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

LUẬN ÁN TIẾN SĨ MÁY TÍNH

Hà Nội - 2024

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

LUẬN ÁN TIẾN SĨ MÁY TÍNH
Ngành: Khoa học máy tính
Mã số: 9 48 01 01

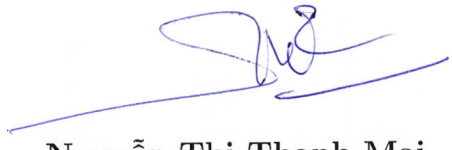
Xác nhận của Học viện Khoa học và Công nghệ	Người hướng dẫn 1	Người hướng dẫn 2
KT. GIÁM ĐỐC PHÓ GIÁM ĐỐC HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM Nguyễn Thị Trung	*(Ký, ghi rõ họ tên)* PGS.TS. Nguyễn Đức Dũng	*(Ký, ghi rõ họ tên)* PGS.TS. Lương Chi Mai

Hà Nội - 2024

LỜI CAM ĐOAN

Tôi xin cam đoan luận án: "Nâng cao độ chính xác xác thực người nói sử dụng học sâu với tài nguyên hạn chế" là công trình nghiên cứu của chính mình dưới sự hướng dẫn khoa học của tập thể hướng dẫn. Luận án sử dụng thông tin trích dẫn từ nhiều nguồn tham khảo khác nhau và các thông tin trích dẫn được ghi rõ nguồn gốc. Các kết quả nghiên cứu của tôi được công bố chung với các tác giả khác đã được sự nhất trí của đồng tác giả khi đưa vào luận án. Các số liệu, kết quả được trình bày trong luận án là hoàn toàn trung thực và chưa từng được công bố trong bất kỳ một công trình nào khác ngoài các công trình công bố của tác giả. Luận án được hoàn thành trong thời gian tôi làm nghiên cứu sinh tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Hà Nội, ngày 20 tháng 12 năm 2024



Nguyễn Thị Thanh Mai

LỜI CẢM ƠN

Luận án này đã hoàn thành nhờ sự nỗ lực không ngừng nghỉ của tác giả cùng với sự hỗ trợ tận tâm từ các thầy giáo hướng dẫn, đồng nghiệp, bạn vè và người thân.

Tác giả muốn bày tỏ lòng biết ơn chân thành và sâu sắc đến thầy giáo hướng dẫn là PGS.TS Nguyễn Đức Dũng và PGS.TS Lương Chi Mai. Những lời hướng dẫn, sự động viên và tận tâm của họ dành cho tác giả trong suốt quá trình thực hiện luận án là không thể nào diễn đạt hết.

Tác giả muốn bày tỏ lòng biết ơn sâu sắc đến các giảng viên và cán bộ của phòng quản lý nghiên cứu sinh thuộc Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Sự hỗ trợ nhiệt tình và tạo điều kiện thuận lợi của họ đã đóng góp quan trọng vào việc hoàn thành luận án của tác giả.

Tác giả xin gửi lời cảm ơn tới toàn thể thành viên Phòng Nhận dạng và Công nghệ tri thức, Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, những người đã đồng hành và hỗ trợ tác giả trong suốt quá trình nghiên cứu.

Đặc biệt, tác giả muốn bày tỏ lòng biết ơn sâu sắc đến Bố, Mẹ, Chồng và các con trong gia đình. Họ đã luôn chia sẻ những khó khăn và động viên tác giả trong quá trình nghiên cứu. Luận án cũng là món quà tinh thần mà tác giả trân trọng gửi đến tất cả thành viên trong gia đình.

Hà Nội, ngày 20 tháng 12 năm 2024

Nguyễn Thị Thanh Mai

MỤC LỤC

DANH MỤC THUẬT NGỮ

Thứ tự	**Thuật ngữ**	**Diễn giải**
1	Baseline	Mô hình cơ bản, mô hình làm cơ sở so sánh
2	End-to-end	Mô hình từ đầu vào đến đầu ra
3	Fine-tune	Kỹ thuật tinh chỉnh các tham số học từ mô hình huấn luyện trước
4	Loss	Hàm mất mát
5	Speaker embedding	Mã hóa đặc trưng người nói
6	Score normalization	Chuẩn hóa điểm số

DANH MỤC CÁC CHỮ VIẾT TẮT

Từ viết tắt	Diễn giải	Ý nghĩa
AFMS	α Feature Map Scaling	Chia tỉ lệ bản đồ đặc trưng α
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
CSSL	Contrastive Self-Supervised Learning	Học tự giám sát
DCT	Discrete Cosine Transform	Biến đổi cosin rời rạc
DFT	Discrete Fourier Transform	Biến đổi Fourier rời rạc
DNN	Deep Neuron Network	Mạng nơ-ron học sâu
EER	Equal Error Rate	Tỷ lệ câu bị lỗi
ECAPA-TDNN	**E**mphasized **C**hannel **A**ttention, **P**ropagation and **A**ggregation - Time Delay Neural Network	Mô hình học sâu ECAPA-TDNN
FFT	Fast Fourier Transform	Biến đổi Fourier nhanh
GMM	Gaussian Mixture Model	Mô hình phân phối Gaussian
GMM-UMB	Gaussian mixture modelling with a universal background model	Mô hình GMM-UMB
GPU	Graphical Processing Unit	Bộ xử lý đồ họa
JFA	Joint Factor Analysis	Phân tích nhân tố kết hợp
MFBEs	Mel-Filter Bank Energies	Đặc trưng âm học MFBEs
MFCCs	Mel Frequency Cepstral Coefficients	Đặc trưng âm học MFCCs
MHAS	Self **M**ulti-**H**ead **A**ttention for **S**peaker Recognition	Mô đun MHAS
PLDA	Probabilistic Linear Discriminant Analysis	Phân tích phân biệt tuyến tính xác suất
RNN	Recurrent neural network	Mạng nơ-ron hồi quy
TDNN	Time Delay Neural Network	Mạng nơ-ron trễ thời gian
SAP	Self Attention Pooling	Tổng hợp tự chú ý

DANH SÁCH HÌNH VẼ

DANH SÁCH BẢNG

MỞ ĐẦU

1. Tính cấp thiết của đề tài 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 [13] gồm 6,112 người nói trong khi dữ liệu VLSP2021-SV [17] 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

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 luận án

Mục tiêu nghiên cứu của luận án là nghiên cứu đề xuất một số giải pháp nâng cao chất lượng 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. Đối tượng và phạm vi nghiên cứu

3.1. Đối tượng nghiên cứu

Đối tượng nghiên cứu của luận án là các đặc trưng và các mô hình mạng học sâu cho bài toán xác thực người nói với tài nguyên hạn chế.

3.2. Phạm vi nghiên cứu

Nghiên cứu các phương pháp trích chọn đặc trưng làm đầu vào cho mạng học sâu trong bài toán xác thực người nói, đánh giá trên các tập dữ liệu đã công bố rộng rãi. Nghiên cứu phương pháp nâng cao độ chính xác hệ thống xác thực người nói sử dụng học chuyển giao với tài nguyên hạn chế.

4. Nội dung nghiên cứu

Để đạt được mục tiêu nghiên cứu đề ra, luận án tập trung nghiên cứu một số nội dung chính sau:

- Nghiên cứu lý thuyết tổng quan về nhận dạng người nói, cụ thể là bài toán xác thực người nói;
- Khảo sát, phân tích, đánh giá các tập dữ liệu công bố cho bài toán xác thực người nói;
- Nghiên cứu các phương pháp trích trọn đặc trưng sử dụng các mô hình học sâu hiện đại cho xác thực người nói
- Nghiên cứu các phương pháp nâng cao độ chính xác hệ thống xác thực người nói dựa trên các mô hình học sâu.

5. Phương pháp nghiên cứu

Phương pháp nghiên cứu của luận án là kết hợp giữa nghiên cứu lý thuyết và thực nghiệm.

- **Về lý thuyết**

  Nghiên cứu tổng quan lý thuyết tổng quan về nhận dạng người nói, xác thực người nói. Nghiên cứu các cách tiếp cận mới nhất trong bài toán xác thực người nói. Nghiên cứu, khảo sát các tập dữ liệu công bố thường sử

dụng cho bài toán xác thực người nói. Nghiên cứu các kỹ thuật trích chọn đặc trưng sử dụng mạng học sâu dùng trong xác thực người nói. Nghiên cứu, phân tích các phương pháp nâng cao độ chính xác hệ thống xác thực người nói.

- **Về thực nghiệm**

  Sử dụng ngôn ngữ lập trình Python để thực hiện hóa các thuật toán, xây dựng các mô hình mạng nơ-ron cho bài toán xác thực người nói trên các tập dữ liệu đã công bố. Sử dụng các công cụ trên framework PyTorch [81] để xây dựng mô hình mạng nơ-ron, cài đặt các thuật toán đề xuất, huấn luyện và kiểm tra. Phân tích, đánh giá, so sánh các mô hình được đề xuất trong luận án với các kết quả nghiên cứu được công bố trên thế giới cũng như ở Việt Nam.

6. Những đóng góp mới của luận án

- Đề xuất sử dụng đặc trưng MFBEs (Mel-Filter Bank Energies) với mô hình ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation-Time Delay Neural Network) cho bài toán xác thực người nói tiếng Việt;
- Đề xuất sử dụng mô hình RawNet3 trong học chuyển giao cho bài toán xác thực người nói với tài nguyên hạn chế.

7. Ý nghĩa khoa học và thực tiễn

Việc nghiên cứu các phương pháp nâng cao chất lượng xác thực người nói tiếng Việt sử dụng mạng học sâu của luận án có ý nghĩa trên cả hai phương diện khoa học và thực tiễn.

Ý nghĩa khoa học:

- Kết quả nghiên cứu của luận án góp phần hoàn thiện cơ sở lý thuyết về nâng cao hiệu năng học sâu trong xác thực người nói tiếng Việt;
- Làm cơ sở cho việc triển khai kỹ thuật xác thực, ứng dụng các phương pháp nhằm nâng cao độ chính xác hệ thống xác thực người nói tiếng Việt.

Thực tiễn:

- Các nội dung nghiên cứu trong Luận án phù hợp với xu hướng nghiên cứu trên thế giới và có các thực tiễn sau:
- Kết quả của các nghiên cứu trong luận án là cơ sở khoa học để xây dựng

và cải tiến các ứng dụng xác thực người nói trong thực tế góp phần hiện đại hóa các ứng dụng tại Việt Nam.

- Các nội dung trong luận án có thể là tài liệu tham khảo trong nghiên cứu khoa học, trong giảng dạy tại Học viện và các khóa huấn luyện chuyên ngành.

8. Bố cục của luận án

Trên cơ sở các nội dung nghiên cứu, luận án được trình bày gồm phần mở đầu, 3 chương chính, kết luận, danh mục công trình khoa học đã công bố, danh mục tài liệu tham khảo. Bố cục của 3 chương chính như sau:

- **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 MFBEs 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 MFBEs với mô hình ECAPA-TDNN cho kết quả tốt hơn so với đặc trưng MFCCs.
- **Chương 3**: Nâng cao độ chính xác xác thực người nói tiếng việt 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.

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

1.1.1. Bài toán xác thực người nói

Trong những năm gần đây, học sâu đã tạo ra một cuộc cách mạng lớn trong xác thực người nói. Ưu điểm chính của học sâu so với các phương pháp thông thường chính là khả năng biểu diễn thông tin. Học sâu có thể tạo ra các đặc trưng nhúng chứa đựng thông tin ở mức cao nên đã giải quyết được những vấn đề còn tồn tại theo các phương pháp truyền thống: chất lượng nhận dạng bị ảnh hưởng bởi tiếng ồn môi trường và miền dữ liệu khác nhau.

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).

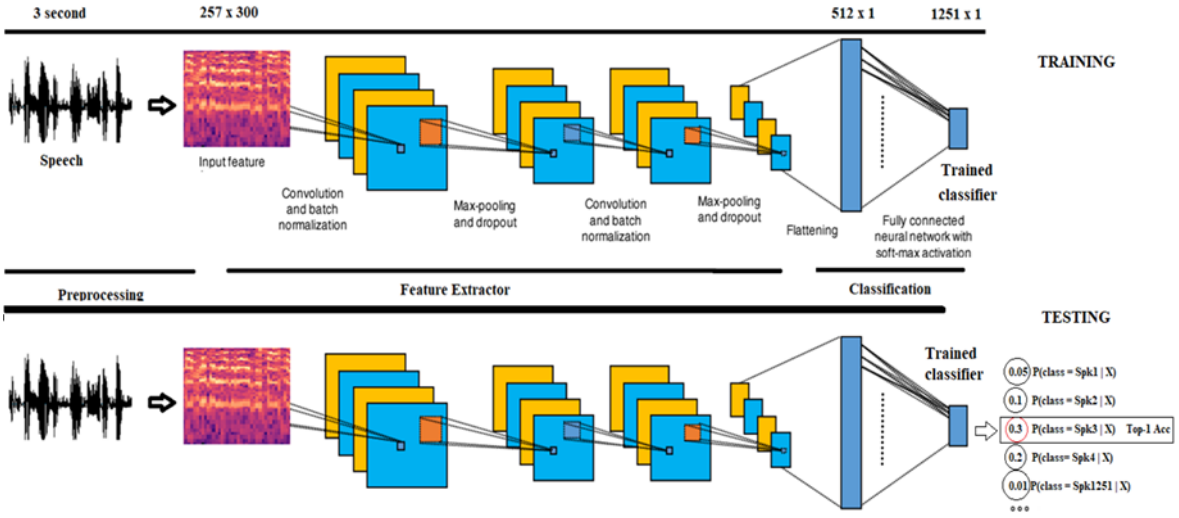
Đầu ra: Một quyết định xác thực để trả lời câu hỏi: ”Người nói kiểm thử có đúng là người đã đăng ký giọng nói mẫu hay không?” Dựa vào kết quả so sánh với ngưỡng, hệ thống sẽ trả về ”đúng” (chấp nhận) hoặc ”sai” (từ chối).

Các mô hình xác thực người nói hiện đại có thể đạt được độ chính xác nhận dạng vượt trội trong điều kiện âm thanh được kiểm soát. Tuy nhiên, trong

môi trường âm thanh thực, xác thực người nói vẫn là một bài toán nhiều thách thức [9][47][85]. Hiệu suất của hệ thống xác thực người nói giảm đáng kể khi tín hiệu giọng nói bị hỏng do các yếu tố gây nhiễu (nhiễu giọng nói, âm thanh nền, độ méo của kênh truyền và âm vang) [3]. Do đó, khả năng chống nhiễu là một yếu tố then chốt khi triển khai hệ thống nhận dạng trong môi trường thực tế.

1.1.2. Những thách thức trong bài toán xác thực người nói

- **Biến đổi giọng nói tự nhiên**: Giọng nói của một người có thể thay đổi theo thời gian do nhiều yếu tố như tuổi tác, sức khỏe, cảm xúc, và môi trường. Điều này gây khó khăn cho các hệ thống xác thực trong việc duy trì độ chính xác. Ngữ điệu và tốc độ nói: Sự biến đổi trong ngữ điệu, tốc độ nói, và ngữ cảnh nói có thể ảnh hưởng đến việc nhận dạng.
- **Chất lượng âm thanh và tiếng ồn**: Tiếng ồn từ môi trường xung quanh, như tiếng gió, tiếng xe cộ, hoặc tiếng người khác, có thể làm giảm chất lượng của tín hiệu giọng nói và ảnh hưởng đến độ chính xác của hệ thống. Sự khác biệt về thiết bị ghi âm (microphone, điện thoại di động) và điều kiện ghi âm có thể tạo ra những biến dạng trong tín hiệu âm thanh.
- **Tính đa dạng của ngôn ngữ và giọng địa phương**: Các hệ thống cần phải xử lý và xác thực giọng nói từ nhiều ngôn ngữ khác nhau, mỗi ngôn ngữ lại có những đặc trưng riêng về âm vị và cấu trúc. Ngay cả trong cùng một ngôn ngữ, phương ngữ và ngữ âm vùng miền có thể rất khác nhau cũng là một thách thức lớn.
- **Bảo mật và chống giả mạo**: Các hệ thống xác thực người nói cần phải đủ mạnh để chống lại các phương thức giả mạo như phát lại ghi âm, biến đổi giọng nói, và tổng hợp giọng nói. Kẻ tấn công có thể tìm cách tạo ra các bản sao giọng nói để đánh lừa hệ thống.
- **Tính khả thi và dễ sử dụng**: Hệ thống cần phải hoạt động tốt trong nhiều tình huống sử dụng thực tế, bao gồm cả các điều kiện môi trường khác nhau và các trạng thái giọng nói khác nhau của người dùng. Hệ thống dễ dàng tích hợp và sử dụng, không gây phiền hà hoặc yêu cầu người dùng phải thực hiện các thao tác phức tạp.
- **Khả năng mở rộng và tính hiệu quả**: Với số lượng người dùng lớn, hệ thống phải đảm bảo khả năng xử lý một lượng lớn dữ liệu giọng nói một cách hiệu quả trong thời gian thực. Hệ thống cần phải tối ưu hóa để giảm thiểu thời gian xử lý và yêu cầu về tài nguyên, đồng thời vẫn đảm bảo độ chính xác cao.



Hình 1.1: *Ví dụ về quá trình huấn luyện và đánh giá hệ thống nhận dạng người nói dựa trên mạng nơ-ron tích chập.*

1.1.3. Ứng dụng của xác thực người nói

Hệ thống xác thực người nói được ứng dụng rộng rãi trong mọi lĩnh vực đời sống:

- **Ngân hàng và tài chính**: Hệ thống xác thực người nói được sử dụng để xác minh danh tính của khách hàng khi họ truy cập vào tài khoản ngân hàng hoặc thực hiện giao dịch tài chính, đảm bảo an ninh và thuận tiện.
- **Chăm sóc sức khỏe**: Xác thực người nói có thể giúp truy cập vào hồ sơ y tế điện tử một cách an toàn, đảm bảo rằng chỉ có những người được ủy quyền mới có thể xem thông tin nhạy cảm của bệnh nhân.
- **Dịch vụ khách hàng**: Các trung tâm chăm sóc khách hàng có thể sử dụng xác thực giọng nói để xác minh danh tính của khách hàng một cách nhanh chóng và hiệu quả, giảm thời gian chờ đợi và tăng hiệu quả chất lượng dịch vụ.

1.2. Các nghiên cứu liên quan

1.2.1. Tình hình nghiên cứu nước ngoài

Nghiên cứu về 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 [18]. 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-vector [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 vectơ 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 dụng trong một số nghiên cứu. Gần đây, mô hình RNN [80][100][126] được phát triển dùng các hệ số MFCCs.

Nhóm tác giả Wang và các cộng sự [118] sử dụng kiến trúc LSTM (Long Short-Term Memory) với đầu vào là đặc trưng MFCCs. 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.

Nhóm tác giả Chung, Nagrani và Zisserman [13] 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 [58].

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-vector [108] làm mô hình cơ

bản sau đó phát triển mở rộng kiến trúc TDNN [82].

Nagrani và các cộng sự [73] đã thực nghiệm trên dữ liệu Voxceleb2 [13], 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 đã diễn ra nhiều cuộc thi tập trung vào kỹ thuật nhận dạng người nói [14][38][73]. 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.

1.2.2. 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 đa nhiệm [84] kết hợp giữa hàm mất mát Triplet [22] 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 [66] 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 [17] 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 đánh giá 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 [21] 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.

1.2.3. Xác thực người nói với tài nguyên dữ liệu 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 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 phương pháp và các nghiên cứu liên quan đến việc giải quyết vấn đề dữ liệu hạn chế trong xác thực người nói:

- **Phương pháp cơ bản**

  Với nguồn dữ liệu huấn luyện hạn chế có thể sử dụng mô hình ECAPA-

TDNN [21], mô hình này có khả năng thích nghi tốt với dữ liệu hạn chế và có nhiễu. Một số nghiên cứu khác sử dụng đặc trưng nhúng người nói từ các mô hình như x-vector [108] và ResNet [74] cho phép chuyển giao các đặc trưng được học từ bài toán khác hoặc dữ liệu tổng quát hơn vào bài toán xác thực người nói với dữ liệu hạn chế.

- **Phương pháp tăng cường dữ liệu**

  Kỹ thuật tăng cường dữ liệu phổ biến SpecAugment [79] 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.

- **Phương pháp học tự giám sát**

  Phương pháp CSSL (Contrastive Self-Supervised Learning) trong học tự giám sát [53] 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. Các mô hình như Wav2vec [8] và HuBERT [35] đã khai thác học tự giám sát trên dữ liệu âm thanh không có nhãn.

- **Phương pháp thích nghi miền**

  Phương pháp thích nghi miền dựa trên mạng nơ-ron sâu sử dụng độ lệch trung bình tối đa và điều chỉnh tính nhất quán cho dữ liệu không gán nhãn trong miền đích. Phép điều chỉnh tính nhất quán khuyến khích các biểu diễn nhúng của người nói trong miền đích bền vững hơn trước những biến đổi bất thường [64] (ví dụ dữ liệu có nhiễu).

- **Phương pháp cải tiến mô hình**

  Các mô hình Siamese [52] 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.

  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.

Với tập dữ liệu người nói tiếng Việt được công bố hiện nay như VLSP2021-SV [17], 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.

1.3. Các cách tiếp cận trong bài toán xác thực người nói

1.3.1. Cách tiếp cận dựa trên thống kê

Trong một thời gian dài, nhận dạng người nói thống trị bởi mô hình GMMs [91] huấn luyện trên tập các vectơ đặc trưng ít chiều. Các mô hình mới nhất gần đây bao gồm cả phân tích hệ số kết hợp và các phương pháp dựa trên mô hình hóa người nói không gian con các kênh và i-vector dự định mô hình cả các không gian con thành không gian nén đơn, ít chiều. Các phương pháp này phụ thuộc vào biểu diễn thấp chiều của âm thanh đầu vào như MFCCs [1]. Tuy nhiên, MFCCs không những bị suy giảm trong môi trường có nhiễu mà nó còn chỉ tập trung vào năng lượng toàn cục của spectral trên các khung ngắn. MFCCs còn thiếu các đặc trưng phân biệt giữa những người nói khác nhau (như thông tin về pitch).

