộng khung là 25ms và độ dịch khung là 10ms. Số khung được chọn là 260 dựa trên dung lượng các file có số lượng khung dao động xung quanh giá trị này.

Sau khi sử dụng bộ lọc di chuyển 3×3 với padding, sẽ có 64 ánh xạ đặc trưng có kích thước 260×260. Tính toán tiếp theo của lớp này là chuẩn hóa theo lô (batch normalization), kích hoạt phi tuyến ELU (Exponential Linear Unit) và max-pooling với cửa sổ 2×2 và cuối cùng là dropout với hệ số 0,5. Thao tác lấy chập được thực

hiện tại các vị trí mà bộ lọc đi qua, mỗi vị trí của bộ lọc sẽ tính được một giá trị theo công thức (4.1). Giả sử *K* là số ảnh đầu vào, *M* là số ánh xạ đặc trưng thì số tham số cần tính cho thao tác lấy chập là *M* × (*K* × kích thước cửa sổ lọc +1).

Đối với mỗi lớp, mục tiêu của chuẩn hóa theo lô là để đạt được sự phân bố ổn định các giá trị kích hoạt trong suốt quá trình huấn luyện và do đó đem lại sự tăng tốc đáng kể trong huấn luyện [211]. ELU làm tăng tốc độ học trong mạng nơron sâu và như thế sẽ giúp độ chính xác phân loại cao hơn [210]. Hàm max-pooling làm giảm số lượng tham số của mô hình cần phải tính toán, từ đó giảm thời gian tính toán [212]. Cuối cùng, dropout được coi là một cách để tránh các mạng nơron bị quá khớp [213].

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

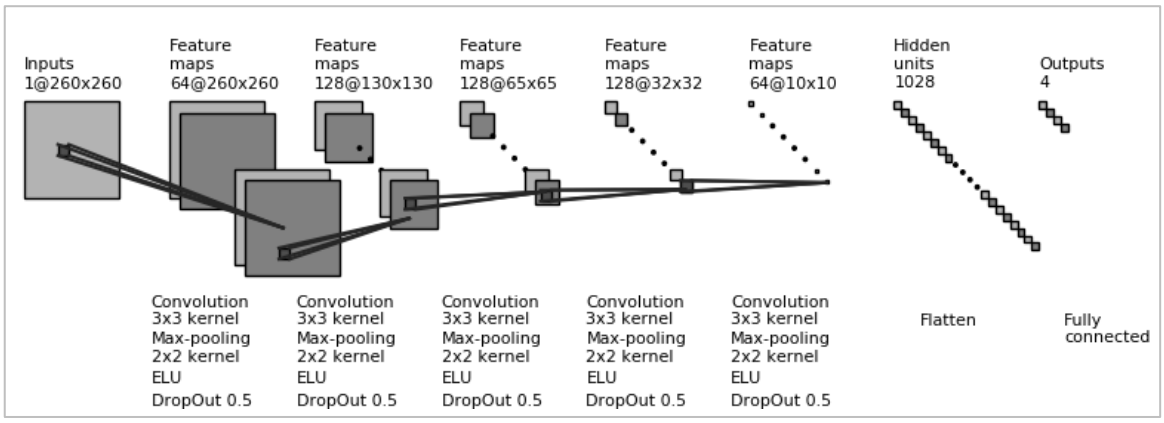
Layer Index	Layer (type)	Output Shape	Param #
	BatchNormalization	(260, 260, 1)	1040
1	Convolution2D (3×3)	(260, 260, 64)	640
	BatchNormalization	(260, 260, 64)	256
	ELU	(260, 260, 64)	0
	MaxPooling2D (2×2)	(130, 130, 64)	0
	Dropout (0,5)	(130, 130, 64)	0
2	Convolution2D(3×3)	(130, 130, 128)	73856
	BatchNormalization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

Các lớp 2, 3 và 4 thực hiện các thao tác tương tự: Lấy chập sử dụng bộ lọc di chuyển kích thước 3×3 có padding và đầu ra của các lớp này là 128 ánh xạ đặc trưng với kích thước lần lượt là 130×130, 65×65, 32×32 tương ứng. Sau khi lấy chập, các lớp này cũng thực hiện chuẩn hóa theo lô, kích hoạt phi tuyến ELU, max-pooling và dropout.

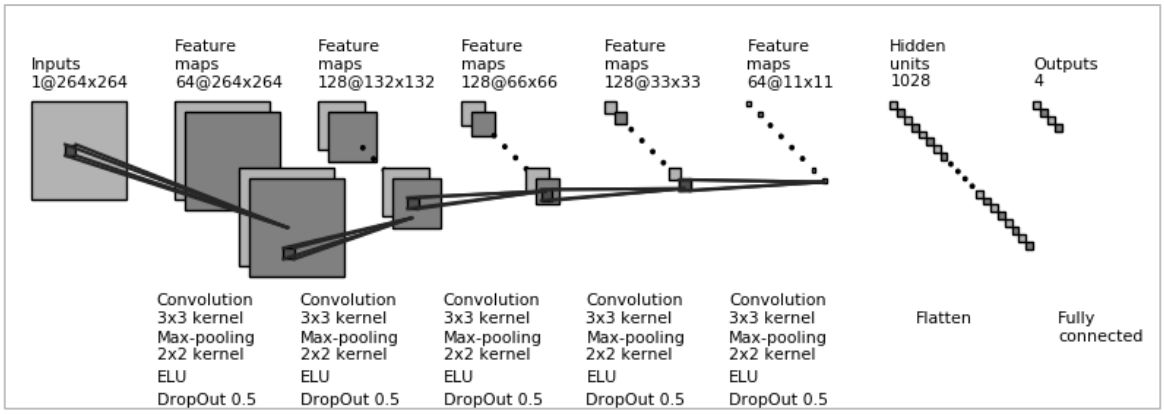
Cuối cùng đối với lớp 5, sau khi chuẩn hóa theo lô, kích hoạt phi tuyến ELU, max-pooling và dropout, ta có lớp kết nối đầy đủ với 256 đầu vào và 4 đầu ra tương ứng với 4 cảm xúc. Hàm truyền của lớp kết nối đầy đủ là Softmax thể hiện phân bố xác suất cho mỗi cảm xúc.

Thông thường, số lớp ẩn của mạng nơron từ 3 trở lên có thể được coi là mạng nơron sâu. Nghiên cứu [214] là một ví dụ về kiến trúc DCNN với 5 lớp song là để phân loại hình ảnh.

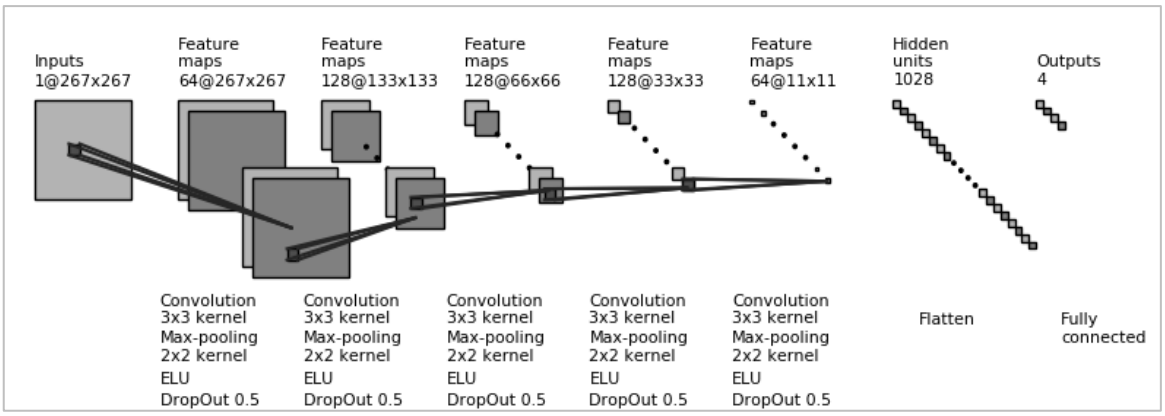
Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt được thực hiện trong luận án với các tập tham số khác nhau được trình bày trên các Hình từ 4.8 đến 4.12. Trong đó, 5 lớp ẩn đều được thực hiện qua các bước lấy chập, chuẩn hóa theo lô, ELU, max-pooling và dropout. Lớp ẩn cuối cùng là lớp kết nối đầy đủ và đầu ra là phân lớp 4 loại cảm xúc.



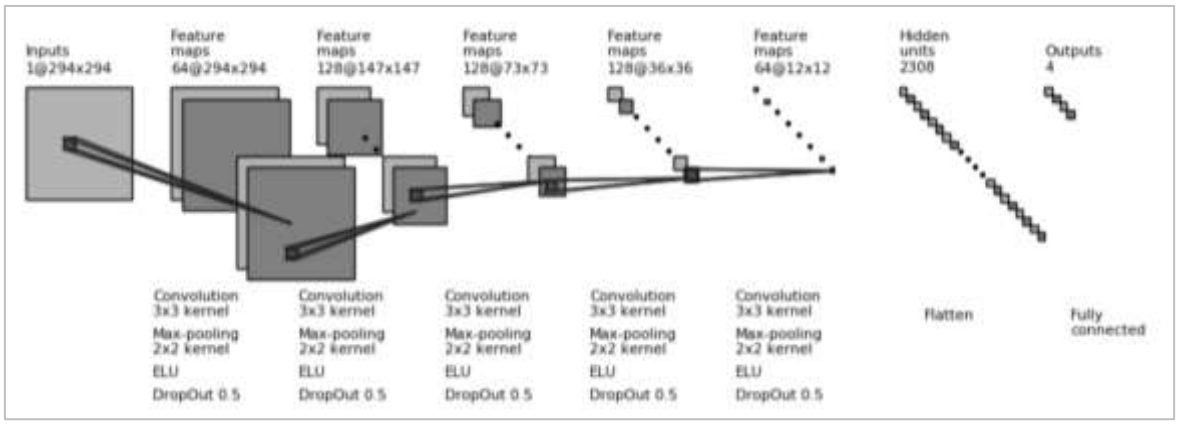
Hình 4.8 *Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt với 260 tham số*



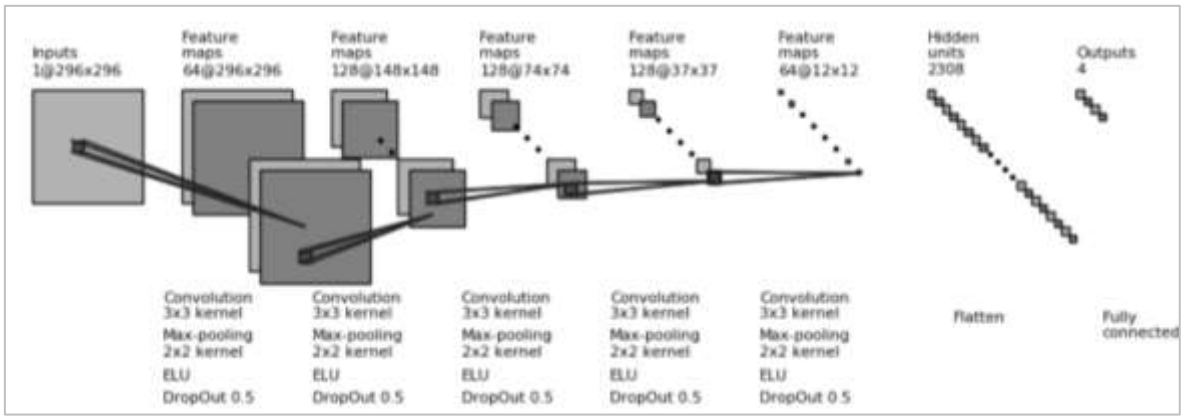
Hình 4.9 *Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt với 264 tham số*



Hình 4.10 *Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt với 267 tham số*



Hình 4.11 *Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt với 294 tham số*



Hình 4.12 *Mô hình DCNN cho nhận dạng cảm xúc tiếng Việt với 296 tham số*

4.3 Ngữ liệu, tham số và công cụ 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. Chi tiết số file ngữ liệu thử nghiệm trong mục này được trình bày trong các Bảng từ 4.2 đến 4.5 sau đây.

