SER: Speech Emotion Recognition using Convolutional Neutral Network

Kieu Qui Hung[1,2], **Tran Tri Duc**[1,3]

[1] Faculty of Information Science and Engineering,
University of Information and Techonology
[2] <22520505@gm.uit.edu.vn> [3] <22520276@gm.uit.edu.vn>

Abstract

Nhận diện cảm xúc từ giọng nói là một nhiệm vụ phức tạp do sự mơ hồ trong định nghĩa cảm xúc. Trong nghiên cứu này, chúng tôi áp dụng phương pháp biến đổi âm thanh thành mel-spectogram và huấn luyện dựa trên các mô hình Convolutional Neutral Network (CNN) để giải quyết nhiệm vụ nhận diện cảm xúc từ giọng nói. Chúng tôi định hình vấn đề như một bài toán phân loại đa nhãn và so sánh hiệu suất của ba loại mô hình. Chúng tôi trích xuất bảy đặc trưng thủ công từ tín hiệu âm thanh từ ba bộ data khác nhau. Trong phương pháp huấn luyện mô hình đầu tiên, các đặc trưng này được sử dụng để huấn luyện với mô hình học sâu CNN tự tạo ra, trong khi phương pháp thứ hai dựa trên mô hình pre-trained wav2vec dành cho dữ liệu âm thanh và phương pháp còn lại là sử dụng bộ phân loại LSTM được huấn luyện trên cùng các đặc trưng. Để giải quyết sự mơ hồ trong giao tiếp, chúng tôi cũng bao gồm các đặc trưng từ văn bản. Chúng tôi báo cáo độ chính xác, f1-score, độ chính xác cho các thiết lập thí nghiệm khác nhau mà chúng tôi đã đánh giá. Tổng thể, chúng tôi cho thấy rằng mô hình học sâu tự tạo nhẹ được huấn luyện trên một vài đặc trưng thủ công có thể đạt được hiệu suất cao hơn so với phương pháp khác trong nhận diện cảm xúc.

Keywords: speech emotion recognition, deep learning, Convolutional Neutral Network.

1 Introduction

Nhận diện cảm xúc từ giọng nói (Speech Emotion Recognition - SER) là một lĩnh vực nghiên cứu đang thu hút sự quan tâm lớn trong các lĩnh vực xử lý ngôn ngữ tự nhiên, học máy và học sâu. Với sự phát triển không ngừng của các hệ thống trí tuệ nhân tạo, SER ngày càng trở nên quan trọng nhờ vào các ứng dụng thực tiễn như cải thiện khả năng tương tác giữa con người và máy móc, hỗ trợ chăm sóc sức khỏe tâm lý, hay phân tích cảm xúc của khách hàng trong các ngành công nghiệp dịch vụ. Mặc dù có tiềm năng ứng dụng cao, nhiệm vụ này lại đối mặt với nhiều thách thức do sự phức tạp trong cách con người biểu đạt cảm xúc. Các yếu tố như ngữ điệu, tốc độ nói, ngữ cảnh, và giọng nói cá nhân đều tạo nên sự mơ hồ và đa dạng trong cách cảm xúc được thể hiện, đòi hỏi các hệ thống nhận diện phải cực kỳ chính xác và linh hoạt. Trong nghiên cứu này, chúng tôi tập trung giải quyết bài toán nhận diện bảy loại cảm xúc cơ bản: vui (happy), buồn (sad), giận dữ (angry), sợ hãi (fear), ngạc nhiên (surprise), ghê tởm (disgust) và trung tính (neutral). Để đạt được điều này, chúng tôi sử dụng và tích hợp ba bộ dữ liệu phổ biến nhất trong lĩnh vực SER: TESS (Toronto Emotional Speech Set), SAVEE (Surrey Audio-Visual Expressed Emotion) và CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset). Sự kết hợp này không chỉ làm tăng số lượng mẫu dữ liệu mà còn đảm bảo tính đa dạng về ngữ điệu, giọng nói, giới tính và ngữ cảnh của các mẫu cảm xúc. Đây là bước quan trọng nhằm tăng cường độ chính xác và khả năng tổng quát hóa của các mô hình nhận diện cảm xúc. Phương pháp của chúng tôi dựa trên việc chuyển đổi tín hiệu âm thanh thành mel-spectrogram, một biểu diễn tần số phổ biến trong phân tích âm thanh, sau đó áp dụng ba hướng tiếp cận chính. Phương pháp đầu tiên sử dụng các đặc trưng thủ công được trích xuất từ tín hiệu âm thanh để huấn luyện mô hình Convolutional Neural Network (CNN) tự xây dựng, cho phép chúng tôi tối ưu hóa mô hình phù hợp với bài toán và dữ liệu. Phương pháp thứ hai dựa trên mô hình pre-trained VGG16, vốn được thiết kế chuyên biệt cho dữ liệu hình ảnh nhưng lại được sử dụng trong bài này cũng vì đặc trưng âm thanh sẽ được chuyển đổi thành dạng biểu đồ riêng biệt, để tận dụng khả năng học sâu từ các mô hình đã được đào tạo trên các tập dữ liệu lớn. Cuối cùng, chúng tôi sử dụng Long Short-Term Memory (LSTM), một mô hình mạnh mẽ trong việc học các chuỗi thời gian, để khai thác đặc trưng âm thanh một cách hiệu quả. Kết quả của nghiên cứu cho thấy rằng mô hình CNN tự xây