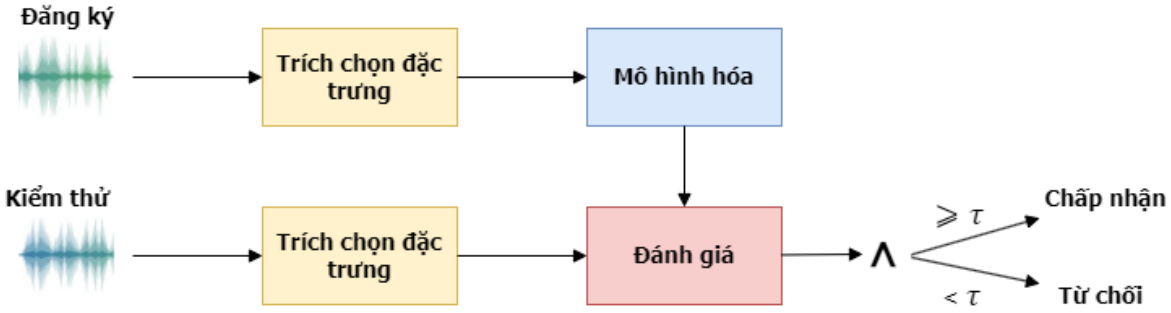
Sau thành công của phương pháp nhúng người nói dựa trên i-vector [18] cho phép triển khai hệ thống xác thực vào các ứng dụng thực tế, nhiều giải pháp đặc trưng nhúng người nói được đề xuất trong [9][85][94]. Thuật ngữ nhúng đề cập đến một "dấu vân tay" duy nhất, là một dạng nén thể hiện mỗi câu nói hoặc đoạn ghi âm và trở thành đặc trưng mức cao cho bài toán phân loại tiếp theo. Đặc trưng nhúng người nói là một phương pháp đơn giản nhưng hiệu quả biểu diễn danh tính của người nói một cách nhỏ gọn dưới dạng vectơ có kích thước cố định.

1.3.2. Cách 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



Hình 1.2: *Sơ đồ tổng quát hệ thống xác thực người nói.*

gian và miền tần số [42] hoặc 1D CNN với tích chập áp dụng cho miền thời gian [29]. 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 vectơ 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 [37]. Trong một số trường hợp, mạng được huấn luyện cho xác thực sử dụng hàm mất mát Contrastive[116] hoặc hàm mất mát triplet [22]. Các độ đo tương tự như cosine [28] hay PLDA [40] 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 xác thực gồm có các thành phần chính: 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, 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, 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 1.2 có thể được áp dụng cho các phương pháp truyền thống và 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

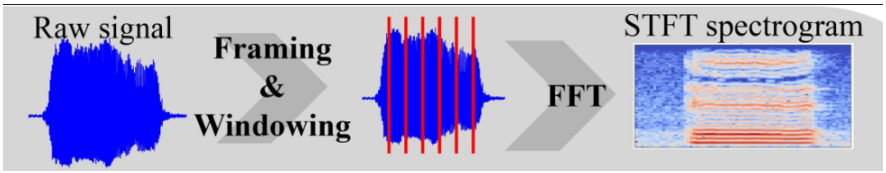
1.4.1.1. Đặc trưng MFCCs

MFCCs là một trong những đặc trưng thường được sử dụng trong nhiều ứng dụng, đặc biệt là trong xử lý tín hiệu giọng nói như nhận dạng người nói, nhận dạng giọng nói, và phân biệt giới tính. MFCC có các bước bao gồm: phân khung tín hiệu, tính toán phổ công suất, áp dụng một bộ lọc Mel lên các phổ công suất đã thu được, tính toán giá trị logarit của tất cả các bộ lọc, và cuối cùng là áp dụng biến đổi cosine rời rạc DCT (Discrete Cosine Transform). MFCCs đặc trưng nổi tiếng nhất đã được các nhà nghiên cứu sử dụng rộng rãi trong các ứng dụng nhận dạng người nói nhờ khả năng nắm bắt tính chất lặp đi lặp lại và hiệu quả của tín hiệu giọng nói [41], [71]. Trong Chương 2, NCS sẽ trình bày chi tiết về kỹ thuật trích chọn đặc trưng MFCCs.

1.4.1.2. Đặc trưng phổ

Ảnh phổ là biểu diễn hình ảnh của phổ tần số tín hiệu âm thanh theo thời gian. Nó cho phép ta quan sát sự thay đổi của các tần số trong tín hiệu âm thanh qua các khoảng thời gian khác nhau. Ảnh phổ thường được sử dụng trong nhiều lĩnh vực như xử lý tín hiệu âm thanh, âm nhạc, và ngôn ngữ học. Hình 1.3 minh họa biến đổi âm thanh thành ảnh phổ.

Quá trình trích chọn phổ âm thanh từ một tín hiệu âm thanh gồm các bước chính sau đây:



Hình 1.3: *Các bước trích chọn ảnh phổ từ âm thanh ban đầu [54].*

1. **Chia tín hiệu thành các khung thời gian**: Tín hiệu âm thanh liên tục được chia thành các đoạn ngắn hơn, gọi là khung. Mỗi khung thường có độ dài từ 20ms đến 40ms, với độ dài phổ biến là 25ms. Mục tiêu của việc chia khung là để đảm bảo sự ổn định của tín hiệu trong mỗi khung, vì các đặc trưng tần số của âm thanh có thể thay đổi theo thời gian.
2. **Áp dụng cửa sổ**: Trước khi thực hiện biến đổi Fourier, một hàm cửa sổ như Hamming hoặc Hanning được áp dụng cho mỗi khung. Điều này giúp giảm các vấn đề như hiện tượng rò rỉ phổ (spectral leakage), khi các tần số không mong muốn xuất hiện trong kết quả.
3. **Biến đổi Fourier nhanh**: Mỗi khung sau đó được biến đổi từ miền thời gian sang miền tần số bằng cách sử dụng biến đổi Fourier nhanh (FFT). Kết quả của FFT là một phổ tần số (spectral magnitude) cho mỗi khung, thể hiện các thành phần tần số có trong tín hiệu và cường độ của chúng.
4. **Tính công suất phổ**: Công suất phổ thường được tính bằng cách lấy bình phương biên độ của phổ tần số. Điều này cung cấp thông tin về năng lượng của các tần số trong mỗi khung.
5. **Biểu diễn log**: Do tai người cảm nhận âm thanh theo thang logarit, các giá trị công suất phổ thường được chuyển đổi sang thang logarit để phản ánh cách mà con người cảm nhận âm thanh. Kết quả này gọi là log-magnitude spectrogram.
6. **Ghép nối khung để tạo ảnh phổ**: Các phổ tần số của các khung được sắp xếp theo thứ tự thời gian để tạo thành một ma trận hai chiều, với trục x đại diện cho thời gian và trục y đại diện cho tần số. Độ sáng hoặc màu sắc trong ma trận này biểu thị cường độ của các thành phần tần số tại mỗi thời điểm, tạo thành một hình ảnh phổ hoàn chỉnh.

1.4.1.3. Tín hiệu thô

Một trong những lợi thế chính của việc học đặc trưng dựa trên dạng sóng thô là nó tránh được nhu cầu thiết kế đặc trưng thủ công. Các phương pháp

truyền thống để trích xuất đặc trưng thường bao gồm các đặc trưng được thiết kế thủ công như MFCCs, GFCCs [112], LPCCs hoặc PLP. Các đặc trưng này được thiết kế để chuyển đổi dạng sóng thô thành biểu diễn gọn nhẹ bằng cách tóm tắt các đặc trưng phổ hoặc thời gian cụ thể. Tuy nhiên, sự chuyển đổi này không thể tránh khỏi việc loại bỏ một số chi tiết và có thể gây ra một mức độ mất thông tin nhất định. Hơn nữa, những đặc trưng này không phải lúc nào cũng tối ưu cho một bài toán nhất định, và quá trình thiết kế và lựa chọn đặc trưng cũng tốn thời gian và dễ gây lỗi [51]. Mạng nơ-ron học cách trích xuất các đặc trưng tối ưu cho một bài toán cụ thể mà không cần đến việc thiết kế đặc trưng thủ công. Điều này có thể dẫn đến kết quả chính xác hơn và sử dụng hiệu quả hơn các tài nguyên tính toán. Một lợi thế khác của việc học đặc trưng này là nó có thể được sử dụng với bất kỳ loại tín hiệu âm thanh nào, bất kể nguồn gốc hay đặc điểm của nó [32]. Kỹ thuật này trở nên linh hoạt và được sử dụng trong một loạt các ứng dụng. Kỹ thuật này cũng còn được gọi là học từ mô hình đầu vào đến đầu ra vì toàn bộ quy trình nhận dạng giọng nói và người nói được thực hiện thông qua các mạng nơ-ron sâu mà không cần trích xuất đặc trưng trước đó.

1.4.2. Mô hình hóa người nói

1.4.2.1. GMM

Mô hình Gaussian hỗn hợp là một mô hình xác suất [91] trong đó các tập dữ liệu được giả định là được hình thành bởi sự pha trộn của một số lượng cố định các phân phối Gaussian với các biến không chắc chắn. Các mô hình hỗn hợp có thể được coi là mở rộng của phương pháp phân cụm k-means để cung cấp thông tin chi tiết về cấu trúc hiệp phương sai của dữ liệu và các trung tâm của các phân phối Gaussian chưa được khám phá. Đây là một hàm gồm nhiều phân phối Gaussian, mỗi phân phối được định nghĩa bởi $k \in \{1, 2, \ldots, K\}$ trong đó K là số lượng cụm trong tập dữ liệu. Các tham số sau đây đặc trưng cho mỗi Gaussian k trong hỗn hợp:

- μ - trung bình với một trung tâm xác định,
- σ - hiệp phương sai xác định độ rộng của nó.Trong trường hợp đa biến, điều này sẽ tương tự như các phép đo của một ellipsoid,
- Xác suất kết hợp xác định kích thước của hàm Gaussian.

Các yếu tố kết hợp là xác suất phải thỏa mãn điều kiện sau:

$$\sum_{k=1}^{K} \pi_k = 1 \tag{1.1}$$

Để làm được điều này, chúng ta phải đảm bảo rằng mỗi Gaussian khớp với các tập dữ liệu trong mỗi cụm. Việc khớp này chính xác là điều mà việc tối đa hóa xác suất đạt được. Hàm mật độ Gaussian được biểu diễn bởi:

$$N(x|\mu, \Sigma) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)}{(2\pi)^{D/2}|\Sigma|^{1/2}} \tag{1.2}$$

trong đó, x đại diện cho các điểm dữ liệu, D là số chiều của mỗi điểm dữ liệu, μ và σ lần lượt là trung bình và hiệp phương sai. Việc tính toán log của công thức 1.3 đã được tìm thấy là quan trọng. Phép tính toán học có thể được cho như sau:

$$\ln N(x|\mu, \Sigma) = -\frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) \tag{1.3}$$

GMM là một phương pháp hữu ích thường được sử dụng cho nhiều bài toán dựa trên phân cụm. Hàm xác suất phổ biến nhất cho nhận dạng giọng nói độc lập với nội dung sử dụng các đặc trưng liên tục, trong đó không có thông tin biết trước về những gì người nói sẽ nói.

1.4.2.2. GMM thích nghi

GMM thích nghi là mô hình phổ biến trong việc trích chọn đặc trưng i-vector [18]. Ý tưởng chính của mô hình này là thích ứng một mô hình hỗn hợp Gaussian (GMM) đã được huấn luyện trước, thường gọi là Universal Background Model (UBM), với dữ liệu cụ thể của người nói. Dưới đây là các bước cơ bản của mô hình thích nghi GMM:

- **Huấn luyện UBM**: Đầu tiên, một UBM được huấn luyện trên một tập dữ liệu lớn bao gồm nhiều người nói khác nhau. UBM biểu diễn các đặc trưng phổ biến trong âm thanh và được sử dụng như một mô hình nền,
- **Thích nghi**: Khi có dữ liệu từ một người nói cụ thể, các thông số của UBM được điều chỉnh để phù hợp hơn với dữ liệu của người nói đó. Quá trình thích ứng này sử dụng kỹ thuật MAP (Maximum A Posteriori) để thay đổi các thông số của các thành phần Gaussian trong mô hình dựa trên dữ liệu mới,
- **Trích chọn đặc trưng**: Sau khi thích nghi, các vectơ đặc trưng, như

i-vector, được trích xuất từ mô hình GMM đã thích ứng. Những vectơ này sau đó có thể được sử dụng để so sánh, nhận dạng, hoặc xác thực người nói.

1.4.2.3. Phân tích nhân tố kết hợp

Mô hình phân tích nhân tố kết hợp (JFA - Joint Factor Analysis) [50] được xây dựng bằng cách kết hợp cả giá trị riêng của câu nói và giá trị riêng của kênh lại với nhau. Mô hình này giả định rằng cả sự biến thiên của người nói và kênh đều nằm trong các không gian con chiều thấp hơn của không gian siêu vectơ GMM. Các không gian con này được trải rộng bởi các ma trận V và U, như trước đây. Mô hình giả định rằng, đối với một phát ngôn được chọn ngẫu nhiên từ người nói s và phiên h, siêu vectơ trung bình của GMM của nó có thể được biểu diễn bởi:

$$\mathbf{m}_{s,h} = \mathbf{m}_0 + \mathbf{U}\mathbf{x}_h + \mathbf{V}\mathbf{y}_s + \mathbf{D}\mathbf{z}_{s,h} \tag{1.4}$$

Trong đó:

- $\mathbf{m}_{s,h}$ là vectơ trung bình của người nói s trong phiên h.
- $\mathbf{m}_0$ là vectơ trung bình toàn cục, đại diện cho thông tin cơ bản hoặc trung bình của toàn bộ người nói trong tập dữ liệu.
- $\mathbf{U}\mathbf{x}_h$: thành phần phụ thuộc vào phiên nói h. U là ma trận và $\mathbf{x}_h$ là vectơ chứa các đặc trưng của phiên h. Thành phần này mô tả sự biến đổi theo phiên.
- $\mathbf{V}\mathbf{y}_s$: thành phần phụ thuộc vào người nói s. V là ma trận và y_s là vectơ chứa các đặc trưng người nói s. Thành phần này mô tả sự biến đổi liên quan đến người nói.
- $\mathbf{D}\mathbf{z}_{s,h}$: thành phần nhiễu, với D là ma trận và $z_{s,h}$ là vectơ chứa các biến ngẫu nhiên đại diện cho các yếu tố nhiễu của người nói và phiên nói.

1.4.2.4. i-vector

i-vector [18] được áp dụng để giảm không gian chiều cao xuống không gian chiều thấp cho các biến đổi của người nói và kênh bằng cách phân tích nhân tố đơn giản. Thay vì sử dụng hai không gian khác nhau, phương pháp này chỉ sử dụng một không gian duy nhất chứa cả người nói và kênh, được gọi là "không gian biến đổi tổng thể". Siêu vectơ GMM mới được định nghĩa bởi: $M = m + Tw$ trong đó m là siêu vectơ độc lập với người nói và kênh, và w là nhân tố tổng

thể với các vectơ phân phối chuẩn chuẩn được gọi là i-vector. Với một câu nói cố định, i-vector được sử dụng để biểu diễn tín hiệu giọng nói thông qua phân phối hậu nghiệm. Sau đó, các thống kê Baum-Welch được trích xuất bằng cách sử dụng UBM để ước lượng i-vector thông qua các thống kê sau:

$$N_c = \sum_{t=1}^{L} P(c \mid y_t, \Omega) \tag{1.5}$$

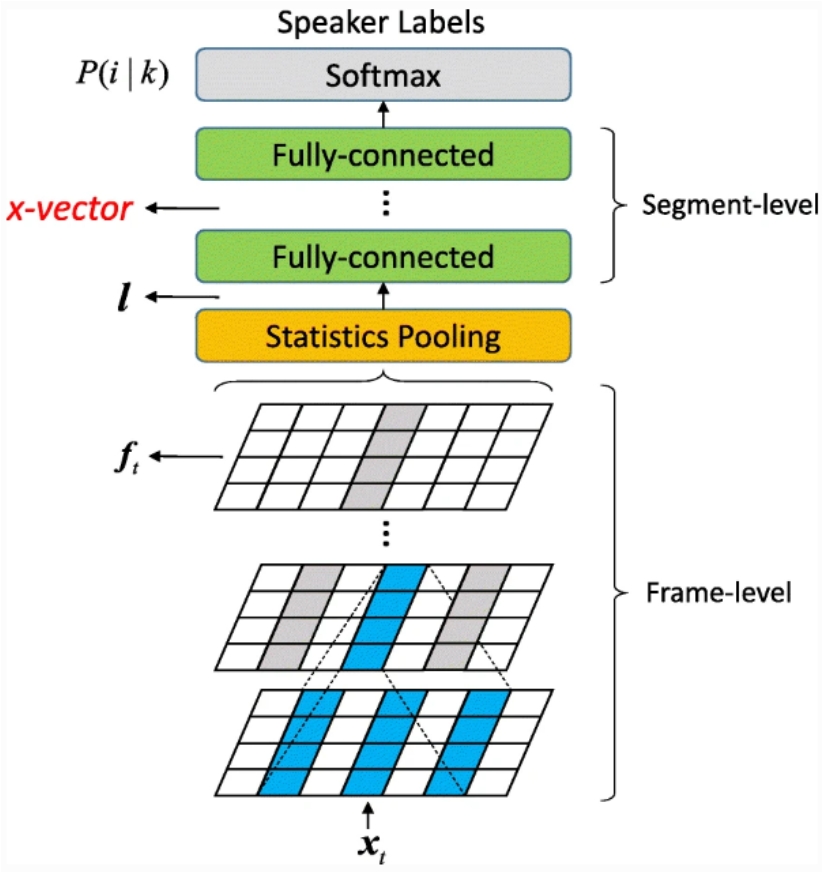
$$F_c = \sum_{t=1}^{L} P(c \mid y_t, \Omega) y_t \tag{1.6}$$

$$\tilde{F}_c = \sum_{t=1}^{L} P(c \mid y_t, \Omega)(y_t - m_c) \tag{1.7}$$

trong đó $[y_1, y_2, \ldots, y_L]$ đại diện cho một chuỗi gồm L khung, được tạo thành bởi C thành phần hỗn hợp trong không gian đặc trưng F. $P(c \mid y_t, \Omega)$ biểu thị xác suất hậu nghiệm của vectơ trong đó $c = 1, 2, ..., C$ là chỉ số Gaussian. Sau đó, các thống kê Baum-Welch bậc nhất được tính toán trên các thành phần hỗn hợp trung bình của UBM và ước lượng i-vector.

1.4.2.5. x-vector

Hệ thống x-vector [108] là một phương pháp dựa trên mạng nơ-ron sâu cho bài toán xác thực người nói, trong đó các mạng nơ-ron được huấn luyện để phân biệt giữa những người nói khác nhau. Mô hình này được phát triển nhằm giải quyết các hạn chế của các phương pháp truyền thống như i-vector. x-vector đã trở thành một kỹ thuật quan trọng và phổ biến nhờ khả năng trích xuất đặc trưng mạnh mẽ, ổn định, phù hợp cho các bài toán nhận dạng và xác thực trong những điều kiện phức tạp. Phương pháp này [107] tuân theo hệ thống từ đầu vào cho đến đầu ra [105] tạo ra các đặc trưng nhúng kết hợp với độ đo tương đồng bằng cách sử dụng mạng nơ-ron sâu có độ trễ thời gian và so sánh chúng bằng một bộ phân loại được huấn luyện riêng biệt như PLDA. Đầu tiên, ngữ cảnh ngắn hạn của khung hiện thời trích xuất theo độ trễ thời gian. Sau đó, một lớp gộp thống kê tổng hợp trên đoạn đầu vào và tính toán trung bình và độ lệch chuẩn. Cuối cùng, tính toán này phân loại đặc trưng mức đoạn cho người nói bằng DNN. Đặc trưng nhúng người nói mức đoạn là x-vector. Hình 1.4 biểu diễn cấu trúc mạng của kiến trúc x-vector.



Hình 1.4: *Kiến trúc x-vector [65].*

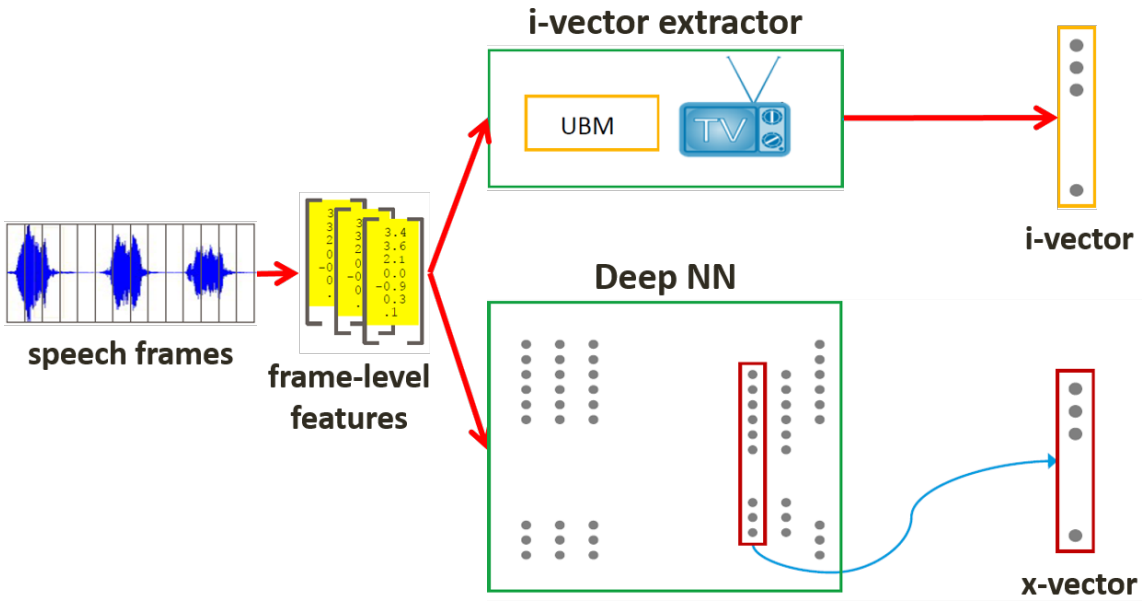
Sự khác biệt chính giữa hai mô hình i-vector và x-vector ở những điểm sau:

- **Phương pháp trích chọn đặc trưng**: i-vector sử dụng GMM, trong khi x-vector sử dụng mạng nơ-ron sâu
- **Khả năng biểu diễn**: x-vector thường biểu diễn đặc trưng phức tạp và chính xác hơn so với i-vector, đặc biệt khi đối mặt với sự biến đổi lớn về âm thanh
- **Hiệu năng**: x-vector thường có hiệu năng tốt hơn trong các ứng dụng nhận diện giọng nói nhờ khả năng học đặc trưng của DNN.

