BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Nguyễn Thị Thanh Mai

NÂNG CAO ĐỘ CHÍNH XÁC XÁC THỰC NGƯỜI NÓI SỬ DỤNG HỌC SÂU VỚI TÀI NGUYÊN HẠN CHẾ

TÓM TẮT LUẬN ÁN TIẾN SĨ MÁY TÍNH

Ngành: Khoa học máy tính

Mã số: 9 48 01 01

Hà Nội - 2024

Công trình được hoàn thành tại: Học viện Khoa học và Công nghệ,
Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Người hướng dẫn khoa học:
Người hướng dẫn 1: PGS.TS. Nguyễn Đức Dũng, Viện Công nghệ thông tin
Người hướng dẫn 2: PGS. TS. Lương Chi Mai, Viện Công nghệ thông tin

Phản biện 1: PGS.TS. Phan Xuân Hiếu
Phản biện 2: PGS.TS. Vũ Hải
Phản biện 3: TS. Đỗ Văn Hải

Luận án được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ cấp Học viện họp tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam vào hồi 09 giờ 00, ngày 02 tháng 12 năm 2024.

Có thể tìm hiểu luận án tại:
1. Thư viện Học viện Khoa học và Công nghệ
2. Thư viện Quốc gia Việt Nam

MỞ ĐẦU

1. Tính cấp thiết của luận án

Trong những năm gần đây, xác thực người nói dựa trên các mô hình học sâu đã đạt được nhiều kết quả vượt trội so với các mô hình học máy truyền thống. Theo cách tiếp cận truyền thống, quá trình trích chọn đặc trưng âm học được thực hiện thủ công và tách biệt khỏi quá trình mô hình hóa đặc trưng người nói. Ví dụ như đặc trưng MFCCs (Mel-Frequency Cepstral Coefficients) đã được sử dụng rộng rãi làm đầu vào cho nhiều hệ thống xử lý giọng nói cũng như hệ thống xác thực người nói. Điểm mạnh của MFCCs nằm ở khả năng biểu diễn tín hiệu giọng nói dạng nén, đặc trưng này nắm bắt những nội dung ngữ âm quan trọng của giọng nói. Tuy nhiên, phần lớn năng lượng của MFCCs thường tập trung vào các hệ số bậc thấp trong khoảng 13 đến 39 hệ số đầu tiên. Nếu sử dụng đặc trưng MFCCs làm đầu vào cho các mô hình học sâu như ResNets thì năng lượng của MFCCs ở các hệ số bậc cao có thể làm giảm hiệu năng hệ thống xác thực. Do vậy, việc nghiên cứu, phát hiện những đặc trưng âm học mới để nâng cao hiệu năng hệ thống xác thực người nói vô cùng quan trọng và cần thiết.

Hơn nữa, việc huấn luyện các mô hình học sâu trong xác thực người nói thường yêu cầu một lượng lớn dữ liệu lớn và đa dạng. Hiện nay, dữ liệu người nói cho tiếng Anh vẫn chiếm ưu thế hơn so với ngôn ngữ tiếng Việt. Cụ thể, tập dữ liệu VoxCeleb2 [14] gồm 6,112 người nói trong khi dữ liệu VLSP2021-SV [18] chỉ có 1,305 người nói. Số giờ thu âm của VoxCeleb2 là 2,442 giờ trong khi dữ liệu tiếng Việt VLSP2021-SV chỉ là 41 giờ (số giờ dữ liệu tiếng Anh lớn gấp 60 lần so với dữ liệu tiếng Việt). Như vậy, khi dữ liệu người nói tiếng Việt còn hạn chế thì có một số giải pháp là thu thập thêm dữ liệu tiếng nói người Việt hoặc sử dụng lại các mô hình được huấn luyện trên các tập dữ liệu lớn tiếng Anh hay tiếng Trung, sau đó huấn luyện tiếp trên dữ liệu tiếng nói người Việt. Tuy nhiên, việc thu thập, bổ sung thêm dữ liệu tiếng Việt có thể rất tốn kém và khó thực hiện. Khi đó, mô hình được huấn luyện trên dữ liệu hạn chế dẫn tới hiện tượng quá khớp và không có khả năng tổng quát hóa với những dữ liệu mới chưa được biết đến. Với phương pháp học chuyển giao có ưu điểm kế thừa được 2 những đặc trưng mức cao từ tập dữ liệu lớn tiếng Anh nên cũng tiết kiệm thời gian huấn luyện mô hình trên dữ liệu người nói tiếng Việt. Cùng với sự phát triển nhanh các mô hình học sâu cho bài toán xác thực người nói, việc lựa chọn đặc trưng nào, mô hình nào phù hợp với dữ liệu tiếng nói hạn chế cũng là một trong những nhiệm vụ mà luận án cần những nghiên cứu, so sánh, thử nghiệm và đánh giá.

Bên cạnh đó, những nghiên cứu về hệ thống xác thực người nói đang rất cần được tích hợp vào các hệ thống thông minh, ứng dụng rộng rãi trong thực tế như:

• Ngăn chặn truy cập trái phép: Hệ thống xác thực người nói giúp đảm bảo rằng chỉ những người được ủy quyền mới có thể truy cập vào các hệ thống, dịch vụ hoặc thông tin nhạy cảm.

- Bảo vệ dữ liệu cá nhân: Trong bối cảnh thông tin cá nhân ngày càng bị đe dọa bởi các hành vi trộm cắp và gian lận, hệ thống xác thực người nói cung cấp một lớp bảo vệ bổ sung, đảm bảo rằng dữ liệu chỉ được truy cập bởi người chủ thực sự.
- Giảm thiểu chi phí quản lý mật khẩu: Việc quản lý và khôi phục mật khẩu truyền thống có thể tốn kém và phức tạp, trong khi xác thực giọng nói có thể giảm thiểu chi phí này.

- Tối ưu hóa quy trình: Hệ thống xác thực giọng nói có thể tự động hóa và đơn giản hóa nhiều quy trình xác thực, từ đó giảm thiểu công việc thủ công và chi phí nhân sự.

Từ những lý do như vậy, luận án lựa chọn đề tài nghiên cứu "Nâng cao độ chính xác xác thực người nói sử dụng học sâu với tài nguyên hạn chế". Đây là một vấn đề cấp thiết và có tính thời sự, ứng dụng cao. Các kết quả nghiên cứu của luận án giúp nâng cao độ chính xác xác thực người nói tiếng Việt.

2. Mục tiêu nghiên cứu

Mục tiêu nghiên cứu của luận án là nghiên cứu đề xuất một số giải pháp nâng cao độ chính xác xác thực người nói với tài nguyên hạn chế. Mục tiêu cụ thể là:

- Nghiên cứu, lựa chọn đặc trưng âm học cho mô hình học sâu nhằm nâng cao độ chính xác xác thực người nói;
- Nghiên cứu, phân tích, so sánh, đánh giá các mô hình học sâu và các phương pháp học chuyển giao áp dụng cho tài nguyên dữ liệu hạn chế nhằm cải thiện độ chính xác xác thực người nói.

3. Bố cục luận án

Chương 1: Tổng quan kiến thức nền tảng và ứng dụng học sâu cho bài toán xác thực người nói

Chương 1 giới thiệu tổng quan về bài toán xác thực người nói theo cách tiếp cận học sâu. Qua đó mô tả hệ thống tổng quan xác thực người nói và định hướng nghiên cứu nâng cao hiệu năng xác thực người nói phù hợp với xu thế hiện nay và thực tiễn.

Chương 2: Nâng cao độ chính xác xác thực người nói tiếng việt sử dụng đặc trưng Mel-Filter bank với mô hình ECAPA-TDNN

Chương 2 tập trung vào khảo sát, đánh giá, thử nghiệm các đặc trưng làm đầu vào cho các mô hình học sâu hiện đại, cụ thể là mô hình ECAPA-TDNN. Từ thực nghiệm với mô hình đề xuất trong luận án cho thấy đặc trưng Mel-Filterbank Energys với mô hình ECAPA-TDNN cho kết quả tốt hơn so với đặc trưng MFCCs (cho ECAPA-TDNN).

Chương 3: Nâng cao độ chính xác xác thực người nói sử dụng học chuyển giao với mô hình Rawnet3

Chương 3 thử nghiệm, đánh giá nâng cao độ chính xác hệ thống xác thực người nói nhờ sử dụng kỹ thuật học chuyển giao. Với các mô hình huấn luyện trước trên các tập dữ liệu lớn, sử dụng mô hình học sâu Rawnet3 với đầu vào là dữ liệu âm thanh thô, sau đó tinh chỉnh và huấn luyện trên dữ liệu tiếng Việt có kết quả tốt hơn so với không học chuyển giao.

Kết luận. Trình bày các đóng góp chính của luận án và chỉ ra các hạn chế và hướng phát triển tiếp theo.

4. Những đóng góp mới của luận án

- **Đề xuất sử dụng đặc trưng MFBEs với mô hình ECAPA-TDNN cho bài toán xác thực người nói tiếng Việt; ([CT1] và [CT2])**
- **Đề xuất sử dụng mô hình RawNet3 trong học chuyển giao cho bài toán xác thực người nói với tài nguyên hạn chế. ([CT3] và [CT4])**

CHƯƠNG 1: TỔNG QUAN KIẾN THỨC VÀ ỨNG DỤNG HỌC SÂU CHO BÀI TOÁN XÁC THỰC NGƯỜI NÓI

Trong Chương 1, phần đầu tiên giới thiệu tổng quan các nghiên cứu liên quan về bài toán xác thực người nói và các vấn đề khó khăn cần giải quyết. Tiếp theo, NCS trình bày tổng quan về tình hình nghiên cứu trong và ngoài nước cũng như cách tiếp cận trong xác thực người nói. Cuối cùng, NCS trình bày tổng quan về hệ thống xác thực người nói: đặc trưng, mô hình, dữ liệu, phương pháp đánh giá, phương pháp cải tiến nâng cao độ chính xác xác thực người nói.

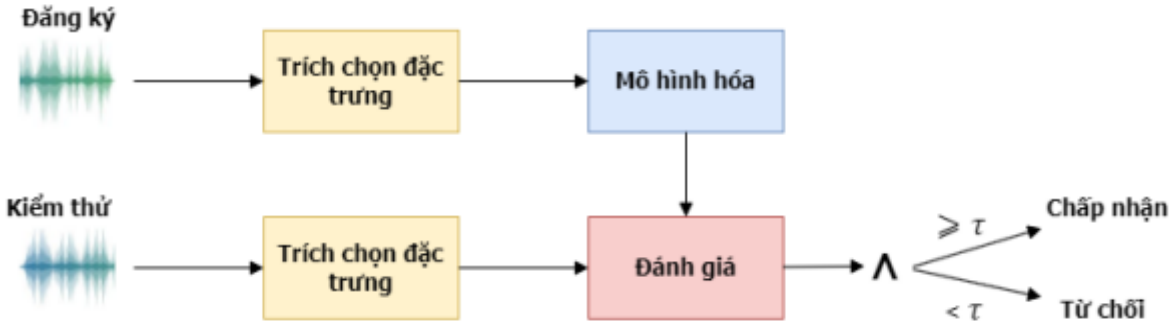
1.1.Giới thiệu

Xác thực người nói là một trong những bài toán thuộc lĩnh vực nhận dạng và xác thực sinh trắc học dựa trên giọng nói. Mục tiêu của bài toán này là kiểm tra xem giọng nói của một người có khớp với giọng nói mẫu đã được đăng ký trước đó hay không.

Bài toán xác thực người nói là một trong những bài toán thuộc lĩnh vực nhận dạng và xác thực sinh trắc học dựa trên giọng nói. Mục tiêu của bài toán này là kiểm tra xem giọng nói của một người có khớp với giọng nói mẫu đã được đăng ký trước đó hay không.

Đầu vào: Một đoạn tín hiệu giọng nói của người dùng muốn xác thực (được gọi là giọng nói kiểm thử), và một mẫu giọng nói đã được lưu trong hệ thống (được gọi là giọng nói đã đăng ký/giọng nói mẫu).

Đầu ra: Một quyết định xác thực để trả lời câu hỏi: "Người nói kiểm thử có đúng là người đã đăng ký giọng nói mẫu hay không?" Dựa vào kết quả so sánh với ngưỡng, hệ thống sẽ trả về "đúng" (chấp nhận) hoặc "sai" (từ chối).