dựng, mặc dù nhẹ và được huấn luyện trên một tập đặc trưng thủ công giới hạn, vẫn có khả năng đạt được hiệu suất vượt trội so với các phương pháp hiện đại khác. Nghiên cứu này không chỉ cung cấp một góc nhìn cải tiến về việc ứng dụng các mô hình học sâu nhẹ trong SER mà còn đóng góp quan trọng vào việc phát triển các hệ thống nhận diện cảm xúc hiệu quả, chính xác và phù hợp hơn với thực tế.

2 Fundamental

2.1 Related works

Nhận diện cảm xúc từ giọng nói đa phương thức đã được nghiên cứu rộng rãi nhằm tích hợp các đặc trưng âm thanh và văn bản. Các nghiên cứu trong lĩnh vực này tập trung vào việc so sánh hiệu quả của các đặc trưng thủ công khi áp dụng trên các mô hình học máy truyền thống và các mô hình học sâu. Với bộ dữ liệu IEMOCAP bao gồm 8 nhãn cảm xúc như giận dữ, hạnh phúc, buồn bã, trung tính, ngạc nhiên, sợ hãi, thất vọng và phấn khích, các mô hình như Random Forest, Gradient Boosting, SVM, Naive Bayes, Logistic Regression, MLP và LSTM đã được áp dụng và đánh giá. Các phương pháp chính bao gồm cân bằng dữ liệu thông qua tăng cường mẫu cho các lớp cảm xúc ít xuất hiện và hợp nhất các nhãn cảm xúc tương đồng để giảm sự chênh lệch về phân bố. Quá trình trích xuất đặc trưng tập trung vào các đặc trưng âm thanh như pitch, harmonics, năng lượng giọng nói, khoảng lặng và các đặc trưng văn bản như TF-IDF. Các mô hình học máy và học sâu được huấn luyện trên các đặc trưng này và được đánh giá bằng các chỉ số như độ chính xác, precision, recall và F-score. Kết quả nghiên cứu cho thấy các mô hình học máy nhẹ thường đạt hiệu suất tương đương hoặc tốt hơn so với các mô hình học sâu, đồng thời yêu cầu ít tài nguyên tính toán hơn. Việc kết hợp các đặc trưng từ nhiều miền (âm thanh và văn bản) giúp cải thiện đáng kể độ chính xác. Tuy nhiên, một số hạn chế vẫn tồn tại, bao gồm phương pháp kết hợp đặc trưng đơn giản (concatenation) chưa tối ưu và giới hạn về quy mô cũng như tính đa dạng của dữ liệu, có thể làm giảm hiệu quả của mô hình. Vậy nên trong đề tài này nhóm chúng tôi sẽ kết hợp giữa nhiều bộ dữ liệu có cùng chất với nhau để tạo nên một bộ dữ liệu mới đa dạng hơn cũng như là tăng cường khả năng nhận diện và độ chính xác của mô hình với 3 model khác nhau.

2.2 Methodology

Bài báo cáo này tập trung vào việc gộp và xử lý dữ liệu từ ba bộ dữ liệu âm thanh chính là TESS, SAVEE và CREMA, nhằm tạo ra một tập dữ liệu tổng hợp đa dạng và phong phú cho bài toán dự đoán cảm xúc của đoạn hội thoại. Đầu tiên, dữ liệu âm thanh từ các bộ này được tiền xử lý và chuyển đổi thành log Mel Spectrogram, nhấn mạnh các đặc trưng âm thanh quan trọng bằng cách sử dụng thang logarit. Các đặc trưng như Zero Crossing Rate (ZCR), Root Mean Square Energy (RMSE) và Mel-Frequency Cepstral Coefficients (MFCC) được trích xuất để đại diện cho kết cấu âm thanh, năng lượng và tính chất phổ của tín hiệu. Để tăng cường dữ liệu, các kỹ thuật như thêm nhiễu Gaussian, kéo dãn hoặc nén thời gian, dịch chuyển tín hiệu, và thay đổi cao độ được áp dụng. Quy trình này không chỉ tăng tính đa dạng của dữ liệu mà còn giúp mô hình học được các biến đổi thực tế trong âm thanh, từ đó cải thiện khả năng khái quát.