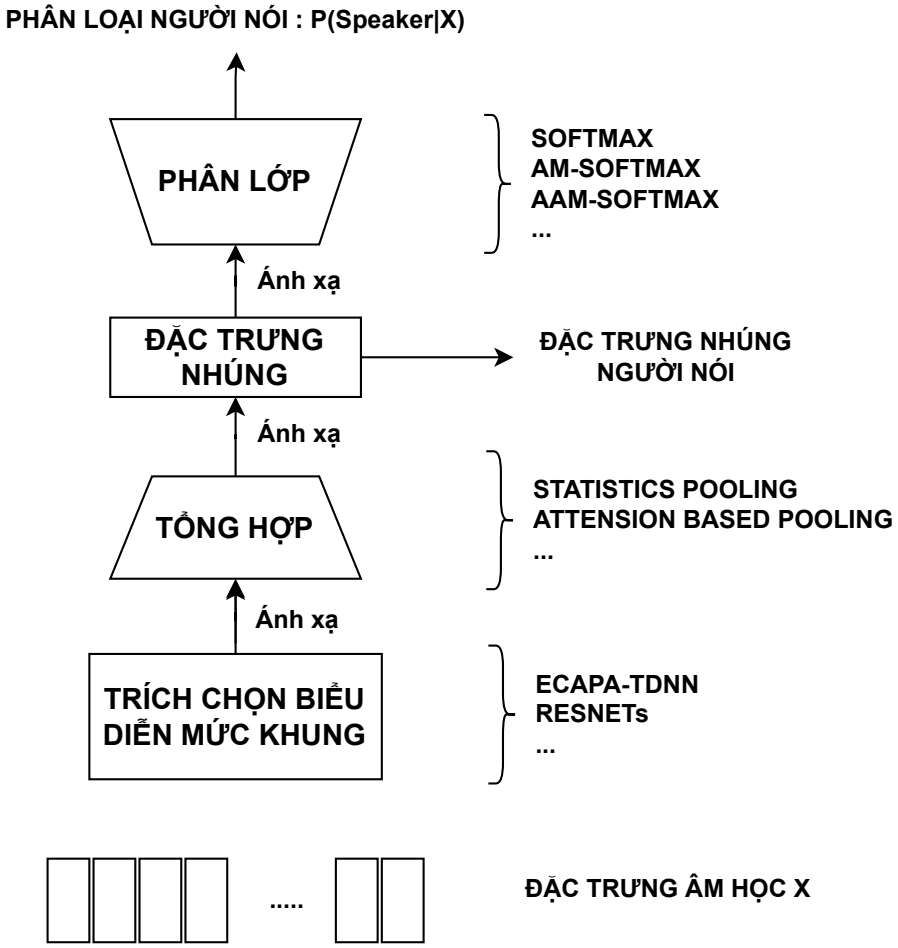
1.4.2.6. Mô hình RawNet

Mô hình RawNet [44] là một kiến trúc mạng nơ-ron sâu được thiết kế đặc biệt cho việc xử lý âm thanh. RawNet được phát triển để trực tiếp xử lý dữ liệu âm thanh "thô", tức là dữ liệu âm thanh được lấy trực tiếp từ microphone mà không cần chuyển đổi thành các đặc trưng thông tin trước.

Điểm đặc biệt của RawNet là khả năng xử lý trực tiếp dữ liệu âm thanh thô mà không cần chuyển đổi thành các biểu diễn khác như MFCCs hay ảnh



Hình 1.5: *Sự khác nhau giữa hai mô hình i-vector và x-vector.*



Hình 1.6: *Mô hình nơ-ron sâu cho đặc trưng nhúng người nói [108].*

phổ. Điều này giúp giảm bớt các bước tiền xử lý và tiêu tốn thời gian của quá trình xử lý âm thanh. Một trong những ứng dụng phổ biến của RawNet là trong lĩnh vực nhận dạng tiếng nói và xử lý ngôn ngữ tự nhiên, nơi mà việc xử lý trực tiếp dữ liệu âm thanh thô có thể cải thiện hiệu năng và độ chính xác của hệ thống.

1.4.2.7. Mô hình nơ-rơn sâu dựa trên đặc trưng nhúng người nói [108].

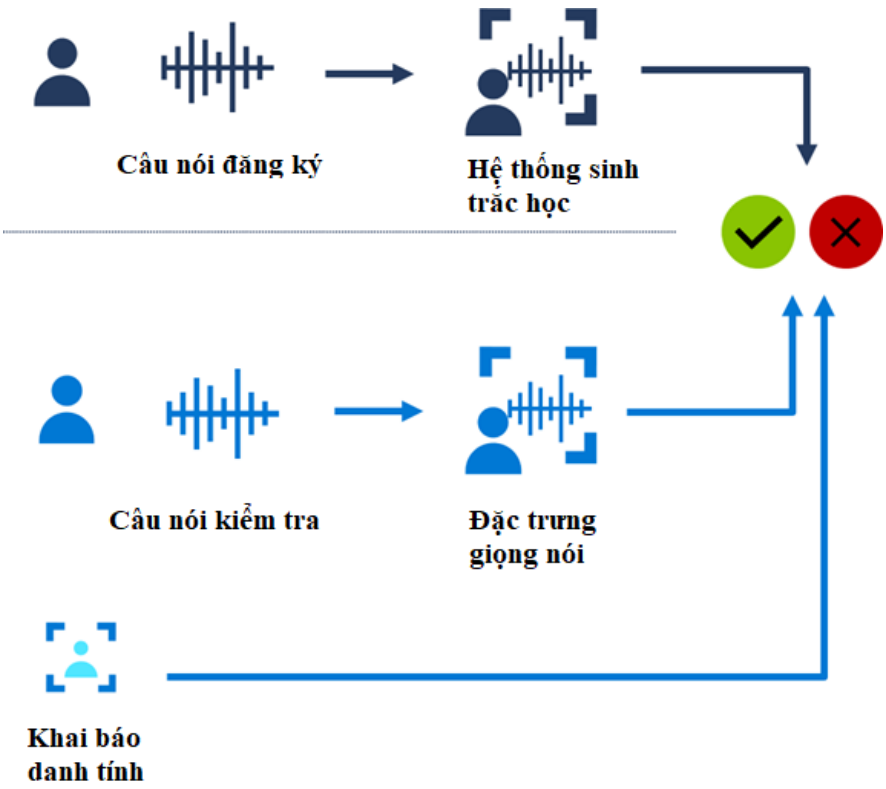
Trong Hình 1.6, đặc trưng nhúng người nói là các vectơ đặc trưng được trích xuất từ đoạn âm thanh của một người nói, các vectơ này đại diện cho đặc điểm giọng nói của người đó. Các vectơ chứa thông tin về các đặc trưng đặc thù của giọng nói, như âm lượng, tần số, và cách phát âm, giúp phân biệt người nói với nhau. Đặc trưng nhúng người nói sử dụng phổ biến trong các hệ thống định danh và xác thực người nói. Khi một đoạn âm thanh mới được nhập vào, hệ thống sẽ trích chọn đặc trưng nhúng từ đoạn âm thanh đó và so sánh nó với các đặc trưng nhúng đã lưu trữ trước đó để xác định người nói.

1.4.3. Đánh giá

Sau quá trình trích chọn đặc trưng và huấn luyện mô hình để có được bộ phân loại người nói khác nhau. Hệ thống xác thực người nói hoạt động theo Hình 1.7. Mô đun đánh giá biến đổi tiếng nói đầu vào thành biểu diễn đặc trưng người nói. Mô hình đánh giá tính toán độ tương tự giữa hai đặc trưng nhúng và xác định xem hai đặc trưng đó có cùng một người nói không. Hiện có hai phương pháp đánh giá điểm số: đánh giá dựa trên phân tích phân biệt tuyến tính xác suất PLDA [40] và đánh giá cosine [28] với giả thiết các đặc trưng nhúng có sự phân biệt.

1.4.3.1. Phân tích phân biệt tuyến tính xác suất

PLDA [40] là một biến thể xác suất của phân tích phân biệt tuyến tính có thể chứa các tập dữ liệu phức tạp hơn. PLDA có một loạt các ứng dụng trong một số lĩnh vực nghiên cứu, cụ thể là thị giác máy tính, nhận dạng giọng nói, ... Ngay cả đối với một ví dụ về một lớp không xác định, PLDA tạo ra một trung tâm lớp bằng cách sử dụng các tham số phi tuyến rời rạc. Các nhà nghiên cứu xem xét các trường hợp khác nhau của một lớp chưa biết trước đây trong phân tích thống kê để xem liệu chúng có liên quan đến cùng một loại hay không - nó cũng phân cụm các nghiên cứu từ các nhóm chưa từng thấy trước đây. PLDA



Hình 1.7: *Giai đoạn xác thực người nói.*

là một lý thuyết tổng hợp bao gồm các tập dữ liệu nhất định được rút ra từ một phân phối. Trong PLDA, các tham số mô hình đại diện tốt nhất cho dữ liệu huấn luyện phải được xác định. Hai yếu tố xác định cách biểu diễn nơi dữ liệu được cho là thu được: Nó phải phản ánh nhiều kiểu dữ liệu khác nhau và việc xử lý tham số phải dễ dàng và nhanh chóng. Gaussian là đại diện phổ biến nhất đáp ứng các yêu cầu này.

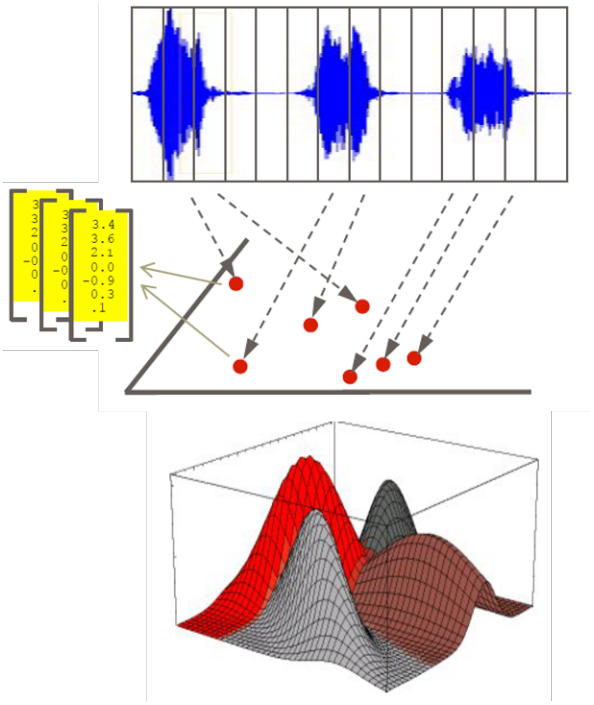
PLDA Gaussian điển hình là i-vector w:

$$\mathbf{w} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{z} \tag{1.8}$$

trong đó m trung bình i-vector, y là biến tiềm ẩn người nói với xác suất chuẩn và residual. Z là phân bố chuẩn với trung bình bằng 0 và ma trận hiệp phương sai đầy đủ $\sum_z$ PLDA dùng thuật toán ước lượng cực đại EM để ước lượng tham số mô hình (V, $\sum_z$)

Tổng quát hóa sau đây, điểm số xác thực mỗi cặp vectơ kiểm thử w_1 và w_2 tính bằng tỉ lệ phù hợp của giả thuyết H_s, mà cả i-vector cùng một người nói và H_d giả thiết là cả hai người khác nhau, công thức toán học biểu diễn là:

$$\text{điểm xác thực} = \log\left(\frac{p(w_1, w_2|H_s)}{p(w_1, w_2|H_d)}\right) \tag{1.9}$$



Hình 1.8: *Mô hình GMM.*

Điểm số PLDA tính bằng:

$$PLDA\_score = \log N\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \begin{bmatrix} m \\ n \end{bmatrix}, \begin{bmatrix} S_T & S_B \\ S_B & S_T \end{bmatrix}\right) - \log N\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \begin{bmatrix} m \\ n \end{bmatrix}, \begin{bmatrix} S_T & 0 \\ 0 & S_T \end{bmatrix}\right) \tag{1.10}$$

trong đó $S_B = W_T$ và $S_T = S_T + \Sigma_z$

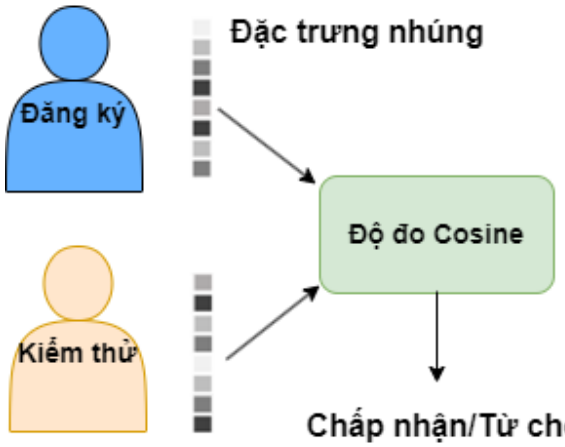
So với đánh giá cosine, PLDA có khả năng học tham số và bổ sung các nhãn người nói trong quá trình huấn luyện. Vì vậy, PLDA thường hiệu quả hơn trong biểu diễn sự phân biệt giữa những người nói khác nhau.

Ưu điểm của PLDA:

- Có thể sinh được tâm dùng hàm phi tuyến liên tục thậm chí cho cả các mẫu không nhìn thấy tên lớp,
- Trong giả thiết đánh giá, có thể so sánh hai mẫu (không biết tên lớp) để xác định chúng có thuộc cùng một lớp hay không,
- Có thể phân cụm các mẫu (mà không nhìn thấy tên lớp) PLDA cho phép suy luận các lớp không xuất hiện trong quá trình huấn luyện.

1.4.3.2. Khoảng cách Cosine

Khoảng cách cosine [86] tính theo độ tương tự cosine. Độ tương tự cosine được định nghĩa là độ tương tự giữa hai vectơ khác không. Nó tính cosine của góc giữa hai vectơ trong không gian đa chiều. Mối quan hệ giữa độ tương tự



Hình 1.9: *Minh họa tính độ tương tự giữa hai đặc trưng nhúng của câu nói đăng ký và câu nói đánh giá.*

cosine với khoảng cách cosine là không cân xứng. Độ tương tự cosine tăng dần trong khi khoảng cách giữa các vectơ 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 phương trình sau:

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

$$\text{cosine}_{\text{distance}} = 1 - \cos(\theta) \tag{1.12}$$

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

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

1.5.1. 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. Để tăng cường hiệu năng xác thực người nói với dữ liệu hạn chế, [99] đã thay đổi tốc độ và cao độ cũng như biến đổi giọng nói để tăng dữ liệu huấn

Bảng 1.1: *So sánh tỉ lệ lỗi trong xác thực người nói sử dụng các đặc trưng đầu vào khác nhau [46].*

Công bố	Đặc trưng đầu vào	EER (%)
Desplanques et al. [21]	MFCC	0.87
Ravanelli et al. [90]	Mel Filter bank	0.69
Kuzmin et al. [59]	Mel Filter bank	0.66
Zhu et al. [135]	âm thanh thô	2.6
Li et al. [63]	âm thanh thô	2.31
Kim et al. [55]	âm thanh thô	1.95
Jee-weon Jung (Stride = 10) [48]	âm thanh thô	1.29
Jee-weon Jung (Stride = 16) [48]	âm thanh thô	0.89

luyện. Bên cạnh đó [119] cũng ghi nhận hiệu quả tăng cường quang phổ trong xác thực người nói.

- Biến đổi trên miền tần số của ảnh phổ,
- Thêm nhiễu âm thanh,
- Độ vang âm thanh,
- Tăng cường dữ liệu cho mô hình nhận dạng người nói.

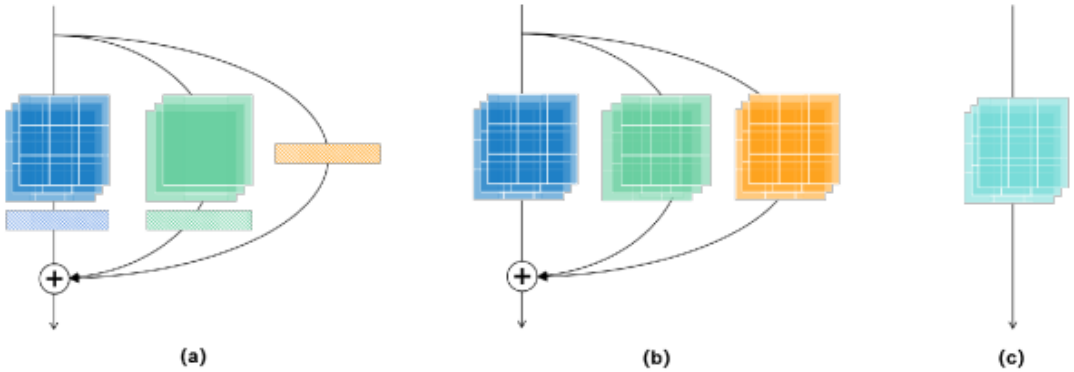
Đóng góp của tăng cường dữ liệu vô cùng quan trọng, nó cho phép đạt được hiệu năng cao tương tự như chúng ta sử dụng các tập dữ liệu lớn có nhiều người nói.

1.5.2. 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 [13], [15], 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



Hình 1.10: *Kiến trúc RepVGG. (a) trạng thái huấn luyện. (b) biểu diễn quá trình conv-bn fusion. (c) trạng thái suy luận. ⊕ hệ số bổ sung [68].*

Bảng 1.2: *Mô tả kiến trúc ResNet34. Đầu vào là $H \times W$, H biểu diễn số chiều đặc trưng, W biểu diễn độ dài frame, S là bước dịch chuyển, K là kích thước nhân, C là số kênh.*

Lớp	Tham số	Đầu ra
stem	$C = 32, K = 3, S = 1$	$32 \times H \times W$
Res1	$C = 32, K = 3, S = 1$	$32 \times H \times W$
Res2	$C = 64, K = 3, S = 2$	$64 \times \frac{H}{2} \times \frac{W}{2}$
Res3	$C = 128, K = 3, S = 2$	$128 \times \frac{H}{4} \times \frac{W}{4}$
Res4	$C = 256, K = 3, S = 2$	$64 \times \frac{H}{8} \times \frac{W}{8}$

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%.

1.5.3. Cải tiến mô hình

1.5.3.1. ***RepVGG***

RepVGG đề xuất xây dựng mạng nơ-ron tích chập. Phương pháp này gọi là kỹ thuật phân loại tham số hóa. Phương pháp này tách rời phần huấn luyện và phần suy luận.

1.5.3.2. ***ResNet-34***

ResNet [31] là một trong những mạng nơ-ron tích chập cổ điển nhất, ResNet chứng tỏ được khả năng mạnh mẽ trong các bài toán xác thực.

Các thực nghiệm nghiên cứu [133] cũng đã khảo sát trên cả các kiến trúc Resnet-34, ResNet-101 và ResNet-152. Kết quả thử nghiệm trong [133] cụ thể như sau:

Bảng 1.3: *Một số thực nghiệm sử dụng các phương pháp tổng hợp khác nhau [67].*

	Hệ thống	**Tỉ lệ lỗi trên Vox1-O (%)**
S1	ResNet34-**MHA**+ASTD-AM-FBANK	0.904
S2	SEResNet34-**MHA**-AM-FBANK	0.93
S3	SEResNet34-**FPN**-AM-FBANK	0.874
S4	TDNN-RES-SE-**TAP**-MFCC	0.970
S5	TDNN-RES-SE-**SAP**-MFCC	0.996
S6	TDNN-RES-SE-**TAP**-PLP	1.032

- Tỉ lệ lỗi giảm từ 3.13% xuống 2.785%.
- Thêm điểm phạt Inter-TopK, EER là 2.58%
- Sử dụng MQMHA (q = 4, h = 16) thay cho MHA, EER là 2.51%
- Áp dụng AS-Norm EER đạt 1.8367%

1.5.4. Cải tiến phương pháp tổng hợp trong mạng học sâu

1.5.4.1. Tổng hợp tự chú ý

Tổng hợp tự chú ý (Self-Attentive Pooling) là một phương pháp giúp mô hình xác định và tập trung vào các phần quan trọng nhất trong một chuỗi dữ liệu (như âm thanh hoặc văn bản) bằng cách tính trọng số cho từng phần tử trong chuỗi đó. Trong bối cảnh xử lý ngôn ngữ tự nhiên và nhận dạng người nói, cơ chế tự chú ý được sử dụng để cải thiện chất lượng các đặc trưng đầu ra, chẳng hạn như việc xác định đặc điểm nổi bật của giọng nói hoặc nội dung lời nói.

Phép tổng hợp SAP ban đầu sẽ xác định các khung quan trọng của đầu vào, sau đó tính toán trung bình có trọng số và độ lệch chuẩn có trọng số theo từng khung quan trọng. Số chiều của tham số chú ý trong các hệ thống sử dụng SAP là 128.