Bảng 4.2 *Phân chia ngữ liệu T1 (phụ thuộc cả người nói và nội dung)*

Cảm xúc	Số file Test	Số file Train	Số file Valid	Tổng số file
Bình thường	350	697	349	1396
Vui	349	699	348	1396
Buồn	351	698	347	1396
Tức	349	698	349	1396
Tổng	1399	2792	1393	5584

Bảng 4.3 *Phân chia ngữ liệu T2 (phụ thuộc người nói và độc lập nội dung)*

Cảm xúc	Số file Test	Số file Train	Số file Valid	Tổng số file
Bình thường	347	700	349	1396
Vui	349	702	345	1396
Buồn	349	699	348	1396
Tức	352	692	352	1396
Tổng	1397	2793	1394	5584

Bảng 4.4 *Phân chia ngữ liệu T3 (độc lập người nói và phụ thuộc nội dung)*

Cảm xúc	Số file Test	Số file Train	Số file Valid	Tổng số file
Bình thường	348	704	344	1396
Vui	352	696	348	1396
Buồn	350	698	348	1396
Tức	350	698	348	1396
Tổng	1400	2796	1388	5584

Bảng 4.5 *Phân chia ngữ liệu T4 (độc lập cả người nói và nội dung)*

Cảm xúc	Số file Test	Số file Train	Số file Valid	Tổng số file
Bình thường	174	346	176	696
Vui	175	350	175	700
Buồn	175	352	176	703
Tức	176	352	176	704
Tổng	700	1400	703	2803

Các tham số sử dụng nhận dạng cảm xúc với mô hình DCNN được thống kê trong Bảng 4.6.

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*
264	- 260 hệ số *MFCC* - Tần số cơ bản *F*0 - 3 biến thể của *F*0: *F0NormMinMax, logF0NormAver, logF0NormMinMax*

Tập tham số	Các tham số sử dụng
267	- 264 tham số - 3 biến thể *F*0: *F0NormAver, F0NormAverStd,* l*ogF0NormAverStd*
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*

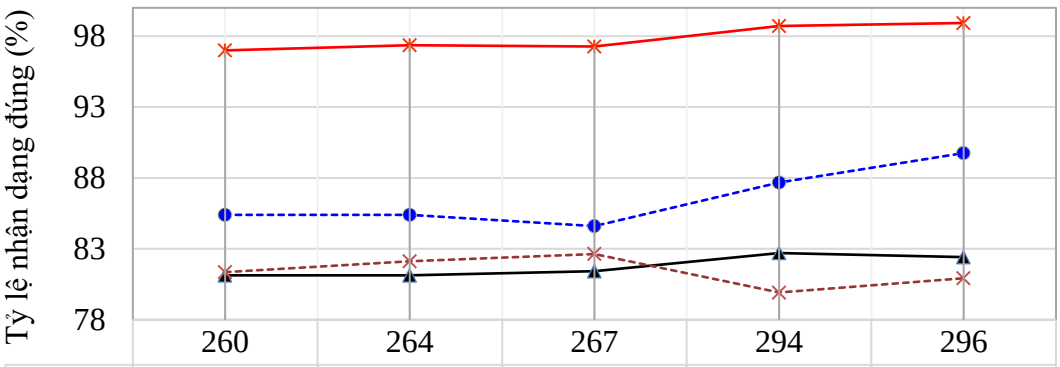
Các thử nghiệm được thực hiện đối với năm tập tham số và bốn tập ngữ liệu (T1, T2, T3, T4).

Để thực hiện các thử nghiệm nhận dạng cảm xúc tiếng Việt với DCNN, luận án đã sử dụng các hàm của Keras [215] và Tensorflow [216]. Keras là một thư viện API (Application Programming Interface) mạng nơron cấp cao, được viết bằng Python và có khả năng chạy trên TensorFlow. TensorFlow là một thư viện phần mềm mã nguồn mở để tính toán số liệu bằng các biểu đồ luồng dữ liệu. TensorFlow ban đầu được phát triển bởi các nhà nghiên cứu và kỹ sư làm việc trong nhóm Google Brain trong tổ chức nghiên cứu trí tuệ máy của Google nhằm mục đích nghiên cứu về học máy và mạng nơron sâu. Hệ thống này có thể áp dụng trong nhiều lĩnh vực khác nhau.

Chương trình thử nghiệm nhận dạng cảm xúc tiếng Việt sử dụng DCNN trong luận án được viết bằng ngôn ngữ Python chạy trên 2 máy có cấu hình CPU Intel Core i7 - 4790 @ 3,60GHz × 8, RAM 16 GB, GPU NVIDIA GeForce GTX 1050 Ti / PCIe / SSE2, RAM 8 GB. Thời gian huấn luyện trung bình với DCNN khoảng 6 ngày cho một trong 4 tập ngữ liệu với một trong 5 tập tham số.

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

Hình 4.13 là kết quả tỷ lệ nhận dạng với 4 tập ngữ liệu khác nhau tương ứng với từng tập tham số. Đối với thử nghiệm $\mathrm{Ti}-260, \mathrm{Ti}-264$ ($\mathrm{i}=1,2,3,4$), tỷ lệ nhận dạng đều tăng theo i theo chiều từ 4 đến 1 (Chú ý cách ký hiệu thử nghiệm được dùng ở đây: T1-260 nghĩa là thử nghiệm dùng tập ngữ liệu T1 với 260 tham số, T2-264 nghĩa là thử nghiệm dùng tập ngữ liệu T2 với 264 tham số…). Trong trường hợp dùng 260 tham số, gia tăng tỷ lệ nhận dạng lớn nhất là 15,86 % ($\mathrm{T1}-260$ so với $\mathrm{T4}-260$), gia tăng tỷ lệ nhận dạng nhỏ nhất là 0,22 % ($\mathrm{T3}-260$ so với $\mathrm{T4}-260$). Trong trường hợp dùng 264 tham số, gia tăng tỷ lệ nhận dạng lớn nhất là 16,22 % ($\mathrm{T1}-264$ so với $\mathrm{T4}-264$), gia tăng tỷ lệ nhận dạng nhỏ nhất là 1,0 % ($\mathrm{T3}-264$ so với $\mathrm{T4}-264$).



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

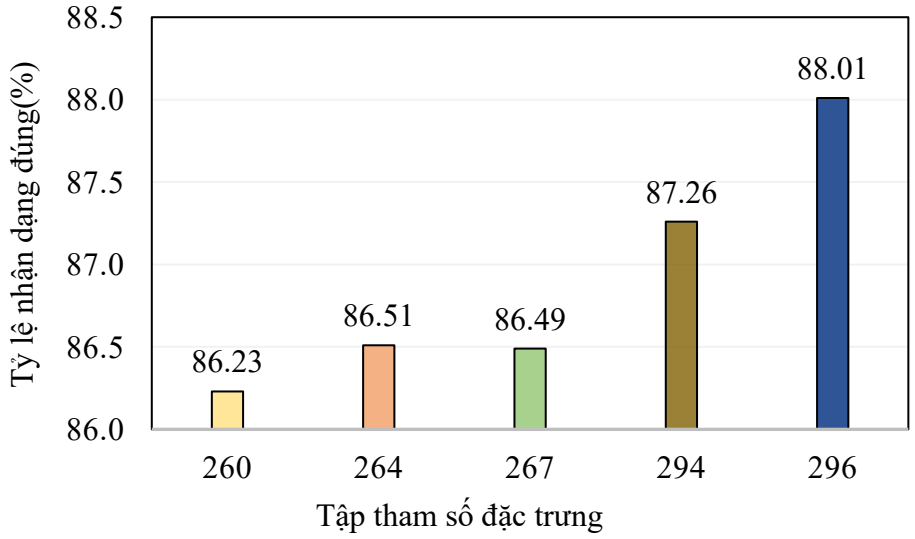
Khi tăng số tham số từ 260 lên 264, các thử nghiệm với tập ngữ liệu T2, T4 đều không tăng tỷ lệ nhận dạng, còn lại các thử nghiệm với tập ngữ liệu T1, T3 đều tăng tỷ lệ nhận dạng.

Đối với thử nghiệm Ti − 267 (i = 1,2,3,4) thì T1, T2, T4 đều tăng tỷ lệ nhận dạng. Gia tăng tỷ lệ nhận dạng lớn nhất là 15,85 % (T1 − 267 so với T4 − 267), gia tăng tỷ lệ nhận dạng nhỏ nhất là 1,21 % (T3 − 267 so với T4 − 267). Riêng trường hợp T2 − 267, tỷ lệ nhận dạng giảm 0,79% so với T2 − 264.

Trong trường hợp dùng 294 tham số, tỷ lệ nhận dạng đều tăng đối với T1, T2, T4. Trong đó, thử nghiệm T2 − 294 tăng lên đáng kể (3,08%) so với T2 − 267. Ngược lại, thử nghiệm T3 − 294 lại giảm 2,71% so với T3 − 267. Khi số tham số tăng lên 296, các tập ngữ liệu T1, T2 đều đạt tỷ lệ nhận dạng cao nhất so với các trường hợp còn lại. Thử nghiệm T1 − 296 đạt 98,93% còn T2 − 296 đạt 89,76%.

