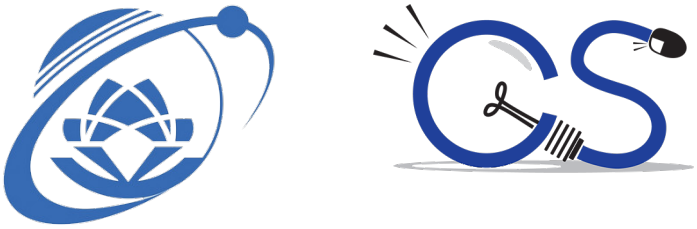
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH



ĐỀ TÀI

PHÂN LOẠI CẢM XÚC TIẾNG NÓI VỚI CÁC LOẠI ĐẶC TRƯNG VÀ MÔ HÌNH

Môn học: Xử lý âm thanh và tiếng nói

Lớp: CS337.O11

GVHD: Th.S Trịnh Quốc Sơn

Nhóm thực hiện:

1. Lê Minh Nguyệt 21521211
2. Nguyễn Thị Thùy 21521514

Thành phố Hồ Chí Minh, Tháng 12 - 2023

Mục lục

Danh sách hình vẽ

Danh sách bảng

Tóm tắt nội dung

Báo cáo này nghiên cứu bài toán phân loại cảm xúc âm thanh. Nhóm thực hiện các thí nghiệm để đánh giá hiệu suất của các bộ đặc trưng và mô hình khác nhau. Kết quả cho thấy bộ đặc trưng MFCC kết hợp với các mô hình SVM, MLP, và CNN cho kết quả tốt nhất.

Nhóm cũng nhận thấy rằng sự lựa chọn và kết hợp linh hoạt giữa các đặc trưng và mô hình là quan trọng để tối ưu hóa hiệu suất phân loại cảm xúc âm thanh. Mục tiêu cuối cùng của nghiên cứu là đánh giá kết quả và nhận biết những điểm mạnh và yếu của từng loại đặc trưng và mô hình, từ đó đề xuất hướng phát triển tương lai trong lĩnh vực phân loại cảm xúc âm thanh.

Chương 1

GIỚI THIỆU

1.1 Lý do chọn đề tài

Bài toán phân loại phân loại cảm xúc là một bài toán quan trọng trong lĩnh vực xử lý âm thanh và có nhiều ứng dụng trong thực tiễn. Mặc dù đã được nghiên cứu và tiếp cận trong nhiều năm, nhưng đối mặt với sự phức tạp của dữ liệu âm thanh và đa dạng của cảm xúc, việc phát triển phương pháp phân loại hiệu quả là vẫn đang là một nhiệm vụ đòi hỏi sự sáng tạo và nghiên cứu chuyên sâu.

Âm thanh có thể được miêu tả bằng nhiều nhóm đặc trưng khác nhau trong đó, đặc trưng miền phổ (spectral features) và ngôn điệu (prosodic features) là hai nhóm đặc trưng được sử dụng phổ biến nhất cho việc phân loại cảm xúc của âm thanh. Các đặc trưng spectral mô tả các đặc tính tần số của âm thanh, trong khi các đặc trưng prosodic mô tả các đặc tính thời gian của âm thanh. Cùng với đó ngày càng có nhiều mô hình phân loại được giới thiệu là cho hiệu quả tốt trong bài toán phân loại cảm xúc trong âm thanh. Mỗi nhóm đặc trưng, mỗi loại đều có những ưu điểm và nhược điểm riêng, nên việc lựa chọn đặc trưng và mô hình phù hợp sẽ ảnh hưởng rất nhiều đến hiệu quả phân loại cuối cùng.

Với những lý do trên, nhóm chọn đề tài "Phân loại cảm xúc âm thanh với các loại đặc trưng và mô hình". Đồ án của nhóm sẽ thực hiện các thực nghiệm trên nhiều loại đặc trưng và mô hình khác nhau để đánh giá hiệu quả của chúng trong việc phân loại cảm xúc của âm thanh.

1.2 Mục tiêu đồ án

- Hiểu được các bước giải quyết một bài toán với dữ liệu là âm thanh
- Nắm được kỹ thuật xử lý dữ liệu âm thanh cơ bản và cách trích xuất đặc trưng từ dữ liệu âm thanh phù hợp với bài toán phân loại cảm xúc giọng nói.
- Giải quyết được bài toán phân loại cảm xúc giọng nói theo một số hướng tiếp cận như máy học truyền thống hay mô hình học sâu
- Khám phá sự kết hợp linh hoạt giữa các đặc trưng và mô hình và xác đinh được các đặc trưng và mô hình hiệu quả nhất cho việc phân loại cảm xúc.

- Đánh giá kết quả áp dụng các phương pháp trên bài toán phân loại cảm xúc, từ đó, nhận biết được ưu nhược điểm của từng loại đặc trưng và mô hình

1.3 Nghiên cứu liên quan

Trong những năm qua, đã có nhiều nghiên cứu và thử nghiệm xoay quanh các mô hình và đặc trưng trên bài toán phân loại cảm xúc tiếng nói.

Các nghiên cứu của nhóm Abbaschian et al. 2021[1] hay nhóm M. G. de Pinto, M. Polignano at al. 2020 [2] đã đề xuất các hướng tiếp cận dựa trên mạng nơ-ron học sâu một cách tương đối hiệu quả.

Nghiên cứu tổng hợp của Akçay, at al. 2020 [3] cũng đã cho thấy cái nhìn tổng quát hơn về các đặc trưng có thể được trích xuất đối với dữ liệu âm thanh tiếng nói và những mô hình đã được thử nghiệm từ các công trình khác trên các đặc trưng đó.

Tuy nhiên, trong phạm vi thực nghiệm của các công trình trên, với mục tiêu là tìm ra giải pháp tối ưu hóa cho bài toán phân loại cảm xúc tiếng nói, rất khó để thấy được sự so sánh giữa các nhóm đặc trưng và mô hình khác nhau khi không có sự thử nghiệm đầy đủ trong một nghiên cứu.