Hình 1.1: Sơ đồ tổng quan hệ thống xác thực người nói.

1.2.Các nghiên cứu liên quan

Tình hình nghiên cứu nước ngoài

Nghiên cứu nhận dạng và xác thực người nói hiện vẫn được coi là một mục tiêu theo đuổi hướng tới việc nâng cao độ chính xác nhận dạng. Ví dụ, nghiên cứu ban đầu bị hạn chế ở những bài toán bị ràng buộc phụ thuộc văn bản và tập trung vào việc giải quyết các biến thể gây ra bởi cách phát âm ngẫu nhiên, trong đó Mô hình Markov ẩn HMM (Hidden Markov Model) [87] là mô hình phổ biến nhất trong các phương pháp xác thực người nói độc lập văn bản và phải xử lý các biến thể

ngữ âm, đã làm bùng nổ mô hình GMM-UBM (Gaussian mixture modelling with a universal background) [92]. Nghiên cứu sâu hơn đã cố gắng giải quyết sự thay đổi giữa các phiên do kênh và phong cách nói, trong đó kiến trúc i-vector/PLDA (Probabilistic Linear Discriminant Analysis) là phổ biến nhất thành công [19]. Gần đây, các nhà nghiên cứu tập trung giải quyết các biến thể phức tạp trong các tình huống tự nhiên và các phương pháp học sâu đã được chứng minh rất mạnh [106][108][114].

Các phương pháp học sâu để nhận dạng người nói có thu hút được nhiều sự chú ý nhờ những tiến bộ trong khả năng tính toán và sự sẵn có của các bộ dữ liệu lớn trong tự nhiên [72]. Một số lượng lớn các nghiên cứu sử dụng mô hình DNN cho trích trọn nhúng người nói đã được thực hiện trong vài năm vừa qua. Hầu hết các nghiên cứu nổi bật đã sử dụng các kiến trúc CNN (Convolutional Neural Network) như ResNet [121] cho kết quả tốt trong một vài năm gần đây. Mặt khác, các mô hình thành công khác như x-vectors [108] đã sử dụng TDNN trích chọn đặc trưng nhúng từ MFCC. Phần lớn các mô hình DNN được sử dụng trong nhận dạng người nói như một câu nói duy nhất làm đầu vào và cung cấp kích thước cố định vectơ như là nhúng lời nói cho một phát âm. Một quá trình khác sau đó sẽ tính toán độ tương tự giữa hai véc tơ nhúng (câu nói đăng kí và câu nói kiểm thử) để xác định người nói.

Mạng nơ-ron hồi quy RNN (Recurrent Neural Network) cũng được sử 10 dụng trong một số nghiên cứu. Gần đây, mô hình RNN [80][100][126] được phát triển dùng các hệ số MFCCs.

Trong [118] kiến trúc LSTM (Long Short-Term Memory) cũng áp dụng trên MFCC, kết quả nhúng sử dụng định danh người nói tính theo trung bình khoảng cách cosin. Mặt khác, các mô hình này cũng sử dụng kiến trúc LSTM như một công cụ trích chọn i-vectors. Một số nghiên cứu khác sử dụng kết hợp CNN và RNN dựa trên các lớp tích chập giữa các MFCC đầu vào và RNN.

Trong công bố [14] của nhóm nghiên cứu đã sử dụng mô hình CNN huấn luyện trên dữ liệu khoảng 6,000 giọng nói tiếng Anh. Với cách tiếp cận này thì mỗi đoạn tiếng nói có độ dài 3 giây sẽ biến đổi thành ảnh phổ. Các ảnh này sẽ là đầu vào cho mạng CNN và hệ thống cho kết quả khá tốt với tỉ lệ lỗi 3.95 % trên dữ liệu kiểm thử [60].

Một hướng nghiên cứu nhận dạng người nói trên dữ liệu câu nói có độ dài ngắn nhỏ hơn 2 giây [126] cũng đã thu hút được sự quan tâm của cộng đồng nghiên cứu. Nhóm nghiên cứu này cũng sử dụng x-vectors [108] làm mô hình cơ bản sau đó phát triển mở rộng kiến trúc TDNN [82].

Trong bài báo đã công bố [73] thực nghiệm trên dữ liệu Voxceleb2 [14], nhóm nghiên cứu tại Mỹ cho kết quả đánh giá tỉ lệ lỗi là 3.82%, công ty AI Trung Quốc cho kết quả 3.81%, nhóm IDLab tại Bỉ cho kết quả tốt nhất 3,73%.

Từ năm 2019 đến 2023 [15][40][73], đã diễn ra nhiều cuộc thi tập trung vào kỹ thuật nhận dạng người nói. Các cuộc thi này nhằm mục đích thúc đẩy nghiên cứu trong lĩnh vực nhận dạng người nói, đồng thời cung cấp các hệ thống nhận dạng cơ sở, dữ liệu huấn luyện, cùng với các tiêu chí đánh giá. Các bài toán trong cuộc thi bao gồm: xác thực người nói, định danh người nói và tách rời người nói.

Tình hình nghiên cứu trong nước

Tại Việt Nam, nghiên cứu và ứng dụng về nhận dạng người nói cũng là một lĩnh vực thu hút được sự quan tâm của các nhà nghiên cứu và phát triển trong vài năm trở lại đây. Các nhóm và hướng nghiên cứu có thể kể đến như sau: Tại cuộc thi Zalo AI Challenge 2020, bài toán nhận dạng người nói đạt tỉ lệ lỗi là 5%. Mô hình huấn luyện trên 400 giọng nói Việt và được đánh giá trên dữ liệu của Ban tổ chức. Bên cạnh đó, một nhóm nghiên cứu khác cũng đã sử dụng mô hình học 11 đa nhiệm [84] kết hợp giữa hàm mất mát Triplet [23] cho bài toán xác thực giọng nói. Mô hình huấn luyện trên dữ liệu tiếng Anh, sau đó tinh chỉnh trên một lượng dữ liệu nhỏ cho tiếng Việt. Kết quả đánh giá trên 65 giọng nói Cơ sở dữ liệu tiếng Việt VIVOS [67] có tỉ lệ lỗi 4.3%. Trong cộng đồng nghiên cứu xử lý ngôn ngữ tự nhiên thì bài toán nhận dạng người nói cũng là một bài toán đang được quan tâm.

Hội thảo VLSP 2021 [18] cũng đã đưa vào cuộc thi nhận dạng người nói tiếng Việt với cơ sở dữ liệu công bố khoảng hơn 1,300 giọng nói. Cuộc thi cũng đã thu hút được cộng đồng nghiên cứu và các nhóm tham gia cuộc thi và kết quả test tốt nhất từ Ban tổ chức có tỉ lệ lỗi 1.9%. Một trong những mô hình mà các đội tham gia đã thử nghiệm là mô hình ECAPA-TDNN. Mô hình ECAPA-TDNN [22] cũng được ứng dụng rộng rãi trong các bài toán như nhận dạng ngôn ngữ, nhận dạng cảm xúc, ...

Nhóm nghiên cứu [111] sử dụng đặc trưng log Mel-filterbanks làm đầu vào cho mạng học sâu ResNet. Kết quả thực nghiệm đánh giá trên dữ liệu do nhóm tự thu thập trên kênh YouTube gồm 580 người nói với 5,000 câu nói. Kết quả thực nghiệm cho thấy sử dụng mô hình huấn luyện có sẵn trên tập dữ liệu tiếng Anh, sau đó tinh chỉnh trên dữ liệu tiếng Việt cho kết quả tốt hơn nếu chỉ huấn luyện dữ liệu tiếng Việt.

Nhóm tác giả Học viện Bưu chính Viễn thông [76] cũng đã thử nghiệm so sánh giữa đặc trưng MFCCs và đặc trưng GFCCs [112] trên tập dữ liệu tiếng Việt hạn chế với số lượng dữ liệu huấn luyện 20 người nói tự thu âm.

Nhóm tác giả thực nghiệm so sánh tỉ lệ lỗi của hai mô hình GMMs và mô hình ResNet. Kết quả cho thấy mô hình ResNet sử dụng đặc trưng đầu vào là GFCCs cho tỉ lệ lỗi thấp hơn so với mô hình GMMs truyền thống.

Nhóm nghiên cứu tại Đại học Bách khoa Hà Nội cũng đã xây dựng cơ sở dữ liệu nhận dạng người nói tiếng Việt Vietnam-Celeb [83] với số lượng 1,000 người nói. Đây là tập dữ liệu mới nhất và lớn nhất dùng cho bài toán nhận dạng người nói tiếng Việt. NCS sẽ trình bày chi tiết cơ sở dữ liệu này trong phần sau.

1.3.Xác thực người nói với dữ liệu tài nguyên hạn chế

Trong thời đại số hóa hiện nay, xác thực người nói đã trở thành một phần quan trọng trong các ứng dụng bảo mật và nhận diện, như trong các hệ thống thanh toán điện tử, truy cập an toàn vào thông tin nhạy cảm và nhận diện giọng nói trong các trợ lý ảo. Tuy nhiên, một trong những thách thức lớn nhất trong 12 việc phát triển các hệ thống xác thực người nói hiệu quả là việc thiếu hụt tài nguyên dữ liệu, đặc biệt là dữ liệu có nhãn. Việc thu thập dữ liệu giọng nói với nhãn (ví dụ: danh tính người nói) thường tốn kém và mất thời gian. Khi số lượng mẫu có nhãn hạn chế, việc huấn luyện các mô hình học máy trở nên khó khăn, dẫn đến hiệu suất kém trong việc xác thực người nói.

Mỗi người nói có các đặc điểm giọng nói riêng biệt, và sự biến đổi giữa các cá nhân có thể rất lớn. Khi dữ liệu có nhãn không đủ, các mô hình không thể học được các đặc trưng chính xác để phân biệt giữa các người nói khác nhau. Mô hình được huấn luyện trên một tập dữ liệu nhỏ có thể không đủ khả năng tổng quát khi được áp dụng vào các tình huống thực tế, nơi tồn tại sự đa dạng về âm thanh, điều kiện môi trường và ngữ điệu của người nói.

Một số hướng nghiên cứu chính và các công trình liên quan đến việc giải quyết vấn đề dữ liệu hạn chế trong xác thực người nói:

- Các mô hình như Wav2vec [8] và HuBERT [37] đã khai thác học tự giám sát trên dữ liệu âm thanh không có nhãn, cho phép mô hình học các đặc trưng giọng nói phong phú mà không cần nhãn trực tiếp. Các mô hình này đã được chứng minh là có hiệu quả trong việc xác thực và nhận diện người nói khi có dữ liệu hạn chế.
- Phương pháp CSSL (Contrastive Self-Supervised Learning) trong học tự giám sát [55] sử dụng đối chiếu giữa các đoạn âm thanh khác nhau của cùng một người nói để xây dựng các đặc trưng, từ đó giảm thiểu nhu cầu về dữ liệu có nhãn trong xác thực người nói.
- SpecAugment [79] là một kỹ thuật tăng cường dữ liệu phổ biến được sử dụng trên biểu đồ phổ bằng cách biến đổi các đoạn âm thanh với nhiều mức độ khác nhau, như thay đổi tần số và thời gian. Kỹ thuật này được ứng dụng để tăng cường khả năng tổng quát của các mô hình xác thực người nói.
- Nghiên cứu về mạng Prototypical [103] cho thấy khả năng nhận dạng người nói chỉ với một số ít mẫu. Phương pháp này học các biểu diễn đặc trưng đại diện cho mỗi lớp người nói, từ đó cho phép phân loại chính xác ngay cả với số lượng mẫu huấn luyện hạn chế.
- Các mô hình Siamese [54] cũng đã được áp dụng cho các bài toán nhận dạng và xác thực người nói với số lượng mẫu hạn chế, giúp cải thiện đáng kể độ chính xác của hệ thống.
- Học chuyển giao: Các nghiên cứu sử dụng đặc trưng nhúng người nói từ các mô hình như x-vector [108] và ResNet [74] cho phép chuyển giao các 13 đặc trưng đã được học từ bài toán khác hoặc dữ liệu tổng quát hơn vào bài toán xác thực người nói với dữ liệu hạn chế. Các mô hình Res2Net [94] và ECAPA-TDNN [22] đã đạt được thành công trong nhận dạng người nói và xác thực người nói bằng cách tận dụng các lớp khung xương có khả năng học các đặc trưng chuyên sâu và quy mô từ các mẫu dữ liệu hạn chế.
- Học đặc trưng thủ công và học kết hợp: các đặc trưng thủ công truyền thống như MFCCs và biểu đồ phổ được kết hợp với các đặc trưng học từ mô hình học sâu giúp tận dụng cả hai loại đặc trưng và cải thiện hiệu suất mô hình khi dữ liệu hạn chế [12]. Với tập dữ liệu người nói tiếng Việt được công bố hiện nay như VLSP2021- SV [18], Vietnam-Celeb [83], NCS tập trung vào các phương pháp lựa chọn đặc trưng thủ công kết hợp với mạng học sâu và phương pháp học chuyển giao nhằm nâng cao độ chính xác xác thực người nói.