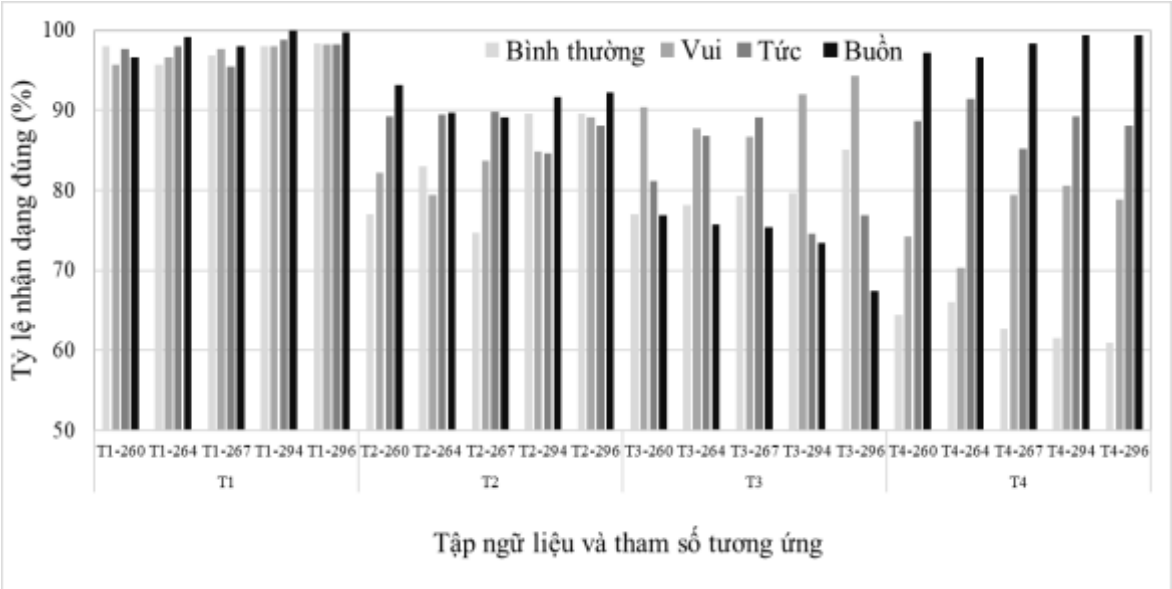
Như vậy trong các thử nghiệm với 5 tập tham số khác nhau, tỷ lệ nhận dạng đạt cao nhất ứng với tập ngữ liệu T1, T2 khi sử dụng 296 tham số. Đối với T3, tỷ lệ nhận dạng là cao nhất khi sử dụng 267 tham số, còn đối với T4 cao nhất khi sử dụng 294 tham số.

Tỷ lệ nhận dạng trung bình của tất cả các thử nghiệm đối với từng tập tham số được trình bày trên Hình 4.14. Hình 4.14 cho thấy tỷ lệ nhận dạng trung bình của các thử nghiệm đạt cao nhất khi sử dụng 296 tham số và nhỏ nhất khi sử dụng 260 tham số. Điều này cũng cho thấy, việc thêm các đặc trưng về năng lượng, phổ, tần số cơ bản $F0$ và biến thể của $F0$, bốn formant và dải thông tương ứng đã làm cho tỷ lệ nhận dạng tăng lên. Các đặc trưng này là các tham số có ảnh hưởng tốt đến khả năng phân biệt các cảm xúc và đã được phân tích đánh giá bằng phương pháp one-way ANOVA và kiểm định T đã được trình bày ở mục 2.6 của Chương 2. 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 tập 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*

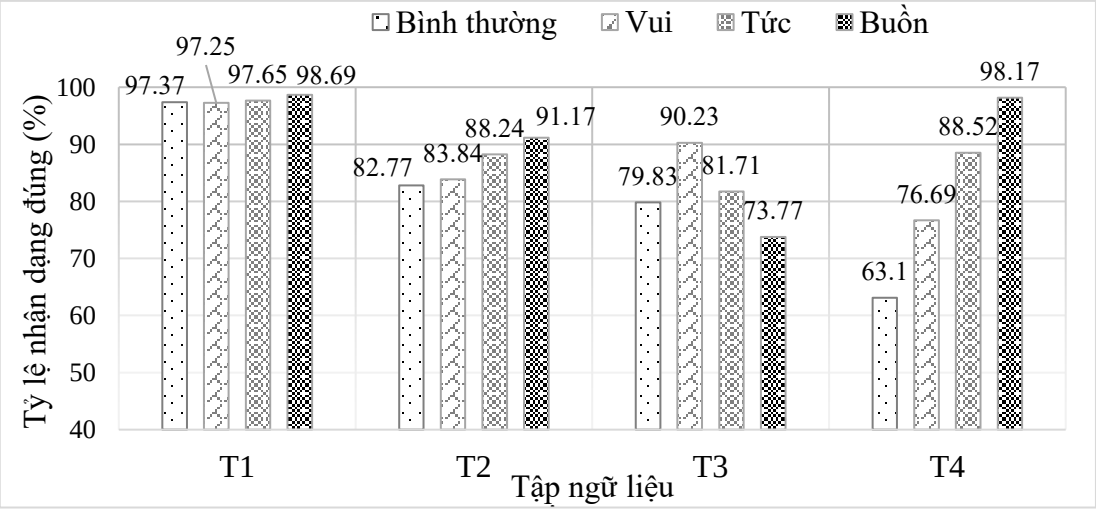
Hình 4.15 cho thấy, nhìn chung cảm xúc buồn đạt tỷ lệ nhận dạng cao nhất so với các cảm xúc khác trong các thử nghiệm sử dụng các tập ngữ liệu T1, T2, T4. Điều này cũng được thể hiện trên Hình 4.16 trong đó đã thống kê tỷ lệ nhận dạng trung bình của từng cảm xúc đối với các thử nghiệm. Đối với thử nghiệm dùng T1, tỷ lệ nhận dạng đạt cao nhất và không chênh lệnh nhiều giữa các cảm xúc (nằm trong khoảng từ 95,42% - 100%). Còn với thử nghiệm dùng T4, các cảm xúc khác có tỷ lệ nhận dạng chênh lệnh nhiều hơn so với cảm xúc buồn.



Hình 4.15 *Tỷ lệ nhận dạng đúng cao nhất của từng cảm xúc đối với từng thử nghiệm*

Nếu xét tỷ lệ nhận dạng trung bình của tất cả các thử nghiệm với từng tập tham số, tỷ lệ nhận dạng trung bình của cảm xúc buồn cao nhất (90,54%), tiếp theo là cảm xúc tức (89,03%), cảm xúc vui (87%) và thấp nhất là cảm xúc bình thường (80,77%). Tỷ lệ nhận dạng trung bình của từng thử nghiệm với từng cảm xúc được thống kê trên Hình 4.16.

Đối với tập 296 tham số, có 14 hệ số đáp ứng xung của bộ lọc đảo của tuyến âm được thêm vào tập tham số và chiếm một lượng đáng kể. Các hệ số này mang thông tin về tuyến âm. Tuy nhiên, có thể thấy rằng sự gia tăng đáng kể các hệ số này không làm tăng đáng kể độ chính xác nhận dạng so với việc tăng độ chính xác nhận dạng khi tăng số lượng tham số liên quan đến $F0$. Điều này cho thấy tầm quan trọng của $F0$ đối với nhận dạng cảm xúc và nên lưu ý khi xem xét mối liên quan giữa số lượng tham số, độ chính xác trung bình, thời gian huấn luyện và chi phí huấn luyện.



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

Kết quả của ANOVA và kiểm định T đã cho thấy khả năng phân biệt tốt các cảm xúc khi sử dụng các tham số đặc trưng tín hiệu tiếng nói đối với bộ ngữ liệu BKEmo. Kết quả nhận dạng bốn cảm xúc sử dụng DCNN cũng phù hợp với các kết quả đánh giá ANOVA và kiểm định T cho bộ ngữ liệu này. Tham số đặc trưng được sử dụng cho DCNN bao gồm các tham số đặc trưng cho nguồn âm và tuyến âm. Đối với các tham số đặc trưng, tần số cơ bản $F0$ là tham số nguồn âm và rất quan trọng đối với tiếng Việt vì đây là ngôn ngữ có thanh điệu. Mặt khác, quy luật biến đổi của tần số cơ bản cũng phụ thuộc vào cảm xúc cần được thể hiện. Nhìn chung, độ chính xác của nhận dạng cảm xúc tiếng Việt đạt được với DCNN là rất tích cực và kết quả thử nghiệm cho thấy thông tin về tần số cơ bản $F0$ đã cải thiện độ chính xác trong nhận dạng cảm xúc.

Luận án đã thử nghiệm nhận dạng 4 cảm xúc vui, buồn, tức và bình thường sử dụng mô hình DCNN và GMM. Việc so sánh tỷ lệ nhận dạng của hai mô hình này để đơn thuần kết luận mô hình nào có tỷ lệ nhận dạng cao hơn là không hợp lý. Bởi vì, các tham số đặc trưng được sử dụng cho hai mô hình này là khác nhau. Mặc dù vậy, cả DCNN và GMM đều đã cải thiện tỷ lệ nhận dạng khi thêm thông tin về các biến thể $F0$ và $F0$. Có thể so sánh một cách sơ bộ giữa GMM và ANN như sau. Đối với GMM, điều thiết yếu là xây dựng được mô hình λ theo M như đã trình bày ở mục 3.1.

Trong khi đó, có thể nói rằng bên cạnh khả năng học, ANN có triển vọng đầy hứa hẹn để xây dựng một hệ thống nhận dạng nói chung do tính đa dạng trong kiến trúc của mạng nơron dẫn đến có rất nhiều tiềm năng để khai thác.

Hướng tiếp theo của luận án là mở rộng kho ngữ liệu với các dạng cảm xúc khác cho tiếng Việt và thực hiện nhận dạng cho các dạng cảm xúc này. Một vấn đề đối với các hệ thống nhận dạng bao gồm cả nhận dạng cảm xúc là ngữ liệu trong môi trường thực có thể không thuộc tập huấn luyện hoặc tập thử nghiệm. Thông thường trong những trường hợp như vậy, độ chính xác nhận dạng có thể bị giảm. Để tiếp cận vấn đề này, đã có những nghiên cứu sử dụng transfer learning [217], [218], [219], [220] và đây cũng sẽ là một trong những hướng nghiên cứu sắp tới để nhận dạng cảm xúc tiếng Việt.

4.5 Kết chương 4

Chương 4 đã trình bày kết quả nhận dạng bốn cảm xúc sử dụng mô hình DCNN. Kết quả nhận dạng cũng cho thấy tỷ lệ nhận dạng cao với cả bốn cảm xúc. Năm tập tham số đặc trưng đã được sử dụng làm đầu vào của mạng nơron trong đó bao gồm các đặc trưng phổ theo thang Mel, các đặc trưng của tuyến âm, tần số cơ bản $F0$ và các biến thể của $F0$. Các thử nghiệm được tiến hành trên năm tập tham số này cho bốn trường hợp ngữ liệu tùy thuộc vào phụ thuộc hay độc lập về nội dung và phụ thuộc hay độc lập về người nói.