Chương 2

BỘ DỮ LIỆU

2.1 Mô tả bộ dữ liệu

Dữ liệu đóng vai trò vô cùng quan trọng trong bài toán phân loại cảm xúc tiếng nói. Chất lượng và nhãn cảm xúc được gán cho dữ liệu ảnh hưởng trực tiếp đến kết quả của quá trình nhận diện cảm xúc qua âm thanh giọng nói. Có nhiều loại dữ liệu được nghiên cứu trong ngữ cảnh của bài toán này, như dữ liệu mô phỏng (simulated speech emotion database) hay dữ liệu giọng nói tự nhiên (natural speech emotion database),... Trong phạm vi của đồ án, nhóm tiến hành thử nghiệm trên sự kết hợp giữa hai bộ dữ liệu mô phỏng cảm xúc, đó là The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) và Toronto Emotional Speech Set (TESS).

2.1.1 Bộ dữ liệu RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) là một bộ dữ liệu cảm xúc tiếng nói thuộc loại mô phỏng, tức dữ liệu trong bộ sẽ được những diễn viên có kinh nghiệm trong việc đọc thoại thu âm lại một vài câu nói đơn giản với những cảm xúc khác nhau. Loại dữ liệu này thường dễ xây dựng, song, cũng vì vậy mà không thể đảm bảo đầy đủ khả năng truyền tải những cảm xúc trong ngữ cảnh thực tế.

RAVDESS là một bộ dữ liệu khá lớn, với tất cả hơn 7000 tệp dữ liệu dạng âm thanh và video, với mỗi dạng âm thanh hoặc video, bộ dữ liệu chứa tệp giọng nói và giọng hát. Bộ dữ liệu được thu âm và ghi hình bởi 24 diễn viên với ngôn ngữ duy nhất là Tiếng Anh.

Trong quá trình thực hiện đồ án, nhóm đã sử dụng 1440 tệp âm thanh giọng nói từ bộ RAVDESS. Tuy nhiên, sau những kết quả quan sát được và phân tích trên bộ dữ liệu này, nhóm kết luận rằng bộ dữ liệu gồm 1440 mẫu là quá nhỏ và thiếu thông tin, dẫn đến việc tăng cường dữ liệu cũng không thể cải thiện hiệu suất của các mô hình trên tập dữ liệu này. Do đó, sau khi tham khảo một số bài báo và nghiên cứu khác, nhóm quyết định lựa chọn phương án kết hợp giữa các bộ dữ liệu có điểm tương đồng.

Nhóm đã sử dụng toàn bộ dữ liệu âm thanh của RAVDESS cho đồ án, bao gồm 1440 tệp âm thanh giọng nói và 1012 tệp âm thanh giọng hát. Việc sử dụng thêm dữ liệu giọng hát có thể gây một số khó khăn cho các mô hình trong quá trình học tập, nhưng sẽ không làm ảnh hưởng đến mục tiêu của đồ án này, do nhóm không tập trung vào việc tối ưu hóa giải pháp cho bài toán phân loại cảm xúc tiếng nói mà chỉ sử dụng nó như một nền

tảng để khảo sát các mô hình máy học và học sâu thường thấy trong ngữ cảnh xử lý âm thanh và tiếng nói.

Về các nhãn xuất hiện trong bộ dữ liệu, dữ liệu âm thanh giọng nói của RAVDESS gồm 8 loại nhãn, còn dữ liệu âm thanh giọng hát gồm 6 loại nhãn như trong Bảng 2.1.

	Neutral	**Calm**	**Happy**	**Sad**	**Angry**	**Fearful**	**Disgust**	**Surprised**
RAVDESS (Speech)	✓	✓	✓	✓	✓	✓	✓	✓
RAVDESS (Song)	✓	✓	✓	✓	✓	✓		
TESS	✓		✓	✓	✓	✓	✓	✓

Bảng 2.1: Danh sách nhãn trong các bộ dữ liệu được sử dụng

2.1.2 Bộ dữ liệu TESS

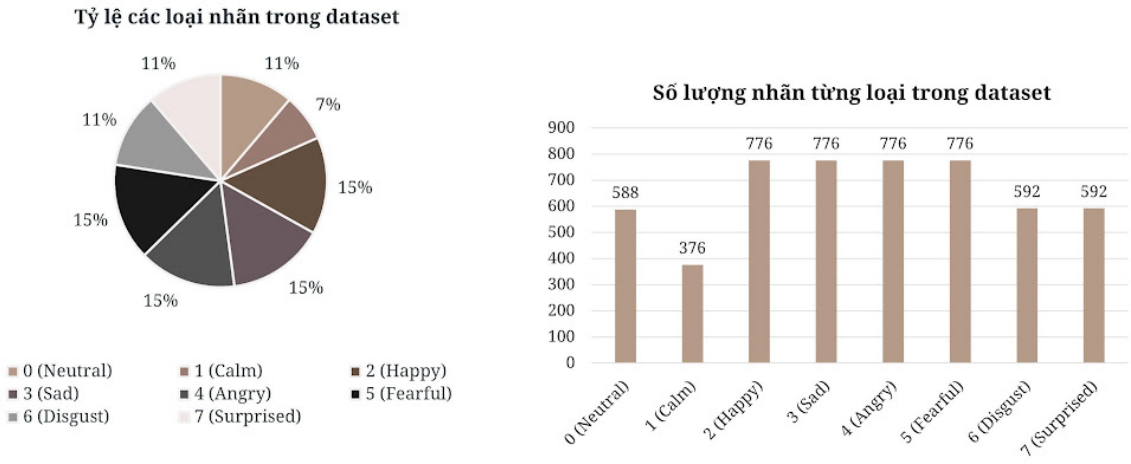
Tương tự như RAVDESS, Toronto Emotional Speech Set (TESS) cũng là một bộ dữ liệu cảm xúc tiếng nói thuộc loại mô phỏng và ngôn ngữ duy nhất là Tiếng Anh, với sự tham gia chỉ của 2 diễn viên nữ trong 2 độ tuổi khác nhau. Bộ dữ liệu gồm 2800 mẫu âm thanh tiếng nói, và 7 nhãn cảm xúc như Bảng 2.1. Điểm hạn chế của bộ dữ liệu này không chỉ ở loại mô phỏng mà còn ở sự tham gia thiếu sự đa dạng của các giọng nói khác nhau. Và cũng như bộ dữ liệu RAVDESS, mỗi mẫu âm thanh có thời lượng khá ngắn, thường chỉ gồm một câu nói đơn giản nên khó có thể áp dụng trong ngữ cảnh thực tế.