1.5.4.2. Multi Head Attention Pooling

The Multi Head Attention Pooling (MHA) [39] đầu tiên chia số chiều đặc trưng thành N head, tính toán trọng số trung bình của các head khác nhau, sau đó ghép nối trọng số trung bình như là thông tin tổng hợp mức phát âm. Kết quả thực nghiệm cho thấy phương pháp tổng hợp MHA cho kết quả tốt nhất trên tập dữ liệu Vox1-O. Như vậy việc thực nghiệm các phương pháp tổng hợp

Bảng 1.4: *Kết quả thực nghiệm trên các hàm mất mát khác nhau. Dữ liệu test Voxceleb1 [72].*

Config.	Loss	Aug.	BN	EER %	MinDCF
FR-34	AP	×	×	2.22	-
Sys 1	Softmax †	✓	×	-	-
Fusion	-	-	-	-	-
Sys A5	AM-Softmax	✓	×	-	-
Fusion	-	-	-	-	-
Q / SAP	AM-Softmax	✓	✓	2.20	0.139
	AAM-Softmax	✓	✓	2.19	0.138
	AP	✓	✓	2.02	0.133
	AP+Softmax	✓	✓	1.85	0.119
H / ASP	AM-Softmax	✓	✓	1.64	0.113
	AAM-Softmax	✓	✓	1.59	0.115
	AP	✓	✓	1.50	0.126
	AP+Softmax	✓	✓	1.18	**0.086**
H / ASP	AM-Softmax	✓	✓	1.49	0.097
	AAM-Softmax	✓	✓	1.48	**0.086**
	AP	✓	✓	1.43	0.113
	AP+Softmax	✓	✓	1.25	0.087

khác nhau trên các tập dữ liệu khác nhau cũng là một trong các cách để làm tăng hiệu năng trong hệ thống xác thực người nói.

1.5.5. Cải tiến hàm mất mát

Hàm mất mát vô cùng quan trọng trong quá trình học đặc trưng sâu. Các loại hàm mất mát sử dụng rộng rãi trong nhận dạng người nói như : Additive margin softmax [117], Additive angular margin softmax (AAM-softmax) [20]. Hàm mất mát AM-softmax và AAM-softmax cũng được dùng trong nhận dạng khuôn mặt và hiệu quả trong nhận dạng người nói. Các hàm mất mát này đưa ra khái niệm "margin" giữa các lớp mà margin biến thiên nội tại trong các lớp. Cả hai hàm mất mát AM-Softmax và AAM-Softmax sử dụng margin 0.2 và thang 30 đạt kết quả tốt nhất trên tập dữ liệu đánh giá VoxCeleb1 [72]. Trong bảng kết quả thực nghiệm cho thấy sử dụng hàm mất mát AP và softmax cho tỉ lệ lỗi thấp nhất 1.18% trên tập dữ liệu đánh giá VoxCeleb1. Kết quả trên đây cũng là cơ sở NCS khảo sát, thực nghiệm trên dữ liệu khác như VLSP2021-SV và VLSP2022-SV.

1.5.6. Cải tiến trong giai đoạn đánh giá

Chuẩn hóa điểm số

Mục tiêu của việc chuẩn hóa điểm số là giảm sự thay đổi trong quá trình đánh giá nhằm nâng cao hiệu suất, chuẩn hóa tốt hơn và thiết lập ngưỡng đáng tin cậy hơn [70]. Phần này mô tả phần lớn các kỹ thuật chuẩn hóa điểm số. Dưới đây, điểm số giữa câu nói đăng kí e và câu nói đánh giá t được ký hiệu là $s(e,t)$.

z-score

Chuẩn hóa điểm số Zero dùng phân bố điểm số mạo danh cho đoạn âm thanh đăng kí. Nó sử dụng cohort $\mathcal{E} = \{\epsilon_i\}_{i=1}^{N}$ với N người nói mà chúng tôi giả thiết là khác với người nói trong câu nói e và t. Điểm số cohort là :

$$S_e = \{s(e,\varepsilon_i)\}_{i=1}^{N} \tag{1.13}$$

$$s(e,t)_{\text{z-norm}} = \frac{s(e,t) - \mu(S_e)}{\sigma(S_e)} \tag{1.14}$$

trong đó $\mu(S_e)$ và $\sigma(S_e)$ là trung bình và độ lệch chuẩn của S_e

T-norm

T-norm (Test score normalization) tương tự như Z-norm nhưng khác ở điểm chuẩn hóa phân bố điểm số mạo danh cho câu nói đánh giá. T-norm biểu diễn như sau :

$$S_t = s(t,\varepsilon_i)_{i=1}^{N} \tag{1.15}$$

$$s(e,t)_{t-norm} = \frac{s\,(e,t) - \mu(S_t)}{\sigma(S_t)} \tag{1.16}$$

trong đó $\mu(S_t)$ và $\sigma(S_t)$ là trung bình và độ lệch chuẩn của S_t

ZT-norm

ZT-norm hay TZ-norm có thể dùng cohort khác nhau trong mỗi bước. Bằng cách này, điểm số chuẩn hóa tương ứng với cả câu nói đăng kí và câu nói đánh giá.

s-norm

S-norm (Symetric normalization) tính trung bình điểm số chuẩn hóa từ Z-norm và T-norm. S-norm là đối xứng s(e,t) = s(t,e) trong khi các phương pháp chuẩn hóa đề cập ở trên phụ thuộc vào thứ tự câu nói e và t.

$$s\,(e,t)_{s-norm} = \frac{1}{2}.\left(s\,(e,t)_{z-norm} + s\,(e,t)_{t-norm}\right) \tag{1.17}$$

Bảng 1.5: *Tỉ lệ lỗi trong thực nghiệm không chuẩn hóa và có chuẩn hóa dùng as-norm.*

	Hệ thống		**Tập đánh giá công khai**		**Tập đánh giá kín T1**	
	Nhúng	Testing	No-norm	As-norm	No-norm	As-norm
1	TDNN x-vector [113]	PLDA	2.0	1.92	3.71	3.5
2	TDNN x-vector [113]	Cosine	5.3			
3	Resnet34 x-vector [113]	PLDA	1.52		2.49	
4	Resnet34 x-vector [113]	Cosine	1.5	1.61	2.43	2.53
	Fusion (1 + 4) [113]		1.35	1.22	1.79	1.75

$$= \frac{1}{2}.(\frac{s\left(e,t\right)-\mu(S_e)}{\sigma(S_e)}+\frac{s\left(e,t\right)-\mu(S_t)}{\sigma(S_t)}) \tag{1.18}$$

Chuẩn hóa điểm số thích nghi

Trong t-norm và top-norm thích nghi, chỉ một phần cohort được lựa chọn để tính trung bình và phương sai cho chuẩn hóa. Có hai lựa chọn cohort thích nghi : cohort thích nghi chọn từ X file gần nhất (hầu hết điểm số dương) với câu đăng kí E có top là e hoặc file đánh giá E top t. Chúng ta chú ý các cohort như vậy khác nhau trong mỗi phát âm đăng kí e hoặc phát âm đánh giá t tương ứng. Điểm số cohort dựa trên lựa chọn như vậy cho câu nói đăng kí là và câu nói đánh giá là :

$$S_e\left(\mathcal{E}_e^{top}\right)=s\left(e,\varepsilon\right)_{\forall\varepsilon\in\mathcal{E}_e^{top}},\quad S_e\left(\mathcal{E}_t^{top}\right)=s\left(e,\varepsilon\right)_{\forall\varepsilon\in\mathcal{E}_t^{top}} \tag{1.19}$$

Hai dạng biến thể của S-norm : điểm số chuẩn hóa thứ nhất gọi là adaptive S-norm1 và dạng thứ hai là adaptive S-norm2 định nghĩa là :

$$s(e,t)_{\text{as-norm1}}=\frac{1}{2}\left(\frac{s(e,t)-\mu(s_e(\epsilon_t^{\text{top}}))}{\sigma(s_e(\epsilon_t^{\text{top}}))}+\frac{s(e,t)-\mu(s_t(\epsilon_e^{\text{top}}))}{\sigma(s_t(\epsilon_e^{\text{top}}))}\right) \tag{1.20}$$

$$s(e,t)_{\text{as-norm2}}=\frac{1}{2}\left(\frac{s(e,t)-\mu(s_e(\epsilon_t^{\text{top}}))}{\sigma(s_e(\epsilon_t^{\text{top}}))}+\frac{s(e,t)-\mu(s_t(\epsilon_e^{\text{top}}))}{\sigma(s_t(\epsilon_e^{\text{top}}))}\right) \tag{1.21}$$

Nhóm tác giả Dinh Van Hung, Mai Van Tuan, Dam Ba Quyen, Nguyen Quoc Bao trong bài báo [113] đã sử dụng phương pháp chuẩn hóa adaptive symmetric score normalization (as-norm) và cho kết quả tốt hơn khi không chuẩn hóa.

Phương pháp chuẩn hóa làm giảm tỉ lệ lỗi giảm khoảng 5% so với không chuẩn hóa. Theo mô tả trên bảng với dữ liệu đánh giá công khai thì tỉ lệ lỗi không chuẩn hóa (1.35%) so với có chuẩn hóa là (1.22%), với dữ liệu đánh giá

Bảng 1.6: *Thống kê một số tập dữ liệu cho xác thực người nói.*

Tập dữ liệu	**Số người nói**	**Số câu nói**	**Số giờ**	**Ngôn ngữ**
VoxCeleb1 [72]	1,251	145,000	352	Phần lớn tiếng Anh
VoxCeleb2 [13]	6,112	1,128,246	2442	Phần lớn tiếng Anh
VLSP-2021 [17]	1,305	31,600	41.43	Tiếng Việt
Vietnam-Celeb [83]	1,000	87,140	187	Tiếng Việt
CN-Celeb1 [23]	1,000	130,109	274	Trung Quốc
CN-Celeb2 [62]	1,136	420,055	669	Trung Quốc

kín T1 thì tỉ lệ lỗi giảm từ 1.79% xuống còn 1.75%. Như vậy, việc sử dụng các phương pháp chuẩn hóa như as-norm sẽ làm giảm tỉ lệ lỗi trong bài toán xác thực người nói.

1.6. Dữ liệu và độ đo đánh giá

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

Bảng 1.6 thống kê một số tập dữ liệu phổ biến trong ứng dụng nhận dạng người nói.

1.6.1.1. 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, VoxCeleb1 có các đặc điểm sau:

- Cơ sở dữ liệu chủ yếu tập trung vào các mẫu giọng nói từ các người nổi tiếng, bao gồm diễn viên, ca sĩ, vận động viên, nhà báo và nhiều người nổi tiếng khác.
- Dữ liệu trong VoxCeleb1 được thu thập từ các video trên YouTube, bao gồm cả những cuộc phỏng vấn, buổi phát biểu, chương trình trò chuyện và các video khác mà người nổi tiếng xuất hiện và nói chuyện.
- Dữ liệu trong VoxCeleb1 không bị giới hạn về ngôn ngữ hoặc nền văn hóa, nó bao gồm các mẫu giọng nói từ nhiều quốc gia và vùng lãnh thổ khác

Bảng 1.7: *Thống kê chi tiết tập dữ liệu VoxCeleb1.*

Tổng số người nói	1,251
Số giọng nam	690
Số giọng nữ	561
Số video nhiều nhất/trung bình/ít nhất của một người	36/18/8
Số câu nói nhiều nhất/trung bình/ít nhất của một người	250/123/45
Độ dài câu nói dài nhất/trung bình/ngắn nhất (giây)	145.0/8.2/4.0

Bảng 1.8: *Phân chia dữ liệu cho bài toán xác thực người nói trên VoxCeleb1.*

	Huấn luyện	**Đánh giá**
Tổng số người nói	1,211	40
Tổng số video	21,819	677
Tổng số câu nói	148,642	4,874

nhau trên thế giới.

- Dữ liệu âm thanh trong VoxCeleb1 cung cấp dưới dạng các tệp âm thanh có định dạng chuẩn như WAV hoặc MP3, cung cấp thông tin về giọng nói của các người nổi tiếng.

Trong tập dữ liệu VoxCeleb1, người nói thu thập đa dạng ở nhiều dân tộc, nghề nghiệp, độ tuổi, quốc tịch, giới tính khác nhau, ở trong điều kiện môi trường đa dạng như văn phòng yên tĩnh, bên ngoài nhà ga. VoxCeleb1 có tỉ lệ giới tính cân bằng với 55% giọng nam và 45% giọng nữ. Quốc tịch và giới tính của mỗi giọng nói thu thập từ Wikipedia. Bảng 1.8 là thống kê chi tiết về cách phân chia tập huấn luyện và tập đánh giá của VoxCeleb1.

1.6.1.2. VoxCeleb2

Tập dữ liệu VoxCeleb2 [13] 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. Những người nói thuộc nhiều sắc tộc, giọng nói, ngành nghề và lứa tuổi khác nhau. Các video có trong tập dữ liệu được quay trong nhiều môi trường thị giác và thính giác là những thách thức. Tập dữ liệu gồm các cuộc phỏng vấn từ thảm đỏ, sân vận động ngoài trời và trường quay trong nhà yên tĩnh, các bài phát biểu trước đông đảo khán giả, các đoạn trích từ các video đa phương tiện được quay chuyên nghiệp và thậm chí

Bảng 1.9: *Thống kê dữ liệu trên hai tập VoxCeleb1 và VoxCeleb2. VoxCeleb2 lớn gấp năm lần so với VoxCeleb1.*

Dữ liệu	**VoxCeleb1**	**VoxCeleb2**
Tổng số người nói	1251	6112
Tổng số giọng nam	690	3761
Tổng số video	22,496	150,480
Tổng số giờ	352	2442
Tổng số câu nói	153,516	1,128,246
Trung bình số video của một người	18	25
Trung bình số câu nói của một người	116	185
Trung bình độ dài câu nói	8.2	7.8

Bảng 1.10: *Phân chia tập phát triển và tập đánh giá VoxCeleb2.*

Dữ liệu	**Huấn luyện**	**Đánh giá**	**Tổng**
Số người nói	5,994	118	6,112
Số video	145,569	4,911	150,480
Số câu nói	1,092,009	36,237	1,128,246

cả các video thô sơ được quay trên thiết bị cầm tay. Dữ liệu trong VoxCeleb2 có định dạng chuẩn như WAV hoặc MP3, ngoài ra còn có thêm thông tin về người nói và nguồn gốc của mẫu giọng nói. VoxCeleb2 dùng cho mục đích nghiên cứu trong nhận dạng giọng nói, xác thực người nói và các lĩnh vực liên quan khác. Bảng 1.9 đưa ra số liệu thống kê chung, so sánh tập dữ liệu VoxCeleb2 và VoxCeleb1. Trong bảng thống kê ở Bảng 1.10, tập phát triển của VoxCeleb2 không trùng với người nói trong VoxCeleb1.

1.6.1.3. CN-Celeb

Tập dữ liệu CN-Celeb1 [23] và CN-Celeb2 [62] là hai tập dữ liệu chủ yếu tập trung vào những người nổi tiếng Trung Quốc. Tập dữ liệu chứa hình ảnh hoặc video của nhiều nhân vật nổi tiếng khác nhau của Trung Quốc, bao gồm diễn viên, ca sĩ, vận động viên và các nhân vật khác. Tập dữ liệu này được thiết kế riêng cho các bài toán liên quan đến nhận dạng khuôn mặt, phân tích thuộc tính khuôn mặt, xác thực người nói, thị giác máy tính khác có liên quan đến người Trung Quốc.

Bộ dữ liệu có thể bao gồm một số lượng lớn mẫu, có khả năng trải dài ở nhiều độ tuổi, giới tính và ngành nghề khác nhau trong bối cảnh người nổi tiếng Trung Quốc. Mỗi mẫu thường đi kèm với siêu dữ liệu, chẳng hạn như tên, giới

Bảng 1.11: *Phân bố tập dữ liệu CN-Celeb1 [23].*

Thể loại	**Số người nói**	**Số câu nói**	**Số giờ**
Advertisement	17	120	0.18
Drama	160	7247	6.43
Entertainment	483	22,064	33.67
Interview	780	59,317	135.77
Live Broadcast	129	8747	16.35
Movie	62	2749	2.20
Play	69	4245	4.95
Recitation	41	2747	4.98
Singing	318	12,551	28.83
Speech	122	8401	36.22
Vlog	41	1894	4.15
Tổng số	**1000**	**130,109**	**273.73**

Bảng 1.12: *Phân bố tập dữ liệu CN-Celeb2 [62].*

Thể loại	**Số người nói**	**Số câu nói**	**Số giờ**
Advertisement	66	1542	3.86
Drama	268	13,116	16.32
Entertainment	616	31,982	60.84
Interview	519	34,024	81.28
Live Broadcast	388	167,019	439.95
Movie	133	4449	5.77
Play	127	14,992	22.04
Recitation	218	58,231	129.18
Singing	394	42,157	75.19
Speech	394	36,680	82.58
Vlog	488	125,293	177.00
Tổng số	**2000**	**529,485**	**1090.01**

tính, tuổi của người nổi tiếng và có thể cả các thuộc tính hoặc nhãn bổ sung. Các nhà nghiên cứu và nhà phát triển có thể sử dụng bộ dữ liệu CN-Celeb2 để huấn luyện và đánh giá các mô hình học máy cho các bài toán như nhận dạng khuôn mặt, phân loại giới tính, ước tính độ tuổi, phân tích biểu hiện khuôn mặt, nhận dạng người nói, xác thực người nói.

1.6.1.4. VLSP2021-SV

Gần đây, hội thảo VLSP 2021 [17] 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

Bảng 1.13: *Phân bố độ dài câu nói trong tập dữ liệu Vietnam-Celeb.*

Độ dài (giây)	Số câu nói	Tỉ lệ (%)
<2	4,889	5.8
5-10	23,112	26.4
10-20	14,155	16.1
20-30	4,518	5.2
>30	2,297	2.7

Bảng 1.14: *Thống kê phương ngữ theo vùng miền của Vietnam-Celeb.*

Phương ngữ	**Bắc**	**Trung**	**Nam**
YouTube			
# người nói	409	32	383
TikTok			
# câu nói	31,220	2,141	27,284
# giờ	61.13	5.15	62.81
# người nói	102	8	64
# câu nói	15,902	1,009	9,584
# giờ	33.64	2.31	22.33

từ nhiều nguồn khác nhau, bao gồm từ cuộc thi ZaloAI, VLSP2020-SV, VIVOS [66] 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. VLSP2021-SV sẽ được NCS mô tả chi tiết hơn trong phần thực nghiệm Chương 2 và Chương 3.

1.6.1.5. 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. Các mẫu âm thanh cũng đại diện cho điều kiện thực tế, nó chứa cả nhiễu nền như lời thoại, âm nhạc, tiếng cười, reo hò. Bảng 1.13 cho thấy các câu nói ngắn chiếm phần lớn trong bộ dữ liệu và nó đại diện cho âm thanh cho bài toán nhận dạng người nói thực tế, trong đó đầu vào âm thanh chủ yếu là các đoạn ngắn. Bộ dữ liệu cân bằng giới tính với 552 giọng nam chiếm 55.2% tổng số giọng. Với phân bố theo phương ngữ, Bảng 1.14 biểu diễn thống kê phương ngữ trên mỗi loại dữ liệu nguồn.

- Vietnam-Celeb-E: Tập đánh giá dễ của Vietnam-Celeb. Các cặp phủ định

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

trong tập – là các cặp khác người nói (lấy ngẫu nhiên).

- Vietnam-Celeb-H: Tập đánh giá khó của Vietnam-Celeb, có thêm cả thông tin về giới tính và phương ngữ. Để tạo các cặp phủ định cần phải đảm bảo mỗi cặp có cùng nhãn về giới tính và phương ngữ.

1.6.2. Độ đo đánh giá hệ thống xác thực người nói

1.6.2.1. EER

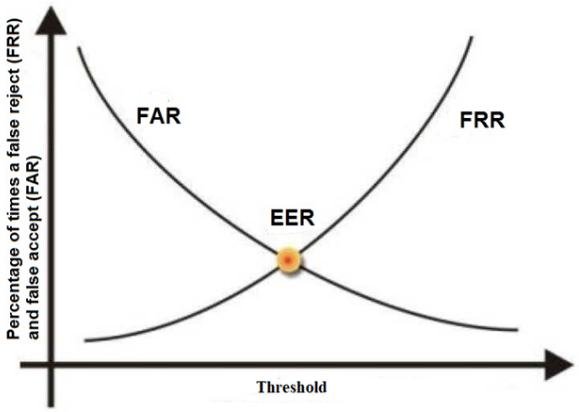
EER là điểm mà tỉ lệ chấp nhận sai (FAR) bằng với tỉ lệ từ chối sai (FRR) trong hệ thống xác thực. Ý nghĩa của chỉ số này được giải thích như sau : FRR càng cao thì hệ thống càng an toàn. Tuy nhiên lại xảy ra nhiều xác nhận của người dùng hợp pháp bị từ chối. Như vậy người dùng phải thực hiện nhiều lần xác thực để có thông điệp thành công dẫn đến cảm giác trải nghiệm của người dùng không được tốt. 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. Trong Hình 1.11 minh họa phương pháp xác định tỉ lệ lỗi EER trong xác thực người nói.

Công thức tính FAR và FRR:

$$FAR = \frac{\text{Số lần chấp nhận sai}}{\text{Tổng số lần thử nghiệm người không phải là chủ nhân}} \tag{1.22}$$

$$FRR = \frac{\text{Số lần từ chối sai}}{\text{Tổng số lần thử nghiệm người chủ nhân}} \tag{1.23}$$

$$EER = FAR = FRR \tag{1.24}$$



Hình 1.11: *Phương pháp xác định tỉ lệ lỗi EER trong xác thực người nói.*

1.6.2.2. Minimum Detection Cost Function (MinDCF)

Để đánh giá hiệu năng hệ thống xác thực người nói, ngoài tỉ lệ lỗi bằng nhau (EER) chúng ta còn độ đo MinDCF được giới thiệu chi tiết trong đánh giá hệ thống nhận dạng người nói NIST 2018 [95]. Các tham số sẽ được thiết lập như sau : $C\_Miss = 1$, $C\_FalseAlarm = 1$ và $P\_Target = 0.05$. Định nghĩa Minimum Detection Cost Function :

$$C_{\text{Det}}(\theta) = C_{\text{Miss}} \times P_{\text{Target}} \times P_{\text{Miss}}(\theta) + C_{\text{FalseAlarm}} \times (1 - P_{\text{Target}}) \times P_{\text{FalseAlarm}}(\theta) \quad (1.25)$$

$$\text{minDCF} = \arg\min_{\theta} \{0.05 \times P_{\text{Miss}}(\theta) + 0.55 \times P_{\text{FalseAlarm}}(\theta)\} \quad (1.26)$$

trong đó các tham số hàm giá trị là C_Miss(tỉ lệ bỏ sót) và $C\_FalseAlarm$(tỉ lệ bắt nhầm), và P_Target (xác suất biết trước người nói đích cụ thể). DCF càng nhỏ thì hệ thống càng tốt.