Tính trung bình, độ chính xác nhận dạng tối đa đạt được là 97,86% đối với trường hợp 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 $F0$ góp phần đáng kể vào sự gia tăng độ chính xác của nhận dạng cảm xúc tiếng Việt. Đối với thử nghiệm nhận dạng sử dụng mô hình DCNN, cảm xúc buồn cho tỷ lệ cao hơn các cảm xúc còn lại.

Việc so sánh độ chính xác nhận dạng giữa các mô hình chỉ mang tính tương đối khi các tập tham số đặc trưng và cách sử dụng các tập tham số đặc trưng này cho các mô hình là không hoàn toàn giống nhau. Trên thực tế có thể nói, về mặt lý thuyết chưa chỉ ra được cấu hình mạng nơron nào sẽ tỏ ra thích hợp nhất với loại bài toán nhận dạng nào. Vì thế việc thử nghiệm nhận dạng với mạng nơron còn tiềm tàng rất nhiều khả năng để khai thác như kiến trúc mạng nơron, các tham số đặc trưng được sử dụng và số lượng các tham số đặc trưng đó. Trong khuôn khổ có hạn, luận án đã thực hiện nghiên cứu thử nghiệm cho cấu hình của mạng nơron tương ứng với các tham số đặc trưng đã sử dụng. Trong tương lai sẽ cần có thời gian để khai thác các kiến trúc đa dạng của mạng nơron với các tham số đặc trưng cảm xúc của tiếng nói.

Các kết quả nghiên cứu chính của chương 4 đã được công bố ở bài báo số 8 trong danh mục công trình nghiên cứu của luận án:

8. "*Deep Convolutional Neural Network for Emotion Recognition of Vietnamese*", International Journal of Machine Learning and Computing (IJMLC), ISSN: 2010-3700, DOI: 10.18178/IJMLC, Indexing: SCOPUS (đã được chấp nhận đăng trên tạp chí).

KẾT LUẬN VÀ ĐỊNH HƯỚNG PHÁT TRIỂN

1. Kết luận

Với sự tiến bộ của khoa học và công nghệ, ngày nay nhiều thiết bị máy móc được tạo ra có thể thay thế con người trong các lĩnh vực khác nhau của đời sống xã hội. Máy có thể nghe và hiểu con người đang nói gì, máy có thể đọc, có thể nhìn và nhận dạng được hình ảnh rất tốt. Kỷ nguyên của máy móc có cảm xúc cũng đang đến gần, các nhà khoa học đã chế tạo ra những máy móc có khả năng biểu cảm cảm xúc qua nét mặt và mong muốn hơn nữa là máy có thể cảm nhận và thể hiện cảm xúc qua giọng nói như con người. Để làm được điều này, các hệ thống cần tích hợp vào đó cảm xúc để tương tác giữa người-máy được tự nhiên như giữa con người với nhau. Từ bối cảnh này, luận án đã thực hiện đề tài nghiên cứu ***"Nhận dạng cảm xúc cho tiếng Việt nói"***.

Cảm xúc của con người là đa dạng, phong phú và không phải lúc nào cũng mạch lạc rõ ràng. Vì vậy, từ những nghiên cứu chung về cảm xúc, luận án nghiên cứu thử nghiệm nhận dạng với 4 cảm xúc được các nhà nghiên cứu cho rằng cơ bản nhất, đó là cảm xúc vui, buồn, tức và bình thường.

Các kỹ thuật để nhận dạng cảm xúc cũng có nhiều phương pháp như dựa trên phân tích ngữ nghĩa văn bản, phân tích tín hiệu tiếng nói, phân tích biểu cảm qua gương mặt, phân tích tín hiệu điện não hay kết hợp của nhiều phương diện. Về phương diện phân tích tín hiệu tiếng nói để nhận dạng cảm xúc tiếng Việt, hãy còn rất ít công trình được công bố và luận án đã thực hiện nghiên cứu nhận dạng cảm xúc theo phương diện này.

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 đề xuất bộ 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.

Với các nghiên cứu thử nghiệm nhận dạng cảm xúc tiếng Việt được thực hiện dựa trên các mô hình GMM, DCNN, luận án đề xuất bộ tham số bao gồm hệ số *MFCC*, tần số $F0$, các biến thể của $F0$, cường độ, năng lượng, formant và các dải thông tương ứng, các đặc trưng phổ. Đây là các tham số có ảnh hưởng tốt đến nhận dạng về cảm xúc. Kết quả nhận dạng được cải thiện tốt khi kết hợp *MFCC* với $F0$ và các biến thể của nó dựa trên mô hình GMM và mô hình DCNN. Bên cạnh đó, luận án cũng thực hiện các thử nghiệm phân lớp, nhận dạng cảm xúc với một số bộ phân lớp như LDA, SMO, IBk, Trees J48 để đánh giá bộ ngữ liệu tiếng Việt.

Luận án có những đóng góp khoa học như sau:

(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

Trong khuôn khổ có hạn của luận án, nội dung nghiên cứu trước hết dành cho bốn cảm xúc cơ bản bao gồm vui, buồn, tức và bình thường. Trên thực tế, các hình thái cảm xúc rất đa dạng và phong phú nên cần có nhiều thời gian và nỗ lực nghiên cứu của cộng đồng nghiên cứu trong lĩnh vực nhận dạng cảm xúc tiếng nói nói chung và đặc biệt là cảm xúc tiếng Việt nói nói riêng. 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 như: Xây dựng ngữ liệu cảm xúc tiếng Việt với các giọng ở vùng miền khác nhau và cho các cảm xúc khác như ngạc nhiên, sợ hãi, hồi hộp …; thực hiện thử nghiệm đánh giá kết quả.
- 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.

DANH MỤC CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ CỦA LUẬN ÁN

1. Phạm Ngọc Hưng, Trịnh Văn Loan, Đào Thị Thu Diệp, Phạm Quốc Hùng, Chu Bá Thành, Đào Thị Lệ Thủy (2014), "*Nghiên cứu và thử nghiệm nhận dạng phương ngữ tiếng Việt*", Tạp chí Khoa học và Công nghệ, ĐHSPKT Hưng Yên, số 4, ISSN 2354-0575, trang 96-101.
2. Đào Thị Lệ Thủy, Lê Xuân Thành, Trịnh Văn Loan, Nguyễn Hồng Quang (2015), "*Emotion recognition and corpus for Vietnamese emotion recognition*", Tạp chí Khoa học và Công nghệ, ĐHSPKT Hưng Yên, số 7, ISSN 2354-0575, trang 51-56.
3. Lê Xuân Thành, Đào Thị Lệ Thủy, Trịnh Văn Loan, Nguyễn Hồng Quang (2016), "*Cảm xúc trong tiếng nói và phân tích thống kê ngữ liệu cảm xúc tiếng Việt*", Chuyên san Các công trình Nghiên cứu, Phát triển và Ứng dụng Công nghệ Thông tin, Tạp chí Bưu chính Viễn thông, tập V-1, số 15 (35), ISSN 1859-3526, trang 86-98.
4. Lê Xuân Thành, Đào Thị Lệ Thủy, Trịnh Văn Loan, Nguyễn Hồng Quang (2016), "*So sánh hiệu năng một số phương pháp nhận dạng cảm xúc tiếng Việt nói*," Kỷ yếu Hội nghị khoa học công nghệ quốc gia lần thứ IX, Nghiên cứu cơ bản và ứng dụng công nghệ thông tin, Cần Thơ, ISBN 978-604-913-472-2, trang 656-662.
5. Lê Xuân Thành, Trịnh Văn Loan, Nguyễn Hồng Quang, Đào Thị Lệ Thủy, Đinh Đồng Lưỡng (2017), "*Tổng hợp tiếng Việt có cảm xúc*", Chuyên san Các công trình nghiên cứu phát triển Công nghệ Thông tin và Truyền thông, Tạp chí Bưu chính Viễn thông, Tập V-2, Số 18 (38), ISSN 1859-3526, trang 67-77.
6. Đào Thị Lệ Thủy, Trịnh Văn Loan, Nguyễn Hồng Quang và Lê Xuân Thành (2017), "*Ảnh hưởng của đặc trưng phổ tín hiệu tiếng nói đến nhận dạng cảm xúc tiếng Việt*", Kỷ yếu Hội nghị khoa học công nghệ quốc gia lần thứ X, Nghiên cứu cơ bản và ứng dụng công nghệ thông tin, Đà Nẵng, ISBN 978-604-913-614-6, trang 36-43.
7. Dao Thi Le Thuy, Trinh Van Loan, Nguyen Hong Quang (2017), "*GMM for emotion recognition of Vietnamese*", Journal of Computer Science and Cybernetics, V.33, N.3, ISSN 1813-9663, pp.229-246.
8. Dao Thi Le Thuy, Trinh Van Loan, Nguyen Hong Quang (2019), "*Deep Convolutional Neural Network for Emotion Recognition of Vietnamese*", International Journal of Machine Learning and Computing (IJMLC), ISSN: 2010-3700, DOI: 10.18178/IJMLC, Indexing: SCOPUS (đã được chấp nhận đăng trên tạp chí).

TÀI LIỆU THAM KHẢO

[1] Nguyễn Tôn Nhan, Phú Văn Hẳn (2013), "*Từ điển Tiếng Việt*", Nhà xuất bản từ điển Bách Khoa.

[2] Rao, K. Sreenivasa, Koolagudi, Shashidhar G. (2013), "*Emotion Recognition using Speech Features*", Springer.

[3] Schubiger M. (1958), "*English intonation: its form and function*", Tubingen, Tubingen, Germany: Niemeyer.

[4] Connor J. and Arnold G. (1973), "*Intonation of Colloquial English*", London: UK: Longman, Second edition.

[5] Robert Plutchik, Henry Kellerman (1989), "*Emotion: Theory, research and experience*", vol. 4, New York, USA: Academic Press.

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

[7] Ekman P. (1999), "*Handbook of Cognition and Emotion*", ch. Basic Emotions, Sussex, UK: JohnWiley and Sons Ltd.

[8] Cowie, Roddy, et al., "*Emotion recognition in human-computer interaction*", IEEE Signal processing magazine 18.1, vol. 12, pp. 32–80.

[9] William J. (1984), "*What is an emotion?*", Mind, vol. 9, pp. 188–205.

[10] Craig A. D. (2009), "*Handbook of Emotion, ch. Interoception and emotion: A neuroanatomical perspective*", New York: September: The Guildford Press, ISBN 978-1-59385-650-2.