1.4.Các cách tiếp cận học sâu trong bài toán xác thực người nói

Có hai cách tiếp cận trong xác thực người nói: tiếp cận dựa trên thống kê và tiếp cận dựa trên học sâu. Trong Luận án này NCS tập trung vào các tiếp cận dựa trên học sâu.

Mạng nơ-ron sâu rất thành công trong trích chọn đặc trưng để học các đặc trưng nhúng phân biệt trong cả thị giác máy tính và tiếng nói. Các phương pháp thường kết hợp các bộ phân lớp và huấn luyện độc lập. Trong khi các phương pháp ghép nối có hiệu quả cao, khi DNN không huấn luyện từ đầu đến cuối và vẫn cần các kỹ thuật trích chọn đặc trưng. Ngược lại, kiến trúc CNN có thể dùng trực tiếp từ ảnh phổ thô và huấn luyện đầu cuối. Hệ thống học sâu từ mô hình đầu vào cho đến đầu ra cho nhận dạng người nói thường sử dụng ba giai đoạn:

- Trích chọn đặc trưng sử dụng DNN
- Tổng hợp đặc trưng mức khung
- Tối ưu hóa hàm mất mát cho mục tiêu phân lớp.

Kiến trúc thân DNN thường dùng 2D CNN với tích chập cho cả miền thời gian và miền tần số [44] hoặc 1D CNN với tích chập áp dụng cho miền thời gian [31]. Một số nghiên cứu cũng sử dụng kiến trúc đầu cuối dựa trên LSTM [98]. Đầu ra bộ trích chọn đặc trưng phụ thuộc độ dài phát âm đầu vào. Lớp tổng hợp dùng và tổng hợp véc tơ đặc trưng mức khung thu được đặc trưng nhúng độ dài cố định hướng dẫn sự mở rộng phương pháp trong độ lệch chuẩn như trung bình. Phương pháp này gọi là tổng hợp thống kê. Không giống như các phương pháp mà thông tin từ tất cả các khung với trọng số như nhau đã phát triển mô hình chú ý phân trọng số cho các khung phân biệt. Ở đây kết hợp các mô hình chú ý và mô hình thống kê cho tổng hợp thống kê chú ý. Giai đoạn tổng hợp cuối cùng này được quan tâm là LDE. Phương pháp này gần với lớp NetVLAD [5] thiết kế cho truy vấn ảnh.

Các hệ thống như vậy được huấn luyện đầu cuối cho phân lớp dùng hàm softmax hoặc một trong các tùy biến như Angular softmax [39]. Trong một số trường hợp, mạng được huấn luyện cho xác thực sử dụng hàm mất mát Contrastive[116] hoặc hàm mất mát triplet [23]. Các độ đo tương tự như cosine [30] hay PLDA [42] thường dùng để sinh ra điểm số các cặp so sánh sau cùng.

1.4.Sơ đồ tổng quan hệ thống xác thực người nói

Hệ thống gồm có các thành phần chính (Hình 1.1): trích chọn đặc trưng, mô hình hóa người nói và đánh giá. Trích chọn đặc trưng biến đổi tín hiệu âm thanh thành tập các đặc trưng phân biệt giữa từng người nói riêng biệt, hay còn gọi là đặc trưng nhúng người nói. Trong giai đoạn đăng ký (enrollment), mô hình người nói dùng đặc trưng đầu vào để xây dựng mô hình thống kê, mô hình này biểu diễn những đặc điểm duy nhất của mỗi người nói cụ thể. Mô hình này thường gọi là mô hình người nói hoặc mô hình giọng nói dùng để suy luận trong quá trình xác thực xác định mẫu giọng nói đã cho có thuộc người nói đã đăng kí hay không. Quyết định xác thực dựa trên mô đun đánh giá (scoring), đánh giá đặc trưng của người nói mới với đặc trưng giọng nói đã đăng ký. Nếu điểm đánh giá lớn hơn hoặc bằng ngưỡng τ đã định nghĩa trước, khi đó quá trình xác thực thành công và xác thực người dùng.

Ngược lại, quá trình sẽ không thành công tức là mẫu giọng nói đã cho không thuộc về giọng nói đã đăng ký.

Các mô đun được đề cập ở trên là những mô đun cơ bản của hệ thống xác thực người nói và ảnh hưởng trực tiếp đến hiệu quả nói chung của hệ thống xác thực người nói. Sơ đồ cơ bản trong Hình 4 có thể được áp dụng cho các phương pháp dựa trên truyền thống và dùng cho cả học sâu.

Trong mục này NCS sẽ phân tích các mô đun trích trọn đặc trưng, mô hình hóa người nói và đánh giá trong ba mô hình học sâu hiện đại cho bài toán xác thực người nói: VGGVox, ECAPA-TDNN và RawNet.

1.4.1. Trích chọn đặc trưng

Học sâu đã được chứng minh là một kỹ thuật mạnh để trích chọn các đặc trưng mức cao từ các thông tin ở mức thấp. Các đăc trưng trích chọn từ các lớp ẩn của các mô hình học sâu khác nhau được gọi là các đặc trưng sâu. Các đặc trưng sâu có thể được trích chọn từ bất kỳ mô hình học sâu nào như mạng nơ-ron tích chập (CNN), mạng nơ-ron sâu (DNN), mạng nơ-ron hồi quy (RNN), mạng bộ nhớ ngắn hạn một chiều (LSTM)), bộ nhớ ngắn hạn hai chiều (BLSTM) và các mô hình tương tự khác.

Các đặc trưng sâu được trích xuất từ mạng nơ-ron sâu (DNNs). Các MFCC hoặc đặc trưng âm thanh liên quan nào khác cung cấp làm đầu vào cho DNN. Các đặc trưng sâu phụ thuộc vào độ sâu của mạng nơ-ron. Nếu chúng ta có mạng nơ-ron nông, các đặc trưng sâu được cung cấp bởi các lớp thấp hơn có thể được coi là các đặc trưng thích nghi người nói. Và từ các lớp trên, các đặc điểm phân biệt dựa trên lớp có thể được trích xuất. Các đặc trưng sâu cũng có thể được trích chọn từ lớp bottleneck của DNN.

1.4.2. Mô hình hóa người nói

Trong luận án này, NCS sẽ trình bày một số mạng học sâu phổ biến nhất xác thực người nói như i-vector, d-vector, x-vector, resnets, ECAPA-TDNN, SincNet, RawNets, ...

Cấu trúc cơ bản của một mạng học sâu thường bao gồm các thành phần sau:

Input Layer (Lớp đầu vào): Lớp này nhận dữ liệu đầu vào và chuyển tiếp nó vào mạng. Số lượng neuron trong lớp này phụ thuộc vào kích thước của dữ liệu đầu vào.

- Hidden Layers (Các lớp ẩn): Các lớp này chứa các neuron và thực hiện các phép biến đổi và tính toán trên dữ liệu đầu vào. Mỗi lớp ẩn có thể có nhiều neuron và thường được xác định bởi số lượng và loại các neuron, cũng như cách chúng kết nối với nhau.
- Weights and Biases (Trọng số và sai số): Mỗi kết nối giữa các neuron trong các lớp liên tiếp có một trọng số, thể hiện sức mạnh của kết nối đó. Ngoài ra, mỗi neuron có một sai số (bias) để điều chỉnh và thích ứng với dữ liệu đầu vào.
- Activation Functions (Hàm kích hoạt): Mỗi neuron trong mạng thường áp dụng một hàm kích hoạt để đưa ra đầu ra phi tuyến tính. Các hàm kích hoạt phổ biến bao gồm ReLU (Rectified Linear Unit), Sigmoid, Tanh, và Leaky ReLU.
- Output Layer (Lớp đầu ra): Lớp này tạo ra đầu ra dự đoán của mạng. Số lượng neuron trong lớp này phụ thuộc vào loại bài toán, ví dụ, một neuron cho mỗi lớp trong bài toán phân loại nhị phân, hoặc một neuron cho mỗi lớp trong bài toán phân loại nhiều lớp.
- Loss Function (Hàm mất mát): Hàm này tính toán sự mất mát giữa dự đoán của mạng và giá trị thực tế. Nó đo lường hiệu suất của mô hình và được sử dụng trong quá trình huấn luyện để điều chỉnh các tham số của mạng.
- Optimizer (Bộ tối ưu hóa): Bộ tối ưu hóa được sử dụng để cập nhật trọng số và sai số của mạng dựa trên giá trị mất mát tính toán được từ dữ liệu huấn luyện. Các phương pháp tối ưu phổ biến bao gồm Gradient Descent và các biến thể của nó.

1.4.3. Đánh giá

Trong Luận án NCS sử dụng khoảng cách Cosine để so sánh độ tương đồng giữa hai véc-tơ.

Khoảng cách cosine tính theo độ tương tự cosine. Độ tương tự cosine được định nghĩa là độ tương đồng giữa hai véc tơ khác không. Nó tính cosine của góc giữa hai véc tơ trong không gian đa chiều. Mối quan hệ giữa độ tương tự cosine với khoảng cách cosine là không cân xứng. Độ tương tự cosine tăng dần trong khi khoảng cách giữa các véc tơ giảm dần và ngược lại. Phương trình sau đây tính độ tương tự cosine và khoảng cách cosine tương ứng. Các hàm được biểu diễn bằng công thức sau:

$$\cos(\theta) = \frac{A.B}{\|A\|.\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1.11)$$

$$cosine_{distance} = 1 - \cos(\theta) \quad (1.12)$$

Trong đó A, B là các véc tơ khác không và $\cos(\theta)$là độ tương tự cosine.

1.5. Các tập dữ liệu thử nghiệm cho bài toán xác thực người nói

VoxCeleb1

Tập dữ liệu VoxCeleb1 [72] là một tập dữ liệu lớn chứa các mẫu giọng nói từ những người nổi tiếng, được thu thập từ các video trên YouTube. VoxCeleb1 chứa hơn 100,000 câu nói của 1,251 người nói. Cơ sở dữ liệu này được công bố năm 2017, nó là tập dữ liệu lớn cho bải toán nhận dạng người nói.

VoxCeleb2

Cơ sở dữ liệu VoxCeleb2 [14] là một tập dữ liệu mở lớn chứa các mẫu giọng nói từ nhiều người nổi tiếng, được thu thập từ các video trên YouTube. VoxCeleb2 là phiên bản mở rộng của VoxCeleb1 và được công bố sau đó với các cải tiến và mở rộng đáng kể. VoxCeleb2 chứa hơn 1 triệu câu nói của hơn 6.000 người nổi tiếng, được trích từ các video tải lên YouTube. Tập dữ liệu khá cân bằng về giới tính, với 61% người nói là nam giới.

VLSP2021-SV

Gần đây, hội thảo VLSP 2021 [18] đã công bố bộ dữ liệu xác thực và nhận dạng người nói tiếng Việt trong môi trường có nhiễu chứa 50 giờ nói của hơn 1.300 người nói (NCS gọi bộ dữ liệu này là VLSP2021-SV). Dữ liệu thu thập từ nhiều nguồn khác nhau, bao gồm từ cuộc thi ZaloAI, VLSP2020-SV, VIVOS và thu thập dữ liệu từ các chương trình truyền hình và kênh YouTube trong môi trường có nhiễu nền đa dạng như tiếng trò chuyện nhỏ, tiếng cười, tiếng ồn đường phố, trường học và âm nhạc.

Vietnam-Celeb

Bộ dữ liệu Vietnam-Celeb [83] bao gồm 1,000 người nói và hơn 87,000 câu nói. Tổng thời lượng của tập dữ liệu là 187 giờ, các câu nói được lấy mẫu tại 16.000 Hz. Dữ liệu bao gồm tất cả các tình huống như phỏng vấn, trò chơi truyền hình, chương trình trò truyện và các loại video giải trí khác.