1.7. 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-FILTER BANK ENERGIES 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.

2.2. Tầm quan trọng của trích chọn đặc trưng

Trích chọn đặc trưng đóng một vai trò quan trọng trong hiệu suất và hiệu quả của các mô hình học máy, bao gồm cả những mô hình được sử dụng trong xác thực người nói. Dưới đây là cái nhìn chi tiết về vai trò và tầm quan trọng của nó:

- **Nâng cao hiệu suất của mô hình**

  Mức độ liên quan: Trích chọn đặc trưng bao gồm việc chọn các tính năng phù hợp nhất từ dữ liệu thô, điều này có thể cải thiện đáng kể hiệu suất của mô hình. Trong bối cảnh xác thực người nói, các tính năng như hệ số tần số Mel (MFCCs), cao độ và đặc điểm quang phổ có thể rất quan trọng.

  Khả năng phân biệt: các đặc trưng được thiết kế tốt có thể giúp mô hình phân biệt tốt hơn giữa các lớp khác nhau. Trong xác thực người nói, điều này có nghĩa là các đặc trưng nắm bắt được những đặc điểm duy nhất trong giọng nói của một cá nhân, giúp phân biệt giữa các người nói khác nhau dễ dàng hơn.

- **Giảm kích thước**

  Đơn giản hóa: Việc giảm số lượng tính năng có thể đơn giản hóa mô hình, giúp quá trình huấn luyện và suy diễn nhanh hơn. Các kỹ thuật như phân tích thành phần chính (PCA) hoặc phân tích phân biệt tuyến tính (LDA) có thể được sử dụng để giảm kích thước trong khi vẫn giữ được thông tin quan trọng.

- **Cải thiện tốc độ hội tụ và tốc độ huấn luyện**

  Chuẩn hóa: Việc chuẩn hóa các đặc trưng có thể giúp hội tụ nhanh hơn trong quá trình đào tạo mô hình. Điều này đặc biệt quan trọng đối với mạng nơ-ron mà các đặc trưng cần có sự tương đồng chia tỉ lệ để đảm bảo huấn luyện hiệu quả.

  Biến đổi đặc trưng: Các kỹ thuật như chuyển đổi log, chuyển đổi lũy thừa hoặc các tính năng đa thức có thể làm cho các mối quan hệ phức tạp trở nên tuyến tính hơn, hỗ trợ quá trình học.

- **Xử lý các giá trị bị thiếu và nhiễu**

  Tính toán: Việc xử lý đúng cách các giá trị bị thiếu thông qua các phương pháp tính toán đảm bảo rằng tập dữ liệu hoàn chỉnh và có thể sử dụng được. Điều này rất quan trọng để duy trì tính toàn vẹn của quá trình đào tạo.

Giảm nhiễu: Các kỹ thuật như làm mịn, lọc hoặc giúp giảm nhiễu trong dữ liệu, dẫn đến các mô hình hiệu quả hơn.

- **Kết hợp kiến thức về miền**

  Phân tích chuyên sâu: Trích chọn đặc trưng cho phép kết hợp kiến thức miền vào mô hình. Trong quá trình xác minh người nói, điều này có thể có nghĩa là tận dụng những hiểu biết sâu sắc về ngữ âm, âm học và ngôn ngữ học để tạo ra các tính năng giúp nắm bắt tốt hơn các sắc thái trong lời nói của con người.

  Khả năng tùy biến: Việc tạo các đặc trưng tùy biến dựa trên kiến thức cụ thể về miền bài toán có thể cung cấp thêm thông tin hữu ích mà có thể bỏ lỡ các đặc trưng chung.

- **Tăng khả năng diễn giải mô hình**

  Bằng các trích chọn đặc trưng có thể hiểu được, việc hiểu các quyết định của mô hình trở nên dễ dàng hơn. Điều này đặc biệt quan trọng trong các ứng dụng đòi hỏi tính minh bạch, chẳng hạn như hệ thống bảo mật.

- **Tăng cường hiệu năng và tổng quát hóa**

  Các kỹ thuật như tăng cường dữ liệu có thể giúp tạo ra các đặc trưng tốt hơn từ các biến thể của dữ liệu gốc. Điều này giúp mô hình tổng quát hóa tốt hơn đối với dữ liệu mới, chưa được nhìn thấy.

Tóm lại, trích chọn đặc trưng là một bước quan trọng trong việc xây dựng các mô hình học máy hiệu quả. Việc tạo ra các đặc trưng phân biệt giữa các người nói khác nhau giúp cho mô hình xác thực chính xác và hiệu quả hơn.

2.3. MFCCs và các ứng dụng trong xác thực người nói

Trong nhiều nghiên cứu, các đặc trưng MFCCs nói chung dùng trong nhận dạng lời nói, người nói, cảm xúc, ngôn ngữ, ... Ngoài ra, hầu hết các nghiên cứu đều so sánh phương pháp được đề xuất của họ với các MFCCs thông thường. Quá trình trích chọn cơ bản của MFCCs là phân tích tín hiệu thời gian ngắn, tổng hợp sau đó chuẩn hóa. Đặc trưng MFCCs là đặc trưng phổ biến nhất trong nhận dạng tiếng nói và nhận dạng người nói. MFCCs là đặc trưng thường dùng biểu diễn âm thanh tiếng nói.

Phản ứng của tai người với quang phổ âm thanh là phi tuyến. Đặc trưng MFBEs cũng là một loại đặc trưng tương tự như thính giác của tai người. Trong kiến trúc x-vector trong xác thực người nói thiết kế bởi [108], đặc trưng MFCCs là đầu vào mô hình TDNN [82]. Mô hình này có khả năng kết hợp với lớp tổng

Bảng 2.1: *Hạn chế của MFCC trong xác thực người nói.*

Cách tiếp cận	Hạn chế
GMM	MFCC không cung cấp mô tả đầy đủ về hệ thống giọng nói của người nói [75],[33]
HMM	Một số tham số của MFCC cần được điều chỉnh dựa trên bài toán cụ thể [33] [7]
Phân tích thống kê	MFCC nhạy cảm với nhiễu, tuy nhiên vấn đề này có thể được giải quyết bằng cách lấy căn bậc ba hoặc thay đổi bộ lọc hoặc áp dụng phép biến đổi Hilbert trước khung MFCC [137], [134], [25]
Học sâu	MFCCs không ổn định khi chất lượng âm thanh suy giảm [12] [6]
SVM	Số lượng hệ số MFCC thu được, cửa sổ độ dốc và pha của phổ công suất đều ảnh hưởng đến hiệu suất của MFCC [11][97]

hợp theo thời gian.

Nhóm tác giả Phạm Việt Thanh và các cộng sự [83] sử dụng đặc trưng MFCCs cho cả mô hình huấn luyện và đánh giá. MFCCs thích ứng tốt với nhiễu, cải thiện các đặc điểm thống kê hoặc liên kết gần hơn với nhận thức của con người.

Có nhiều nghiên cứu sử dụng MFCCs trong việc xác thực người nói. MFCCs là một trong những đặc trưng âm thanh phổ biến nhất được áp dụng trong lĩnh vực nhận dạng và xác thực người nói nhờ khả năng mô phỏng quá trình nghe của con người. Dưới đây là một số nghiên cứu nổi bật về việc sử dụng MFCCs trong xác thực người nói:

Nghiên cứu của Reynolds và cộng sự [92] là một nghiên cứu cổ điển trong lĩnh vực xác thực người nói, trong đó tác giả sử dụng MFCCs để trích xuất đặc trưng giọng nói và áp dụng mô hình GMM để thực hiện xác thực người nói.

HTK [125] là một công cụ phổ biến cho nghiên cứu về nhận dạng giọng nói, bao gồm xác thực người nói. Quyển sách này trình bày cách sử dụng MFCCs kết hợp với mô hình Markov ẩn (HMM) để thực hiện các bài toán về nhận dạng và xác thực giọng nói.

Nhóm nghiên cứu Li và cộng sự [60] áp dụng MFCCs như là một trong các đặc trưng âm thanh chính cho hệ thống nhúng người nói, kết hợp với các mô hình học sâu để cải thiện hiệu suất xác thực người nói.

2.4. 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 [2] 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.

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.

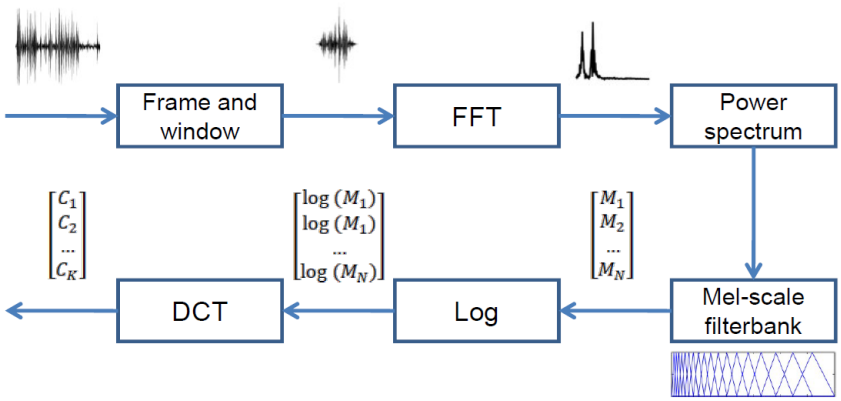
Ngoài ra, Jahangir và cộng sự [41] đã đề xuất các đặc trưng kết hợp dựa trên MFCC và các đặc trưng dựa trên thời gian. Các đặc trưng kết hợp này được đưa vào DNN để nhận dạng người nói. Kết quả cho thấy rằng hạn chế của đặc trưng MFCCs có thể được giải quyết bằng phương pháp này.

2.4.1. Đặc trưng MFBEs

MFBEs có thể được tính toán thông qua bốn bước liên tiếp, cụ thể là chia khung tín hiệu, tính toán phổ công suất, áp dụng bộ lọc Mel lên phổ công suất đã thu được và cuối cùng tính toán giá trị logarit của tất cả các bộ lọc. Hình 2.1 minh họa các bước trích chọn đặc trưng MFBEs.

2.4.1.1. Pre-emphasis

Đây là bước đầu tiên để nhấn mạnh các tần số cao của tín hiệu âm thanh, nhằm cải thiện chất lượng phân tích. Nhấn mạnh miền tần số là một trong



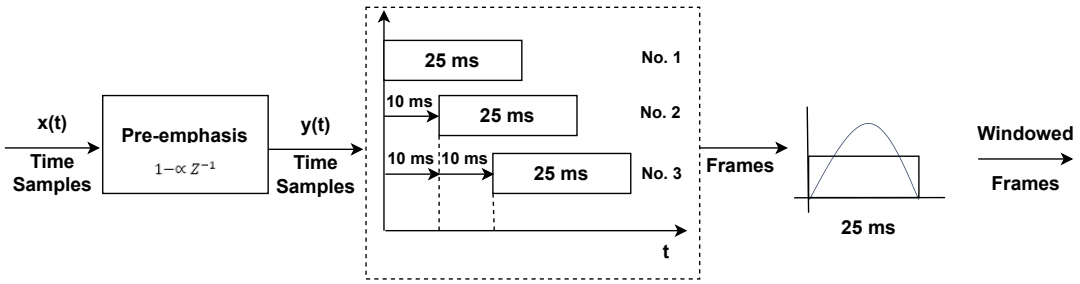
Hình 2.1: *Trích chọn đặc trưng MFBEs (Mel-scale filterbank energies).*

những bước tiền xử lý phổ biến trong lĩnh vực xử lý tín hiệu, nó dùng để bù đắp cho tần số cao của tín hiệu đã bị triệt tiêu trong quá trình tạo tín hiệu. Nhấn mạnh tín hiệu là bước đầu tiên trong quá trình thích nghi MFCC, có thể được thực hiện đơn giản bằng cách áp dụng bộ lọc thông cao. Quá trình lọc này thay đổi phân bố năng lượng trên các tần số, cũng như mức năng lượng tổng thể [114].

2.4.1.2. Chia khung và cửa sổ

Ý tưởng việc chia tín hiệu thành các "khung" riêng biệt là để phân chia tín hiệu dữ liệu thô thành các khung mà trong đó tín hiệu có xu hướng ổn định hơn. Để có được các đặc trưng âm thanh ổn định, giọng nói cần được xem xét trong khoảng thời gian đủ ngắn. Đối với tín hiệu giọng nói, khoảng thời gian 20-30ms được coi là một đoạn ổn định vì thời gian giữa hai lần đóng nắp thanh môn khoảng 20ms. Tuy nhiên, các nguyên âm được ghi nhận là có thể được bắt trong khoảng 40ms-80ms [61]. Do đó, các phép đo phổ ngắn hạn thường được thực hiện trong các cửa sổ 20ms, và mỗi khung được chồng lên nhau 10ms với khung tiếp theo. Sự chồng chéo của các khung 10ms cho phép theo dõi các đặc trưng thời gian của tín hiệu giọng nói. Với việc chồng chéo các khung, việc biểu diễn âm thanh sẽ được tập trung xấp xỉ tại một số khung nhất định.

Trên mỗi khung, một cửa sổ được áp dụng để thu hẹp tín hiệu về phía biên của khung. Nói chung, cửa sổ Hanning và Hamming [88] là những loại cửa sổ phổ biến nhất. Những cửa sổ này có thể tăng cường các hài âm, làm mịn các cạnh, và giảm hiệu ứng cạnh khi thực hiện DFT trên tín hiệu.



Hình 2.2: *Sơ đồ phân tích khung.*

2.4.1.3. Phổ công suất

Phổ công suất có thể được mô tả là sự phân bố năng lượng của các thành phần tần số cấu thành tín hiệu [24]. Theo truyền thống, biến đổi Fourier rời rạc (DFT) được sử dụng để tính phổ công suất. Phổ công suất của mỗi khung tín hiệu thu được phải được xác định dựa trên công thức 2.1 dưới đây:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{2\pi jnk}{N}} \quad k = 0, 1, 2, \ldots, N-1 \tag{2.1}$$

trong đó $x(n)$ là tín hiệu rời rạc và N là độ dài của tín hiệu.

2.4.1.4. Ngân hàng bộ lọc Mel

Bộ lọc thông dải Mel là một ngân hàng các bộ lọc được xây dựng dựa trên cảm nhận cao độ. Bộ lọc Mel ban đầu được phát triển cho phân tích giọng nói và giống như tai người khi cảm nhận giọng nói, nó nhắm đến việc trích xuất biểu diễn phi tuyến của tín hiệu giọng nói. Ngân hàng bộ lọc Mel thông thường được cấu tạo từ 40 bộ lọc hình tam giác [124]. Hàm truyền (TF) của mỗi bộ lọc thứ m có thể được tính toán thông qua công thức 2.2.

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k < f(m) \\ 1 & k = f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \le f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{2.2}$$

trong đó $f(m)$ là tần số trung tâm của bộ lọc hình tam giác và $\sum_m^{M-1} H_m(k) = 1$ Thang Mel đối với tần số đáp ứng và ngược lại được tính bằng các công thức

2.3 và 2.4:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.3}$$

$$f = 700\left(10^{\frac{m}{2595}} - 1\right) \tag{2.4}$$

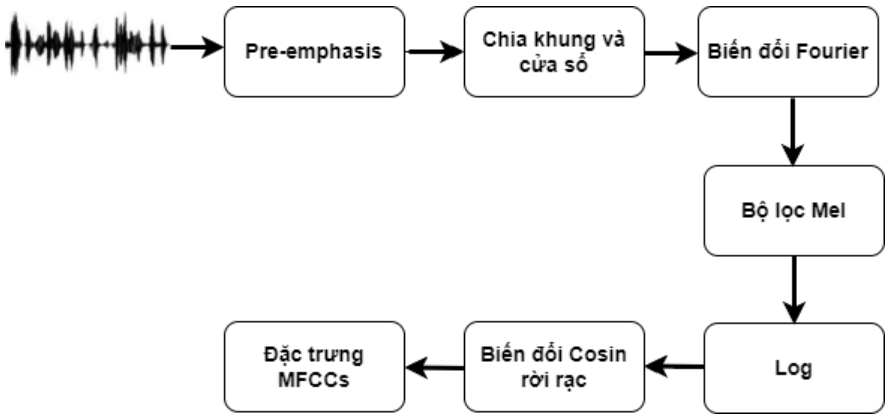
Trong các hệ thống học sâu, ngày càng có nhiều sự quan tâm đến việc sử dụng các đặc trưng cơ bản hơn, chẳng hạn như các đặc trưng MFBEs thay thế cho MFCCs trong các bài toán nhận dạng tiếng nói. Điều này là do MFBEs chứa nhiều thông tin hữu ích hơn (như các tương quan ngắn hạn) có thể hữu ích trong việc học đặc trưng. Tuy nhiên, các đặc trưng MFBEs cũng có thể gặp phải những hạn chế vì xử lý tín hiệu trong khoảng thời gian ngắn (âm thanh thường được chia thành các khung với cửa sổ khoảng 25ms). Trong khi đó, âm thanh tự nhiên phức tạp hơn nhiều. Các đơn vị giọng nói như âm vị hoặc từ có độ dài khác nhau và có thể ngắn hơn hoặc dài hơn cửa sổ phân tích. Do đó, có thể tồn tại các tương quan khác nhau giữa các thành phần quang phổ của giọng nói. Trong trường hợp có hiện tượng dội âm, năng lượng của giọng nói có thể lan tỏa ra xa hơn nhiều so với độ dài cửa sổ, làm tăng thêm các tương quan này.

2.4.2. Đặc trưng MFCCs

MFCCs là một trong những đặc trưng thường được sử dụng trong nhiều ứng dụng, đặc biệt là trong xử lý tín hiệu giọng nói như nhận dạng người nói, nhận dạng giọng nói và phân biệt giới tính [92]. Đặc trưng MFCCs có thể được tính toán thông qua năm bước liên tiếp, bao gồm phân khung tín hiệu, tính phổ công suất, áp dụng ngân hàng bộ lọc Mel lên các phổ công suất thu được, tính giá trị logarithm của tất cả các bộ lọc, và cuối cùng áp dụng phép biến đổi Cosine rời rạc (DCT). Hình 2.3 minh họa các bước trong quá trình tính toán MFCCs. Sau khi có được đặc trưng MFBEs, áp dụng phép biến đổi Cosine rời rạc để có đặc trưng MFCCs. Trong quá trình trích chọn MFCCs, DCT được áp dụng trên ngân hàng bộ lọc Mel để chọn các hệ số có gia tốc lớn nhất hoặc để tách mối quan hệ trong các biên độ phổ logarit từ ngân hàng bộ lọc [109]. DCT được tính bằng công thức 2.5 dưới đây:

$$X(k) = \sum_{n=0}^{N-1} x_n \cos\left(2\pi j \frac{nk}{N}\right), \quad k = 1, 2, 3, \ldots, N-1 \tag{2.5}$$

trong đó x_n là tín hiệu rời rạc và N là độ dài của tín hiệu.



Hình 2.3: *Các bước trích chọn đặc trưng MFCCs.*

2.4.3. 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. 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.

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.

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

2.5.1. Tiền xử lý và trích chọn đặc trưng MFBEs

Để trích chọn đặc trưng MFBEs cũng cần có những bước tiền xử lý tín hiệu như sau:

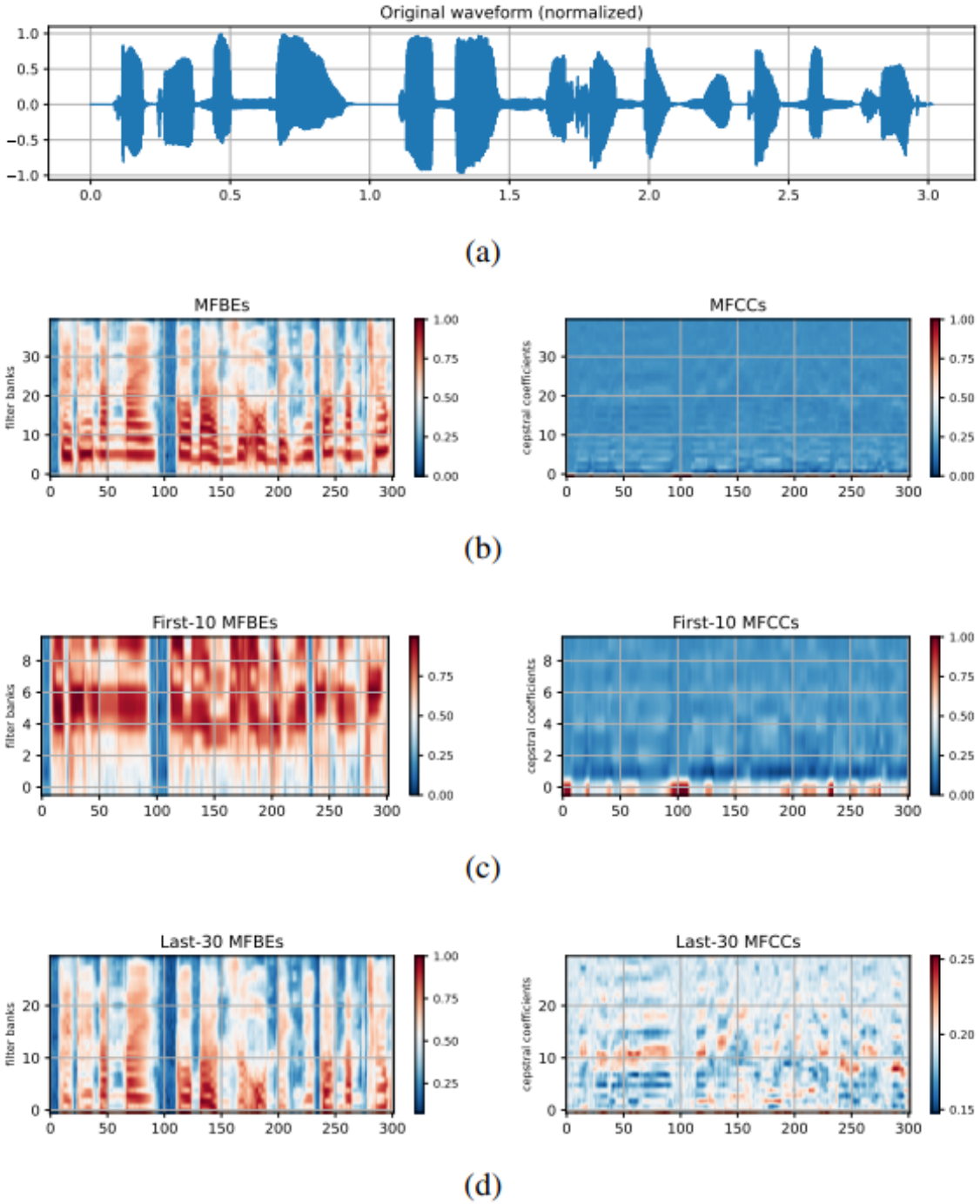
- Chia tín hiệu thành các khung nhỏ: Tín hiệu giọng nói được chia thành các khung nhỏ (frames) với độ dài 25 ms, mỗi khung được xử lý riêng biệt;
- Áp dụng Windowing: Áp dụng hàm cửa sổ (window function) như Hàm Hamming hoặc Hàm Hanning để giảm thiểu các hiệu ứng biên của khung và tăng tính liên tục của tín hiệu
- Biến đổi Fourier: Biến Đổi Fourier (FFT): Áp dụng biến đổi Fourier nhanh (Fast Fourier Transform - FFT) cho mỗi khung để chuyển đổi tín hiệu từ miền thời gian sang miền tần số.
- Áp dụng Mel-filter bank: Chuyển đổi tần số thành Mel: Các kết quả từ FFT sau đó được đưa qua Mel-filter bank, phân tách phổ thành các dải tần số theo thang độ Mel.