Nhận xét chung: Cả hai bộ dữ liệu đều có những hạn chế nhất định, không phải là một bộ dữ liệu quá tốt cho các mô hình học để giải quyết bài toán phân loại cảm xúc giọng nói. Việc kết hợp hai bộ dữ liệu khác nhau cũng được nhóm cân nhắc có thể xảy ra rủi ro, song đó là điều nhóm chấp nhận để thử nghiệm trong đồ án này.

2.2 Phân phối dữ liệu

Trong mỗi bộ dữ liệu trên, từng nhãn được phân phối khá đồng đều. Như đã đề cập, khi kết hợp các bộ dữ liệu, nhóm chấp nhận nguy cơ thiếu hụt dữ liệu tại một số lớp nhất định, để đánh đổi sự đầy đủ và đa dạng của bộ dữ liệu thử nghiệm. Trong đồ án này, 8 nhãn (lớp) cảm xúc sẽ được sử dụng để phân loại đó là: Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised.

Phân phối của các nhãn trong bộ dữ liệu như Hình 2.1:



(a) Tỷ lệ các loại nhãn trong bộ dữ liệu *(b) Số lượng nhãn từng loại trong bộ dữ liệu*

Hình 2.1: Phân phối các nhãn trong bộ dữ liệu

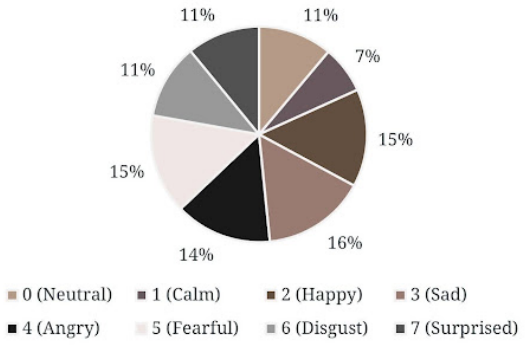
2.3 Phân chia bộ dữ liệu

Với tổng cộng 5252 mẫu dữ liệu tổng hợp được, nhóm thực hiện phân chia thành các tập train, dev và test phục vụ cho quá trình huấn luyện và đánh giá mô hình. Theo những tham khảo từ các nghiên cứu trước đó và quá trình tự nhìn nhận đánh giá, nhóm quyết định lựa chọn tỷ lệ 70:15:15 lần lượt cho 3 tập dữ liệu kể trên. Nhóm cảm thấy đây là một tỷ lệ ổn định, đảm bảo cho quá trình huấn luyện, với lượng dữ liệu huấn luyện không quá lấn át lượng dữ liệu kiểm thử và kiểm tra. Lượng dữ liệu trong tập dev và test được nhóm đánh giá là tương đối đầy đủ để xem xét hiệu suất của các mô hình trên đó.

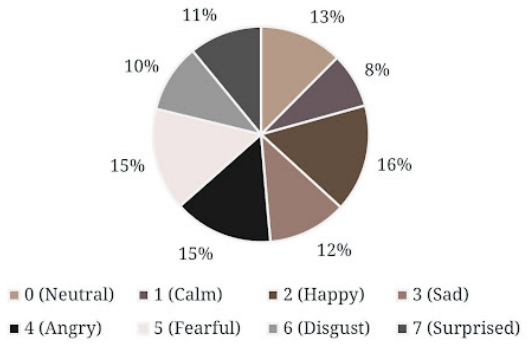
Bên cạnh đó, để đảm bảo tính công bằng trong bộ dữ liệu, nhóm thực hiện phân chia các nhãn cho từng bộ một cách tương đối đồng đều, tránh trường hợp mô hình thiếu hụt dữ liệu trên một số nhãn trong quá trình học. Sau cùng, kết quả phân phối các nhãn trên 3 tập dữ liệu như Bảng 2.2 và Hình 2.2:

	Neutral	**Calm**	**Happy**	**Sad**	**Angry**	**Fearful**	**Disgust**	**Surprised**	**Total samples**
Train (70%)	408	262	539	570	533	548	414	402	3676
Dev (15%)	98	65	127	93	117	121	81	86	788
Test (15%)	82	49	110	113	126	107	97	104	788
Total labels	588	376	776	776	776	776	592	592	5252

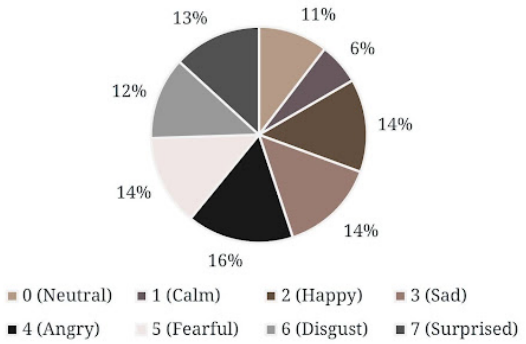
Bảng 2.2: Số lượng nhãn trên từng tập dữ liệu



(a) Tỷ lệ nhãn trên tập train



(b) Tỷ lệ nhãn trên tập dev



(c) Tỷ lệ nhãn trên tập test

Hình 2.2: Tỷ lệ nhãn trên từng tập dữ liệu

2.4 Tiền xử lý

Nhóm lựa chọn phương pháp chuẩn hóa dữ liệu là Standardization và mã hóa các nhãn dữ liệu bằng phương pháp One-hot Encoding (Bảng 2.3).