Sau khi trích xuất đặc trưng, các mô hình CNN (Convolutional Neural Network) được áp dụng để thực hiện phân loại. CNN được chọn vì khả năng xuất sắc trong việc học các đặc trưng không gian và trích xuất mẫu từ dữ liệu hình ảnh như spectrogram. Các mô hình này được huấn luyện trên tập dữ liệu tổng hợp, với đầu vào là ma trận đặc trưng từ tín hiệu gốc và tín hiệu đã tăng cường. Việc kết hợp các kỹ thuật tiền xử lý, tăng cường dữ liệu, và áp dụng CNN đảm bảo mô hình có khả năng phân loại cảm xúc một cách chính xác và hiệu quả, đồng thời giảm thiểu overfitting và tăng cường hiệu suất trên các tập dữ liệu chưa từng thấy.

3 Dataset

3.1 Toronto Emotional Speech Set (TESS)

TESS (Toronto Emotional Speech Set) là một bộ dữ liệu được thiết kế để nghiên cứu cách con người nhận diện cảm xúc từ giọng nói, với trọng tâm là giọng nói nữ. Bộ dữ liệu này được tạo ra bởi hai nữ diễn viên chuyên nghiệp thuộc hai độ tuổi khác nhau (26 và 64 tuổi). Tổng cộng có 2.800 tệp âm thanh, mỗi tệp đại diện cho một câu ngắn, ví dụ: "The dog is sitting by the door." Các câu này được diễn đạt với 7 cảm xúc khác nhau: Angry, Happy, Sad, Neutral, Fear, Disgust, và Surprise. Các tệp âm thanh được lưu dưới định dạng WAV với chất lượng cao, đảm bảo phù hợp để nghiên cứu cảm xúc từ giọng nói trong môi trường phòng thí nghiệm. TESS đặc biệt hữu ích trong việc phân tích cách cảm xúc ảnh hưởng đến giọng nói của nữ giới và

là một nguồn dữ liệu đáng tin cậy cho các bài toán SER tập trung vào ngữ cảnh ngôn ngữ tự nhiên.

3.2 SAVEE (Surrey Audio-Visual Expressed Emotion)

SAVEE (Surrey Audio-Visual Expressed Emotion) là một bộ dữ liệu được xây dựng nhằm nghiên cứu biểu đạt cảm xúc từ giọng nói và hình ảnh khuôn mặt, tuy nhiên trong bài toán SER, chỉ phần âm thanh được sử dụng. Bộ dữ liệu này được thu âm bởi 4 nam diễn viên người Anh, đảm bảo sự đồng nhất về ngôn ngữ và ngữ âm. SAVEE bao gồm 480 tệp âm thanh ngắn, mỗi tệp thể hiện một câu thoại được diễn đạt với 7 cảm xúc: Neutral, Happy, Sad, Angry, Fear, Disgust, và Surprise. Các tệp này được lưu dưới định dạng WAV và có chất lượng âm thanh tốt, dễ dàng cho việc phân tích tín hiệu âm thanh. SAVEE đặc biệt phù hợp cho các nghiên cứu liên quan đến giọng nói nam giới, cung cấp một tập dữ liệu đáng tin cậy cho các hệ thống SER cần phân tích và nhận diện cảm xúc từ giọng nói trong ngữ cảnh tiếng Anh.

3.3 CREMA-D (Crowd-sourced Emotional Multimodel Actors Dataset)

CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset) là một trong những bộ dữ liệu lớn và đa dạng nhất được sử dụng trong nghiên cứu SER. Được tạo ra bởi 91 diễn viên chuyên nghiệp, bao gồm 48 nam và 43 nữ, bộ dữ liệu này không chỉ đa dạng về giới tính mà còn phong phú về độ tuổi và dân tộc, mang đến tính đại diện cao. Bộ dữ liệu chứa hơn 7.400 tệp âm thanh WAV, mỗi tệp là một câu thoại được diễn đạt với một trong 7 cảm xúc: Angry, Happy, Neutral, Sad, Fear, và Disgust. Điểm đặc biệt của CREMA-D là các diễn viên biểu diễn cảm xúc với các mức cường độ khác nhau (low, medium, high), giúp tăng tính phong phú và chân thực của dữ liệu. Bộ dữ liệu này được thu âm trong điều kiện phòng thí nghiệm tiêu chuẩn, đảm bảo chất lượng âm thanh cao, và phù hợp với các nghiên cứu yêu cầu dữ liệu thực tế. CREMA-D là lựa chọn lý tưởng cho các dự án SER đa ngữ cảnh, đặc biệt là những ứng dụng cần tích hợp nhiều yếu tố như nhân khẩu học, cường độ cảm xúc, và tính tự nhiên trong giọng nói.

4 The proposed method

Bài báo cáo này được chạy bằng Kaggle để giảm thời gian huấn luyện và tiết kiệm chi phí vì các bộ dữ liệu khi được gộp lại lên tới 10722 dữ liệu dạng wav với các nhãn cảm xúc được đánh như Neutral, Happy, Sad, Angry, Fear, Disgust, và Surprise cũng như là để tiết kiệm tài nguyên máy. Cùng với đó là sử dụng các thư viện cơ bản như seaborn, numpy, metplotlib, pandas cũng như là các thư viện chuyên dụng hơn dùng cho việc nhận diện và huấn luyện mô hình như scikit-learn, tensorflow và keras.