[11] Jin X. and Wang Z (2005), "*An emotion space model for recognition of emotions in spoken chinese*", in In ACII (J. Tao, T. Tan, and R. Picard, eds.), LNCS 3784, Verlag Berlin Heidelberg., Springer, pp. 397–402.

[12] Williams C. E. and Stevens K. N. (1981), "*Vocal correlates of emotional states. Speech Evaluation in Psychiatry*", in Grune and Stratton Inc.

[13] Cahn J. (1990), "*The generation of affect in synthesized speech*", Journal of American Voice Input/Output Society, vol. 8, pp. 1–19.

[14] Makhoul J. (1975), "*Linear prediction: A tutorial review*", Proceedings of the IEEE, vol. 63, no. 4, pp. 561–580.

[15] Scherer K.R. (2000), "*Psychological models of emotion. The neuropsychology of emotion*", New York, USA: Oxford University Press, pp. 137–162.

[16] Gmytrasiewicz PJ, Lisetti CL (2000), "*Using decision theory to formalize emotions for multi-agent system applications*", in 4th international conference on multi-agent sys-tems, Boston, USA.

[17] Anscombe E, Geach P (1970), "*Descartes philosophical writings*". The Open University, NelsonBaken RJ, Orlikoff RF (2000) Clinical measurements of speech and voice. sSingular Thomson learning, San Diego, USA.

[18] Cornelius, R. R. (2000), "*Theoretical approaches to emotion*", in ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Belfast, United Kingdom.

[19] Plutchik R. (1994), "*The psychology and biology of emotion*", New York, USA: HarperCollins College Publishers.

[20] Nisimura R., Omae S., Kawahara H., Irino T. (2006), "*Analyzing dialogue data for real-worldemo-tional speech classification*", in International conference on speech and language processing (ICSLP), Pittsburgh, USA.

[21] Schlosberg H. (1954), "*Three dimensions of emotion*", Psychol Rev, pp. 61(2):81–88.

[22] Ekman P. (1992), "*An argument for basic emotions*", Cogn Emot, vol. 6(3–4), pp. 169–200.

[23] Fujisawa T., Cook N.D. (2004), "*Identifying emotion in speech prosody using acoustical cues ofharmony*", in International conference on speech and language processing (ICSLP), Jeju, Korea.

[24] H. Miwa, T. Umetsu, A. Takanishi, H. Takanobu (2000), "*Robot personalization based on the mental dynamics*", in IEEE/RSJ Conference on Intelligent Robots and Systems, Takamatsu.

[25] Devillers L., Vasilescu I., Lamel L. (2002), "*Annotation and detection of emotion in a task-oriented human-human dialog corpus*", in ISLE Workshop on dialogue tagging, Edinburgh, United Kingdom.

[26] Power M., Dalgleish T. (1997), "*Cognition and emotion From order to disorder*. Pschology Press, Hove, United Kingdom.

[27] Koolagudi S. G. , Kumar N. and Rao K. S. (2011), "*Speech emotion recognition using segmental level prosodic analysis*", in In International

Conference on Devices and Communication, (Mesra, India), Birla Institute of Technology, IEEE Press.

[28] Cowie R, Schr¨oder M. (2004), "*Piecing together the emotion jigsaw*", in Workshop onmultimodal interaction and related machine learning algorithms (MLMI04), Martigny, Switzerland, pp. 305–317.

[29] J. Nicholson, K. Takahashi, R. Nakatsu (2000), "*Emotion recognition in speech using neural networks*", in Neural Comput, Appl. 9, pp. 290–296.

[30] B.Schuller ,G.Rigoll, M.Lang (2004), "*Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture*", in Proceedings of the ICASSP.

[31] L. Fu, X. Mao, L. Chen (2008), "*Speaker independent emotion recognition based on svm/hmms fusion system*", in International Conference on Audio, Language and Image Processing (ICALIP 2008).

[32] J. Hansen, D. Cairns (1995), "*Icarus: source generator based real-time recognition of speech in noisy stressful and Lombard effect environments*", Speech Commun, vol. 16, no. 4, pp. 391–422.

[33] J. Ma, H. Jin, L. Yang, J. Tsai (2006), "*Ubiquitous Intelligence and Computing*", in Third International Conference, UIC 2006, Wuhan, China, September 3–6, 2006, Proceedings (Lecture Notes in ComputerScience), Springer-Verlag, NewYork, Inc., Secaucus, NJ, USA.

[34] R. Banse, K. Scherer (1996), "*Acoustic profiles in vocal emotion expression*", J. Pers. Soc. Psychol, vol. 70, no. 3, pp. 614–636.

[35] V. Hozjan, Z. Kacic (2003), "*Context-independent multilingual emotion recognition from speech signal*", Int. J. Speech Technol, vol. 6, pp. 311–320.

[36] P.R. Kleinginna Jr., A.M. Kleinginna (1981), "*A categorized list of emotion definitions, with suggestions for aconsensual definition*", Motivation Emotion, vol. 5, no. 4, pp. 345–379.

[37] R. Fernandez (2004), "*A computational model for the automatic recognition of affect in speech*".Ph.D. Thesis, Massachusetts Institute of Technology.

[38] J. Liscombe (2007), "*Prosody and speaker state: paralinguistics, pragmatics, andproficiency*", Ph.D. Thesis, Columbia University.

[39] Reeves B., Nass C. (1996), "*The media equation: How people treat computers, television, newmedia like real people and places*", United Kingdom: Cambridge University Press, Cambridge.

[40] Schroder M., Cowie R. (2006), "*Developing a consistent view on emotion-orientedcomputing*". In: Renals S, Bengio S (eds) MLMI 2005, LNCS 3869, Springer, Heidelberg.

[41] Peter C., Beale R. (eds) (2008), "*Affect and emotion in human-computer interaction: Fromtheory to applications*. (Lecture notes in computer science), Berlin, Germany: Springer.

[42] Holzapfel H., F¨ugen C., Denecke M., Waibel A. (2002), "*Integrating emotional cues into aframe-work for dialogue management*", in Proceedings of the international conference on multimodal interfaces, Pittsburgh, USA.

[43] Brown P., Levinson S.C. (1987), "*Politeness - Some universals in language Use*", Cambridge, United Kingdom: Cambridge University Press.

[44] Walker MA, Cahn JE, Whittaker SJ (1997a), "*Improvising linguistic style: social and affectivebases of agent personality*", in Johnson WL, Hayes-Roth B (eds) Proceedings of the first international conference on autonomous agents (Agents'97), Marina del Rey, USA.

[45] Rodrigues, L. M. L., & Carvalho, M. (2003), "*Emotional and motivational ITS architecture*", in In Proceedings 3rd IEEE International Conference on Advanced Technologies (pp. 467), July. IEEE, Joensuu, Finland.

[46] Sun Y, Willett D, Brueckner R, Gruhn R, B ¨uhler D (2006), "*Experiments on Chinese speechrecognition with tonal models and pitch estimation using the Mandarin Speecon data*", in International conference on speech and language processing (ICSLP), Pittsburgh, USA.

[47] Tepperman J., Traum D., Narayanan S. (2006), "*YeahRight: Sarcasm recognition for spokendialogue systems*", in International conference on speech and language processing (ICSLP), Pittsburgh,USA.

[48] Rank E., Pirker H. (1998), "*Generating emotional speech with a concatenative synthesizer*", in Inter-national conference on speech and language processing (ICSLP), Sydney, Australia.

[49] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao (2009), "*IITKGP-SESC: Speech Database for Emotion Analysis*", in Communications in Computer and Information Science, JIIT University, Noida, India: Springer, issn: 1865-0929 ed., August 17–19.

[50] M. Schroder and R. Cowie (2006), "*Issues in emotion-oriented computing – toward a shared understanding*", Workshop on Emotion and Computing. HUMAINE.

[51] Koike K, Suzuki H, Saito H (1998), "*Prosodic parameters in emotional speech*", in International conference on speech and language processing (ICSLP), Sydney, Australia.

[52] Martin J-C, Caridakis G, Devillers L, Karpouzis K, Abrilian S (2006), "*Manual annotation andautomatic image processing of multimodal emotional behaviours: Validating the annotation of TV interviews*", in International conference on language resources and evaluation (LREC),Genova, Italy.

[53] Picard R.W. (2000), "*Toward computers that recognize and respond to user emotion*", IBM systems journal, 39(3.4), Picard, R. W. (2000). Toward computers that recognize and respond to user emotion.

[54] Bosma W, André E (2004), "*Exploiting emotions to disambiguate dialogue acts*", in International conference on intelligent user interfaces (IUI), Funchal, Portugal.

[55] Kim K.H., Bang S.H., Kim S.R. (2004), "*Emotion recognition system using short-termmonitoring of physiological signals*, Medical and biological engineering and computing, 42(3), pp. 419–427.

[56] Cowie R, Douglas-Cowie E, Apolloni B, Taylor J, Romano A, Fellenz W (1999a), "*What aneural net needs to know about emotion words*", in Proceedings of the 3rd IMACS international multiconference on circuits, systems, communications and computers (CSCC'99), Athens, Greece.

[57] Le Thi Xuyen (1989), "*Etude contrastive de l'intonation expressive en français et en vietnamien*, Thèse de Doctorat "Nouveau Régime", Université Paris 3.

[58] Do Tien Thang (2009), "*Primary examination of Vietnamese intonation*, Hanoi National University Publishing House.

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

[60] Dang-Khoa Mac, Do-Dat Tran (2015), "*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, pp. 273-287.

[61] Mac Dang Khoa (2012), "*Génération de parole expressive dans le cas des langues à tons*", MICA, INPG.

[62] Thi Duyen Ngo, The Duy Bui (2012), "*A study on prosody of Vietnamese emotional speech*", in Proceedings of the Fourth International Conference on Knowledge and Systems Engineering (KSE 2012), IEEE, Danang city, Vietnam, Aug 17-19.

[63] Christopher M. Bishop F.R.Eng (2006), "*Pattern Recognition and Machine Learning*", Springer Science +Business Media, LLC.

[64] Jay L. Devore (2010), "*Probability and Statistics for Engineering and the Science*s", Eighth Edition, Brooks/Cole Edition.

[65] Trevor Hastie, Robert Tibshirani, Jerome Friedman (2013), "*The Elements of Statistical Learning*", 10th Edition, USA: Springer.

[66] Y. Ephraim, N. Merhav (2002), "*Hidden Markov processes*", IEEE Trans. Inf. Theory, vol. 48, no. 6, pp. 1518–1569.

[67] L. Rabiner, B. Juang (1986), "*An introduction to hidden Markov models*", IEEE ASSP Mag, vol. 3, no. 1, pp. 4-16.

[68] A. Dempster, N. Laird, D. Rubin (1977), "*Maximum likelihood from incomplete data via the EM algorithm*", J. R. Stat. Soc, vol. 39, pp. 1-38.