Bảng 1.15: Thống kê các tập con của Vietnam-Celeb.

Tập con	Sô người nói	Số câu nói	Số cặp
Vietnam-Celeb-T	880	82,907	-
Vietnam-Celeb-E	120	4,207	55,015
Vietnam-Celeb-H	120	4,217	55,015

Các phương pháp nâng cao độ chính xác của hệ thống xác thực người nói

Tăng cường dữ liệu

Huấn luyện trong nhiều điều kiện quy mô lớn là một cách hiệu quả tăng khả năng xác thực người nói trong môi trường nhiễu. Đặc biệt hiệu năng của hệ thống xác thực người nói dựa trên học sâu phụ thuộc nhiều vào lượng dữ liệu huấn luyện. Một phương pháp để chuẩn bị lượng dữ liệu nhiễu lớn chính là tăng cường dữ liệu. Trong [108], các tác giả đã sử dụng tiếng ồn cộng thêm và độ vang trên dữ liệu huấn luyện gốc cho tăng cường dữ liệu x-vector rất hiệu quả. Trong [136] đã áp dụng chiến lược học kết hợp để cải tiến bộ trích chọn x-vector.

Lựa chọn đặc trưng

Trong xác thực người nói, lựa chọn đặc trưng thủ công làm đầu vào cho mạng nơ-ron sâu là một phương pháp phổ biến để kết hợp các đặc trưng âm thanh truyền thống với khả năng học sâu của mạng nơ-ron. Điều này giúp tận dụng cả thông tin âm thanh sẵn có từ đặc trưng thủ công và khả năng phân tích mẫu phức tạp của mạng nơ-ron sâu. Một số đặc trưng thủ công phổ biến làm đầu vào bao gồm: MFCCs, ảnh phổ, âm thanh thô, FBank. Lựa chọn đặc trưng thủ công làm đầu vào cho mạng nơ-ron sâu không chỉ giúp cải thiện hiệu suất của mô hình mà còn làm tăng khả năng tận dụng các đặc trưng quan trọng từ tín hiệu giọng nói, đặc biệt trong các trường hợp có dữ liệu hạn chế. Đặc trưng thủ công có thể được kết hợp với các đặc trưng học từ các tầng của mạng nơ-ron sâu, như CNN [14], [16], để tạo ra các đặc trưng lai, qua đó tối ưu hóa độ chính xác trong xác thực người nói. Bảng 1.1 cho thấy đặc trưng đầu vào Mel-filter Bank cho kết quả tốt nhất trên dữ liệu VoxCeleb1 và VoxCeleb2 với EER 0.66%. Nếu chỉ xét riêng đặc trưng từ sóng âm thanh thì mô hình RawNet3 cho kết quả tốt nhất với tỉ lệ lỗi EER 0.89%.

Độ đo đánh giá hệ thống xác thực người nói

EER

EER là điểm mà tỉ lệ chấp nhận sai (FAR) bằng với tỉ lệ từ chối sai (FRR) trong hệ thống xác thực. Ý nghĩa của chỉ số này được giải thích như sau : FRR càng cao thì hệ thống càng an toàn. Tuy nhiên lại xảy ra nhiều xác nhận của người dùng hợp pháp bị từ chối. Như vậy người dùng phải thực hiện nhiều lần xác thực để có thông điệp thành công dẫn đến giảm cảm giác trải nghiệm của người dùng. Do đó, độ nhạy và sự tiện lợi của hệ thống còn kém. Ngược lại, nếu tỉ lệ từ chối sai (FRR) quá nhỏ thì FAR thường rất cao. Kết quả dẫn tới hệ thống chấp nhận nhiều xác thực người dùng không hợp lệ hay người dùng dễ dàng xác thực thành công. Điều này ảnh hưởng đến an ninh hệ thống. Điểm mà FAR = FRR được gọi là tỉ lệ lỗi bằng nhau (EER). Tại thời điểm này, hệ thống cân bằng giữa độ an toàn và độ nhạy cảm, tiện lợi. Vì vậy, EER là thường được sử dụng làm độ đo cho các hệ thống xác thực. EER càng nhỏ thì chất lượng xác thực người nói của hệ thống càng tốt.

Kết luận chương 1

Trong Chương 1 NCS đã trình bày những kiến thức tổng quan về hệ thống nhận dạng và xác thực người nói dựa trên mô hình học sâu. Cũng như cách tiếp cận truyền thống, hệ thống xác thực bao gồm ba mô-đun chính: trích chọn đặc trưng, mô hình hóa đặc trưng, đánh giá. Qua chương này NCS cũng khảo sát được các tập dữ liệu công bố trên thế giới cho bài toán xác thực, các cách tiếp cận mới nhất hiện nay cũng như những thách thức cho bài toán này. Cụ thể:

- Các tập dữ liệu dùng cho huấn luyện và đánh giá mô hình xác thực: VoxCeleb1, VoxCeleb2, Cn-Celeb2, VLSP2021-SV, Vietnam-Celeb.
- Các mô hình học sâu hiện đại với sự đa dạng đặc trưng đầu vào áp dụng cho bài toán xác thực người nói: ResNets, x-vector, ECAPA-TDNN, RawNets.
- Với các mô hình học sâu, dữ liệu huấn luyện đóng vai trò quan trọng đối với độ chính xác của mô hình.

Như vậy cần phải giải quyết vấn đề hạn chế về dữ liệu huấn luyện theo hai cách: Ứng dụng học chuyển giao và lựa chọn đặc trưng âm học phù hợp cho các mô hình xác thực người nói tiên tiến nhất nhằm nâng cao hiệu quả xác thực người nói tiếng Việt. Những nghiên cứu tổng quan trong chương này làm cơ sở cho NCS đề xuất các giải pháp nâng cao hệ thống xác thực người nói với tài nguyên hạn chế trong các chương tiếp theo.

Một số kết quả nghiên cứu ban đầu về hệ thống xác thực tiếng nói cơ sở được công bố trong công trình **[CT4]** trong phần "Danh mục các công trình của tác giả"

CHƯƠNG 2: NÂNG CAO ĐỘ CHÍNH XÁC XÁC THỰC NGƯỜI NÓI TIẾNG VIỆT SỬ DỤNG ĐẶC TRƯNG MEL-FILTERBANK ENERGYS VỚI MÔ HÌNH ECAPA-TDNN

Trong Chương 2, NCS tập trung vào việc lựa chọn đặc trưng âm học làm đầu vào mạng học sâu hiện đại trong hệ thống xác thực người nói với dữ liệu tiếng Việt. Qua chương này, NCS trình bày quá trích chọn đặc trưng MFBEs và MFCCs đồng thời phân tích những hạn chế của MFCCs trong xác thực người nói. Từ đó có những so sánh sự khác biệt giữa hai đặc trưng MFBEs và MFCCs và thực nghiệm đánh giá so sánh hai đặc trưng này trong hệ thống xác thực người tiếng Việt.

2.1. Bài toán xác thực người nói tiếng Việt và các đặc trưng tiếng nói

Xác thực người nói là một lĩnh vực quan trọng trong công nghệ nhận dạng giọng nói, đặc biệt trong bối cảnh ngày càng gia tăng nhu cầu về an ninh và bảo mật thông tin cá nhân. Tại Việt Nam, bài toán xác thực người nói tiếng Việt đang thu hút sự quan tâm lớn từ cả giới nghiên cứu và các nhà phát triển công nghệ, do những đặc điểm ngôn ngữ và văn hóa độc đáo của tiếng Việt. Tiếng Việt có hệ thống thanh điệu phong phú, với sáu thanh điệu khác nhau, cùng với sự đa dạng về ngữ âm giữa các vùng miền Bắc, Trung, và Nam. Sự đa dạng này tạo ra những thách thức lớn cho các hệ thống xác thực, yêu cầu phải phát triển các phương pháp có khả năng nhận diện và phân biệt chính xác giữa các giọng nói khác nhau, kể cả trong các điều kiện môi trường khác nhau. Xác thực người nói tiếng Việt hiện đang gặp nhiều thách thức như còn thiếu dữ liệu huấn luyện, kết quả đánh giá

trên tập dữ liệu mới nhất hiện nay Vietnam-Celeb [83] có tỉ lệ lỗi EER lớn hơn 10% nên vẫn cần có những nghiên cứu thực nghiệm nhằm nâng cao chất lượng hệ thống xác thực. Với các cách tiếp cận hiện đại dựa trên học sâu thì việc lựa chọn đặc trưng âm học và lựa chọn mô hình huấn luyện là một trong những giải pháp nâng cao độ chính xác xác thực người nói trên dữ liệu người nói tiếng Việt.

Những hạn chế của MFCCs

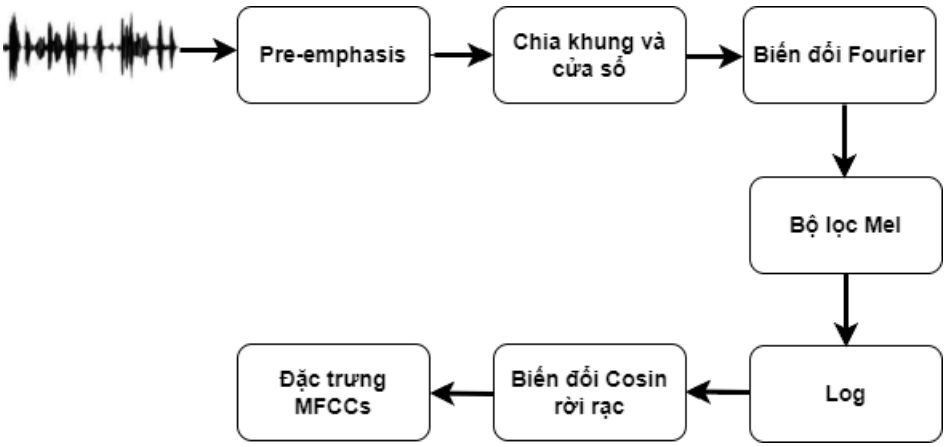
Các đặc trưng cho nhận dạng giọng nói tự động đã được đánh giá dựa trên phương pháp mờ bao gồm MFCC, DTW và FFT. Kết quả cho thấy MFCC cải thiện hiệu suất của mô hình Fuzzy so với đặc trưng FFT [102]. Al-Ali và cộng sự đã cải tiến việc xác thực giọng nói pháp y dựa trên sự kết hợp của các đặc trưng MFCC và DWT, trong đó mô hình của họ được đánh giá trong môi trường nhiễu [2]. Mặc dù MFCC có khả năng nắm bắt các đặc điểm của người nói, hiệu suất của MFCC suy giảm trên các tập dữ liệu giọng nói phức tạp và trong môi trường nhiễu. Ví dụ, nghiên cứu [4] cho thấy rằng việc nhận dạng người nói sử dụng MFCC và k-NN giảm đáng kể trong môi trường nhiễu và kết luận rằng làm sạch tín hiệu đầu vào có thể cải thiện kết quả hơn khi sử dụng các MFCC cao nhất. Để khắc phục vấn đề này, trong [10], một khung huấn luyện đa kênh trong mạng nhúng người nói sâu đã được đề xuất cho xác thực người nói trong môi trường vang dội và nhiễu. Phương pháp này nhận thông tin về thời gian, tần số và không gian từ đầu vào đa kênh để cải thiện quá trình nhúng người nói mạnh mẽ hơn. Công trình kết luận rằng việc tăng nhẹ các tham số của mô hình có thể giúp phương pháp này vượt trội đáng kể so với hệ thống i-vector với MFCC có sử dụng tăng cường tín hiệu ở đầu vào. Ngoài ra, Jahangir và cộng sự đã đề xuất các đặc trưng kết hợp dựa trên MFCC và các đặc trưng dựa trên thời gian. Các đặc trưng kết hợp này được đưa vào DNN để nhận dạng người nói. Kết quả cho thấy rằng hạn chế của đặc trưng MFCC có thể được giải quyết bằng phương pháp này [43].