Phương pháp phổ biến để phân tích tín hiệu âm thanh và giọng nói là biểu diễn dưới dạng 2D. Phân tích thời gian - tần số thường được sử dụng trong xử lý âm thanh. Chúng tôi chuyển đổi tín hiệu giọng nói thành biểu diễn 2D bằng phép biến đổi thành biểu đồ log mel-spectogram. Sau đó, biểu diễn 2D này được phân tích thông qua các kiến trúc Mạng nơ-ron tích chập (CNNs) và Bộ nhớ dài ngắn hạn (LSTM). Học sâu (deep learning) bao gồm các biểu diễn phân cấp với mức độ trừu tượng ngày càng tăng. Bằng cách lần lượt đi qua các mạng được xây dựng theo trình tự, kết quả tương ứng với từng khung âm thanh được chọn sẽ được phân loại bằng cách tính tổng các xác suất.

4.1 Preprocessing

Quá trình tiền xử lý dữ liệu âm thanh bao gồm việc tích hợp và chuẩn bị dữ liệu từ nhiều nguồn khác nhau. Cụ thể, dữ liệu từ các tập tin CREMA-D, TESS, và SAVEE được gộp lại thành một khung dữ liệu duy nhất để tiện cho việc sử dụng và chia sẻ.

Table 1: Distribution of Emotions

Emotion	Count
Disgust	1731
happy	1731
Sad	1731
Fear	1731
Angry	1731
Neutral	1607
Surprise	460

Tiếp theo, chuyển đổi tín hiệu âm thanh thô (waveform) sang dạng biểu diễn đặc trưng phù hợp cho mô hình học máy hoặc học sâu. Dữ liệu âm thanh được biểu diễn dưới dạng Mel spectrogram với các thông số như số lượng kênh Mel (n_mels=128) và tần số tối đa ($f_{max} = 8000$). Dải tần Mel là một thang đo phi tuyến tính, được tính toán dựa trên công thức:

$$m = 2595 * \log_{10}(1 + f/700) \quad (1)$$

Sau đó, Mel-Spectrogram được chuyển đổi thành Log-Mel-Spectrogram bằng cách áp dụng hàm log-

arit, với công thức

$$LogMel(n, m) = \log\left(M(n, m) + \epsilon\right) \quad (2)$$

Trong đó:

- f: là tần số gốc (Hz)
- m: là tần số Mel tương ứng

Bước này nén dải động (dynamic range), giảm tác động của các giá trị năng lượng cao, và làm nổi bật sự khác biệt ở các giá trị năng lượng thấp, vốn thường chứa thông tin cảm xúc quan trọng.

4.2 Data Augmentation

Nhằm vào mục đích tăng cường độ phong phú của tập dữ liệu đào tạo, nhất là đối với các bài toán yêu cầu dữ liệu lớn như Speech Emotion Recognition (SER), phương pháp này giúp mô hình học được từ những biến đổi đa dạng của dữ liệu thực tế, giảm nguy cơ overfitting và tăng khả năng tổng quát hóa. Các kỹ thuật data augmentation phổ biến bao gồm thêm nhiễu (add noise), giãn tín hiệu (stretch), dịt tín hiệu (shift), và thay đổi cao độ (pitch).

Thêm nhiễu là một kỹ thuật được áp dụng phổ biến bằng cách thêm nhiễu Gaussian (đặc biệt là white noise) vào tín hiệu âm thanh. Phương pháp này giúp giảm thiểu tác động của nhiễu ồn trong môi trường thực tế, giúp mô hình trở nên mạnh mẽ hơn trước nhiễu trong môi trường thực tế, nơi âm thanh thường bị lẫn tạp âm. Nó mô phỏng các điều kiện ghi âm khác nhau và làm giảm hiện tượng overfitting (quá khớp). Công thức tính như sau:

$$y[n] = x[n] + noise[n] \quad (3)$$

Trong đó:

- $y[n]$: Tín hiệu âm thanh sau khi dịch chuyển
- $x[n]$: Tín hiệu âm thanh gốc
- $shift$: Số lượng mẫu dịch chuyển.

Giãn tín hiệu (stretch) thay đổi tốc độ tín hiệu phát lại của âm thanh bằng cách kéo dài hoặc rút ngắn tín hiệu theo thời gian mà không thay đổi cao độ. Kỹ thuật này giúp tăng khả năng nhận diện của mô hình trước các biến đổi tự nhiên về tốc độ, làm cho mô hình không bị phụ thuộc vào tốc độ của người đọc.

Dịch chuyển tín hiệu (shift) thực hiện bằng cách dịch chuyển tín hiệu âm thanh trong miền thời gian. Phương pháp này giúp mô phỏng các biến đổi về vị trí thời gian xuất hiện tín hiệu trong dữ liệu, hữu ích trong các tình huống tự nhiên hoặc bất ngờ, những đoạn âm thanh vị cắt xén hoặc không bắt đầu chính xác tại điểm mong muốn. Công thức như sau:

$$y[n] = x[n - shift] \quad (4)$$

Trong đó:

- $y[n]$: Tín hiệu âm thanh sau khi dịch chuyển
- $x[n]$: Tín hiệu âm thanh gốc
- $shift$: Số lượng mẫu dịch chuyển.