	Neutral	**Calm**	**Happy**	**Sad**	**Angry**	**Fearful**	**Disgust**	**Surprised**
Neutral	**1**	0	0	0	0	0	0	0
Calm	0	**1**	0	0	0	0	0	0
Happy	0	0	**1**	0	0	0	0	0
Sad	0	0	0	**1**	0	0	0	0
Angry	0	0	0	0	**1**	0	0	0
Fearful	0	0	0	0	0	**1**	0	0
Disgust	0	0	0	0	0	0	**1**	0
Surprised	0	0	0	0	0	0	0	**1**

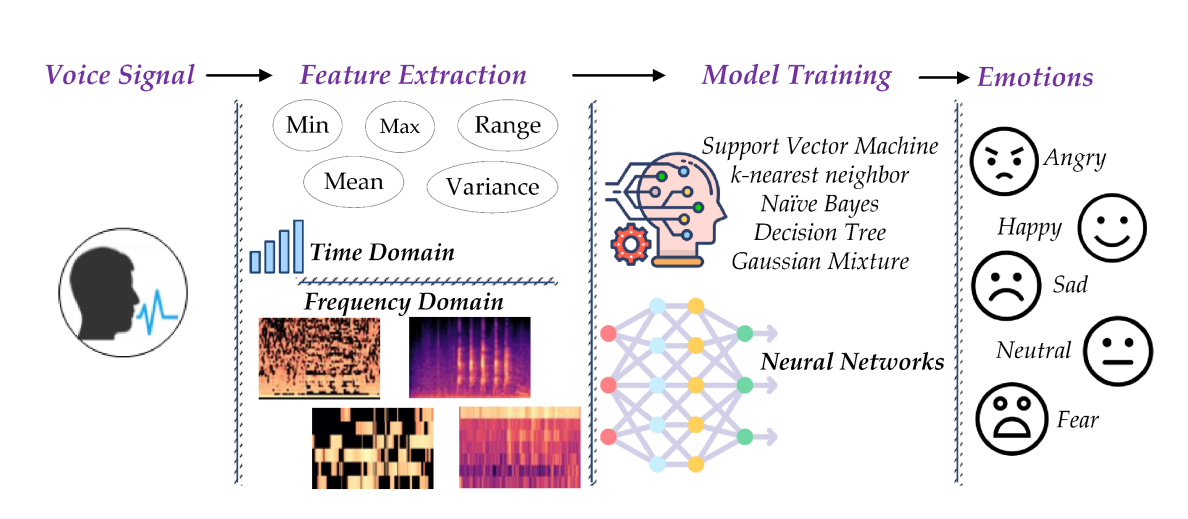
Bảng 2.3: Mã hóa các nhãn bằng phương pháp One-hot Encoding

Chương 3

PHƯƠNG PHÁP TIẾP CẬN

3.1 Tổng quan hướng tiếp cận

Đối với mỗi trường hợp thử nghiệm (một mô hình và một nhóm đặc trưng), bài toán phân loại cảm xúc sẽ được giải quyết qua một sơ đồ tổng quát như sau:



Hình 3.1: Sơ đồ tổng quan các bước giải quyết bài toán phân loại cảm xúc tiếng nói

Từ một dữ liệu âm thanh, các đặc trưng trong nhóm thử nghiệm sẽ được trích xuất theo một số điều kiện, đi cùng với quá trình tiền xử lý và được đưa vào một mô hình để thực hiện huấn luyện. Kết thúc quá trình, mô hình học được các nhãn cảm xúc và phân loại chúng dựa trên tệp âm thanh. Đồ án của nhóm tiếp cận bài toán phân loại cảm xúc tiếng nói một cách thử nghiệm với hai nhóm mô hình Máy học Truyền thống (Traditional Machine Learning) và nhóm mô hình Học sâu (Deep Learning). Cụ thể, các mô hình mà nhóm thử nghiệm ở bảng 3.1:

Bảng 3.1: Các mô hình thử nghiệm

Nhóm Máy học truyền thống	**Nhóm Học sâu**
SVM (Support Vector Machine)	MLP (Multi-Layer Perceptron)
Naive Bayes	CNN (Convolutional Neural Network)
Decision Tree	LSTM (Long Short-Term Memory)
Random Forest	CNN và LSTM

3.2 Đặc trưng

Không chỉ thử nghiệm trên các mô hình khác nhau, mục tiêu của nhóm còn là tìm hiểu các đặc trưng thường được sử dụng trong xử lý âm thanh. Đó là hai nhóm đặc trưng miền phổ (spectral features) và ngôn điệu (prosodic features) với các loại đặc trưng cụ thể thuộc mỗi nhóm như sau.

3.2.1 Nhóm đặc trưng miền phổ (spectral features)

- **Mel-frequency cepstral coefficients (MFCCs):** Các hệ số (coefficients) cepstral được trích chọn từ thang đo tần số Mel trên miền phổ được chuyển đổi từ dữ liệu âm thanh giọng nói.
- **Mel-scaled spectrogram:** Thể hiện sự phân phối cường độ của âm thanh theo thời gian và tần số. Có thể thấy, Mel-scaled spectrogram và MFCCs tốt trong việc nhận diện và theo dõi các biến độ giọng nói, nhưng không biểu diễn rõ ràng về cao độ của âm thanh.

Có thể thấy, Mel-scaled spectrogram và MFCCs tốt trong việc nhận diện và theo dõi các biến độ giọng nói, nhưng không biểu diễn rõ ràng về cao độ của âm thanh.

- **Chromagrams:** Biểu diễn của sự phân bố các âm thanh theo các pitch class trong âm nhạc.
- **Spectral Contrast:** Đo lường cường độ của âm thanh ở các băng tần số cận kề, cung cấp thông tin về sự độ chênh lệch giữa các băng tần số.
- **Tonnetz Coefficients:** Tập hợp các giá trị số đại diện cho các quan hệ tần số âm thanh trong âm nhạc.

Như vậy, kết hợp cả 5 sẽ trích xuất được nhiều đặc trưng spectral đa dạng hơn cho mỗi sample.

3.2.2 Nhóm đặc trưng ngôn điệu (prosodic features)

- **Pitch:** Cao độ cơ bản của giọng nói, thường liên quan đến giọng điệu hoặc cảm xúc trong tiếng nói.
- **Duration:** Mức độ kéo dài của một từ hoặc âm tiết, trong tiếng nói tự nhiên,

duration có thể truyền đạt sự nhấn mạnh hoặc tâm trạng của người nói. Ví dụ như độ kéo dài lớn thường liên quan đến sự tức giận, trong khi độ dài ngắn có thể diễn đạt sự hứng khởi hoặc lo sợ.

- **Energy:** Mức năng lượng (hay "độ mạnh") của tín hiệu giọng nói, năng lượng cao thường diễn đạt sự tức giận hay hứng khởi, trong khi năng lượng thấp có thể diễn đạt sự buồn bã.

Những đặc trưng này cung cấp thông tin quan trọng về ngữ điệu về cách tiếng nói được diễn đạt.