So sánh đặc trưng MFCCs và MFBEs

Sự khác nhau chính giữa đặc trưng MFBEs và MFCCs ở chỗ sử dụng phép biến đổi cosin rời rạc DCT [109]. Các đặc trưng MFBEs có thể đồng nhất hoặc không đồng nhất DCT tùy thuộc vào cài đặt cụ thể, trong khi MFCCs luôn có liên quan đến DCT để nén thông tin thành tập hệ số nhỏ hơn. Cả đặc trưng MFCCs và MFBEs đều có ảnh hưởng đến biểu diễu tín hiệu âm thanh trong các ứng dụng xử lý tiếng nói và xử lý âm thanh. MFCCs cung cấp thông tin chuỗi thời gian của năng lượng theo tần số từ nguồn ân thanh. Việc hiệu chỉnh từ chuỗi năng lượng dựa trên DFT thô phục vụ cho hai mục đích:

- Thay đổi thang tuyến tính (của tần số và năng lượng) từ DFT thô thành thang logarit (log scale). Điều này phù hợp với thính giác của con người (và hầu hết các động vật) trong việc cảm nhận âm thanh.
- Việc nén lượng lớn dữ liệu thành các đặc trưng nhỏ hơn mà vẫn đảm bảo phân biệt sự khác nhau giữa các âm thanh. Điều này đặc biệt có ích ở miền có tần số cao cho hầu hết các ứng dụng nhận dạng tiếng nói, phát hiện sự khác nhau giữa các mức năng lượng ở 1001 Hz và 999 Hz.

Ưu điểm việc dùng biến đổi cosin rời rạc so với biến đổi Fourie rời rạc là loại bỏ bớt nhiễu trong tín hiệu tiếng nói.



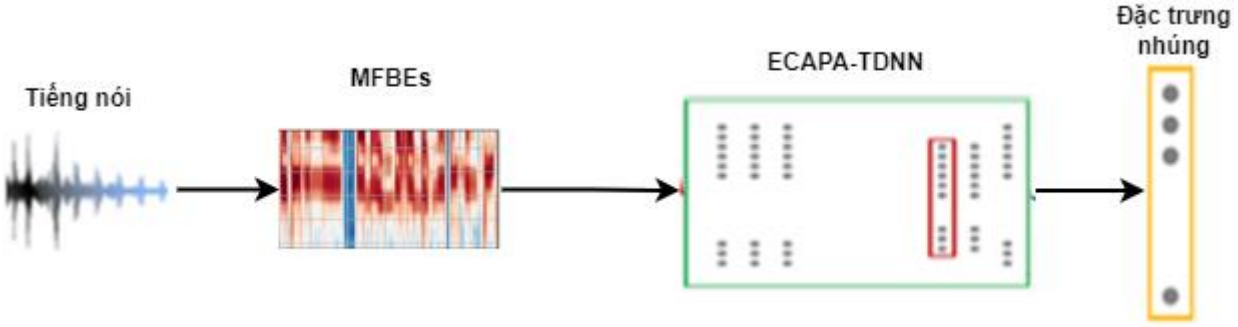
Hình 2.1: Các bước trích chọn đặc trưng MFCCs.

Sự phân tích biểu diễn đặc trưng đầu vào khác nhau nhằm mục tiêu lựa chọn đầu vào phù hợp cho mô hình âm học nơ-ron sâu. MFCCs kém hơn biến đổi DCT, MFCCs làm các mô hình nơ-ron sâu loại bớt các thông tin về người nói.

2.2. Đề xuất sử dụng đặc trưng MFBEs với mô hình ECAPA-TDNN trong xác thực người nói tiếng Việt

Giai đoạn huấn luyện

Khác với mô hình VGGVox [74], ECAPA-TDNN và x-vector không dùng ảnh phổ hai chiều (2D) làm đầu vào mà dùng mảng MFCCs một chiều (1D) với độ dài khung là 25 ms (x-vector sử dụng MFCC có 24 chiều, chuẩn hóa trung bình (mean-normalized) trên cửa sổ 3 giây, bước nhảy 10 ms; ECAPA-TDNN dùng MFCC có 80 chiều, chuẩn hóa trung bình trên cửa sổ 2 giây, bước nhảy 10 ms). Mô hình x-vector dựa trên mạng nơ-ron sâu với các lớp TDNN được trích chọn đặc trưng ở mức khung, sau đó tổng hợp thông tin thành biểu diễn có số chiều cố định. Sau đó, lớp kết nối đầy đủ sẽ sinh ra mã hóa đặc trưng người nói cuối cùng. ECAPA-TDNN tập trung vào trích chọn đặc trưng mức khung và cải tiến ở mức tổng hợp đặc trưng trên mô hình gốc x-vector và các biến thể. Trong mô hình gốc x-vector, lớp ngữ cảnh tạm thời bị giới hạn 15 khung. Để cải thiện trích chọn đặc trưng hiệu quả và giảm số lượng tham số của mô hình, ECAPA-TDNN tích hợp mô đun Res2Net [94] (giảm số chiều) với khối SE [38] trở thành SE-Res2Block trong quá trình trích chọn đặc trưng mức khung. Các đặc trưng đầu vào của mô hình ECAPA-TDNN là MFBEs 80 chiều từ cửa sổ 25 ms với độ dịch chuyển khung 10 ms. Các vector đặc trưng MFBEs dài hai giây được chuẩn hóa thông qua việc trừ trung bình cepstral. Để tăng lượng dữ liệu huấn luyện, NCS có sử kỹ thuật tăng cường dữ liệu trong quá trình huấn luyện như thêm nhiễu môi trường.



Hình 2.5: Đề xuất sử dụng đặc trưng MFBEs với mô hình ECAPA-TDNN trong xác thực người nói tiếng Việt.

Giai đoạn xác thực

Trong giai đoạn xác thực sẽ gồm các bước:

- Trích xuất đặc trưng từ âm thanh đầu vào,
- So sánh đặc trưng nhúng của âm thanh cần xác thực với âm thanh đã đăng ký,
- Dựa vào độ tương đồng giữa hai đặc trưng nhúng để xác định liệu hai đoạn âm thanh có cùng đến từ một người hay không.

Các điểm số đánh giá được tính toán dựa trên khoảng cách Cosine [30] giữa các đặc trưng nhúng của người nói (xem Hình 2.5). Sau đó, tất cả các điểm số này được chuẩn hóa thông qua phương pháp adaptive s-norm [17], [51].

2.4. Thực nghiệm

Bộ dữ liệu

VLSP2021-SV

Trong chương này, NCS sử dụng bộ dữ liệu VLSP2021-SV của tác giả Vi Thanh Dat và cộng sự [18]. Bộ dữ liệu gồm các đoạn tiếng nói ngắn được trích chọn từ các cuộc phỏng vấn trên YouTube. Trong các đoạn video này, các giọng nói đều là giọng nói tự nhiên. Bộ dữ liệu dùng huấn luyện có hơn 31,000 câu nói của 1,305 người nói.

Vietnam-Celeb

Trong thực nghiệm, NCS còn sử dụng bộ dữ liệu Vietnam-Celeb [83]. Bộ dữ liệu này có hơn 1,000 người nói với hơn 87,000 câu nói. Tổng số giờ thu âm khoảng 187 giờ, các câu nói thu âm lấy mẫu ở tần số 16,000 Hz.

Môi trường thực nghiệm

Trong phần thực nghiệm, NCS đánh giá sự kết hợp của hai phương pháp trích chọn đặc trưng với hai mô hình học sâu: Sự kết hợp cụ thể như sau:

- Mô hình ECAPA-TDNN với MFBEs và MFCC,
- Mô hình ResnetSE-34 với MFBEs và MFCC.

Đặc trưng: NCS sử dụng MFCC (80 chiều) và MFBEs (80 chiều) làm đầu vào cho mô hình ECAPA-TDNN (hoặc ResnetSE-34); các đặc trưng trích chọn sử dụng cửa sổ Hamming độ rộng 25 ms, bước nhảy 10 ms từ âm thanh. Các đặc trưng trong dữ liệu huấn luyện chia thành độ dài 2 s, sau đó chuẩn hóa. Trong thực nghiệm, NCS có sử dụng kỹ thuật tăng cường dữ liệu.

Kiến trúc mô hình: trong mô hình ECAPA-TDNN, các lớp tích chập có số kênh là 1024. Các nút ở lớp liên kết cuối cùng có số chiều là 192, tổng số người nói huấn luyện là 1,305. Trong mô hình ResnetSE-34, các nút ở lớp cuối cùng thiết lập là 512 chiều, tổng số người nói được huấn luyện là 1,305.

NCS thực nghiệm dựa trên nền tảng PyTorch. Các mô hình huấn luyện trên máy chủ NVIDIA A100 GPU với 80GB bộ nhớ và sử dụng tối ưu Adam. NCS sử dụng tốc độ học ban đầu là 0.001 và

giảm dần 10% sau 2 epoch. Mô hình huấn luyện 200 epoch với kích thước mini-batch là 100. Quá trình huấn luyện mô hình mất khoảng 6 giờ.

Kết quả thực nghiệm và phân tích

So sánh giữa hai mô hình ResnetSE-34 và ECAPA-TDNN: Bảng 2.8 cho thấy tỉ lệ lỗi của các hệ thống khác nhau. Mô hình ECAPA-TDNN có tỉ lệ lỗi nhỏ nhất đạt 11.37 % (Vietnam-Celeb-E) và 12.74% (Vietnam-Celeb-H). Điều đó có nghĩa mô hình ECAPA-TDNN có hiệu năng tốt hơn mô hình ResnetSE-34 với cùng điều kiện thực nghiệm và huấn luyện.

Bảng 2.8: Kết quả thực nghiệm của các đặc trưng khác nhau, đánh giá trên tập dữ liệu Vietnam-Celeb-E và VietnamCeleb-H.

Mô hình	Dữ liệu đầu vào	Dữ liệu huấn luyện	Kết quả dánh giá (%EER)	
			Vietnam-Celeb-E	Vietnam-Celeb-H
ECAPA-TDNN [83]	80 MFCCs	VLSP2021-SV	11.58	14.3
ECAPA-TDNN	80 MFBEs	VLSP2021-SV	**11.37**	**12.74**
ResNetSE-34 [16]	80 MFCC	VLSP2021-SV	12.98	14.31
ResNetSE-34	80 MFBEs	VLSP2021-SV	**11.84**	**13.16**

So sánh giữa hai đặc trưng MFCC và MFBEs: NCS phân tích hiệu năng của các đặc trưng khác nhau và các kết quả liệt kê trong Bảng 21. Trong Bảng 21, đặc trưng MFBEs cho kết quả tốt hơn đặc trưng MFCC. Kết quả chỉ rõ trong Bảng 21 đánh giá trên tập dữ liệu Vietnam-Celeb-E và Vietnam-Celeb-H. Các hệ thống ở điều kiện đánh giá học đặc trưng nhúng người nói từ năng lượng MFBEs trong thời gian ngắn và MFCC. Dựa trên thực nghiệm ở Bảng 21, có thể thấy rằng đặc trưng nhúng người nói mô hình hóa từ MFCC cho kết quả thấp hơn so với MFBEs với cùng điều kiện tương tự. Đặc biệt, trong mô hình ECAPA-TDNN (80 MFCCs) thì tỉ lệ lỗi lần lượt là 11.58% và 14.3% trong khi mô hình ECAPA-TDNN (80 MFBEs) là 11.37% và 12.74%. Mô hình DNN học từ các thông tin người nói từ dữ liệu đầu vào năng lượng MFBEs hiệu quả hơn MFCCs là do áp dụng phép biến đổi cosin rời rạc lên đặc trưng MFBEs để tạo ra đặc trưng MFCCs. Hơn nữa, với những thách thức trong nhận dạng người nói, nhiều nhà nghiên cứu cũng chuyển sự chú ý nhiều hơn sang MFBEs thay vì MFCCs. Nguyên nhân chính là đặc trưng MFCCs bị mất một vài thông tin trong quá trình biến đổi MFBEs sang MFCCs.

2.7. Kết luận chương 2

Trong Chương 2, NCS đã có những nghiên cứu, phân tích, thực nghiệm so sánh việc sử dụng các đặc trưng MFCCs và MFBEs làm đầu vào cho hai mô hình ECAPA-TDNN và ResNetSE-34.