Cuối cùng, thay đổi cao độ (pitch) thay đổi cao độ tín hiệu mà không thay đổi tốc độ. Kỹ thuật này giúp mô phỏng giọng nói của người có cao độ giọng nói khác nhau, giúp mô hình nhận diện không bị phụ thuộc vào đặc điểm giọng nói cá nhân.

Nhờ những kỹ thuật data augmentation này, tập dữ liệu trở nên phong phú hơn, giúp mô hình cải thiện hiệu quả trong việc nhận diện cảm xúc đầu ra cũng như là giúp tăng khả năng chịu đựng các dữ liệu bị nhiễu và đảm bảo khả năng tổng quát hóa tốt hơn trong các tình huống khác nhau.

4.3 Feature Extraction

Trong quá trình chuyển đổi giọng nói thành mel-spectogram thì có những trường hợp nhiễu xảy ra ví dụ như giọng nói theo những tone khác nhau và đôi khi lại có những âm thanh nhiễu khiến cho sự nhận diện trong bài này không được mạch lạc và có phần bị lệch hướng. Vì vậy phần trích xuất đặc trưng (Feature Extraction) là một bước quan trọng trong xử lý tín hiệu âm thanh, đặc biệt trong bài toán nhận diện cảm xúc từ giọng nói (Speech Emotion Recognition). Báo cáo này sẽ phân tích ba đặc trưng chính được sử dụng trong đoạn mã là Zero-Crossing Rate (ZCR), Root Mean Square Energy (RMSE) và Mel Frequency Cepstral Coefficients (MFCC). Kết hợp ba đặc trưng Zero-Crossing Rate (ZCR), Root Mean Square Energy (RMSE), và Mel Frequency Cepstral Coefficients (MFCC) với các kỹ thuật tăng cường dữ liệu để đảm bảo đầu vào phong phú và phù hợp cho mô hình nhận diện cảm xúc từ giọng nói. Quy trình này giúp cải thiện độ chính xác của mô hình thông qua việc nắm bắt được các thông tin quan trọng từ tín hiệu âm thanh.

4.3.1 Zero-Crossing Rate (ZCR)

Zero-Crossing Rate (ZCR) là một phép đo đếm số lần tín hiệu vượt qua giá trị 0 (đổi dấu) trong một khung (frame). Đặc trưng này thường dùng

để phân biệt giữa âm thanh có âm điệu và không có âm điệu. Tín hiệu âm thanh được chia thành các khung nhỏ (frame), sau đó đếm số lần tín hiệu thay đổi dấu trong mỗi khung. Kết quả được chuẩn hóa theo số mẫu trong khung. Đặc trưng này được sử dụng nhiều trong việc phân loại giọng nói và các loại âm thanh khác, đặc biệt hữu ích trong việc phân biệt âm thanh có âm điệu (voiced) như giọng nói bình thường và không có âm điệu (unvoiced) như tiếng thì thầm hoặc tiếng ồn, giúp phân loại trạng thái căng thẳng hoặc bình tĩnh.

$$ZCR = \frac{1}{N-1} \sum_{n=1}^{N-1} 1\{x[n] \cdot x[n-1] < 0\} \quad (5)$$

Trong đó:

- N: Số lượng mẫu trong một khung.
- $x[n]$: Giá trị tín hiệu tại thời điểm n.
- $1\{\cdot\}$: Hàm chỉ thị, trả về 1 nếu điều kiện đúng, và 0 nếu sai.

4.3.2 Root Mean Square Energy (RMSE)

Root Mean Square Energy (RMSE) là phép đo năng lượng trung bình của tín hiệu trong một khung, phản ánh độ lớn (amplitude) của tín hiệu. Với việc tín hiệu được chia thành các khung, sau đó tính bình phương của mỗi giá trị tín hiệu trong khung, trung bình các giá trị này, và lấy căn bậc hai của kết quả.

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^{N} x[n]^2} \quad (6)$$

Trong đó:

- N: Số lượng mẫu trong một khung.
- $x[n]$: Giá trị tín hiệu tại thời điểm n.

RMSE thường được dùng để đo cường độ tín hiệu âm thanh. Trong bài toán nhận diện cảm xúc, RMSE giúp phân biệt cảm xúc mạnh (như tức giận hoặc phấn khích) và cảm xúc nhẹ (như buồn hoặc bình tĩnh).