Hình 2.4 minh họa kết quả trích trọn đặc trưng MFBEs từ tín hiệu tiếng nói đầu vào.

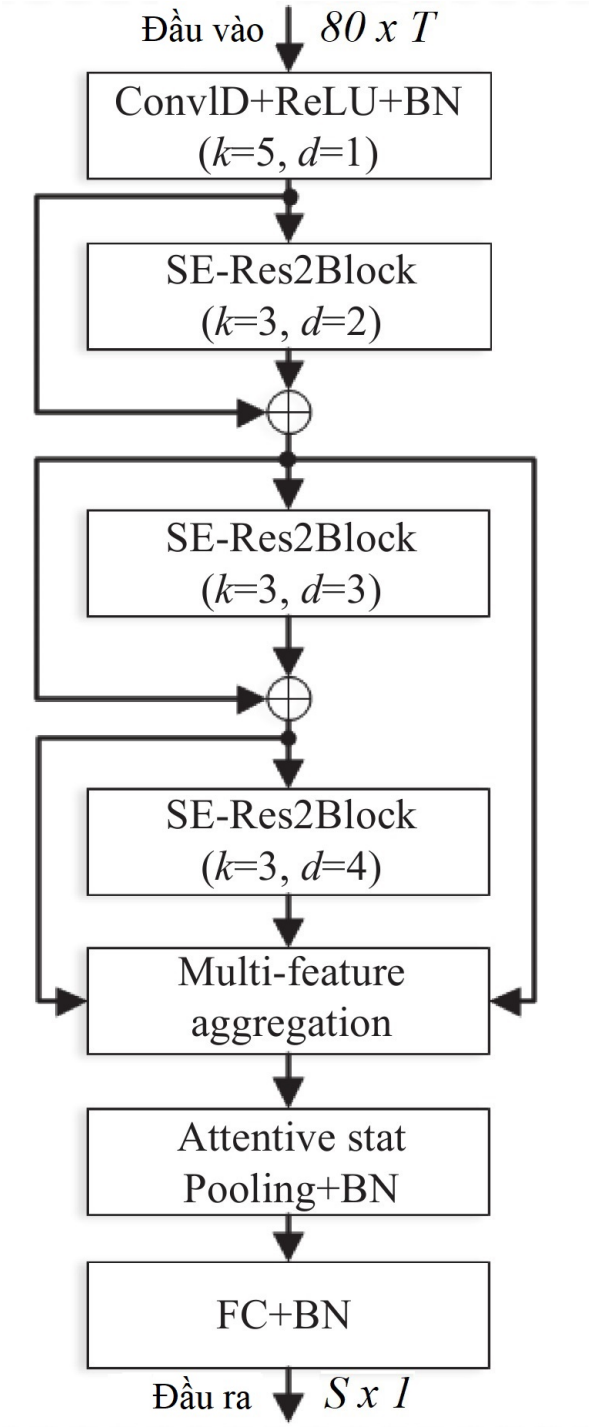
2.5.2. Mô hình học sâu ECAPA-TDNN

Mô hình ECAPA-TDNN được cải tiến từ mô hình x-vector [108] truyền thống, mô hình x-vector thường được sử dụng trong bài toán nhận dạng người nói. Mô hình sử dụng đầu vào là đặc trưng (MFCCs hoặc F-Banks) với 80 chiều và độ dài T. Trong đó: k đại diện cho kích thước kernel(k=3, tức là tích chập một chiều), d là khoảng giãn nở của các lớp Conv1D hoặc SE-Res2Blocks, C là số kênh của bản đồ đặc trưng trung gian, T là chiều thời gian, S là số lượng người nói. Mô hình ECAPA-TDNN trong hình 2.5 gồm các khối sau:

- Khối Conv1D+ReLu+BN: được thiết kế để xử lý dữ liệu chuỗi âm thanh đầu vào một chiều (1D).
- Khối SE-Res2Block, mỗi khối này bao gồm:
  - Res2 Dilated Conv1D + ReLU + BN: Lớp tích chập giãn với hàm kích hoạt ReLU và chuẩn hóa lớp (Batch Normalization). Các hệ số giãn được sử dụng trong ba khối SE-Res2Block đầu tiên lần lượt



Hình 2.4: *Trực quan hóa năng lượng các dải lọc Mel (MFBEs) và các hệ số Mel-frequency cepstral (MFCCs): (a) Sóng âm ban đầu (đã chuẩn hóa), (b) Năng lượng của 40 dải lọc và 40 hệ số cepstral (đã chuẩn hóa), (c) Năng lượng của 10 dải lọc đầu tiên và 10 hệ số đầu tiên, (d) Năng lượng của 30 dải lọc cuối cùng và 30 hệ số cuối cùng.*



Hình 2.5: *Kiến trúc mô hình ECAPA-TDNN [21].*

là 2, 3 và 4. Kích thước kênh được sử dụng trong ba khối trên là 512 hoặc 1024, kích thước nhân k là 3.

  - SE-Block (Squeeze-Excitation): khối chú ý kênh (Channel Attention) giúp mô hình tập trung vào các đặc trưng quan trọng của tín hiệu. Các khối này giúp mô hình học được các đặc trưng không gian phụ thuộc theo nhiều cấp độ khác nhau.

- Muti-Feature Aggregation: trong khối này, các đặc trưng từ ba khối SE-Res2 được nối lại và kết hợp để tạo ra một tập hợp đặc trưng phong phú hơn.
- Attentive Statistic Pooling: phương pháp tổng hợp thống kê chú ý, giúp mô hình chọn lọc các đặc trưng quan trọng và tính toán trung bình có trọng số dựa trên cơ chế chú ý.
- Khối FC+BN: khối này kết hợp các đặc trưng đã qua xử lý để đưa ra dự đoán cuối cùng. Lớp FC+Softmax được dùng để thực hiện phân loại, còn lớp FC+BN (Batch Normalization) dùng để chuẩn hóa đầu ra. S là số người nói đầu ra.

Mô hình ECAPA-TDNN có ba cải tiến so với mô hình TDNN truyền thống:

- Khối SE-Res2Block một chiều

  Thành phần đầu tiên của Squeeze–Excitation là phép toán ”nén” toán tử tạo ra các mô tả cho từng kênh và chỉ bao gồm việc tính toán vectơ trung bình của các đặc trưng mức khung trên miền thời gian: các đặc trưng mức khung cho mỗi khung được lấy trung bình theo thời gian, và các đầu vào đặc trưng là $[N, C, L]$ được nén thành $[N, C, 1]$ nhờ tổng hợp trung bình:

$$Z = \frac{1}{T} \sum_{t=1}^{T} h_t \tag{2.6}$$

  trong đó N là kích thước lô (batch size), L là số đặc trưng mức khung, C là số kênh, Z là vectơ trung bình đặc trưng mức khung theo miền thời gian và h_t là giá trị kích hoạt lớp sau cùng ở thời gian t.

  Trọng số mỗi kênh tính toán nhờ phép toán trong Z:

$$s = \sigma\left(W_2 f(W_1 Z + b_1) + b_2\right) \tag{2.7}$$

  trong đó $\sigma(\cdot)$ là hàm sigmoid , $f(\cdot)$ là hàm phi tuyến, $W_1 \in \mathbb{R}^{R \times C}$, $W_2 \in \mathbb{R}^{C \times R}$, R là kích thước đã được giảm, và C số kênh đầu vào.

  vectơ s thu được từ phép toán kích thích là các trọng số trong khoảng 0

và 1. Các trọng số này được nhân với đầu vào ban đầu sau mỗi lần tính toán kênh để thu được ước lượng đầu ra:

$$\tilde{h}_c = s_c h_c \tag{2.8}$$

trong đó s_c là trọng số mỗi kênh, h_c là đầu vào gốc, và $\tilde{h}_c$ là ước lượng đầu ra.

- Phép tổng hợp và cộng gộp đặc trưng đa lớp
  Dùng phương pháp bổ sung đa lớp, sử dụng tất cả đầu ra của các SE-Res2Block và lớp tích chập ban đầu làm đầu vào cho từng khối lớp khung, điều này được thực hiện bằng cách định nghĩa sự kết hợp dư thừa của mỗi SE-Res2Block là tổng của các đầu ra của tất cả các khối trước đó, chọn tổng các ánh xạ đặc trưng thay vì kết hợp nối tiếp để giới hạn số lượng tham số của mô hình.
- Tổng hợp thống kê phụ thuộc ngữ kênh và ngữ cảnh
  Các trọng số khác nhau được gán cho từng khung thông qua cơ chế chú ý.

$$e_{t,c} = \mathbf{v}_c^\top f(\mathbf{W}\mathbf{h}_t + \mathbf{b}) + k_c \tag{2.9}$$

trong đó $\mathbf{h}_t$ biểu diễn khung hiện tại t, $\mathbf{W} \in \mathbb{R}^{R\times C}$, $\mathbf{b} \in \mathbb{R}^{R\times 1}$, R là kích thước giảm, C là số kênh đầu vào, $e_{t,c}$ là phần tử vô hướng, $f(\cdot)$ là hàm phi tuyến, $\mathbf{v}_c$ là trọng số của mỗi khung, và k_c trọng số độ chệch.

$$\alpha_{t,c} = \frac{\exp(e_{t,c})}{\sum_t \exp(e_{t,c})} \tag{2.10}$$

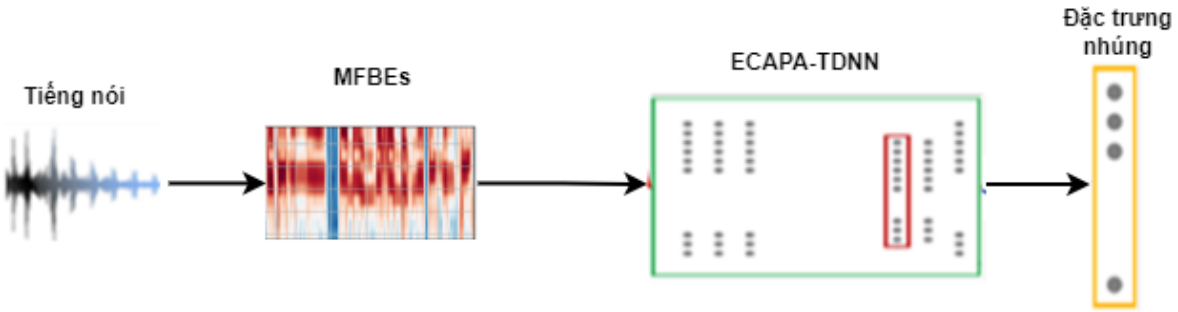
trong đó $\alpha_{t,c}$ là điểm số tự chú ý và $e_{t,c}$ là điểm số vô hướng. Giá trị trung bình tính như sau:

$$\tilde{\mu}_c = \sum_t \alpha_{t,c} h_{t,c} \tag{2.11}$$

trong đó $\alpha_{t,c}$ điểm số tự chú ý, $\tilde{\mu}_c$ là vectơ trung bình trọng số, và $h_{t,c}$ là giá trị kích hoạt tại thời điểm t.
Độ lệch chuẩn được tính như sau:

$$\tilde{\sigma}_c = \sqrt{\sum_t \alpha_{t,c} h_{t,c}^2 - \tilde{\mu}_c^2} \tag{2.12}$$

trong đó $\tilde{\sigma}_c$ là độ lệch chuẩn có trọng số, $\alpha_{t,c}$ là điểm số tự chú ý, $\tilde{\mu}_c$ là



Hình 2.6: *Đề xuất sử dụng đặc trưng MFBEs với mô hình ECAPA-TDNN trong xác thực người nói tiếng Việt.*

vectơ trung bình có trọng số and $h_{t,c}$ giá trị kích hoạt tại thời điểm t. Đầu ra cuối cùng của lớp gộp thu được bằng cách kết nối các vectơ trung bình có trọng số μ và độ lệch chuẩn có trọng số σ.

2.5.3. 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 [36] 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 vectơ 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 (Mục 2.6.2).

Bảng 2.2: *Thống kê chi tiết tập dữ liệu VLSP2021-SV.*

Tập dữ liệu	Ngôn ngữ	Số giờ	Tập huấn luyện Số câu nói (Số người nói)	Tập đánh giá Số cặp đánh giá (Số người nói)
VLSP2021-SV	Tiếng Việt	41.43	31,600 (1305)	20,000 (114)

2.5.4. 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 [28] giữa các đặc trưng nhúng của người nói (xem Hình 2.6). Sau đó, tất cả các điểm số này được chuẩn hóa thông qua phương pháp adaptive s-norm [16], [49].

2.6. Thực nghiệm

2.6.1. 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 [17]. 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. Bảng 2.2 là thống kê chi tiết của bộ dữ liệu dùng để thực nghiệm trong luận án. Trong cơ sở dữ liệu này, file âm thanh có độ dài trung bình là 5.2 giây. Môi trường giọng đọc ở trong văn phòng ít nhiễu. Giọng nói nam, nữ từ các vùng miền khác nhau: Bắc, Trung và Nam. Nội dung âm thanh thuộc các chủ đề kinh tế, xã hội, giáo dục. Các thống kê chi tiết mô tả trong Bảng 2.4.

Vietnam-Celeb

Trong thực nghiệm, NCS còn sử dụng bộ dữ liệu Vietnam-Celeb [83] và cũng được mô tả trong Chương 1. Bộ dữ liệu này có 1,000 người nói với 87,000 câu

Bảng 2.3: *Thống kê dữ liệu VLSP2021-SV.*

	Huấn luyện	**Đánh giá công khai**	**Đánh giá kín T1**	**Đánh giá kín T2**
Tổng số người nói	1,305	114	125	125
Tổng số giờ	41.43	4.35	4.91	4,91
Tổng số câu nói	31,600	2,941	3,983	3,983
Tổng số cặp	-	20,000	40,000	40,000

Bảng 2.4: *Thống kê chi tiết các câu nói trong dữ liệu huấn luyện VLSP2021-SV.*

Tổng số người nói	1305
Số câu nói nhiều nhất của một người	283
Số câu nói trung bình của một người	18
Số câu nói ít nhất của một người	1
Câu nói có độ dài lớn nhất	47.23 giây
Câu nói có độ dài trung bình	5.2 giây
Câu nói có độ dài nhỏ nhất	0.81 giây

nói. Tổng số giờ thu âm là 187 giờ, các câu nói thu âm lấy mẫu ở tần số 16,000 Hz.

2.6.2. Tăng cường dữ liệu

Để cải thiện hiệu quả của mô hình bằng cách tăng kích thước và đa dạng hóa bộ dữ liệu huấn luyện mà không cần thu thập thêm dữ liệu mới, NCS có sử dụng hai tập dữ liệu MUSAN [104] và RIS [57].

- Thay đổi tốc độ, bổ sung nhiễu, tăng âm là kỹ thuật tăng cường dữ liệu phổ biến trong [104] và thực hiện trực tuyến;
- Tỉ lệ làm nhiễu tốc độ ở các mẫu ngẫu nhiên là 1.0, 1.1, 0.9 và tăng tốc độ trên mỗi nhiễu loạn (tỉ lệ 1.0 là không có nhiễu loạn tốc độ). Sự nhiễu loạn tốc độ sẽ thay đổi cao độ của âm thanh và sẽ xem xét âm thanh được xử lý từ một người mới.

2.6.3. 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 mô hình ECAPA-TDNN với MFCCs,

Bảng 2.5: *Cài đặt siêu tham số trong thực nghiệm.*

Siêu tham số	**Giá trị thiết lập**
Tốc độ học	1×10^{-3}
Kích thước lô	100
Số chiều đặc trưng nhúng	512
Tỷ lệ bỏ học	0.2
Biên độ	0.3
Suy giảm trọng số	1×10^{-4}
Bộ tối ưu	Adam
Ngữ cảnh thời gian	2 giây
Phương pháp gộp	Thống kê
Dữ liệu tăng cường	MUSAN
Số vòng lặp	200

- Mô hình ResNetSE-34 với MFBEs và mô hình ResNetSE-34 với MFCCs.

Đặc trưng: NCS sử dụng MFCCs (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 giây, 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 [81]. Các mô hình huấn luyện trên máy chủ NVIDIA A100 GPU với 80GB bộ nhớ và sử dụng tối ưu Adam [56]. 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.

2.6.4. Độ đo

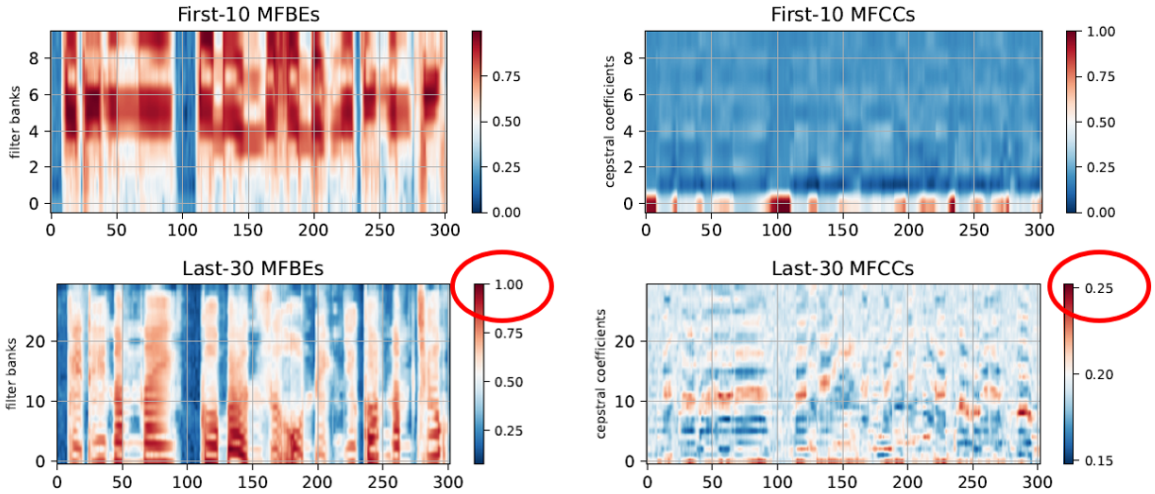
Trong luận án này, NCS sử dụng độ đo EER. EER chính là điểm mà tỉ lệ FAR = FRR. EER nhỏ hơn cho kết quả tốt hơn.

2.6.5. Kết quả thực nghiệm và phân tích

Trên tập dữ liệu Voxceleb2, đặc trưng MFBEs cho kết quả tốt hơn đặc trưng MFCCs [90]. (Kết quả thống kê trong Bảng 2.6). Mô hình ECAPA-TDNN

Bảng 2.6: *Thống kê kết quả thực nghiệm trên dữ liệu VoxCeleb2.*

Mô hình	Đặc trưng đầu vào	Huấn luyện	Đánh giá	Tỉ lệ lỗi (% EER)
ECAPA-TDNN [21] (C=512)	80 MFCCs	VoxCeleb2	VoxCeleb1-O	1.01
ECAPA-TDNN [21] (C=1024)	80 MFCCs	VoxCeleb2	VoxCeleb1-O	0.87
ECAPA-TDNN [90]	80 MFBEs	VoxCeleb2	VoxCeleb1-O	0.81



Hình 2.7: *Hiển thị trực quan sự khác biệt giữa hai đặc trưng MFBEs và MFCCs.*

trong bài báo gốc [21] lấy ngẫu nhiên đoạn âm thanh đầu vào có độ dài 2 giây trong khi mô hình ECAPA-TDNN của nhóm tác giả [90] sử dụng đoạn âm thanh có độ dài 3 giây.

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.