Đóng góp chính của Chương 2 bao gồm:

- Lý thuyết: Phân tích, đánh giá, so sánh sự khác nhau của hai phương pháp trích chọn đặc trưng MFCCs và MFBEs sử dụng mạng học sâu trong bài toán xác thực người nói tiếng Việt;
- Thực nghiệm: Kết quả cho thấy đặc trưng MFBEs với mô hình ECAPA-TDNN cho kết quả tốt đặc trưng MFCCs trên hai tập dữ liệu Vietnam-Celeb-E và Vietnam-Celeb-H. Kết quả nghiên cứu này vô cùng có ích cho các nghiên cứu tiếp theo trong xác thực người nóivà

Vietnam-Celeb-H. Kết quả nghiên cứu này vô cùng có ích cho các nghiên cứu tiếp theo trong nhận dạng người nói.

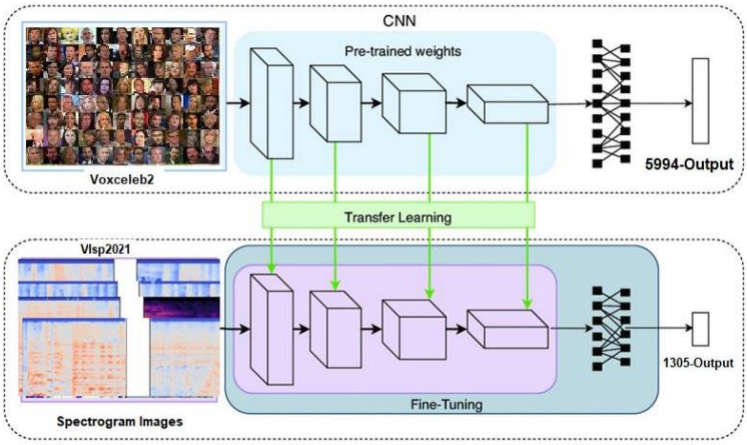
Kết quả của đề xuất dùng đặc trưng MFBEs làm đầu vào cho mô hình ECAPA-TDNN đã được công bố tại công trình **[CT1], [CT2]** trong phần "Danh mục các công trình của tác giả".

CHƯƠNG 3: NÂNG CAO ĐỘ CHÍNH XÁC XÁC THỰC NGƯỜI NÓI TIẾNG VIỆT SỬ DỤNG HỌC CHUYỂN GIAO VỚI MÔ HÌNH RAWNET3

3.1. Giới thiệu về học chuyển giao trong bài toán xác thực người nói

Học chuyển giao được chứng minh là các tiếp cận hiệu quả trong việc cải tiến hiệu năng nói chung cũng như giảm chi phí huấn luyện các mô hình học sâu. Hiệu quả học chuyển giao bị ảnh hưởng bởi một số các yếu tố như lượng dữ liệu có sẵn của bài toán nguồn và bài toán đích, lựa chọn mô hình.

Học chuyển giao đã trở thành kỹ thuật phổ biến và hiệu quả trong cải thiện khả năng tổng quát hóa mô hình và là giảm chi phí quá trình huấn luyện mô hình. Trong học sâu, học chuyển giao thường liên quan đến việc sử dụng mạng nơ-ron đã huấn luyện trước, mô hình này được huấn luyện trên bộ dữ liệu lớn của một bài toán cụ thể và làm cơ sở để huấn luyện mô hình mới cho bài toán liên quan. Quá trình chuyển giao thường được thực hiện bằng cách sử dụng các trọng số đã học từ mô hình huấn luyện trước làm giá trị khởi tạo sau đó tinh chỉnh các trọng số này trên dữ liệu mới. Học chuyển giao làm giảm thời gian huấn luyện và nâng cao khả năng tổng quát hóa của các mô hình học sâu. Một ví dụ điển hình là mô hình phổ biến CNN sâu trong nhận dạng ảnh như VGG [101], ResNet [33], Inception [110]. Các mô hình này huấn luyện trên tập dữ liệu ImageNet, tập dữ liệu nhận dạng hình ảnh lớn gồm hàng chục triệu hình ảnh được có nhãn và sắp xếp theo hệ thống phân cấp ngữ nghĩa của WordNet [20]. Các mô hình huấn luyện trước này không những sử dụng trong bài toán phát hiện đối tượng [28], phân đoạn ngữ nghĩa, phân loại ảnh y tế mà còn sử dụng 62 cho cả nhận dạng người nói [74], [122]. Nhưng mô-đun đó và các biến thể được huấn luyện và cài đặt trong các thư viện phổ biến như TensorFlow và PyTorch [81].



Hình 3.1: Học chuyển giao từ mô hình huấn luyện có trước trên tập dữ liệu VoxCeleb2, tinh chỉnh mô hình trên tập dữ liệu VLSP2021-SV.

Sự đa dạng của nhiều mô hình huấn luyện trước xuất phát từ thực tế. Thứ nhất, bất kỳ mô hình học sâu nào cũng có điểm mạnh và điểm yếu và vì không ai biết trước về sự phù hợp của một mô hình nhất định cho một bài toán cụ thể. Ví dụ, VGG nổi tiếng vì cấu trúc đơn giản và hiệu năng cao nhưng cũng có nhiều tham số gây tốn kém chi phí huấn luyện và triển khai mô hình. ResNet và Inception, cùng với các biến thể của chúng, là những kiến trúc quan trọng giúp giải quyết các thách thức khác nhau trong lĩnh vực học sâu. Ưu điểm của mô hình ResNet là khả năng huấn luyện mạng sâu một cách hiệu quả, trong khi Inception nhấn mạnh vào việc học các đặc trưng đa dạng và tối ưu. Hạn chế của các mô hình này xoay quanh độ phức tạp tính toán, độ phức tạp về kiến trúc và những thách thức tiềm ẩn trong việc tối ưu hóa mô hình. Thực tế thứ hai là các kiến trúc mạng nơ-ron khác nhau được thiết kế cho các mục đích khác nhau hoặc để giải quyết những vấn đề của một bài toán cụ thể. Ví dụ: MobileNet [36] kết hợp tích chập theo chiều sâu và tích chập theo điểm để giảm số lượng tính toán và tham số trong khi vẫn giữ được những đặc trưng cần thiết từ dữ liệu đầu vào. Kiến trúc MobileNet giúp triển khai các mô hình học sâu trên các thiết bị có tài nguyên tính toán hạn chế.

Việc lựa chọn mô hình huấn luyện có trước phù hợp là một trong những yếu tố quan trọng để đảm bảo việc học chuyển giao đạt hiệu quả.

Học chuyển giao có thể đặc biệt có lợi trong các tình huống mà dữ liệu được gắn nhãn để nhận dạng người nói bị hạn chế, vì nó cho phép tận dụng kiến thức từ các bộ dữ liệu lớn hơn và đa dạng hơn, có khả năng dẫn đến cải thiện hiệu suất và hội tụ nhanh hơn trong quá trình huấn luyện.

Học chuyển giao với tài nguyên dữ liệu hạn chế

Trong bài toán xác thực người nói, các mô hình đã được huấn luyện trước thường được phân thành hai loại: mô hình huấn luyện có giám sát và mô hình học tự giám sát [131]. Các mô hình huấn luyện có giám sát được huấn luyện với các bộ dữ liệu lớn có gán nhãn người nói như VoxCeleb2 [14], CnCeleb2 [64].

Các mô hình này thường được sử dụng để khởi tạo các mô hình xác thực người nói ([130], [127], [128]) hoặc áp dụng các kỹ thuật thích ứng miền ([129], [93]) nhằm chuyển giao tính bền vững và khả năng tổng quát hóa của chúng cho các tình huống xác thực người nói với tài nguyên dữ liệu hạn chế. Xác thực người nói với tài nguyên dữ liệu hạn chế liên quan đến việc phát triển các hệ thống xác thực người nói khi nguồn dữ liệu có giới hạn, chẳng hạn như trong trường hợp các ngôn ngữ ít phổ biến hoặc các ứng dụng xác thực người nói từ xa. Giới hạn về tài nguyên dữ liệu có thể do số lượng người nói ít, sự thiếu đa dạng trong nội dung lời nói, hoặc không có đủ các môi trường âm học khác nhau trong tập dữ liệu. Để khắc phục vấn đề thiếu hụt dữ liệu, một phương pháp đơn giản là tăng cường dữ liệu âm thanh từ các nguồn tài nguyên thấp. Các phương pháp thông thường bao gồm thêm tiếng ồn [104], tiếng vang [29], thay đổi tốc độ [123], và SpecAug [79]. Một số phương pháp sinh cũng đã được áp dụng để thực hiện tăng cường cho xác thực người nói với dữ liệu hạn chế [120], [128]. Trong chương này, NCS nghiên cứu tiềm năng của mô hình lớn đã được huấn luyện trước ResNetSE-34, ECAPA-TDNN, RawNet3 (trên dữ liệu VoxCeleb2 [14]) để nâng cao độ chính xác xác thực người nói với lượng dữ liệu nhỏ.

3.2. Lựa chọn dữ liệu cho bài toán xác thực người nói

VoxCeleb2 [14] là một trong những bộ dữ liệu lớn nhất thường được sử dụng cho các bài toán nhận dạng người nói và xác thực người nói. Nó là phần mở rộng của bộ dữ liệu VoxCeleb1 [72] ban đầu và được thiết kế nghiên cứu về nhận dạng, xác thực người nói cũng như các lĩnh vực liên quan. VoxCeleb2 đã được sử dụng rộng rãi trong việc phát triển và đánh giá các mô hình và thuật toán nhận dạng người nói khác nhau. Bộ dữ liệu này đã được mô tả chi tiết trong Chương 1 của luận án. Bảng 3.1: Tóm tắt ba tập dữ liệu công khai cho bài toán xác thực người nói. VoxCeleb2 được dùng để huấn luyện trước ba mô hình học sâu và tập dữ liệu VLSP2021-SV được dùng để tinh chỉnh và đánh giá mô hình.

Dataset	Language	Recording hours	Training set # of utterances (# of speaker)	Test set # of test pair (# of speakers)
VoxCeleb2 [14]	Phần lớn tiếng Anh	2,300	1,128,246 (6112)	37,720 (40) 581,480 (1251) 552,536 (1190)
CN-CELEB2 [64]	Trung Quốc	1,090	529.485 (2000)	3,586,776 (-)
VLSP2021-SV [18]	Việt nam	41	31,600 (1305)	20,000 (114)

Bên cạnh tiếng Anh, một số bộ dữ liệu đã được xây dựng cho các ngôn ngữ ngữ khác như tiếng Trung và tiếng Việt. Ví dụ: CN-CELEB1 [25] là bộ dữ liệu nhận dạng người nói lớn được thu thập 'trong tự nhiên' chứa hơn 130.000 câu nói của 1.000 người nổi tiếng Trung Quốc. Kết quả công bố của nhóm nghiên cứu [25] cho thấy hiệu năng của các phương pháp nhận dạng người nói mới nhất có kết quả kém xa so với nhóm nghiên cứu VoxCeleb2 [14]. Gần đây, hội thảo VLSP 2021 đã công bố bộ dữ liệu xác thực và nhận dạng, tập dữ liệu này là VLSP2021-SV. Một số thống kê của ba bộ dữ liệu được báo cáo trong Bảng 3.1. Chúng ta có thể thấy rằng VoxCeleb2 hiện lớn hơn nhiều so với hai bộ dữ liệu nhận dạng người nói tiếng Việt và tiếng Trung. Nó cho phép phát triển và đánh giá các thuật toán tiên tiến, đồng thời góp phần vào sự phát triển của lĩnh vực này bằng cách cung cấp bộ sưu tập bản ghi âm lớn và đa dạng. Đặc biệt, VoxCeleb2 được sử dụng làm nguồn dữ liệu huấn luyện trước để học chuyển giao. Các nhà nghiên cứu và nhà phát triển có thể tinh chỉnh các mô hình được huấn luyện trước trên VoxCeleb2 với lượng dữ liệu nhỏ được gắn nhãn từ mục đích ứng dụng riêng, thích nghi hiệu quả mô hình mới cho bài toán cụ thể.

3.3. Các mô hình huấn luyện trước cho học chuyển giao

ECAPA-TDNN sử dụng mảng đặc trưng âm học mức khung (frame) tương tự như i-vector truyền thống, trong khi VGGVox dùng ảnh phổ 2 chiều và RawNet sử dụng tín hiệu tiếng nói thô làm đầu vào. Các mô hình khác nhau sử dụng các cách trích chọn đặc trưng mức khung (frame) khác nhau để biểu diễn đặc trưng người nói riêng biệt hoặc cố định số chiều véc tơ nhúng người nói trong bài toán nhận dạng và xác thực người nói.