Lý do mà nhóm lựa chọn các đặc trưng trên để thử nghiệm, một phần là do trong quá trình nghiên cứu và tham khảo các bài báo khác nhau, nhóm nhận thấy các đặc trưng này thường được đề cập như những đặc trưng nổi bật trong ngữ cảnh xử lý âm thanh và tiếng nói. Tuy nhiên, gần như chưa có một nghiên cứu đầy đủ nào thử nghiệm trên toàn bộ các đặc trưng đó trong bài toán phân loại cảm xúc giọng nói, mà thường chỉ đề cập như một nội dung có thể thử nghiệm trong tương lai. Đó cũng là động lực thúc đẩy nhóm thực hiện đồ án này.

Từ hai nhóm đặc trưng spectral và prosodic, nhóm tiến hành thử nghiệm trên 4 set đặc trưng lần lượt như Bảng 3.2:

Set	Spectral Features					Prosodic Features			Total Features
	MFCC	Mel-spectrogram	Chromagrams	Spectral Contrast	Tonnetz Coefficients	Pitch	Duration	Energy	
1	40	0	0	0	0	0	0	0	**40**
2	40	128	12	7	6	0	0	0	**193**
3	0	0	0	0	0	65	1	1	**67**
4	40	128	12	7	6	65	1	1	**260**

Bảng 3.2: Các đặc trưng thử nghiệm

Trong đó, set 1 gồm 40 đặc trưng MFCCs, set 2 gồm 193 đặc trưng từ 5 loại thuộc nhóm spectral, set 3 gồm 67 đặc trưng từ 3 loại của nhóm prosodic và set 4 là sự kết hợp của set 2 và set 3.

3.3 Mô hình

3.3.1 Học máy

Support Vector Machine

Support Vector Machine(SVM) là một mô hình máy học có khả năng phân loại các dữ liệu thành các lớp khác nhau bằng cách tìm ra ranh giới quyết định tốt nhất giữa các điểm dữ liệu. Nó sử dụng một hàm kernel để chuyển đổi dữ liệu từ không gian ban đầu thành một không gian có số chiều lớn hơn, giúp tạo ra các ranh giới phân chia phức tạp hơn.

RBF (Radial Basis Function) là một trong những hàm kernel phổ biến nhất được sử dụng trong SVM, có khả năng phân loại phi tuyến tính, phù hợp với dữ liệu của bài toán phân loại cảm xúc giọng nói.

Gaussian Naive Bayes

Gaussian Naive Bayes là một biến thể của mô hình Naive Bayes, là một mô hình thống kê dựa trên định lý Bayes, được sử dụng khi các biến đặc trưng là liên tục và được giả định tuân theo phân phối Gaussian (hoặc phân phối chuẩn). Trong bài toán nhận diện cảm xúc, nó có thể được áp dụng để tính xác suất của một mẫu giọng nói thuộc về một lớp cảm xúc cụ thể dựa trên các đặc trưng của nó.

Decision Tree

Decision Tree (DT) là một mô hình quyết định tạo ra một cây quyết định từ dữ liệu đào tạo. Trong ngữ cảnh nhận diện cảm xúc trong giọng nói, cây quyết định có thể học cách phân chia dữ liệu để đưa ra quyết định về trạng thái cảm xúc của mẫu giọng nói.

Random Forest

Random Forest (RF) là một mô hình quyết định tạo ra một cây quyết định từ dữ liệu đào tạo. Trong ngữ cảnh nhận diện cảm xúc trong giọng nói, cây quyết định có thể học cách phân chia dữ liệu để đưa ra quyết định về trạng thái cảm xúc của mẫu giọng nói.

3.3.2 Học sâu

Multi-Layer Perceptrone

Multi-Layer Perceptron (MLP) Mô hình học sâu với nhiều lớp ẩn (hidden layers) được sử dụng trong các bài toán phân loại. MLP học các quan hệ tuyến tính giữa các đặc trưng. Trong ngữ cảnh bài toán này, MLP có thể học các mối quan hệ phức tạp giữa các đặc trưng của giọng nói và trạng thái cảm xúc tương ứng.

Convolutional Neural Network

Convolutional Neural Network (CNN) là một mô hình mạng nơ-ron thích hợp cho xử lý dữ liệu không gian như hình ảnh. Trong bài toán nhận diện cảm xúc từ giọng nói, CNN có thể được áp dụng để học các đặc trưng không gian từ biểu đồ tần số của âm thanh.

Long Short-Term Memory

Long Short-Term Memory (LSTM) à một kiểu mô hình mạng nơ-ron hồi quy được thiết kế đặc biệt để xử lý dữ liệu chuỗi, giúp nắm bắt thông tin trạng thái cảm xúc từ giọng nói qua thời gian.

Kết hợp CNN và LSTM

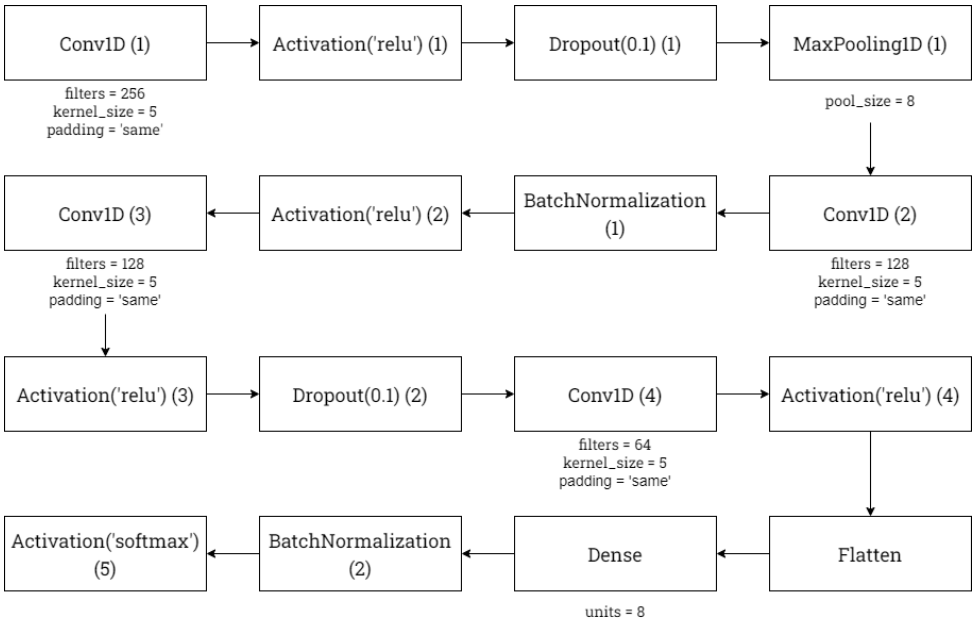
Việc kết hợp này trong một mô hình được gọi là mạng nơ-ron học sâu 2D-3D, giúp mô hình xử lý cả thông tin không gian và thời gian, nắm bắt cả các đặc trưng cấp cao và mối quan hệ chuỗi thời gian trong dữ liệu giọng nói. Điều này có thể dẫn đến hiệu suất tốt trong bài toán nhận diện cảm xúc từ giọng nói.