[69] F. Dellert, T. Polzin, and A. Waibel (1996), "*Recognizing emotion in speech*", in 4th International Conference on Spoken Language Processing, Philadelphia, PA, USA.

[70] Cheng X., Duan Q. (2012), "*Speech Emotion Recognition Using Gaussian Mixture Model*", in The 2nd International Conference on Computer Application and System Modeling.

[71] Utane A., Nalbalwar S. (2013), "*Emotion Recognition Through Speech Using Gaussian Mixture Model And Hidden Markov Model*", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 4, pp. 742-746, April.

[72] C. Breazeal, L. Aryananda (2002), "*Recognition of affective communicative intent in robot-directed speech*", Autonomous Robots 2, pp. 83–104.

[73] N. Vlassis, A. Likas (1999), "*A kurtosis-based dynamic approach to Gaussian mixture modeling*", in IEEE Trans. Syst. Man Cybern. 29.

[74] M. Slaney, G. McRoberts (2003), "*Babyears: a recognition system for affective vocalizations*", Speech Commun. 39, pp. 367–384.

[75] B. Schuller (2002), "*Towards intuitive speech interaction by the integration of emotional aspects*", in IEEE International Conference on Systems, Man and Cybernetics.

[76] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss (2005), "*A database of German emotional speech*", in Proceedings of the Interspeech, Lissabon, Portugal.

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

[78] Kalyana Kumar Inakollu, Sreenath Kocharla (2013), "*Gender Dependent and Independent Emotion Recognition System for Telugu Speeches Using Gaussian Mixture Models*", International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, no. 11, pp. 4172-4175.

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

[80] Thurid Vogt, Elisabeth André (2006), "*Improving Automatic Emotion Recognition from Speech via Gender Differentiation*", in In Proceedings of Language Resources and Evaluation Conference LREC.

[81] T. Nwe, S. Foo, L. De Silva (2003), "*Speech emotion recognition using hidden Markov model*", Speech Commun, vol. 41, pp. 603–623.

[82] O. Kwon, K. Chan, J. Hao, T. Lee (2003), "*Emotion recognition by speech signal*", EUROSPEECH Geneva, pp. 125-128.

[83] C. Lee, S. Yildrim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, S. Narayanan (2004), "*Emotion recognition based on phoneme classes*", in Proceedings of ICSLP.

[84] Petrushin (2000), "*Emotion recognition in speech signal: experimental study, development and application*", in Proceedings of the ICSLP 2000.

[85] Petrushin, V. A. (2000), "*Emotion recognition in speech signal: experimental study, development, and application*", in In Sixth International Conference on Spoken Language Processing.

[86] Haytham M. Fayek, Margaret Lech, Lawrence Cavedon (2017), "*Evaluating deep learning architectures for Speech Emotion Recognition*".Neural Networks.

[87] Rawat et al. (2015), "*Emotion Recognition through Speech Using Neural Network*", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 5, pp. 422-428.

[88] A. Razak, R. Komiya, M. Abidin (2005), "*Comparison between fuzzy and nn method for speech emotion recognition*", in 3rd International Conference on Information Technology and Applications ICITA 2005.

[89] O. Pierre-Yves (2003), "*The production and recognition of emotions in speech: features and algorithms,* Int. J. Human–Computer Stud. 59, pp. 157–183.

[90] B. Schuller, G. Rigoll, M. Lang (2003), "*Hidden Markov model-based speech emotion*", in International Conference on Multimedia and Expo (ICME).

[91] V. Petrushin (2000), "*Emotion recognition in speech signal: experimental study, development and application*", in Proceedings of the ICSLP.

[92] B. Schuller, M. Lang, G. Rigoll (2005), "*Robust acoustic speech emotion recognition by ensembles of classifiers*", in Proceedings of the DAGA'05, 31, Deutsche Jahrestagung fur Akustik, DEGA, 2005.

[93] M. Lugger, B. Yang (2009), "*Combining classifiers with diverse feature sets for robust speaker independent emotion recognition*", in Proceedings of EUSIPCO.

[94] L.I.Kuncheva (2004), "*Combining Pattern Classifiers: Methodsand Algorithms*".Wiley.

[95] J. Wu, M.D. Mullin, J.M. Rehg (2005), "*Linear asymmetric classifier for cascade detectors*", in 22th International Conference on Machine Learning.

[96] D. Mashao, M. Skosan (2006), "*Combining classifier decisions for robust speaker identification*", Pattern Recognition 39 (1), pp. 147–155.

[97] L.I. Kuncheva (2002), "*A theoretical study on six classifier fusion strategies*", in IEEE Trans. Pattern Anal. Mach. Intell. 24.

[98] M. Lugger, B. Yang (2008), "*Psychological motivated multi-stage emotion classification exploiting voice quality features*", Speech Recognition, no. ISBN 978-953-7619-29-9, pp. 395-410.

[99] H. Schlosberg (1954), "*Three dimensions of emotion*", Psychological Rev61 (2), pp. 81–88.

[100] K. Stevens, H. Hanson (1994), "*Classification of glottal vibration from acoustic measurements*", Vocal Fold Physiol, pp. 147–170.

[101] Lugger M. and Yang B. (2007), "*The relevance of voice quality features in speaker independent emotion recognition*", in ICASSP, Honolulu, Hawai, USA.

[102] Rajisha T. M.a, Sunija A. P.b, Riyas K. S (2015), "*Performance Analysis of Malayalam Language Speech Emotion Recognition System using ANN/SVM*", in International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST), Kerala, India.

[103] Viet Hoang Anh, Manh Ngo Van, Bang Ban Ha, Thang Huynh Quyet (2012), "*A real-time model based Support Vector Machine for emotion recognition through EEG*", in International Conference on Control, Automation and Information Sciences (ICCAIS), Ho Chi Minh city, Vietnam, Nov 26-29, Ho Chi Minh city, Vietnam, Nov 26-29.

[104] La Vu Tuan, Huang Cheng-Wei, Ha Cheng, Zhao Li (2013), "*Emotional Feature Analysis and Recognition from Vietnamese Speech*", Journal of Signal Processing, China,, vol. 20, no. 10, pp. 1423-1432.

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

[106] Pao T. L., Chen Y. T., Yeh J. H. , and Liao W. Y. (2005), "*Combining acoustic features for improved emotion recognition in mandarin speech*", In ACII (J. Tao, T. Tan, and R. Picard, eds.), (LNCS 3784)", Springer-Verlag Berlin Heidelberg, pp. 279–285.

[107] Schroder M., Cowie R., Douglas-Cowie E., Westerdijk M. and Gielen S. (2001), "*Acoustic correlates of emotion dimensions in view of speech synthesis*", in EUROSPEECH 2001 Scandinavia, 2nd INTERSPEECH Event, September 3–7. 7th European Conference on Speech Communication and Technology, Aalborg, Denmark.

[108] Williams C. and Stevens K. (1972), "*Emotionsandspeech:some acoustical correlates*", Journal of Acoustic Society of America, vol. 52, no. no. 4 pt 2, p. 1238–1250.

[109] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005), "*A database of germanemo-tional speech*", in European conference on speech and language processing (EUROSPEECH), Lisbon, Portugal.

[110] Batliner A., Buckow J., Niemann H., N¨oth E. and VolkerWarnke (2000), "*Verbmobile Foundations of speech to speech translation*", ISBN 3540677836, 9783540677833: Springer.

[111] Campbell N, Devillers L, Douglas-Cowie E, Auberg´e V, Batliner A, Tao J (2006), "*Resources for the processing of affect in interactions*", in ELRA (ed) International conference on language resources and evaluation (LREC), Genova, Italy.

[112] Campbell N (2000), "*Databases of emotional speech*", in Proceedings of ISCA.

[113] Johannes Pittermann, Angela Pittermann, Wolfgang Minker (2010), "*Handling Emotions in Human-Computer Dialogues*", Springer.

[114] University of Pennsylvania Linguistic Data Consortium (2002), "*Emotional prosody speech and transcripts*", July 2002.

[115] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss (2005), "*A database of German emotional speech*", in Proceedings of the Interspeech, Lissabon.

[116] Engberg, I. S., & Hansen, A. V. (1996), "*Documentation of the Danish emotional speech database (DES)* ". Internal AAU report, Center for Person Kommunikation, Denmark, 22.

[117] D. Morrison, R. Wang, L. De Silva (2007), "*Ensemble methods for spoken emotion recognition in call-centres*", Speech Commun, vol. 2, no. 49, pp. 98-112.

[118] T. Nwe, S. Foo, L. De Silva (2003), "*Speech emotion recognition using hidden Markov models*", Speech Commun, no. 41, pp. 603–623.

[119] V. Hozjan, Z. Moreno, A. Bonafonte, A. Nogueiras (2002), "*Interface databases: design and collection of a multilingual emotional speech database*", in Proceedings of the 3rd InternationalConference on Language Resources and Evaluation (LREC'02) Las Palmas de Gran Canaria, Spain.

[120] C. Breazeal, L. Aryananda (2002), "*Recognition of affective communicative intent in robot-directed speech*", Autonomous Robots 2, pp. 83-104.

[121] M. Slaney, G. McRoberts (2003), "*Babyears: a recognition system for affective vocalizations*", Speech Commun, no. 39, pp. 367–384.

[122] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, G. Rigoll (2005), "*Speaker independent speech emotion recognition by ensemble*

classification", in IEEE International Conference on Multimedia and Expo, ICME.

[123] B. Schuller (2002), "*Towards intuitive speech interaction by the integration of emotional aspects*", in IEEE International Conference on Systems, Man and Cybernetics, vol. 6, 2002, pp. 6.

[124] E. Kim, K. Hyun, S. Kim, Y. Kwak (2007), "*Speech emotion recognition using eigen-fft in clean and noisy environments*", in The 16th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN.

[125] J. Zhou, G. Wang, Y. Yang, P. Chen (2006), "*Speech emotion recognition based on rough set and SVM*", in 5th IEEE International Conference on Cognitive Informatics, ICCI.

[126] N. Amir, S. Ron, N. Laor (2002), "*Analysis of an emotional speech corpus in Hebrew based on objective criteria*", Speech Emotion, pp. 29–33.

[127] H. Hu, M. Xu, W. Wu (2000), "*Dimensions of emotional meaning in speech*", in Proceedings of the ISCA ITRW on Speech and Emotion.

[128] Lê Xuân Thành (2018), "*Tổng hợp tiếng Việt với các chất giọng khác nhau và có biểu lộ cảm xúc*", Luận án, Đại học Bách khoa Hà Nội.

[129] Joseph Picone (1995), "*Fundamentals of speech recognition: a short course*", Department of Electrical and Computer Engineering, Mississippi State University.

[130] Makhoul J. (1975), "*Linear prediction: A tutorial review*", Proceedings of the IEEE, Vols. 63, no. 4, pp. 561–580.