VGGVox

VGGVox sử dụng phổ cường độ ngắn 2D làm đầu vào mạng CNN sâu. Ví dụ phổ kích thước (512,300) sinh ra từ đoạn tín hiệu tiếng nói có độ dài 3s với cửa sổ trượt dùng cửa sổ Hamming có độ rộng 25 ms, bước nhảy 10 ms.

Trong bài báo (Nagrani et al., 2019), tác giả trích chọn đặc trưng đoạn tiếng nói với bốn backbone CNN sâu: VGG-M, ResNet-34, ResNet-50 và Thin-ResNet. Trong giai đoạn huấn luyện, VGGVox lấy ngẫu nhiên đoạn 3s trong mỗi câu nói dữ liệu huấn luyện. Trong giai đoạn kiểm thử, sử dụng các chiến lược khác nhau đánh giá hiệu năng bài toán xác thực người nói.

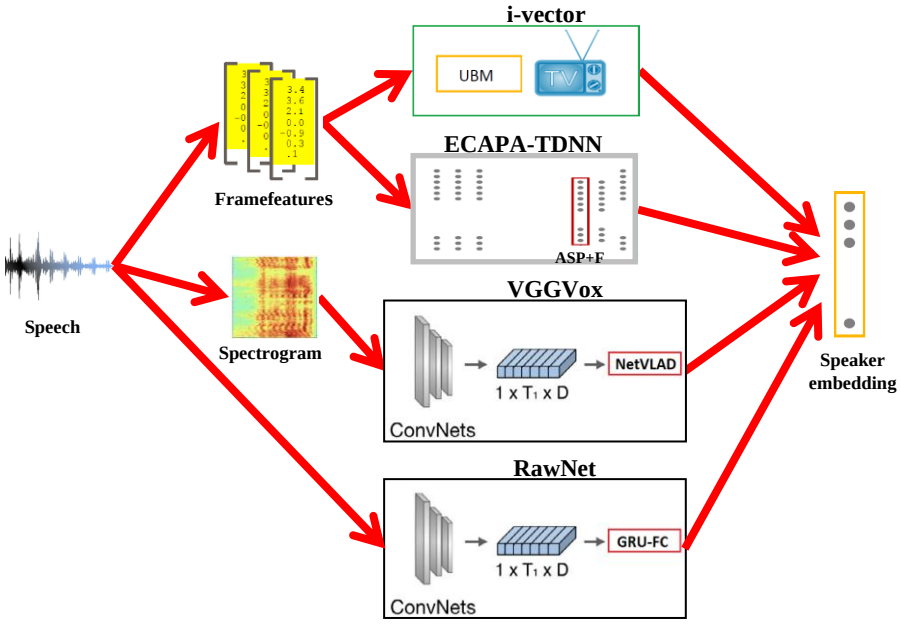
RawNet3

RawNet3 là mô hình cải tiến từ RawNet2 [47] và ECAPA-TDNN [22]. Như minh họa ở Hình 3.5, RawNet3 bao gồm một lớp ngân hàng bộ lọc phân tích tham số hóa (ParamFbank) [78] và khối Res2Dilated với việc mở rộng bản đồ đặc trưng α và tổng hợp (max pooling) (AFMS-Res2MP). Khi tín hiệu sóng một chiều được đưa vào RawNet3, nó được lớp ParamFbank biến đổi thành bản đồ đặc trưng hai chiều. Lớp ParamFbank học các bộ lọc tham số hóa với giá trị thực, được mở rộng từ lớp SincNet [89]. Bản đồ đặc trưng sau khi được xử lý qua lớp ParamFbank, sẽ được truyền qua ba khối AFMS-Res2MP để thu nhận thông tin người nói ở các mức độ khác nhau.

Ý tưởng chính bên trong RawNet là sử dụng mạng nơ-ron tích chập (CNN) trực tiếp trên âm thanh thô, tương tự như cách áp dụng CNN cho hình ảnh. Một trong những ưu điểm chính của RawNet là khả năng tìm hiểu các tính năng phân biệt trực tiếp từ tín hiệu âm thanh thô, điều này có thể giúp cải thiện hiệu suất so với trích chọn đặc trưng truyền thống.

Trong các phiên bản cải tiến, RawNet2 [42] và RawNet3 [50] sử dụng hàng loạt các kỹ thuật nâng cao trích chọn đặc trưng mức frame như sử dụng mô đun khung Res2Net và chia tỉ lệ bản đồ đặc trưng có trọng số quan trọng hoặc tập trung vào những đặc trưng quan trọng trong bản đồ đặc trưng mức khung. Bên cạnh đó, lớp tích chập đầu tiên thay bằng lớp sinc-conv cho kết quả tốt hơn.

Cả ba mô hình huấn luyện trên tập dữ liệu VoxCeleb2 (37,720 cặp câu nói từ 40 người nói; 1,092,009 câu nói của 5,994 người nói).



Hình 3.2: Minh họa sự khác nhau giữa phương pháp i-vector truyền thống và ba mô hình học sâu cho mô hình hóa người nói.

Bảng 3.4 cho thấy với cùng dữ liệu đầu vào thô, RawNet3 đạt tỉ lệ lỗi EER là 0,89% tốt hơn so với 3,00% của RawNet2. Sự cải thiện này chính là việc tích hợp những đặc điểm mạnh của mô hình ECAPA-TDNN. Những đặc điểm này bao gồm các khối xương sống (backbone) có kết nối dư, tổng hợp đặc trưng nhiều lớp, thay thế GRU bởi kênh và tổng hợp thống kê phụ thuộc ngữ cảnh, dùng hàm mục tiêu AAM-softmax trong quá trình huấn luyện. Số chiều đặc trưng nhúng người nói của RawNet3 là 256 nhỏ hơn nhiều 1024 của RawNet2. Số chiều đặc trưng nhúng của ECAPA-TDNN là 192 và của ResNet-34 là 512. ECAPA-TDNN đạt tỉ lệ lỗi EER là 0,96% kém hơn RawNet3 một chút nhưng tốt hơn nhiều so với RawNet2. Kết quả này cho thấy tính hiệu quả của mô hình ECAPA-TDNN với đặc trưng đầu vào MFBEs 80 chiều. VGGVox (ResNet-34) cũng có tỉ lệ lỗi EER là 1.18% không tốt như RawNet3. Có thể do việc chuyển đổi tín hiệu giọng nói 1D thành hình ảnh phổ 2D, sau đó áp dụng khung xương (backbone) trích chọn đặc trưng đã được chứng minh hiệu quả trong lĩnh vực hình ảnh. Việc áp dụng các kỹ thuật học máy hiện đại trên VoxCeleb sử dụng NetVLAD để tổng hợp đặc trưng mức khung, kết hợp so sánh độ tương tự Cosin và phân loại cho quá trình xác thực.

Bảng 3.4: So sánh sự khác nhau giữa các mô hình.

Mô hình	Số lượng tham số (M)	Dữ liệu đầu vào	Số chiều đặc trưng nhúng	Tỉ lệ lỗi đánh giá (%)
VGGVox (ResNet-34) [16]	22.0	2D spectrum image	512	2.22
ECAPA-TDNN [22] (C=512)	6.2	1D array of MFCCs	192	1.01
ECAPA-TDNN [22] (C=1024)	14.9	1D array of MFCCs	192	0.87
RawNet2 [47]	13.2	59.049 samples	1024	3.00
RawNet3 [50]	16.3	59.049 samples	256	0.89

3.4. Thực nghiệm

Bộ dữ liệu

VLSP2021-SV

Trong thực nghiệm này này, NCS sử dụng bộ dữ liệu VLSP2021-SV (trình bày chi tiết trong Chương 1) của tác giả Vi Thanh Dat và cộng sự [18]. Bộ dữ liệu gồm các đoạn câu nói ngắn được trích chọn từ các cuộc phỏng vấn trên YouTube. Trong các đoạn video này, các giọng nói đều là giọng nói tự nhiên. Bộ dữ liệu dùng huấn luyện có 31,600 câu nói của 1,305 người nói.

Tinh chỉnh trên mô hình huấn luyện trước

Mục tiêu chung của học chuyển giao là tận dụng kiến thức hoặc dữ liệu hiện có để nâng cao hiệu năng hệ thống xác thực người nói. Cách đơn giản nhất là sử dụng lại mô hình huấn luyện trước và trích chọn đặc trưng mức cao trên đoạn âm thanh đầu vào, tạo ra đặc trưng nhúng. Các đặc trưng này có thể làm đầu vào mô hình xác thực người nói hoặc các bài toán khác. Trong phần này, NCS tinh chỉnh các mô hình đã huấn luyện trước với cùng tập dữ liệu phát triển VLSP2021-SV, thay thế lớp phân loại sau cùng bằng số người nói trong tập dữ liệu VLSP2021-SV. Cụ thể, các thiết lập sau áp dụng trong các thực nghiệm :

- ResNetSE-34 : NCS dùng MFBEs có 64 chiều, độ dài cửa sổ 25ms, bước nhảy 10ms, kích thước FFT là 512 giới hạn miền tần số 20-7600 Hz ;
- ECAPA-TDNN : NCS dùng 14.85 triệu tham số không dùng phần phân loại (classification), đầu vào mô hình là đặc trưng MFCCs có 80 chiều. Các tham số khác giống như mô hình gốc ECAPA-TDNN [22].
- Rawnet3: mô hình có 16.28 triệu tham số, dùng kích thước trượt là 10 cho lớp bộ lọc phân tích tham số hóa. Mô hình sử dụng dữ liệu đầu vào là âm thanh thô.

Các mô hình trên huấn luyện trên NVIDIA Telsa P100 GPUs, 16GB bộ nhớ, sử dụng tối ưu Adam, tốc độ học khởi tạo là 0.0001. Mỗi mô hình huấn luyện 500 epoch, tốc độ học giảm 5% sau 10 epoch. NCS có sử dụng kỹ thuật tăng cường dữ liệu trực tuyến trong quá trình huấn luyện. NCS sử dụng tập dữ liệu MUSAN [104] và RIR[59] để bổ sung dữ liệu huấn luyện.

Kết quả thực nghiệm

Trong Bảng 3.6, NCS trình bày kết quả thực nghiệm của ba mô hình học sâu cho bài toán xác thực người nói với tài nguyên hạn chế: không học chuyển giao và có học chuyển giao. Có thể thấy mô hình RawNet3 mặc dù kế thừa từ hai mô hình ECAPA-TDNN và mô hình RawNet2 nhưng kết quả tỉ lệ lỗi là 4.07% không tốt hơn so với mô hình ECAPA-TDNN có tỉ lệ lỗi là 3.92% (các mô hình huấn luyện từ đầu). Với kỹ thuật học chuyển giao, RawNet3 đặt được tỉ lệ lỗi 1,61% trên bộ đánh giá của VLSP 2021. Đây là mức chênh lệch lớn khi so sánh với tỉ lệ lỗi 2,21% của mô hình ECAPA-TDNN. Học chuyển giao cũng giúp nâng cao hiệu suất của VGGVox (ResnetSE-34) khi tỉ lệ lỗi giảm từ 5,98% xuống 2,22%, rất gần với 2,21% của ECAPA-TDNN. Từ sự so sánh và phân tích về mặt phương pháp trong mục 3.3 và kết quả thử nghiệm trong phần này, chúng ta có thể lập luận rằng việc chuyển đổi từ tín hiệu giọng nói thô sang hình ảnh phổ 2D, sau đó áp dụng các mô hình CNN sâu thông thường để nhận dạng hình ảnh có thể không phải là một cách tiếp cận lý tưởng cho bài toán xác thực người nói. VGGVox đã thử nghiệm với các mạng xương sống khác nhau như