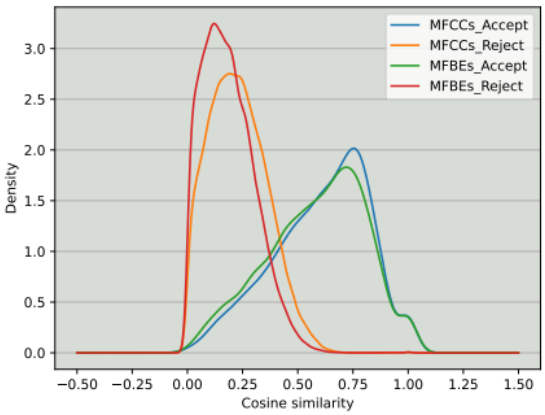
Trong Hình 2.7 cho thấy với 30 đặc trưng MFBEs sau cùng, năng lượng

Bảng 2.7: *Tỉ lệ lỗi trên tập dữ liệu đánh giá Vietnam-Celeb của hai mô hình ResNetSE-34 và mô hình ECAPA-TDNN trên hai đặc trưng MFCCs và MFBEs.*

Model / Dataset	**ResNetSE-34**		**ECAPA-TDNN**	
	MFCCs	***MFBEs***	***MFCCs***	***MFBEs***
Vietnam-Celeb-E	12.98	**11.84**	11.58	**11.37**
Vietnam-Celeb-H	14.31	**13.16**	14.30	**12.74**

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.*

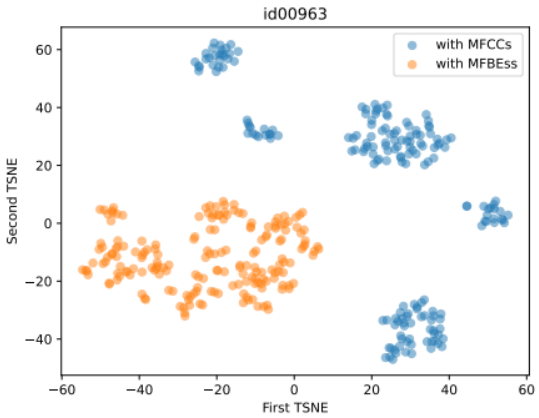
Mô hình	**Đặc trưng đầu vào**	**Huấn luyện**	**Tỉ lệ lỗi (%EER)**	
			Vietnam-Celeb-E	Vietnam-Celeb-H
ECAPA-TDNN [83]	80 MFCCs	VLSP2021-SV	11.58	14.3
ECAPA-TDNN	80 MFBEs	VLSP2021-SV	**11.37**	**12.74**
ResNetSE-34 [15]	80 MFCCs	VLSP2021-SV	12.98	14.31
ResNetSE-34	80 MFBEs	VLSP2021-SV	**11.84**	**13.16**



Hình 2.8: *Ước lượng mật độ của các độ tương đồng Cosine giữa 55.015 cặp nhúng trong tập dữ liệu kiểm tra Vietnam-Celeb-H. Các giá trị trung bình và độ lệch chuẩn của các độ tương đồng là 0.605 ± 0.214 và 0.229 ± 0.130 cho các cặp được chấp nhận và bị từ chối với đầu vào MFCCs. Các giá trị tương ứng với đầu vào MFBEs là 0.583 ± 0.222 và 0.186 ± 0.118*

dàn đều đối với các hệ số (tương ứng với các filter banks ở các tần số khác nhau). Tuy nhiên, ở 30 đặc trưng MFCCs sau cùng: năng lượng tập trung ở một số ít hệ số thấp (0.25) và sự khác biệt của các giá trị ở hệ số cao là không đáng kể.

So sánh giữa hai đặc trưng MFCCs và MFBEs: NCS phân tích hiệu năng của các đặc trưng khác nhau và các kết quả liệt kê trong Bảng 2.8. Trong Bảng 2.8, đặc trưng MFBEs cho kết quả tốt hơn đặc trưng MFCCs. Kết quả chỉ rõ trong Bảng 2.8 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 trong thời gian ngắn MFBEs và MFCCs. Dựa trên thực nghiệm ở Bảng 2.8, 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



Hình 2.9: *Trực quan hóa 2D các vectơ nhúng của người nói có nhãn id00963.*

Bảng 2.9: *Thống kê độ tương đồng Cosine trung bình từng cặp và khoảng cách Euclidean giữa các vectơ nhúng của cùng một người nói trong tập dữ liệu kiểm tra Vietnam-Celeb (120 người nói)*

Độ đo	**MFCCs**		**MFBEs**	
	mean (std)	**min - max**	**mean (std)**	**min - max**
Cosine↑	0.69 (0.14)	0.38 - 1.00	0.99 (0.01)	0.95 - 1.00
Euclidean↓	0.69 (0.23)	0.00 - 1.10	0.13 (0.05)	0.00 - 0.27

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.

Trong Hình 2.8, NCS trình bày ước lượng mật độ của các độ tương đồng Cosine giữa tất cả 55.015 cặp nhúng trong tập dữ liệu kiểm tra Vietnam-Celeb-H. Giá trị trung bình và độ lệch chuẩn của độ tương đồng giữa các cặp nhúng được chấp nhận và bị từ chối lần lượt là 0.605 ± 0.214 và 0.229 ± 0.130. Các giá trị tương ứng với đầu vào MFBEs là 0.583 ± 0.222 và 0.186 ± 0.118. Điều này có nghĩa là MFBEs giúp giảm độ tương đồng giữa các câu nói từ những người nói khác nhau tốt hơn nhiều so với MFCCs.

Theo Hình 2.9 độ tương đồng Cosine trung bình từng cặp là 0.3786 ± 0.2015 với đầu vào MFCCs và 0.9876 ± 0.0085 với đầu vào MFBEs tương ứng.

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 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 SỬ DỤNG HỌC CHUYỂN GIAO VỚI MÔ HÌNH RAWNET3

Trong Chương 3, NCS tập trung giải quyết vấn đề hạn chế về dữ liệu huấn luyện theo cách ứng dụng học chuyển giao trên các mô hình xác thực người nói tiên tiến nhất nhằm nâng cao hiệu quả xác thực người nói tiếng Việt. Chương này sẽ trình bày về các mô hình xác thực hiện đại nhất cũng như việc lựa chọn tập dữ liệu nào cho học chuyển giao. Cuối cùng là thực nghiệm quá trình học chuyển giao và những phân tích đánh giá.

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, cũng như sự 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 [31], 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 có nhãn và sắp xếp theo hệ thống phân cấp ngữ nghĩa của WordNet [19]. Các mô hình huấn luyện trước này không những sử dụng trong bài toán phát hiện đối tượng [26], phân đoạn ngữ nghĩa, phân loại ảnh y tế mà còn sử dụng

cho cả nhận dạng người nói [74], [122]. Những mô-đun đó và các biến thể được huấn luyện và cài đặt trong các thư viện phổ biến như TensorFlow và PyTorch [81].

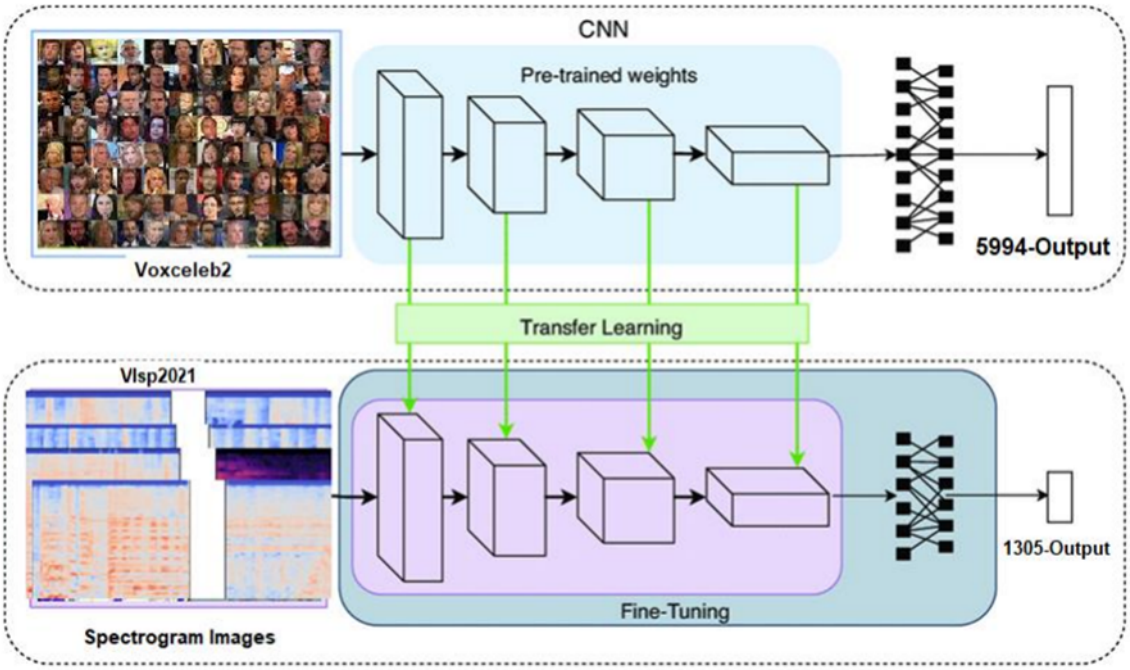
Sự đa dạng của nhiều mô hình huấn luyện trước xuất phát từ thực tế. Thứ nhất, bất kỳ mô hình học sâu nào cũng có điểm mạnh và điểm yếu và vì không ai biết trước về sự phù hợp của một mô hình nhất định cho một bài toán cụ thể. Ví dụ, VGG nổi tiếng vì cấu trúc đơn giản và hiệu năng cao nhưng cũng có nhiều 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 [34] 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. Trong các phần dưới đây, NCS mô tả các bước chính và phân tích thực nghiệm đánh giá sự hiệu quả của việc học chuyển giao trong bài toán xác thực người nói tiếng Việt.

Học chuyển giao đặc biệt có lợi trong các tình huống mà dữ liệu được gắn nhãn sử dụng cho bài toán xác thực người nói còn hạn chế. Nó cho phép tận dụng kiến thức từ các bộ dữ liệu lớn hơn và đa dạng hơn, có khả năng cải thiện hiệu năng và hội tụ nhanh hơn trong quá trình huấn luyện.

3.1.1. 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 [13], CnCeleb2 [62].



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.*

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 [27], 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 [13]) để nâng cao độ chính xác xác thực người nói với lượng dữ liệu nhỏ.

3.1.2. Các bước học chuyển giao trong xác thực người nói

- **Huấn luyện mô hình cơ sở**:

  Bài toán nguồn: Huấn luyện mô hình trên một tập dữ liệu lớn với một nhiệm vụ liên quan, chẳng hạn như định danh người nói hoặc nhận dạng giọng nói. Mô hình này sẽ học được các đặc trưng cơ bản của âm thanh và giọng nói.

  Mạng nơ-ron sâu: Thường được sử dụng vì khả năng học các đặc trưng phức tạp từ dữ liệu.
- **Chuyển giao tri thức** Chuyển giao trọng số: Sử dụng các trọng số đã học từ mô hình cơ sở làm điểm khởi đầu cho mô hình xác thực người nói. Tinh chỉnh: Tinh chỉnh mô hình bằng cách tiếp tục huấn luyện trên tập dữ liệu cụ thể cho nhiệm vụ xác thực người nói. Giai đoạn này thường sử dụng một lượng dữ liệu nhỏ hơn so với giai đoạn huấn luyện ban đầu.
- **Các bước thực hiện học chuyển giao**

  Trích chọn đặc trưng: Sử dụng các lớp mạng đã học để trích xuất đặc trưng từ dữ liệu âm thanh. Các đặc trưng này sau đó được sử dụng để huấn luyện mô hình xác thực người nói.

  Tinh chỉnh mô hình từ đầu vào đến đầu ra: Huấn luyện toàn bộ mô hình, bao gồm cả các lớp đã học từ trước và các lớp mới được thêm vào, trên tập dữ liệu đích.

Tinh chỉnh thường liên quan đến việc điều chỉnh mô hình được đào tạo trước để phù hợp hơn với một nhiệm vụ cụ thể bằng cách thực hiện các điều chỉnh có mục tiêu đối với kiến trúc của nó. Quá trình này thường tập trung vào các lớp cuối cùng của mạng, chịu trách nhiệm trực tiếp thực hiện phân loại. Hơn nữa, việc tinh chỉnh không chỉ giới hạn ở những thay đổi về kiến trúc; nó cũng bao gồm việc tối ưu hóa các siêu tham số. Các siêu tham số chính, chẳng hạn như tốc độ học tập và kích thước lô, cùng với các tham số khác, đóng một vai trò quan trọng trong việc xác định mức độ nhanh chóng và hiệu quả của mô hình có thể thích ứng với dữ liệu mới [43].

3.1.3. Lợi ích của học chuyển giao trong xác thực người nói

Hiệu suất cao hơn: Bằng cách bắt đầu từ một mô hình đã được huấn luyện trên một lượng lớn dữ liệu, mô hình xác thực người nói có thể đạt được hiệu suất cao hơn ngay cả khi dữ liệu huấn luyện cho nhiệm vụ đích hạn chế.

Tiết kiệm tài nguyên: Giảm thời gian và tài nguyên cần thiết cho việc

huấn luyện mô hình từ đầu. Điều này đặc biệt quan trọng khi làm việc với các mô hình lớn và phức tạp.

Cải thiện khả năng tổng quát hóa: Mô hình có thể học được các đặc trưng chung của giọng nói từ nhiệm vụ nguồn, giúp nó tổng quát hóa tốt hơn cho các tình huống khác nhau trong bài toán đích.

Học chuyển giao trong xác thực người nói là một phương pháp hứa hẹn mang lại nhiều lợi ích và cải thiện hiệu năng của hệ thống. Việc áp dụng đúng cách có thể giúp vượt qua các hạn chế về dữ liệu và tài nguyên, đồng thời nâng cao độ chính xác và khả năng ứng dụng của các hệ thống xác thực người nói. 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. Trong các phần tiếp theo, NCS mô tả các bước chính và phân tích thực nghiệm, đánh giá sự hiệu quả của việc học chuyển giao trong bài toán xác thực người nói tiếng Việt.

3.1.4. Những thách thức trong học chuyển giao áp dụng cho bài toán xác thực người nói.

Việc áp dụng phương pháp học chuyển giao để xác thực người nói đặt ra một số thách thức có thể ảnh hưởng đến tính hiệu quả và độ tin cậy của hệ thống thu được. Dưới đây là một số thách thức chính:

Khác biệt miền dữ liệu

Sự khác biệt về phân phối dữ liệu: Phân phối dữ liệu trong giai đoạn tiền huấn luyện có thể khác biệt lớn so với miền đích. Ví dụ: một mô hình huấn luyện trước trên dữ liệu sạch, chất lượng âm thanh tốt có thể gặp khó khăn với dữ liệu âm thanh thực tế có nhiễu.

Các biến thể về ngôn ngữ và giọng nói: Mô hình huấn luyện có sẵn có thể đã được huấn luyện trên những người nói một ngôn ngữ hoặc giọng cụ thể, điều này có thể không tổng quát tốt với các ngôn ngữ hoặc giọng khác.

Quá khớp và chưa khớp dữ liệu : quá khớp dữ liệu đích nhỏ: Việc tinh chỉnh mô hình được huấn luyện trước trên tập dữ liệu đích nhỏ có thể dẫn đến tình trạng trang quá khớp, trong đó mô hình hoạt động tốt trên dữ liệu huấn luyện nhưng kém trên dữ liệu không nhìn thấy. Chưa khớp dữ liệu: Nếu các đặc trưng của mô hình được huấn luyện trước không liên quan đến bài toán đích thì mô hình có thể không phù hợp, không nắm bắt được các chi tiết cần thiết để xác thực người nói hiệu quả.

Mất tri thức

Mất tri thức đã được huấn luyện có sẵn: Trong quá trình tinh chỉnh, mô hình có thể mất đặc trưng có ích mà nó đã học được trong giai đoạn huấn luyện có trước, đặc biệt nếu tập dữ liệu tinh chỉnh nhỏ hoặc nhiễu.

Các ràng buộc về tính toán và tài nguyên

Tinh chỉnh chuyên sâu về tài nguyên: Tinh chỉnh các mô hình lớn được đào tạo trước có thể tốn kém về mặt tính toán và tốn thời gian, đòi hỏi nguồn lực đáng kể. Các mô hình lớn cũng có thể gặp khó khăn khi triển khai trong môi trường hạn chế về tài nguyên, chẳng hạn như hệ thống di động hoặc hệ thống nhúng.

Lựa chọn mô hình huấn luyện có sẵn

Chọn mô hình phù hợp: Việc chọn một mô hình được đào tạo trước phù hợp chặt chẽ với bài toán đích là rất quan trọng. Sự không phù hợp có thể dẫn đến hiệu suất kém. Khả năng tương thích của mô hình: Đảm bảo rằng kiến trúc và biểu diễn đặc trưng của mô hình được đào tạo trước tương thích với các yêu cầu cụ thể của hệ thống xác thực người nói.

Đánh giá hiệu năng

Việc đánh giá hiệu năng của các mô hình học chuyển giao trong xác thực người nói có thể gặp khó khăn do tiêu chuẩn đáng giá cũng như dữ liệu dùng để đánh giá.

Khả năng tổng quát với dữ liệu chưa được thấy: việc đảm bảo rằng mô hình có khả năng khái quát tốt với dữ liệu mới, chưa từng thấy là vô cùng quan trọng. Bằng cách giải quyết những thách thức này bằng các chiến lược phù hợp, việc học chuyển giao có thể được áp dụng một cách hiệu quả cho các hệ thống xác thực người nói, nâng cao hiệu suất và độ mạnh mẽ của chúng.

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

VoxCeleb2 [13] 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.

Bên cạnh tiếng Anh, một số bộ dữ liệu đã được xây dựng cho các ngôn

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 và tập dữ liệu VLSP2021-SV được dùng để tinh chỉnh và đánh giá mô hình.*

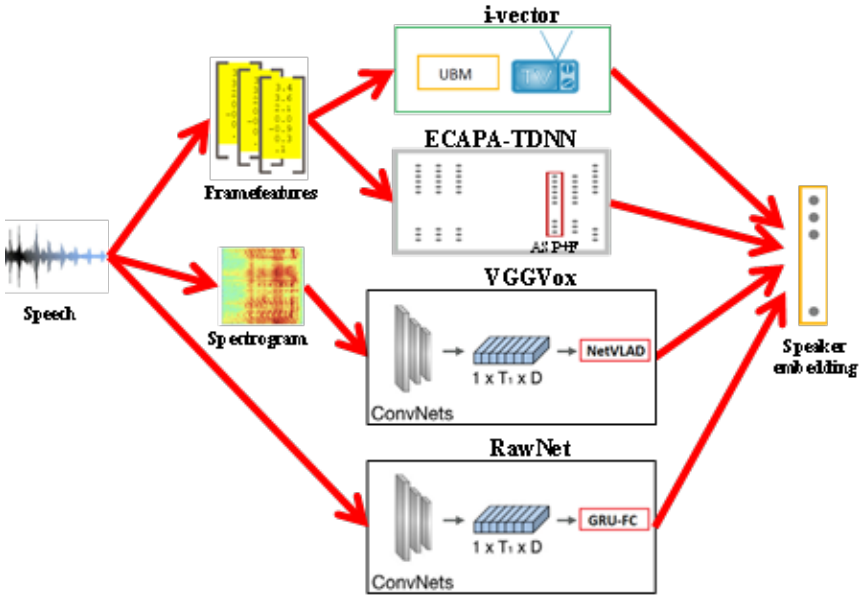
Tập dữ liệu	Ngôn ngữ	Số giờ thu âm	Tập huấn luyện Số câu nói (Số người nói)	Tập đánh giá Số cặp (Số người nói)
VoxCeleb2 [13]	Phần lớn tiếng Anh	2,300	1,128,246 (6112)	581,480 (1251) 552,536 (1190) 37,720 (40)
CN-CELEB2 [62]	Trung Quốc	1,090	529,485 (2000)	3,586,776 (-)
VLSP2021-SV [17]	Việt Nam	41	31,600 (1305)	20,000 (114)

ngữ khác như tiếng Trung và tiếng Việt. Ví dụ: CN-CELEB1 [23] 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 [23] 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 [13]. 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. Học chuyển giao từ các mô hình đã được huấn luyện trước

3.3.1. Mô hình ECAPA-TDNN

Desplanques và cộng sự [21] đề xuất một kiến trúc mới để xác thực người nói bằng cách sử dụng ECAPA-TDNN. ECAPA-TDNN cải tiến từ mô hình



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.*

TDNN và có những thành phần:

- **Emphasized Channel Attention**:
  Cơ chế này tập trung vào các kênh (tính năng) phù hợp nhất cho nhiệm vụ, nâng cao khả năng phân biệt giữa các loa khác nhau của mạng. Nó thường liên quan đến các cơ chế chú ý gán các trọng số khác nhau cho các kênh khác nhau, nhấn mạnh vào những kênh có nhiều thông tin hơn.
- **Lan truyền**
  Đề cập đến cách các tính năng được truyền bá qua mạng. Điều này bao gồm nhiều kết nối và biến đổi khác nhau để nắm bắt động lực thời gian và mối tương quan của tín hiệu giọng nói.
- **Tổng hợp**
  Thành phần này tổng hợp thông tin theo thời gian, điều này rất quan trọng đối với nhiệm vụ xác minh người nói. Nó thường bao gồm các hoạt động tổng hợp hoặc cơ chế dựa trên sự chú ý kết hợp các tính năng từ các bước thời gian khác nhau để tạo thành một bản trình bày nhỏ gọn về người nói.
- **TDNN** (Time Delay Neural Network) [82]
  TDNN là một loại mạng thần kinh đặc biệt phù hợp với dữ liệu tuần tự như giọng nói. Nó có thể nắm bắt được sự phụ thuộc về thời gian bằng cách xem xét nhiều bước thời gian trong quá khứ (và đôi khi là tương lai).
  ECAPA-TDNN sử dụng mảng đặc trưng âm học mức khung tương tự như

Bảng 3.2: *Kiến trúc VGG [74]. Kích thước dữ liệu cột bên phải là kích thước dữ liệu đầu ra của mỗi lớp.*

Layer	Support	Filt dim.	# filts.	Stride	Data size
conv1	7 × 7	1 × 96	96	2 × 2	254 × 148
mpool1	3 × 3	-	-	2 × 2	126 × 73
conv2	3 × 5	96	256	2 × 2	62 × 36
mpool2	3 × 3	-	-	2 × 2	30 × 17
conv3	3 × 3	256	384	1 × 1	30 × 17
conv4	3 × 3	384	256	1 × 1	30 × 17
conv5	3 × 3	256	256	1 × 1	30 × 17
mpool5	**3 × 3**	-	-	2 × 2	**14 × 8**
fc6	1 × n	256	4096	1 × 1	4096 × 1
apool6	**1 × 1**	-	-	1 × 1	**4096 × 1**
fc7	1 × 1	4096	1024	1 × 1	1024 × 1
fc8	1 × 1	1024	1251	1 × 1	1 × 1

i-vector truyền thống, trong khi VGGVox dùng ảnh phổ 2 chiều và RawNet [44] 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 khác nhau để biểu diễn đặc trưng người nói riêng biệt hoặc cố định số chiều vectơ nhúng người nói trong bài toán nhận dạng và xác thực người nói.