[131] Rabiner L. R. and Juang B. H (1993), "*Fundamentals of Speech Recognition, Englewood Cliffs*", New Jersy: Prentice-Hall.

[132] Benesty J., Sondhi M. M., and Huang Y. (2008), "*Springer Handbook on Speech Processing*", Springer Publishers.

[133] Rao, K. S., & Koolagudi, S. G. (2012), "*Emotion Recognition using Speech Features*", Springer Science & Business Media, .

[134] Ananthapadmanabha T. V. and Yegnanarayana B. (1979), "*Epoch extraction from linear prediction residual for identification of closed glottis interval*", IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 27, pp. 309–319.

[135] B.Yegnanarayana, S.R.M.Prasanna, and K. Rao (2002), "*Speech enhancement using excitation source information*", in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Orlando, Florida, USA.

[136] Bajpai A. and Yegnanarayana B. (2008), "*Combining evidence from sub-segmental and segmental features for audio clip classification*", in IEEE Region 10 Conference (TENCON), India, IIIT, Hyderabad.

[137] Wakita H. (1976), "*Residual energy of linear prediction to vowel and speaker recognition*", IEEE Trans. Acoust. Speech Signal Process, vol. 24, pp. 270–271.

[138] Rao K. S., Prasanna S. R.M. and Yegnanarayana B. (2007), "*Determination of instants of significant excitation in speech using hilbert envelope and group delay function*", IEEE Signal Processing Letters, vol. 14, pp. 762–765.

[139] Bajpai A. and Yegnanarayana B. (2004), "*Exploring features for audio clip classi?cation using LP residual and AANN models*", in The international Conference on Intelligent Sensing and Information Processing 2004 (ICISIP 2004), Chennai, India.

[140] Yegnanarayana B., Swamy R. K., and Murty K.S.R. (2009), "*Determining mixing parameters from multispeaker data using speech-specific information*", IEEE Trans. Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1196–1207, ISSN 1558–7916.

[141] G. Bapineedu, B. Avinash, S. V. Gangashetty, and B. Yegnanarayana (2009), "*Analysis of lombard speech using excitation source information*", INTERSPEECH, September, 6–10, Brighton, UK.

[142] K. E. Cummings and M. A. Clements (1995), "*Analysis of the glottal excitation of emotionally styled and stressed speech*", Journal of Acoustic Society of America, vol. 98, pp. 88-98.

[143] Zhen-Hua Ling, Yu Hu, Ren-Hua Wang (2005), "*A novel source analysis method by matchin spectral characters of lf model with straight spectrum*", Springer-Verlag, pp. 441–448.

[144] Atal B. S. (1972), "*Automatic speaker recognition based on pitch contours*", Journal of Acoustic Society of America, vol. 52, no. 6, pp. 1687–1697.

[145] Thevenaz P. and Hugli H. (1995), "*Usefulness of LPC residue in textindependent speaker verification*", Speech Communication, vol. 17, pp. 145–157.

[146] Yegnanarayana B., Murthy P. S., Avendano C., and H. Hermansky (1998), "*Enhancement of reverberant speech using LP residual*", in IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle, WA , USA.

[147] Mubarak O. M. , Ambikairajah E. , and Epps J. (2005), "*Analysis of an mfcc-based audio indexing system for eficient coding of multimedia sources*", in The 8th International Symposium on Signal Processing and its Applications, (Sydney, Australia), 28–31 August.

[148] Pao, T. L., Chen, Y. T., Yeh, J. H., Cheng, Y. M., & Chien, C. S. (2007), "*Feature combination for better differentiating anger from neutral in Mandarin emotional speech*", in In International Conference on Affective Computing and Intelligent Interaction. Springer, Berlin, Heidelberg.

[149] Kamaruddin N. and Wahab A. (2009), "*Features extraction for speech emotion*", Journal of Computational Methods in Science and Engineering, vol. 9, no. 9, pp. 1–12.

[150] Neiberg D., Elenius K. and Laskowski K. (2006), "*Emotion recognition in spontaneous speech using GMMs*", in In INTERSPEECH 2006 - ICSLP, Pittsburgh, Pennsylvania.

[151] Bitouk D., Verma R. and Nenkova A. (2010), "*Class-level spectral features for emotion recognition*", Speech Communication. Article in press.

[152] Sigmund M. (2007), "*Spectral analysis of speech under stress*", IJCSNS International Journal of Computer Science and Network Security, vol. 7, pp. 170–172.

[153] Banziger T. and Scherer K. R (2005), "*The role of intonation in emotional expressions*", vol. 46, Speech Communication, pp. 252–267.

[154] Cowie R. and Cornelius R. R. (2003), "*Describing the emotional states that are expressed in speech*", vol. 40, Speech Communication, pp. 5–32.

[155] Rao K. S. and Yegnanarayana B. (2006), "*Prosody modification using instants of ignificant excitation",* IEEE Trans. Speech and Audio Processing, vol. 14, pp. 972–980.

[156] Werner S. and Keller E. (1994), "*Prosodic aspects of speech*", in Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts State of the Art the Future Challenges (E. Keller, ed.), Chichester: John Wiley, pp. 23–40.

[157] Murray I. R. and Arnott J. L. (1995), "*Implementation and testing of a system for producing emotion by rule in synthetic speech*", Speech Communication, vol. 16, pp. 369–390.

[158] Murray I. R., Arnott J. L. and Rohwer E. A. (1996), "*Emotional stress in synthetic speech: Progress and future directions*", Speech Communication, vol. 20, pp. 85–91.

[159] Scherer K. R. (2003), "*Vocal communication of emotion: A review of research paradigms*", vol. 40, Speech Communication, pp. 227–256.

[160] McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., & Stroeve, S. (2000), "*Approaching automatic recognition of emotion from voice: A rough benchmark*", ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.

[161] T. L. Nwe, S. W. Foo, and L. C. D. Silva (2003), "*Speech emotion recognition using hidden Markov models*", Speech Communication, vol. 41, pp. 603–623, Nov.

[162] Ververidis D. and Kotropoulos C. (2006), "*A state of the art review on emotional speech databases*", in In Eleventh Australasian International Conference on Speech Science and Technology, Auckland, New Zealand.

[163] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura (2003), "*A corpus-based speech synthesis system with emotion*", Speech Communication, vol. 40, pp. 161–187, Apr.

[164] Luengo I., Navas E., Hernáez I., and Sánchez J. (2005), "*Automatic emotion recognition using prosodic parameters*", in INTERSPEECH, Lisbon, Portugal.

[165] Iliou T. and Anagnostopoulos C. N. (2009), "*Statistical evaluation of speech features for emotion recognition*", in Fourth International Conference on Digital Telecommunications, Colmar, France.

[166] Kao Y. hao and Lee L. shan (2006), "*Feature analysis for emotion recognition from mandarin speech considering the special characteristics of chinese language*", in INTERSPEECH - ICSLP, Pittsburgh, Pennsylvania.

[167] Zhu A. and Luo Q. (2007), "*Study on speech emotion recognition system in e learning*", in Human Computer Interaction, Part III, HCII (J. Jacko, ed.), Berlin Heidelberg, Springer Verlag, LNCS:4552, pp. 544–552,.

[168] Wang Y., Du S., and Zhan S. (2008), "*Adaptive and optimal classi?cation of speech emotion recognition*", in Fourth International Conference on Natural Computation.

[169] Zhang S. (2008), "*Emotion recognition in chinese natural speech by combining prosody and voice quality features*", in In Advances in Neural

Networks, Lecture Notes in Computer Science, Volume 5264 (S. et al., ed.), Berlin Heidelberg, Springer Verlag, pp. 457–464.

[170] F. Dellaert, T. Polzin, and A. Waibel (1996), "*Recognising emotions in speech*", in ICSLP 96, Oct.

[171] D. Ververidis, C. Kotropoulos, and I. Pitas (2004), "*Automatic emotional speech classification*", in ICASSP 2004, IEEE, pp. I593 - I596.

[172] Rao K. S., Reddy R., Maity S. and Koolagudi S. G. (2010), "*Characterization of emotions using the dynamics of prosodic features*", in International Conference on Speech Prosody, Chicago, USA.

[173] Jean Vroomen, René Collier, Sylvie Mozziconacci (1993), "*Duration and intonation in emotional speech*", in Proceedings of the Third European Conference on Speech Communication and Technology, Berlin, Germany.

[174] Deepa P. Gopinath, Sheeba P.S, Achuthsankar S. Nair (2007), "*Emotional Analysis for Malayalam Text to Speech Synthesis Systems*", in Proceedings of the Setit 2007 - 4th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, Tunisia.

[175] Pao, T. L., Chen, Y. T., Yeh, J. H., & Liao, W. Y. (2005), "*Combining acoustic features for improved emotion recognition in mandarin speech*", in International Conference on Affective Computing and Intelligent Interaction, Springer, Berlin, Heidelberg.

[176] Yixiong Pan, Peipei Shen, Liping Shen (2012), "*Speech Emotion Recognition Using Support Vector Machine*", International Journal of Smart Home, vol. 6, no. 2, pp. 101-108.

[177] R. Subhashree1, G. N. Rathna (2016), "*Speech Emotion Recognition: Performance Analysis based on Fused Algorithms and GMM Modelling*", Indian Journal of Science and Technology, Vols. Vol 9(11), March , pp. 1-8.

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

[179] Kun Han, Dong Yu, Ivan Tashev, "*Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine*.", in INTERSPEECH 2014, Singapore, 2014.

[180] Mai Ngoc Chu (1997), "*The basics of linguistics and Vietnamese*", Hanoi: Education Publishing House.

[181] "www.praat.org", [Online].

[182] Jean-François Bonastre, Frédéric Wils (2005), "A*lize, a free toolkit for speaker recognition*", in IEEE International Conference.

[183] Ian H. Witten, Eibe Frank (2005), "*Data Mining: Practical machine learning tools and techniques*", Second Edition, Morgan Kaufmann Publishers.

[184] J. C. Platt, (1998), "*Writer Technical Report MSR-TR-98-14*", [Performance]. Microsoft Research.

[185] Quinlan J. R. (1993), "*C4.5: Programs for Machine Learning*", Morgan Kaufmann Publishers.

[186] Eyben, Florian, Martin Wöllmer, and Björn Schuller (2010), "*Opensmile: the munich versatile and fast open-source audio feature extractor*", in Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italy.

[187] Siqing Wua, Tiago H. Falkb, Wai-Yip Chan (2011), "*Automatic speech emotion recognition using modulation spectral features*", Speech Communication, vol. 53, no. 5, pp. 768–785.

[188] S. Lalitha, Abhishek Madhavan, Bharath Bhushan, Srinivas Saketh (2014), "*Speech emotion recognition*", in Proceedings of the International Conference on Advances in Electronics, Computers and Communications, Bangalore, India.