VGG, ResNets, các phương pháp tổng hợp đặc trưng mức khung, các kỹ thuật nâng cao chất lượng nhận dạng và xác thực. Tuy nhiên, VGGVox vẫn thua kém ECAPA-TDNN và RawNet trên bộ dữ liệu VoxCeleb2 và VLSP2021-SV. Nguyên nhân có thể là do toán tử tích chập 2D thông thường không thiết kế phép chiếu các câu nói có độ dài thay đổi thành các đặc trưng nhúng có chiều dài cố định như TDNN trong ECAPA-TDNN hoặc tổng hợp thống kê phụ thuộc vào ngữ cảnh trong cả ECAPA-TDNN và RawNet3. Trong Bảng 3.6 MFBEs(2D) có nghĩa là đặc trưng MFBE được sử dụng làm đầu vào cho mạng tích chập nơ-ron 2 chiều. Các hệ số MFBE được sắp xếp thành một ma trận hai chiều, trong đó một chiều là thời gian và chiều còn lại là các hệ số MFBE. Ma trận này có thể được coi như một hình ảnh với các kênh màu khác nhau, và được sử dụng làm đầu vào cho lớp Conv2D. MFBEs(1D) có nghĩa là đặc trưng MFBE được sử dụng làm đầu vào cho mạng nơ-ron tích chập Conv1D. Các hệ số MFBE được sắp xếp thành một ma trận hai chiều, trong đó một chiều là thời gian và chiều còn lại là các hệ số MFBE. Ma trận này được sử dụng làm đầu vào cho lớp Conv1D. Conv1D sẽ áp dụng các bộ lọc 1D dọc theo trục thời gian của ma trận MFBE. Đầu vào mảng véc-tơ âm học một chiều của ECAPA-TDNN chỉ ra ưu điển cho huấn luyện mô hình xác thực người nói với tập dữ liệu nhỏ. ECAPA-TDNN đạt tỉ lệ lỗi là 3.92% khi chỉ huấn luyện 41 giờ âm thanh (VLSP2021-SV) trong khi RawNet3 chỉ đạt tỉ lệ lỗi là 4.07%. Trên tập dữ liệu lớn hơn VoxCeleb2 (2,300 giờ thu âm), cả RawNet3 và ECAPA-TDNN đều cho kết quả tốt với tỉ lệ lỗi là 0.87% và 0.89%. Sự so sánh này có nghĩa RawNets cần nhiều dữ liệu huấn luyện và nó có hiệu suất tốt hơn ECAPA-TDNN. Ở một khía cạnh khác, học chuyển giao trong miền tín hiệu tiếng nói thô được chứng minh hiệu quả hơn miền đặc trưng âm học. Học chuyển giao giúp ECAPA-TDNN giảm tỉ lệ lỗi từ 3.92% xuống còn 2.21% trong khi RawNet3 có tỉ lệ lỗi giảm từ 4.07% xuống còn 1.61%. Nguyên nhân có thể do trích chọn đặc trưng âm học dùng trong ECAPA-TDNN phụ thuộc vào các tham số nói chung như số chiều đặc trưng MFCCs, kích thước cửa sổ, độ rộng bước nhảy. Việc tối ưu các tham số này trên một tập dữ liệu (nhiều quốc tịch và phương ngữ trong VoxCeleb2) không phải là một lựa chọn tốt cho tập dữ liệu khác (ví dụ Việt Nam). Mô hình RawNet không đòi hỏi siêu tham số, nó học tự động các đặc trưng mức khung từ dữ liệu thô tín hiệu.

Bảng 3.6: So sánh hiệu năng của ba mô hình học sâu cho bài toán xác thực người nói tiếng Việt. (Không học chuyển giao và học chuyển giao).

Mô hình	**Đặc trưng đầu vào**	**Không học chuyển giao**	**Học chuyển giao**	**Số lượng tham số**	**GFLOPs**
		SV EER (%)	SV EER (%)	**(M)**	
ResnetSE-34	MFBEs 64 (2D)	5.98	2.22	22.0	3.82
ECAPA-TDNN	MFBEs 80 (1D)	3.92	2.21	14.9	3.96
RawNet3	Raw waveform	4.07	1.61	16.3	8.13

ECAPA-TDNN và RawNet3 có hiệu suất tính toán cao hơn ResnetSE-34 do thiết kế đặc biệt cho các bài toán như nhận dạng tiếng nói và xử lý âm thanh thô. Giá trị FLOPs phụ thuộc vào kiến trúc mạng cụ thể bao gồm số lớp và kích thước các lớp, kích thước nhân và kích thước dữ liệu đầu vào. Dữ liệu đầu vào mô hình ECAPA-TDNN và RawNet3 là âm thanh thô thường có số chiều cao hơn so với đầu vào của ResnetSE-34 (mảng 2D – 64 MFBEs). Xử lý dữ liệu nhiều chiều cần nhiều

tài nguyên tính toán cũng góp phần làm FLOPs cao hơn. Bảng 3.6 cho thấy RawNet3 có FLOPs cao nhất là 8.13G. Đó là do đầu vào mô hình RawNet3 có kích thước lớn nhất (59,049 mẫu) trong khi hai mô hình còn lại ResnetSE-34 (64 MFBEs) và ECAPA-TDNN (80 MFBEs).

ECAPA-TDNN và RawNet3 có hiệu suất tính toán cáo hơn ResnetSE-34 là do thiết kế đặc biệt cho các bài toán như nhận dạng tiếng nói và xử lý âm thanh thô. Giá trị FLOPs phụ thuộc vào kiến trúc mạng cụ thể bao gồm số lớp và kích thước các lớp, kích thước nhân và kích thước dữ liệu đầu vào. Dữ liệu đầu vào mô hình ECAPA-TDNN và RawNet3 là âm thanh thô thường có số chiều cao hơn so với đầu vào của ResnetSE-34 (mảng 2D – 64 MFBEs). Xử lý dữ liệu nhiều chiều cần nhiều tài nguyên tính toán cũng góp phần làm FLOPs cao hơn. Bảng 25 cho thấy RawNet3 có FLOPs cao nhất là 8.13G. Đó là do đầu vào mô hình RawNet3 có kích thước lớn nhất (59,049 mẫu) trong khi hai mô hình còn lại ResnetSE-34 (64 MFBEs) và ECAPA-TDNN (80 MFBEs).

3.5. Kết luận chương 3

NCS đã nghiên cứu tính hiệu quả của ba mô hình học sâu cho bài toán xác thực người nói tiếng Việt. Kết quả thử nghiệm cho thấy ECAPA-TDNN cho kết quả tốt với tập dữ liệu huấn luyện nhỏ. Mô hình này đạt tỉ lệ lỗi 3,92% khi huấn luyện trên tập dữ liệu gồm 31,600 câu nói (tập dữ liệu VLSP2021-SV), so với 4,07% của RawNet3 và 5,98% của VoxCeleb. Tuy nhiên, khi các mô hình được huấn luyện trên tập dữ liệu lớn hơn nhiều gồm 1.128.246 câu nói (VoxCeleb2), tinh chỉnh RawNet3 đã giảm tỉ lệ lỗi xuống còn 1,61%. Học chuyển giao cũng giúp hệ thống sử dụng mô hình ResnetSE-34 đạt tỉ lệ lỗi 2,22%, gần bằng tỉ lệ lỗi 2,21% của ECAPA-TDNN. Từ việc so sánh và phân tích ba mô hình, NCS kết luận rằng sức mạnh của RawNet3 nằm ở việc kế thừa các đặc điểm của ECAPA-TDNN, chẳng hạn như tổng hợp đặc trưng mức khung và hiệu quả sự kết hợp với các hàm mục tiêu khác nhau. Dữ liệu đầu vào dạng thô cũng khiến Rawnet ít phụ thuộc trích chọn đặc trưng âm học và tận dụng được sự đa dạng của dữ liệu huấn luyện. Một số hướng cũng có thể thử nghiệm trong tương lai. Thứ nhất, tính chất âm học của ngôn ngữ nói khác nhau chưa được nghiên cứu kỹ lưỡng. Ví dụ, Tiếng Việt là ngôn ngữ có thanh điệu và có sáu loại thanh điệu, còn tiếng Anh thì không. Tiếng Anh sử dụng cao độ và ngữ điệu chứ không phải thanh điệu như tiếng Việt. Tinh chỉnh một mô hình được huấn luyện trước trên cùng một ngôn ngữ thanh điệu có thể mang lại lợi ích cho bài toán xác thực người nói tiếng Việt. Hướng thứ hai là áp dụng kiến trúc đã được chứng minh thành công trong các lĩnh vực khác như transformer [115] trong xử lý ngôn ngữ tự nhiên tổng hợp các đặc trưng mức khung biểu diễn người nói tốt hơn. Việc áp dụng các kỹ thuật tiền xử lý khác như phát hiện đoạn tiếng nói và lựa chọn đặc trưng 80 âm học cũng giúp nâng cao hiệu năng của bất cứ hệ thống xác thực người nói nào. Các kết quả về Đề xuất sử dụng mô hình RawNet3 trong học chuyển giao cho bài toán xác thực người nói với tài nguyên hạn chế được công bố tại công trình **[CT3], [CT4]** trong phần "Danh mục các công trình của tác giả".

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Các kết quả chính của luận án

Luận án đã trình bày nghiên cứu hệ thống về vấn đề xác thực người nói cho tiếng nói với tài nguyên hạn chế, đây là bài toán quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên. Luận án tập

trung nghiên cứu đề xuất các đặc trưng với mô hình học sâu nhằm nâng cao độ chính xác hệ thống xác thực người nói tiếng Việt. Bên cạnh đó luận án còn nghiên cứu, đề xuất mô hình học sâu hiện đại trong học chuyển giao cho bài toán xác thực người nói với tài nguyên hạn chế. Kết quả nghiên cứu của luận án có thể được tóm tắt như sau :

1. **Đề xuất sử dụng đặc trưng MFBEs với mô hình ECAPA-TDNN cho bài toán xác thực người nói tiếng Việt. Mô hình ECAPA-TDNN (80 MFBEs) thì tỉ lệ lỗi EER giảm lần lượt là 11.37% (so với 11.58%) và 12.47% (so với 14.3%) ([CT1], [CT2])**
2. **Đề xuất sử dụng mô hình RawNet3 trong học chuyển giao cho bài toán xác thực người nói với tài nguyên hạn chế. Mô hình RawNet3 cho kết quả tốt nhất với tỉ lệ lỗi EER là 1.61%. ([CT3], [CT4])**

Những hạn chế của luận án

1. Cơ sở dữ liệu cho tiếng Trung Quốc và mô hình huấn luyện trên dữ liệu này cũng được công bố rộng rãi, hiện NCS chưa có những thực nghiệm, đánh giá khi áp dụng cho dữ liệu tiếng Việt.
2. NCS cũng cần thử nghiệm, đánh giá sự kết hợp giữa các đặc trưng MFCCs và MFBEs cho bài toán xác thực tiếng nói trên cả tiếng Anh và tiếng Việt Hướng phát triển Thử nghiệm đặc trưng MFBEs/LPCCs hoặc kết hợp MFCCs nâng cao độ chính xác SV; Sử dụng mô hình huấn luyện có sẵn từ dữ liệu tiếng Trung, sau đó tinh chỉnh trên Vietnam-Celeb-T.

Hướng nghiên cứu tiếp theo

Từ những kết quả đạt được của luận án, các vấn đề cần đặt ra trong thời gian tới cần được nghiên cứu:

1. Thử nghiệm đặc trưng MFBEs/LPCCs hoặc kết hợp MFCCs nâng cao độ chính xác hệ thống xác thực người nói;
2. Sử dụng mô hình huấn luyện trên dữ liệu tiếng Trung Quốc, sau đó tinh chỉnh trên tập dữ liệu Vietnam-Celeb-T.

DANH MỤC CÁC BÀI BÁO ĐÃ XUẤT BẢN
LIÊN QUAN ĐẾN LUẬN ÁN

1. **[CT1] T. -T. -M. Nguyen**, D. -D. Nguyen and C. -M. Luong, "Vietnamese Speaker Verification with Mel-Scale Filter Bank Energies and Deep Learning," in IEEE Access, vol. 12, pp. 150114-150122, 2024, doi: 10.1109/ACCESS.2024.3479092. **(SCIE, Q1)**
2. **[CT2] Nguyễn Thị Thanh Mai**, Nguyễn Đức Dũng, "Kết hợp đặc trưng MFCCs và Mel-Filter Bank Energies trong xác thực người nói tiếng Việt", VNICT-2024, Tr. 288-293.
3. **[CT3] Thi-Thanh-Mai Nguyen**, Duc-Dung Nguyen, Chi-Mai Luong (2024). "Transfer Learning for Vietnamese Speaker Verification." Vietnam Journal of Science and Technology. **(Được chấp nhận) (SCOPUS, Q4)**
4. **[CT4] Mai Nguyen Thi Thanh**, Dung Nguyen Duc, "Vietnamese Speaker Verification based on ResNet model", VNICT-2023, pp 377-381.