4.3.3 Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral Coefficients (MFCC) là tập hợp các đặc trưng phổ được tính toán dựa trên thang tần số Mel (phản ánh cảm nhận âm thanh của con người). MFCC sẽ trích xuất các đặc trưng quan trọng của giọng nói, bao gồm âm sắc, ngữ điệu, và nội dung. Trong nhận diện cảm xúc, MFCC giúp phân biệt các trạng thái cảm xúc dựa trên âm sắc và tần số. Đây là đặc trưng phổ biến nhất trong xử lý giọng nói và âm thanh với quy trình tính toán bao gồm các bước sau:

- Chia tín hiệu thành các khung nhỏ để xử lý.
- Áp dụng cửa sổ Hamming để giảm nhiễu tại biên khung.
- Tính phổ năng lượng (Power Spectrum) bằng Fourier Transform.
- Áp dụng bộ lọc Mel để chuyển đổi phổ sang thang tần số Mel.
- Lấy log năng lượng của thang Mel để chuẩn hóa giá trị.
- Áp dụng Discrete Cosine Transform (DCT) để trích xuất các hệ số MFCC cuối cùng.

Hệ số MFCC k được tính từ phổ Mel S:

$$MFCC_k = \sum_{m=1}^{M} \log(S_m) \cos\left[k \cdot (m - 0.5) \cdot \frac{\pi}{M}\right]$$

Trong đó:

- M: Số lượng bộ lọc Mel.
- S_m: Giá trị năng lượng sau khi áp dụng bộ lọc Mel.
- k: Chỉ số hệ số MFCC.

4.4 Training model

4.4.1 Model CNN nhóm tự tạo

CNN là một mạng nơ-ron được thiết kế để phân tích dữ liệu dưới dạng mảng đa chiều. Một mảng đầu vào trong đó các giá trị dữ liệu lân cận có tương quan là một ứng dụng lý tưởng cho CNN. Vì vậy, CNN thường được triển khai trong nhiều lĩnh vực như xử lý hình ảnh, video và biểu diễn thời gian-tần số của âm thanh. Ý tưởng chính của CNN là khai thác các đặc tính của tín hiệu: kết nối cục bộ, chia sẻ trọng số, pooling và sử dụng nhiều lớp.

Tabel 2 trình bày chi tiết kiến trúc của mô hình CNN được đề xuất. Trong nghiên cứu này, chúng tôi sử dụng biểu đồ mel-spectogram để biểu diễn tín hiệu giọng nói dưới dạng 2D với kích thước khung là 256 và chồng lấn 50%, với kích thước đầu vào của dữ liệu là 128 × 128, như được minh

Table 2: Self-generated CNNs Model Architecture

Layers	Output
Conv1D - BN - MP	(None, 1188, 512)
Conv1D - BN - MP - Dr(0.2)	(None, 2376, 512)
Conv1D - BN - MP	(None, 594, 512)
Conv1D - BN - MP - Dr(0.2)	(None, 297, 512)
Conv1D - BN - MP - Dr(0.2)	(None, 149, 512)
Flatten	(None, 9600)
FC - BN	(None, 512)
FC-Softmax	(None, 7)

Table 3: LSTM Model Architecture

Layers	Output
LSTM - Dr(0.2) - BN	(None, 2376, 128)
LSTM - Dr(0.2) - BN	(None, 2376, 128)
LSTM - Dr(0.2) - BN	(None, 2376, 128)
LSTM - Dr(0.2) - BN	(None, 2376, 128)
LSTM - Dr(0.2) - BN	(None, 2376, 128)
LSTM - Dr(0.2) - BN	(None, 2376, 128)
LSTM - Dr(0.2)	(None, 128)
FC-Softmax	(None, 7)

họa trong Table 2. Mô hình bắt đầu bằng việc xếp chồng các lớp tích chập (Conv1D), chuẩn hóa (BN), max-pooling (MP), và dropout (Dr) để học các đặc trưng của tín hiệu. Cụ thể, mô hình bao gồm năm lớp tích chập xen kẽ với max-pooling, và một số lớp có thêm dropout để giảm overfitting. Sau các lớp tích chập, đầu ra được làm phẳng (Flatten) và đi qua một lớp kết nối đầy đủ (FC) với 512 nút, trước khi đến lớp đầu ra (Softmax) để dự đoán một trong bảy lớp cảm xúc.

4.4.2 Model LSTM

Để xử lý dữ liệu tuần tự như tín hiệu âm thanh, việc kết hợp thêm LSTM giúp tăng khả năng học các mối quan hệ thời gian và ngữ cảnh giữa các bước thời gian. Mô hình được trình bày trong Table 3 sử dụng LSTM với 128 nút, kết hợp Dropout để giảm overfitting và Batch Normalization để ổn định quá trình huấn luyện. Tổng cộng, sáu lớp LSTM được xếp chồng để khai thác sâu các đặc trưng thời gian, trong đó lớp cuối cùng trả về chuỗi ẩn cuối cùng thay vì chuỗi đầy đủ để tóm gọn thông tin ngữ cảnh. Cuối cùng, đầu ra từ lớp LSTM được đưa qua một lớp Dense sử dụng hàm Softmax để dự đoán 7 lớp cảm xúc. Việc tích hợp LSTM cùng CNN không chỉ tối ưu hóa việc học các đặc trưng cục bộ (CNN) mà còn tăng cường khả năng nắm bắt ngữ cảnh toàn cục (LSTM), giúp mô hình phù hợp hơn với bài toán nhận diện cảm xúc từ tín hiệu âm thanh.