Chương 4

THỰC NGHIỆM

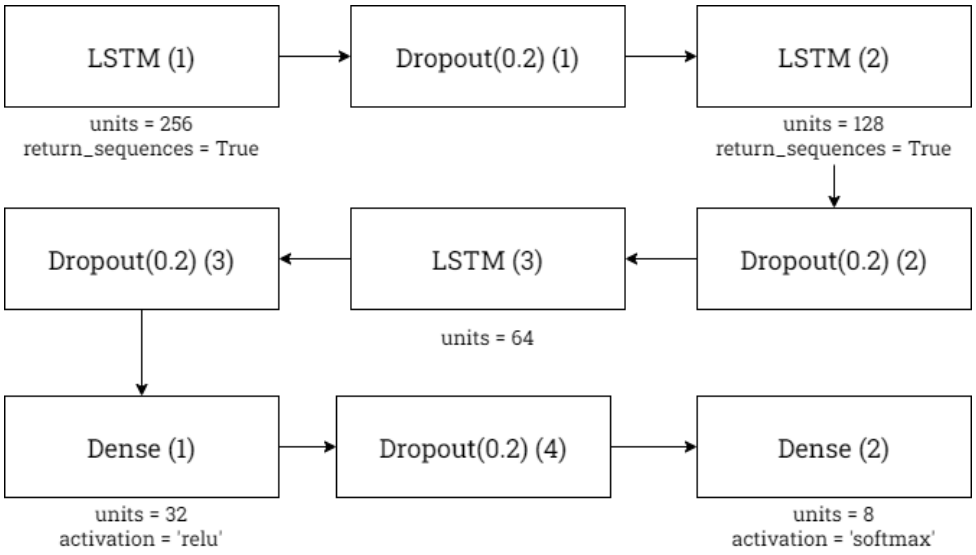
4.1 Cài đặt thực nghiệm

- **Mô hình Naive Nayes (NB):**
  Sử dụng mô hình mặc định của thư viện `sklearn.naive_bayes`
- **Mô hình Decision Tree (DT):**
  `max_depth = 10`, `min_samples_split = 2`
- **Mô hình Random Forest (RF):**
  `criterion = "gini"`, `max_features = "log2"`, `max_leaf_nodes = 100`, `min_samples_leaf = 3`, `min_samples_split = 20`, `n_estimators = 22000`, `max_depth = 10`
- **Mô hình Multi-Layer Perceptron (MLP):**
  `hidden_layer_sizes = 250`, `max_iter = 300`
- **Convolutional Neural Network (CNN):**
  `loss = "categorical_crossentropy"`, `optimizer = "RMSprop"`, `learning_rate = 0.0001`, `weight_decay = 1e-6`, `epochs = 300`, `batch_size = 64`.
  Kiến trúc mô hình xem Hình 4.1.



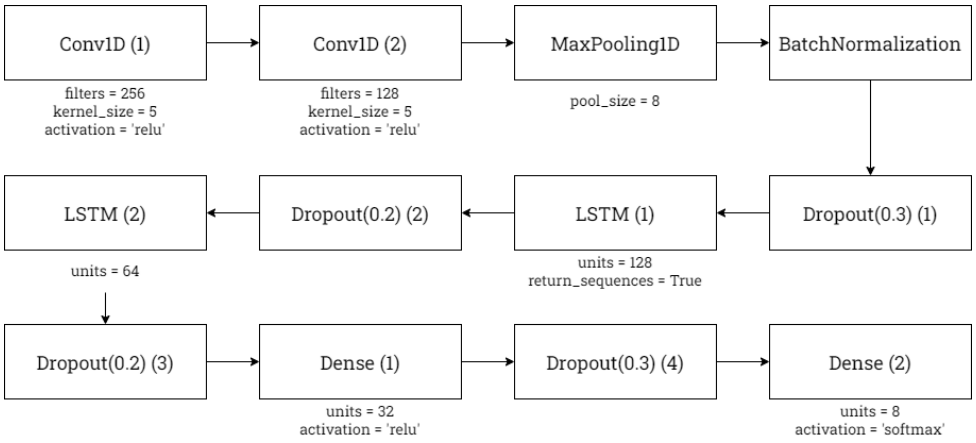
Hình 4.1: Kiến trúc mô hình CNN

- **Mô hình Long Short-Term Memory (LSTM):**
  `loss = "categorical_crossentropy"`, `optimizer = "Adam"`, `learning_rate = 0.0005`, `epochs = 200`, `batch_size = 64`.
  Kiến trúc mô hình xem Hình 4.2.



Hình 4.2: Kiến trúc mô hình LSTM

- **Mô hình CNN kết hợp với LSTM (CNN_LSTM):**
  `loss = "categorical_crossentropy"`, `optimizer = "Adam"`, `learning_rate = 0.001`, `epochs = 200`, `batch_size = 64`.
  Kiến trúc mô hình xem Hình 4.3h



Hình 4.3: Kiến trúc mô hình CNN kết hợp với LSTM

4.2 Kết quả thực nghiệm

Nhóm thực hiện đánh giá trên 3 thang đo: Accurary, F1-Macro score, F1-Weighted score. Đồng thời tính tốc độ thực thi trung bình của các mô hình.