[189] Maria Schubiger (1960), "*English intonation: its form and function*", Language, Vols. 36, No. 4, pp. 544-548.

[190] Ankush Chaudhary, Ashish Kumar Sharma, Jyoti Dalal, Leena Choukiker (2015), "*Speech Emotion Recognition*", Journal of Emerging Technologies and Innovative Research, vol. 2, no. 4, pp. 1169-1171.

[191] Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., and Deller Jr., J. R. (2002), "*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.

[192] Bin MA, Donglai ZHU and Rong TONG (2006), "*Chinese Dialect Identification Using Tone Features Based On Pitch*", in ICASSP.

[193] Bağcı U., Erzin E. (2005), "*Boosting Classifiers for Music Genre Classification*", in In: Yolum., Güngör T., Gürgen F., Özturan C. (eds)

Computer and Information Sciences – ISCIS, Lecture Notes in Computer Science, vol. vol 3733, Berlin, Heidelberg, Springer.

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

[195] Bendjillali R. I.; Beladgham M.; Merit K.; Taleb-Ahmed A (2019), "*Improved Facial Expression Recognition Based on DWT Feature for Deep CNN*", Electronics, 2019, 8(3), 324.", Electronics, Vols. 8(3), 324.

[196] Yue Zhao; Xingyu Jin; Xiaolin Hu (2017), "*Recurrent Convolutional Neural Network for Speech Processing*", in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[197] Abdulsalam W. H., Alhamdani R. S., Abdullah M. N (2019), "*Facial Emotion Recognition from Videos Using Deep Convolutional Neural Networks",* International Journal of Machine Learning and Computing.

[198] Supaporn Bunrit, Thuttaphol Inkian, Nittaya Kerdprasop, and Kittisak Kerdprasop (2019), "*Text-Independent Speaker Identification Using Deep Learning*", International Journal of Machine Learning and Computing, vol. 9, no. 2, pp. 143-148, April .

[199] Stuhlsatz; André et al. (2011), "*Deep neural networks for acoustic emotion recognition: Raising the benchmarks*", in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[200] Aharon satt; Shai Rozenberg; Ron Hoory (2017), "*Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms*.", in INTERSPEECH 2017, Stockholm, Sweden, August 20–24.

[201] Kun Han; Dong Yu; Ivan Tashev 9 (2014), "*Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine*", in INTERSPEECH.

[202] Eduard Frant et al. (2017), "*Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots*", Romanian Journal of Information Science and Technology, 2017, vol. 20, no. 3, pp. 222–240.

[203] Michael Neumann; Ngoc Thang Vu (2017), "*Attentive Convolutional Neural Network Based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech",* in InterSpeech.

[204] Wootaek Lim, Daeyoung Jang, and Taejin Lee (2016), "*Speech emotion recognition using convolutional and recurrent neural networks*", in Asia

Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA).

[205] Abdul Malik Badshah, Jamil Ahmad, and Nasir Rahim (2017), "*Speech emotion recognition from spectrograms with deep convolutional neural network*", in in International Conference on Platform Technology and Service (PlatCon).

[206] Yafeng Niu, Dongsheng Zou, Yadong Niu, Zhongshi He, and Hua Tan (2017), "*A breakthrough in speech emotion recognition using deep retinal convolution neural networks",* arXiv:1707.09917.

[207] Najafabadi, Maryam M., et al. (2015), "*Deep learning applications and challenges in big data analytics*", Journal of Big Data, vol. 2, no. 1, (2015):1.

[208] X. Chen and X. Lin (2014), "*Big Data Deep Learning: Challenges and Perspectives*", IEEE Access, vol. 2, pp. 514-525.

[209] Badshah; Abdul Malik et al. (2017), "*Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network",* in International Conference on Platform Technology and Service (PlatCon).

[210] Djork-Arne Clevert; Thomas Unterthiner & Sepp Hochreiter (2016), "*Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*", in In 4th International Conference on Learning Representations, (ICLR), arXiv:1511.07289 [cs. LG]..

[211] Sergey Ioffe; Christian Szegedy (2015), "*Batch normalization: accelerating deep network training by reducing internal covariate shift.*", in ICML'15 Proceedings of the 32nd International Conference on Machine Learning, France, July 06 – 11, 2015, 37, 448-456. Lille..

[212] Matthew D. Zeiler; Rob Fergus (2013), "*Stochastic Pooling for Regularization of Deep Convolutional Neural Networks*", in Proceedings of the International Conference on Learning Representation (ICLR), arXiv preprint arXiv:1301.3557.

[213] Nitish Srivastava et al (2014), "*Dropout: A Simple Way to Prevent Neural Networks from Overfitting*", Journal of Machine Learning Research 15, pp. 1929-1958.

[214] Anthimopoulos, Marios, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou (2016), "*Lung pattern classification for interstitial lung diseases using a deep convolutional neural network*", IEEE Transactions on Medical Imaging 35, No. 5, pp. 1207-1216.

[215] Chollet F., "*Keras" (2015)*", https://github.com/fchollet/keras, [Online].

[216] Chollet F., "*Tensorflow*" (2015)", https://github.com/topics/tensorflow, [Online].

[217] Gideon, John, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost (2017), "*Progressive neural networks for transfer learning in emotion recognition*", arXiv preprint arXiv:1706.03256 .

[218] Deng, Jun, Sascha Frühholz, Zixing Zhang, and Björn Schuller (2017), "*Recognizing emotions from whispered speech based on acoustic feature transfer learning*", IEEE Access 5 (2017): 5235-5246.

[219] Latif, Siddique, Rajib Rana, Shahzad Younis, Junaid Qadir, and Julien Epps (2018), "*Transfer learning for improving speech emotion classification accuracy*", arXiv preprint arXiv:1801.06353.

[220] Lech Margaret, Melissa Stolar, Robert Bolia, and Michael Skinner (2018), "*Amplitude-Frequency Analysis of Emotional Speech Using Transfer Learning and Classification of Spectrogram Images*", Advances in Science, Technology and Engineering Systems Journal, Vol. 3, No.4, pp. 363-371.

PHỤ LỤC

A. Danh sách các câu được chọn để thể hiện cảm xúc của bộ ngữ liệu thử nghiệm nhận dạng cảm xúc tiếng Việt nói

TT	Nội dung câu
1	Ông nói gì thế tôi không hiểu
2	Anh đã biết chuyện gì chưa
3	Có chuyện gì thế hả
4	Anh đến đón em nhé
5	Chán quá đi cậu ạ
6	Lại phải chờ hả anh
7	Thôi vui lên đi ông
8	Không biết thì dựa cột mà nghe hiểu chưa
9	Cứ lanh chanh, hỏng hết cả việc rồi
10	Anh đừng nói chuyện với em nữa
11	Có lương rồi!
12	Trời đất ơi! Thuốc gì mà hay quá chừng!
13	Ôi dào, người như vậy không thay đổi được đâu
14	Mai là chủ nhật rồi!
15	Làm gì mà lâu thế
16	Hôm nay chẳng được việc gì cả
17	Ông cho con đi với nhé!
18	Bác đi đâu đấy?
19	Chuyển lá thư này cho anh ấy nhé!
20	Không tặng gì cho em à!
21	Sao nhiều thế ạ?
22	Sao lại không được gì?

B. Kết quả thử nghiệm nhận dạng cảm xúc với bộ ngữ liệu tiếng Đức dùng công cụ Alize dựa trên mô hình GMM

Cơ sở ngữ liệu cảm xúc tiếng Đức được thực hiện thu âm với tần số lấy mẫu 16kHz và 16bit/mẫu trong phòng thu của trường Đại học Berlin. Có 800 phát ngôn được thu âm từ 10 nghệ sĩ chuyên nghiệp (5 nghệ sĩ nam, 5 nghệ sĩ nữ) cho 7 cảm xúc gồm tức, vui, buồn, sợ hãi, ghê tởm, chán nản, bình thường.

Mỗi diễn viên nói một số câu với tất cả các cảm xúc. Mỗi cảm xúc có thể được thu âm từ 1 đến 5 lần. Cơ sở ngữ liệu đã được đánh giá và loại bỏ một số phiên bản do

lỗi hoặc do nhiễu. Tổng số file tiếng nói cho 4 cảm xúc vui, buồn, tức và bình thường là 339 tập tin (151 file giọng nam, 188 file giọng nữ). Bộ ngữ liệu này được thống kê trong Bảng B.1.

Bảng B.1. *Bộ ngữ liệu tiếng Đức với bốn cảm xúc vui, buồn, tức và bình thường*

Cảm xúc	**Giọng nam**					**Giọng nữ**					**Tổng số file theo cảm xúc**
	1	***2***	***3***	***4***	***5***	***1***	***2***	***3***	***4***	***5***	
Tức	12	10	14	11	13	13	12	12	16	14	**127**
Vui	2	4	7	8	6	4	11	10	8	11	**71**
Buồn	4	3	7	7	4	4	9	5	10	9	**62**
Bình thường	4	4	11	9	11	9	10	9	7	5	**79**
Tổng số file theo người nói	**22**	**21**	**39**	**35**	**34**	**30**	**42**	**36**	**41**	**49**	**339**

Tham số dùng cho thử nghiệm gồm các hệ số MFCC, năng lượng cùng với đạo hàm bậc nhất, đạo hàm bậc hai của MFCC và năng lượng. Các tham số này được trích chọn từ công cụ Alize. Thử nghiệm được chia làm 2 trường hợp:

- Trường hợp 1: Số file dùng cho huấn luyện (339 file) đều được đưa vào nhận dạng. Kết quả nhận dạng được thống kê trong Bảng B.2. Tỷ lệ nhận dạng đúng trung bình đạt 96.25%.

Bảng B.2. *Kết quả nhận dạng cảm xúc tiếng Đức trong trường hợp 1*

Cảm xúc	**Số file nhận dạng**	**Số file nhận dạng đúng**	**Tỷ lệ nhận dạng**
Vui	71	67	94%
Buồn	62	61	98%
Tức	127	125	98%
Bình thường	79	75	95%
Tổng số file	339	328	

- Trường hợp 2: Dùng một nửa số file để huấn luyện, nửa còn lại dùng cho nhận dạng. Kết quả nhận dạng trong trường hợp này được thống kê trong Bảng B.3. Tỷ lệ nhận dạng đúng trung bình đạt 64,5%.

Bảng B.3. *Kết quả nhận dạng cảm xúc tiếng Đức trong trường hợp 2*

Cảm xúc	**Số file nhận dạng**	**Số file nhận dạng đúng**	**Tỷ lệ nhận dạng**
Vui	35	12	34%
Buồn	31	22	71%
Tức	63	53	84%
Bình thường	39	27	69%
Tổng số file	168	114	