4.4.3 Pre-trained model VGG16

Ngoài việc sử dụng các model trên, chúng tôi sẽ sử dụng thêm pre-trained model như VGG16, một mạng nơ-ron tích chập sâu đã được tiền huấn luyện trên bộ dữ liệu về Image, để tăng cường khả năng học các đặc trưng.

VGG16 với kiến trúc các lớp tích chập xếp chồng (stacked convolutional layers) và max-pooling giúp trích xuất đặc trưng mạnh mẽ từ dữ liệu đầu vào. Việc sử dụng mạng tiền huấn luyện này mang lại hai lợi ích chính: thứ nhất, nó tận dụng các đặc trưng cấp thấp và cấp trung đã được học từ một tập dữ liệu lớn, giúp cải thiện khả năng tổng quát hóa; thứ hai, giảm thời gian huấn luyện nhờ việc tái sử dụng các trọng số đã được tối ưu hóa.

Sau khi trích xuất đặc trưng bằng VGG16, đầu ra được đưa qua các lớp Dense và Softmax để dự đoán 7 lớp cảm xúc. Sự kết hợp VGG16 giúp mô hình đạt hiệu suất tốt hơn trong việc nhận diện cảm xúc, đặc biệt khi dữ liệu đầu vào là hình ảnh hoặc biểu diễn tần số-thời gian của âm thanh.

Table 4: VGG16 Pre-trained Model Architecture

Layers	Output
Conv2D-Conv2D-MP2D-BN	(None, 1188, 1, 128)
Conv2D-Conv2D-MP2D-BN	(None, 594, 1, 256)
Conv2D-Conv2D-Conv2D-MP2D-BN	(None, 297, 1, 512)
Conv2D-Conv2D-Conv2D-MP2D-BN	(None, 148, 1, 512)
Conv2D-Conv2D-Conv2D-MP2D-BN	(None, 71, 1, 512)
Flatten	(None, 37888)
FC - Dr(0.2)	(None, 4096)
FC - Dr(0.2)	(None, 4096)
FC-Softmax	(None, 7)

4.5 Evaluation metrics

Trong bài nghiên cứu này, chúng tôi sử dụng các độ đo phổ biến như Precision, Recall, và F1-Score để đánh giá hiệu quả của mô hình nhận diện cảm xúc từ giọng nói. Đặc biệt, các độ đo được tính toán theo cả hai cách: macro và weighted. Macro average tính trung bình các giá trị Precision, Recall, và F1-Score cho từng lớp cảm xúc mà không quan tâm đến số lượng mẫu trong mỗi lớp, giúp đánh giá hiệu năng mô hình trên từng lớp một cách công bằng. Trong khi đó, weighted average điều chỉnh trung bình các độ đo theo tỷ lệ số lượng mẫu trong từng lớp, đảm bảo kết quả phản ánh chính xác hơn

hiệu suất tổng thể trên tập dữ liệu mất cân bằng. Cách tiếp cận này giúp chúng tôi đánh giá toàn diện khả năng mô hình nhận diện chính xác cả các cảm xúc phổ biến và ít phổ biến hơn trong tập dữ liệu.

5 Experimental Results

Hệ thống nhận diện cảm xúc đa phương thức đã được đánh giá dựa trên 3 mô hình khác nhau. Kết quả thực nghiệm đánh giá hiệu suất của ba mô hình nhận diện cảm xúc giọng nói: CNNs, LSTMs và VGG16 được huấn luyện trước, dựa trên các tiêu chí Precision, Recall, F1-score và Accuracy.

Mô hình CNNs nổi bật với hiệu suất vượt trội, đạt độ chính xác tổng thể 94% và duy trì các chỉ số Precision, Recall và F1-score trên 90% ở hầu hết các cảm xúc, cho thấy khả năng nhận diện toàn diện và ổn định. Ngược lại, mô hình LSTMs cho thấy hạn chế rõ rệt với độ chính xác tổng thể chỉ đạt 53%, đặc biệt yếu trong việc nhận diện các cảm xúc như "Sợ hãi" và "Trung tính". Trong khi đó, mô hình VGG16 đã được huấn luyện trước gây ấn tượng mạnh khi đạt độ chính xác 88%, đồng thời thể hiện sự đồng đều và ổn định trong việc nhận diện hầu hết các cảm xúc.

Tóm lại, VGG16 có độ ổn định khá tốt và hiệu suất nhất quán, tuy nhiên LSTMs gặp nhiều khó khăn trong việc nhận diện một số cảm xúc nhất định, trong khi CNNs chứng minh khả năng vượt trội với độ chính xác cao và sự cân bằng giữa các cảm xúc, trở thành lựa chọn ưu việt trong bài toán này.