Đặc trưng	**MFCC**	**Spectral**	**Prosodic**	**Tổng hợp đặc trưng**
SVM	90.74	84.39	74.49	87.56
Naive Bayes	60.41	48.60	42.01	49.37
MLP	89.72	88.45	73.10	89.09
Decision Tree	64.59	68.65	58.76	70.05
Random Forest	82.23	80.33	68.78	80.20
LSTM	81.35	70.05	65.48	62.69
CNN_LSTM	87.31	83.76	56.22	83.25
CNN	90.99	90.23	61.55	90.86

Bảng 4.1: Kết quả đánh giá Accuracy cho các mô hình và đặc trưng

Đặc trưng	MFCC	Spectral	Prosodic	Tổng hợp đặc trưng
SVM	90.77	84.45	74.66	87.57
Naive Bayes	61.85	49.86	43.19	50.66
MLP	89.72	88.45	73.04	89.09
Decision Tree	64.79	69.00	58.95	70.47
Random Forest	82.39	80.67	69.11	80.50
LSTM	81.46	70.64	65.60	63.04
CNN_LSTM	87.35	83.77	56.43	83.44
CNN	91.01	90.26	61.58	90.85

Bảng 4.2: Kết quả đánh giá F1 - weighted cho các mô hình và đặc trưng

Đặc trưng	MFCC	Spectral	Prosodic	Tổng hợp đặc trưng
SVM	90.49	84.80	73.92	87.33
Naive Bayes	60.90	49.66	42.27	50.39
MLP	89.38	88.46	72.62	88.87
Decision Tree	64.81	68.76	58.98	70.15
Random Forest	82.44	80.65	68.30	80.50
LSTM	81.40	69.92	65.18	62.39
CNN_LSTM	87.41	83.71	55.75	82.63
CNN	90.87	90.39	60.52	90.66

Bảng 4.3: Kết quả đánh giá F1 - Macro cho các mô hình và đặc trưng

Đặc trưng	MFCC	Spectral	Prosodic	Tổng hợp đặc trưng
SVM	3.48	5.30	2.58	7.11
Naive Bayes	0.04	0.15	0.05	0.11
MLP	0.24	0.19	0.30	0.18
Decision Tree	0.01	0.01	0.01	0.02
Random Forest	78.71	75.21	103.30	93.64
LSTM	43.49	277.01	76.19	559.83
LSTM + CNN	18.39	42.87	46.44	107.75
CNN	8.61	11.61	7.16	10.34

Bảng 4.4: Thời gian thực thi trung bình (E-4s) cho các mô hình và đặc trưng

4.3 Nhận xét

Dựa vào bảng kết quả trên, nhóm có đưa ra một số nhận xét sau:

- **Đặc trưng:** Trên các thang đo, đặc trưng MFCC cho kết quả tốt nhất trong hầu hết các mô hình (trừ mô hình Decision Tree), tiếp theo là đặc trưng nhóm spectral. Đặc trưng nhóm prosodic cho kết quả thấp nhất. Điều này cho thấy các đặc trưng MFCC và spectral có thể nắm bắt được thông tin quan trọng hơn để phân loại âm thanh. Ngoài ra, cũng có thể thấy, việc sử dụng quá nhiều đặc trưng đối với các mô hình không quá phức tạp trong đồ án này cũng khiến cho hiệu suất của mô hình giảm đi.
- **Phương pháp:** Nhìn chung với đặc trưng MFCC và spectral, các phương pháp SVM, CNN và MLP cho kết quả tốt nhất, tiếp theo là các phương pháp Random Forest, LSTM, LSTM_CNN, các phương pháp Naive Bayes, Decision Tree cho kết quả thấp hơn.
  - **Naive Bayes và Decision Tree hoạt động không tốt:** NB và DT là các mô hình học máy truyền thống, dựa trên các khái niệm thống kê và logic. Các mô hình này hoạt động tốt trong các bài toán phân loại có số lượng đặc trưng nhỏ và các đặc trưng này có mối quan hệ tuyến tính với nhau. Tuy nhiên, trong bài toán phân âm thanh tiếng nói này, số lượng đặc trưng thường lớn và các đặc trưng này có mối quan hệ phi tuyến với nhau. Do đó, các mô hình NB và DT không thể khai thác được thông tin đầy đủ từ các đặc trưng âm thanh, dẫn đến kết quả phân loại không tốt.
  - **Các mô hình MLP, SVM, CNN hoạt động tốt:** MLP, SVM, và CNN là các mô hình học máy hiện đại có khả năng hiệu quả hóa quá trình khai thác thông tin từ dữ liệu. Chúng không chỉ có khả năng xử lý tốt với dữ liệu phức tạp, đa dạng mà còn linh hoạt đối với nhiều đặc trưng khác nhau. MLP có khả năng học được mối quan hệ phức tạp giữa đặc trưng âm thanh và cảm xúc, nhờ vào cấu trúc nhiều tầng ẩn trong mô hình. SVM có thể xử lý không gian

đặc trưng lớn và tìm ra ranh giới phân loại tối ưu cho các loại cảm xúc khác nhau. CNN được thiết kế để hiệu quả trong việc xử lý các đặc trưng không gian, giúp mô hình nắm bắt được cấu trúc không gian của dữ liệu âm thanh.

  - **Các phương pháp Random Forest, LSTM, LSTM_CNN đều cho kết quả tốt hơn các mô hình NB và DT, nhưng không bằng các mô hình MLP, SVM, CNN:** Random Forest, một phương pháp học máy tổ hợp, tự tin hướng đến sự đa dạng và ổn định trong dữ liệu, giúp giảm nguy cơ overfitting. LSTM, một mô hình mạng nơ-ron tái sinh, được xem xét là một lựa chọn mạnh mẽ cho xử lý dữ liệu chuỗi, trong khi LSTM_CNN, kết hợp giữa LSTM và CNN, khai thác hiệu quả cả thông tin cục bộ và thời gian từ các đặc trưng âm thanh. Tuy nhiên, khi kết hợp hai mô hình không phải lúc nào cũng cho hiệu suất tốt hơn một mô hình đơn lẻ, ví dụ như trong trường hợp này, sự phức tạp của các đặc trưng khiến cho hiệu suất của mô hình kết hợp so với mô hình đơn CNN là không cải thiện.