Emotion	Precision	Recall	F1-score
Angry	0.95	0.95	0.95
Disgust	0.96	0.93	0.94
Fear	0.92	0.94	0.93
Happy	0.96	0.93	0.94
Neutral	0.97	0.92	0.95
Sad	0.9	0.97	0.93
Surprise	0.96	0.96	0.96
Accuracy			0.94
Macro avg	0.94	0.94	0.94
Weighted avg	0.94	0.94	0.94

Table 5: Result of CNNs model

Emotion	Precision	Recall	F1-score
Angry	0.61	0.58	0.6
Disgust	0.54	0.44	0.49
Fear	0.74	0.24	0.36
Happy	0.45	0.61	0.51
Neutral	0.43	0.75	0.55
Sad	0.61	0.49	0.55
Surprise	0.69	0.62	0.66
Accuracy			0.53
Macro avg	0.58	0.53	0.53
Weighted avg	0.57	0.53	0.52

Table 6: Result of LSTMs model

Emotion	Precision	Recall	F1-score
Angry	0.82	0.92	0.87
Disgust	0.89	0.89	0.89
Fear	0.9	0.78	0.84
Happy	0.82	0.9	0.86
Neutral	0.96	0.87	0.91
Sad	0.92	0.9	0.91
Surprise	0.85	0.87	0.86
Accuracy			0.88
Macro avg	0.88	0.88	0.88
Weighted avg	0.88	0.88	0.88

Table 7: Result of Pre-trained model VGG16

6 Conclusion and Future works

Mặc dù mô hình CNNs đã được đào tạo sẵn cho thấy hiệu quả vượt trội trong bài toán nhận diện cảm xúc giọng nói, nghiên cứu này vẫn tồn tại một số hạn chế cần được khắc phục. Hiệu suất thấp của mô hình LSTMs, đặc biệt đối với các cảm xúc bị lệch số lượng nhãn như "Trung tính" và "Bất ngờ", cho thấy cần phải cải thiện cách biểu diễn đặc trưng và tối ưu hóa siêu tham số. Ngoài ra, nghiên cứu chủ yếu tập trung vào ba mô hình, chưa khai thác các mô hình tiên tiến hơn như Transformer hay Wav2Vec, và cũng chưa xem xét sự mất cân đối dữ liệu giữa các cảm xúc, có thể dẫn đến hiệu suất không đồng đều. Hơn nữa, thời gian thực hiện cũng như là tài nguyên bị hạn chế và tính khả thi khi triển khai thực tế chưa được đánh giá, làm hạn chế việc ứng dụng các mô hình này trong thực tiễn.

Trong tương lai, cần tập trung cải thiện hiệu suất của LSTMs bằng cách áp dụng các kỹ thuật như attention mechanism, đồng thời thử nghiệm các mô hình hiện đại để so sánh hiệu quả. Việc cân bằng dữ liệu bằng các kỹ thuật như oversampling hoặc thêm các kĩ thuật data augmentation cũng rất quan trọng để cải thiện độ chính xác. Hướng đi đầy hứa hẹn khác là tích hợp dữ liệu âm thanh với các tín hiệu khác, như nét mặt hoặc ngôn ngữ cơ thể, để xây dựng hệ thống nhận diện cảm xúc đa phương tiện, giúp nâng cao hiệu suất và khả năng ứng dụng. Những cải tiến này không chỉ cải thiện

kết quả nghiên cứu mà còn mở ra cơ hội mới trong các lĩnh vực như chăm sóc khách hàng, giáo dục và y tế.

References

S. Choudhary. 2021. In-depth intuition of k-means clustering algorithm in machine learning. *Analytics Vidhya*.

R. Coyle. 2019. A gentle introduction to k-nearest neighbours for machine learning. URL không còn hợp lệ.

Wenxing Hong, Siting Zheng, Huan Wang, and Jianchao Shi. 2013. A job recommender system based on user clustering. *Journal of Computers*, 8.

M. Horn. 2023. Choosing the right number of principal components. *Baeldung*.

Yuxuan Hu, Ke Li, and Anran Meng. Agglomertive hierarchical clustering using ward linkage.

P. Jain. 2021. Basics of countvectorizer. *Medium*.

J. Karjalainen. 2020. Machine learning pipeline. Valohai.

R. Kumar. 2020. Davies bouldin index. GeeksforGeeks.

M. Przybyla. 2019. Basics of countvectorizer. *Towards Data Science*.

Betty Puspasari, Lany Damayanti, Andy Pramono, and Aang Darmawan. 2021. Implementation k-means clustering method in job recommendation system. pages 1–6.

T. Raschka. 2019. Silhouette coefficient - validating clustering techniques. *Towards Data Science*.

M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, and F. A. Rodrigues. 2016. Clustering algorithms: A comparative approach. *PLOS ONE*, 11(14):e0210236.

Phú Huy Đỗ and Xuân Hoàng Huy Phan. 2024. Xây dựng hệ thống website tìm kiếm việc làm công nghệ thông tin của thành phố Đà nẵng. In *Ngành Công nghệ thông tin*.