- **Sự kết hợp các đặc trưng và phương pháp:** Sự kết hợp các đặc trưng MFCC và spectral với các phương pháp SVM, MLP, CNN cho kết quả tốt nhất. Trong đó hai đặc trưng MFCC và spectral hoạt động khá ổn định trên các mô hình. Tuy nhiên đặc trưng prosodic thì không như vậy, với mô hình CNN hai đặc trưng MFCC và spectral cho kết quả tốt nhất, tuy nhiên kết quả với đặc trưng prosodic thì không được như vậy. Điều này cho thấy model CNN hoạt động không tốt với đăc trưng prosodic. Sự kết hợp các đặc trưng và phương pháp có thể cải thiện kết quả phân loại âm thanh.

- **Thời gian thực thi trung bình:** Đặc trưng MFCC có tốc độ thực thi trung bình nhanh hơn các đặc trưng còn lại. Hầu hết các phương pháp học sâu như CNN, LSTM, LSTM và phương pháp Random Forest có thời gian thực thi trung bình lớn hơn so với các phương pháp khác như Naive Bayes, MLP, SVM, Decision Tree.

Tuy nhiên, cần lưu ý rằng các kết quả trên chỉ mang tính tham khảo, vì chúng phụ thuộc vào nhiều yếu tố như bộ dữ liệu, kích thước tập dữ liệu, cách chuẩn hóa dữ liệu, cách lựa chọn tham số của các mô hình,... Nhóm hiện tại chỉ đang khảo sát khái quát hiệu quả của các loại đặc trưng và mô hình, các các mô hình mà nhóm huấn luyện chưa phải là mô hình tốt nhất của từng phương pháp. Để có kết quả chính xác hơn và đưa ra những kết luận cuối cùng, cần thực hiện thêm các thí nghiệm trên các tập dữ liệu khác nhau và các mô hình tố nhất của từng phương pháp trên bộ dữ liệu đó.

Chương 5

TỔNG KẾT

5.1 Kết luận

Trong đồ án này nhóm đã hiểu được các bước giải quyết một bài toán với dữ liệu là âm thanh. Đồng thời cũng nắm được kỹ thuật xử lý dữ liệu âm thanh cơ bản và cách trích xuất đặc trưng từ dữ liệu âm thanh phù hợp với bài toán phân loại cảm xúc giọng nói.

Nhóm thực hiện rút trích các nhóm đặc trưng và huấn luyện các mô hình theo nhiều phương pháp và thấy rằng sự kết hợp giữa đặc trưng MFCC và các mô hình SVM, MLP, CNN cho kết quả tốt nhất và sấp sỉ nhau, trong đó MLP có thời gian thực thi nhanh nhất.

Mục tiêu chính của nhóm trong đồ án này là muốn đánh giá khái quát hiệu quả khác nhau của các đặc trưng và phương pháp trong bài toán phân loại âm thanh nói chung và phân loại cảm xúc trong âm thanh nói riêng. Nên các mô hình mà nhóm huấn luyện được trên bộ dữ liệu chưa phải là tốt nhất của từng phương pháp. Vì vậy, các kết quả trên chỉ mang tính tham khảo, kết luận cuối cùng còn phụ thuộc vào nhiều yếu tố như tham số của mô hình, kiến trúc của mô hình, bộ dữ liệu, kích thước tập dữ liệu, cách chuẩn hóa dữ liệu,... Để có kết quả chính xác hơn và đưa ra những kết luận cuối cùng, cần thực hiện thêm các thí nghiệm trên các tập dữ liệu khác nhau và các mô hình tố nhất của từng phương pháp trên bộ dữ liệu đó.

5.2 Điểm hạn chế và hướng mở rộng

5.2.1 Điểm hạn chế

- Chỉ mới thực nghiệm trên một bộ dữ liệu
- Chưa thực nghiệm phương pháp thuộc nhóm học sâu
- Các kết quả trên các phương pháp chưa phải là tốt nhất nên chỉ có thế đưa ra kết luận một cách khái quát
- Bộ dữ liệu kết hợp từ hai dataset khác nhau, khiến chất lượng dữ liệu không đồng đều và thiếu hụt một số dữ liệu.

- Việc tăng cường dữ liệu tương đối khó khăn trên bài toán phân loại cảm xúc tiếng nói, do cảm xúc tương quan với các thành phần như cao độ,..
- Hầu hết các mô hình đều hoạt động tốt hơn với các đặc trưng spectral. Kết hợp nhiều đặc trưng cũng làm tăng nguy cơ nhiễu cho dữ liệu.
- Chưa tìm hiểu được các phương pháp tăng cường dữ liệu hiệu quả trên bộ dữ liệu này.

5.2.2 Hướng mở rộng

- Thực nghiệm trên nhiều bộ dữ liệu
- Thực nghiệm thêm nhiều phương pháp học sâu và tìm ra các siêu tham số, kiến trúc tốt nhất tương ứng với các phương để có được kết luận chặt chẽ hơn
- Tìm hiểu các phương pháp tăng cường dữ liệu phức tạp hơn.

Tài liệu tham khảo

[1] Abbaschian, Babak Joze, Daniel Sierra-Sosa, and Adel Elmaghraby. 2021. "Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models"Sensors 21, no. 4: 1249. https://doi.org/10.3390/s21041249

[2] M. G. de Pinto, M. Polignano, P. Lops and G. Semeraro, "Emotions Understanding Model from Spoken Language using Deep Neural Networks and Mel-Frequency Cepstral Coefficients,"2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), Bari, Italy, 2020, pp. 1-5, https://vjol.info.vn/index.php/tnu/article/view/63913

[3] Akçay, Mehmet Berkehan and Kaya Oğuz. "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers." Speech Commun. 116 (2020): 56-76. https://doi.org/10.1016/j.specom.2019.12.001

[4] Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, 101894. https://doi.org/10.1016/j.bspc.2020.101894

[5] Tarunika, K., Pradeeba, R. B., & Aruna, P. (2018, July). Applying machine learning techniques for speech emotion recognition. In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE. https://doi.org/10.1109/ICCCNT.2018.8494104

[6] Mohamed, A., Lee, H. Y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., ... & Watanabe, S. (2022). Self-supervised speech representation learning: A review. IEEE Journal of Selected Topics in Signal Processing. https://doi.org/10.1109/JSTSP.2022.3207050