Tuy nhiên, mô hình ECAPA-TDNN vẫn có những hạn chế. Nó tập trung mô hình hóa đặc trưng cục bộ, thiếu khả năng kết hợp ở mức toàn cục. Nhân tích chập lại có kích thước cố định khiến khả năng nắm bắt đặc trưng toàn cục và mẫu người nói miền tần số không hiệu quả. Điểm yếu này khiến biểu diễn người nói đã trích chọn không có những thông tin lân cận toàn cục. Để giải quyết vấn đề này có nhiều mô hình dựa trên transformer [69], [96], [132], các mô hình này bổ sung thêm mô-đun MHSA (Self **M**ulti-**H**ead **A**ttention for **S**peaker Recognition) [39] để bắt được thông tin phụ thuộc phạm vi rộng. Tuy nhiên, các mô hình dựa trên transformer cũng còn nhiều vấn đề cần cải tiến.

3.3.2. Mô hình 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 [136], 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

Bảng 3.3: *Chi tiết kiến trúc ResNet-34 [74].*

Layer name	**ResNet-34**
conv1	7×7, 64, stride 2
pool1	3×3, max pool, stride 2
conv2_x	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$
conv3_x	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$
conv4_x	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$
conv5_x	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$
fc1	9×1, 512, stride 1
pool_time	$1 \times$ N, avg pool, stride 1

huấn luyện. Trong giai đoạn kiểm thử, sử dụng các chiến lược khác nhau đánh giá hiệu năng bài toán xác thực người nói.

Kiến trúc mạng ResNet [30] tương tự như một CNN nhiều lớp tiêu chuẩn nhưng được bổ sung các kết nối bỏ qua. Kiến trúc này cho phép các lớp thêm các phần dư vào ánh xạ đồng nhất trên đầu ra của kênh. Trong luận án này, NCS thử nghiệm với một biến thể của ResNet, cụ thể là kiến trúc ResNet-34.

3.3.3. Mô hình RawNet

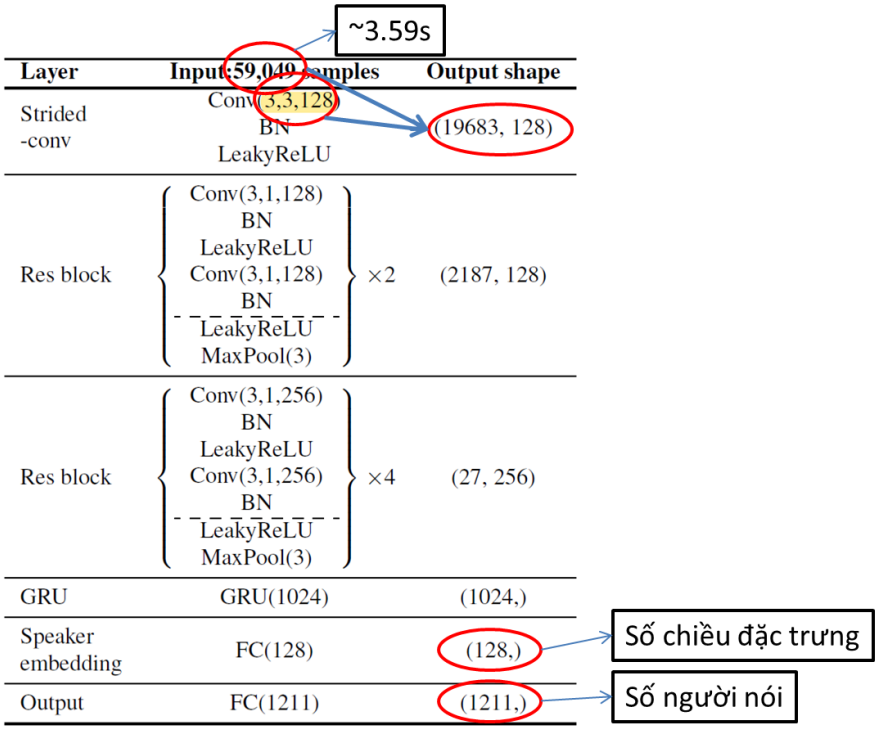
3.3.3.1. RawNet

RawNet [44] là kiến trúc mạng nơ-ron được thiết kế đặc biệt để xử lý âm thanh thô mà không yêu cầu bất kỳ phương pháp trích chọn đặc trưng thủ công nào. Kiến trúc này đặc biệt phù hợp cho các bài toán như nhận dạng giọng nói và phân loại âm thanh, trong đó âm thanh dạng thô chứa thông tin có giá trị mà các phương pháp trích chọn đặc trưng truyền thống có thể bỏ qua.

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.

RawNet có ưu điểm nổi bật là khả năng học các đặc trưng phân biệt trực tiếp từ tín hiệu âm thanh thô, điều này có thể nâng cao hiệu suất so với phương pháp trích xuất đặc trưng truyền thống.

RawNet là kiến trúc cải tiến từ mô hình CNN-LSTM, được tăng cường trước khi huấn luyện bằng cách thêm hàm mục tiêu trong nhúng người nói. Trong kiến trúc này, khối dư được kết nối với lớp tổng hợp được xây dựng trước



Hình 3.3: *Mô hình RawNet [44].*

tiên để xử lý các đặc trưng đầu vào cho việc trích chọn nhúng mức khung nhờ sử dụng leaky ReLU.

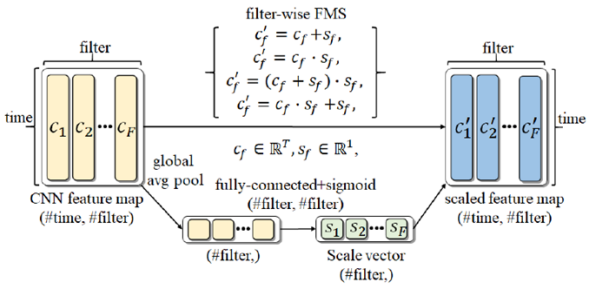
Lớp GRU tổng hợp các đặc trưng nhúng mức câu nói thay cho lớp LSTM. Tâm hàm mất mát L_c và hàm mất mát cơ bản L_BS được bổ sung thêm vào hàm mục tiêu trong RawNet để giảm hiệp phương sai bên ngoài lớp và tăng hiệp phương sai bên trong lớp trong quá trình phân biệt các đặc trưng nhúng.

DNN huấn luyện sử dụng hàm mục tiêu cuối cùng bên cạnh hàm mất mát L_ce thuộc dạng cross-entropy để tăng cường đặc trưng. Lớp tổng hợp đầy đủ khởi tạo mức phát âm cho đặc trưng nhúng người nói làm giảm tham số khung và tăng tính hiệu quả. Mô hình RawNet không trích chọn đặc trưng âm học ở mức khung, mô hình này dùng 59,049 mẫu từ 3.59s âm thanh lấy mẫu ở tần số 16kHz làm đầu vào mô hình huấn luyện. Để trích chọn đặc trưng mức khung, RawNet dùng 6 khối dư. Mỗi khối dư gồm hai lớp tích chập, hai lớp chuẩn hóa (BN), hai lớp ReLU và lớp tổng hợp cực đại. Lớp đầu tiên trong RawNet dùng tích chập một chiều, kỹ thuật này làm giảm kích thước dữ liệu đầu vào (từ 59k còn 19.7K).

3.3.3.2. RawNet2

Mô hình RawNet2 [45] cải tiến từ RawNet ở hai điểm chính (Hình 3.4):

Layer	Input:59049 samples	Output shape
Sinc -conv	Sinc(251,1,128) MaxPool(3) BN LeakyReLU	(19683, 128)
Res block	BN LeakyReLU Conv(3,1,128) BN LeakyReLU Conv(3,1,128) MaxPool(3) FMS ×6	(27, 256)
GRU	GRU(1024)	(1024,)
Speaker embedding	FC(1024)	(1024,)



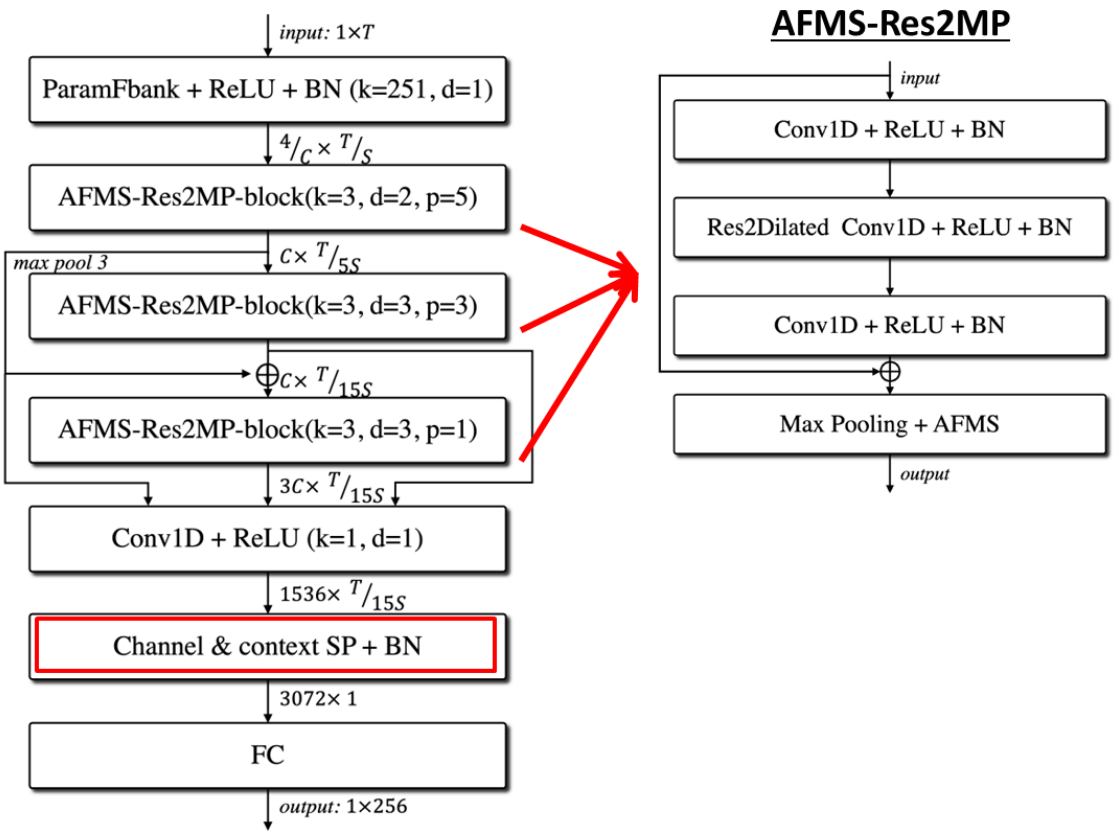
Hình 3.4: *Mô hình RawNet2 [45].*

- Thay thế lớp tích chập đầu tiên bằng lớp sinc-conv. Việc thay thế làm giảm bớt số lượng tham số và tối ưu hóa khi kết hợp cùng với các tham số khác của mạng học sâu.
- Bổ sung kỹ thuật tỷ lệ hóa bản đồ đặc trưng theo từng bộ lọc FMS (Feature Map Scaling). Kỹ thuật này áp dụng một vectơ tỷ lệ để thực hiện quá trình tỷ lệ hóa, trong đó kích thước của vectơ tương đương với số lượng bộ lọc. Các phương pháp dựa trên FMS thực hiện tỷ lệ hóa các bộ lọc trong bản đồ đặc trưng nhằm xây dựng các bản đồ đặc trưng cải tiến, tập trung vào các đặc trưng quan trọng hơn trong bản đồ đặc trưng mức khung thông qua phép cộng, phép nhân, hoặc cả hai.

3.3.3.3. RawNet3

RawNet3 là mô hình cải tiến từ RawNet2 [45] và ECAPA-TDNN [21]. 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.

Mỗi khối xương sống, được minh họa trong Hình 3.5 gọi là khối AFMS-Res2MP-block. Nó tương tự như khối xương sống của ECAPA-TDNN, với hai



Hình 3.5: *Mô hình RawNet3 [48].*

thay đổi:

- Áp dụng kỹ thuật tỷ lệ hóa bản đồ đặc trưng α (Alpha FMS) từ RawNet2 thay cho phương pháp squeeze-excitation [36] dựa trên các kết quả thực nghiệm trước đó,
- Tùy chọn áp dụng quá trình tổng hợp trước khi thực hiện AFMS.

Sau cùng, thông tin người nói được trích xuất thành các vectơ biểu diễn người nói thông qua tích chập, tổng hợp thống kê có trọng số (ASP) [77] với các giá trị trọng số phụ thuộc vào kênh và ngữ cảnh và các tầng tuyến tính.

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). Bảng 3.4 cho thấy với cùng dữ liệu đầu vào tín hiêu một chiều, 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 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 [20] trong quá trình huấn luyện. Số chiều đặc trưng nhúng người nói của RawNet3 là 256 nhỏ hơn nhiều 1024 của RawNet2 [45]. Số chiều đặc trưng

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**	**Tỉ lệ lỗi EER (%)**
VGGVox (ResNet-34) [15]	22.0	2D spectrum image	512	2.22
ECAPA-TDNN [21] (C=512)	6.2	1D array of MFCCs	192	1.01
ECAPA-TDNN [21] (C=1024)	14.3	1D array of MFCCs	192	0.87
RawNet2 [45]	13.2	59,049 samples	1024	3.00
RawNet3 [48]	16.3	59,049 samples	256	0.89

nhúng của ECAPA-TDNN là 192 và của ResNet-34 là 512. ECAPA-TDNN đạt tỉ lệ lỗi EER là 1.01% 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 2.22% 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 [5] để tổng hợp đặc trưng mức khung, kết hợp so sánh độ tương tự Cosin và phân loại cho quá trình xác thực.

Trong Bảng 3.5 là danh sách các mô hình huấn luyện đã được cung cấp nhờ vào một số yếu tố quan trọng. Trước hết, mỗi mô hình học sâu đều có những ưu và nhược điểm riêng, và không thể biết trước được mô hình nào sẽ phù hợp nhất với một nhiệm vụ cụ thể. Chẳng hạn, VGG được biết đến với cấu trúc đơn giản và hiệu suất cao, nhưng số lượng tham số lớn khiến nó tốn kém về mặt tính toán và bộ nhớ trong quá trình huấn luyện và triển khai. Trong khi đó, ResNet và Inception (cùng các biến thể) là những kiến trúc có tầm ảnh hưởng, giải quyết những thách thức khác nhau trong học sâu. ResNet nổi bật ở khả năng huấn luyện các mạng sâu hiệu quả, còn Inception tập trung vào việc học các đặc trưng đa chiều và tối ưu hóa. Tuy nhiên, cả hai đều gặp khó khăn về độ phức tạp trong tính toán và kiến trúc, cũng như trong quá trình tối ưu hóa. Thứ hai, mỗi kiến trúc mạng nơ-ron thường được phát triển cho những mục đích riêng hoặc để giải quyết các thách thức đặc thù của một bài toán cụ thể. Ví dụ, MobileNet sử dụng tích chập theo độ sâu kết hợp với tích chập theo điểm để giảm đáng kể số phép tính và tham số, đồng thời vẫn đảm bảo khả năng trích

Bảng 3.5: *Tóm tắt các thành phần chính của mô hình i-vector truyền thống và ba mô hình học sâu, ResNet-34, ECAPA-TDNN và RawNet trong bài toán xác thực người nói.*

Thành phần	i-vector [18]	ResNet-34 [13]	ECAPA-TDNN [21]	RawNet [48]
Đặc trưng đầu vào	MFCCs	Ảnh phổ	MFCCs	1D tín hiệu thô
Đặc trưng mức frame	MFCCs	Khối dư	1D SE, Khối dư	Khối dư
Đặc trưng mức câu nói	UBM	phân cụm NetVLAD	ASP	GRU
Mô hình người nói	JFA	Lớp kết nối đầy đủ	Lớp kết nối đầy đủ	Lớp kết nối đầy đủ
Huấn luyện và hàm mục tiêu	01/02 giai đoạn: JFA, PLDA/SVM	02 giai đoạn: Frontend: SR với hàm loss CE, Backend: SV với Contrastive loss	01 giai đoạn SR với AAM-Softmax, PLDA	02 giai đoạn: Front-end: SR với CE, Center, BS losses. Backend: SV với CE loss
Mô hình người nói	i-vector	Đặc trưng nhúng người nói và bộ phân loại quan hệ	Giá trị trung bình đặc trưng nhúng theo người nói	Phân loại dựa trên b-vector
Đánh giá	Tương đồng Cosine	Trung bình của tương đồng Cosine và điểm phân loại	Tương đồng cosine được chuẩn hóa	Điểm phân loại

xuất các đặc trưng quan trọng từ dữ liệu đầu vào. Các kiến trúc MobileNet đặc biệt hữu ích khi triển khai mô hình học sâu trên các thiết bị có tài nguyên tính toán hạn chế.

Việc chọn lựa một mô hình huấn luyện có trước phù hợp là yếu tố then chốt để đảm bảo quá trình học chuyển giao diễn ra hiệu quả. Trong các phần sau, NCS trình bày thực nghiệm và phân tích đánh giá về hiệu quả của việc học chuyển giao đối với bài toán xác thực dữ liệu người nói tiếng Việt.

3.4. Thực nghiệm

3.4.1. 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 [17]. 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.

3.4.2. 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. NCS sử dụng mô hình huấn luyện trước trên tập dữ liệu VoxCeleb2, sau đó tinh chỉnh trên tập huấn luyện VLSP2021-SV (1,305 người nói). Dữ liệu đánh giá là tập VLSP2021-SV bao gồm 20,000 cặp câu nói của 114 người nói. Tập dữ liệu đánh giá độc lập với dữ liệu huấn luyện. 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, sử dụng 80 MFCCs (1D), các tham số còn lại giống như mô hình gốc ECAPA-TDNN [21];
- 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 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 [57] để bổ sung dữ liệu huấn luyện. Quá trình tăng cường dữ liệu thực hiện như sau: với mỗi đoạn âm thanh huấn luyện, NCS sử dụng 6 chiến lược tăng cường khác nhau:

- Tiếng nói: Chọn ngẫu nhiên từ ba đến bảy người nói, trộn lẫn và sau đó thêm vào tín hiệu gốc với mức SNR từ 13-20 dB.
- Nhạc: Một tệp nhạc được chọn ngẫu nhiên từ MUSAN và thêm vào âm thanh gốc với mức tỷ lệ tín hiệu trên nhiễu (SNR) từ 5-15 dB. Thời lượng của tiếng nhạc được điều chỉnh sao cho khớp với thời lượng của tín hiệu gốc.
- Tiếng ồn: Tiếng ồn được chọn ngẫu nhiên từ MUSAN và thêm vào đoạn ghi âm gốc với mức SNR từ 0-15 dB.
- Dội âm: Tín hiệu được dội âm nhân tạo bằng phương pháp tích chập với dữ liệu RIRs thực tế.

3.4.3. Độ đo

Trong thực nghiệm này NCS sử dụng độ đo tỷ lệ lỗi bằng nhau (EER) để đánh giá tính độ chính xác của hệ thống xác thực người nói.

3.4.4. 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,

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

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

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 vectơ âm học một chiều của ECAPA-TDNN chỉ ra ưu điểm 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.

ECAPA-TDNN và RawNet3 có hiệu suất tính toán cao 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 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).

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 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 12.74% (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. Tập 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 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 CÔNG TRÌNH

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.

Tài liệu tham khảo

[1] Zrar Kh. Abdul and Abdulbasit K. Al-Talabani (2022), *Mel Frequency Cepstral Coefficient and its Applications: A Review*, IEEE Access, 10, pp. 122136–122158, DOI: 10.1109/ACCESS.2022.3223444.

[2] Ahmed Kamil Hasan Al-Ali, Bouchra Senadji, and Ganesh R. Naik, *Enhanced forensic speaker verification using multi-run ICA in the presence of environmental noise and reverberation conditions*, in: 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2017, pp. 174–179, DOI: 10.1109/ICSIPA.2017.8120601.

[3] Khamis A Al-Karawi (2021), *Mitigate the reverberation effect on the speaker verification performance using different methods*, International Journal of Speech Technology, 24, pp. 143–153, ISSN: 1572-8110, DOI: 10.1007/s10772-020-09780-1, URL: https://doi.org/10.1007/s10772-020-09780-1.

[4] Abdulbasit Al-Talabani and Fatima Faek (2013), *Speaker Recognition from Noisy Spoken Sentences*, International Journal of Computer Applications, 70, pp. 11–14, DOI: 10.5120/12182-8213.

[5] Relja Arandjelovic et al., *NetVLAD: CNN architecture for weakly supervised place recognition*, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5297–5307.

[6] Aweem Ashar, Muhammad Shahid Bhatti, and Usama Mushtaq, *Speaker identification using a hybrid cnn-mfcc approach*, in: 2020 International Conference on Emerging Trends in Smart Technologies (ICETST), IEEE, 2020, pp. 1–4.

[7] Hyun Bae, Ho Lee, and Suk Lee (2016), *Voice Recognition-Based on Adaptive MFCC and Deep Learning for Embedded Systems*, Journal of Institute of Control, Robotics and Systems, 22, pp. 797–802, DOI: 10.5302/J.ICROS.2016.16.0136.

[8] Alexei Baevski et al. (2020), *wav2vec 2.0: A framework for self-supervised learning of speech representations*, Advances in neural information processing systems, 33, pp. 12449–12460.

[9] Zhongxin Bai and Xiao Lei Zhang (2021), *Speaker recognition based on deep learning: An overview*, Neural Networks, 140, pp. 65–99, ISSN: 18792782, DOI: 10.1016/j.neunet.2021.03.004.

[10] Danwei Cai, Xiaoyi Qin, and Ming Li, *Multi-Channel Training for End-to-End Speaker Recognition Under Reverberant and Noisy Environment*, in: Proc. Interspeech 2019, 2019, pp. 4365–4369, DOI: 10.21437/Interspeech.2019-1437.

[11] Shi-Huang Chen and Yu-Ren Luo, *Speaker verification using MFCC and support vector machine*, in: Proceedings of the International multiconference of engineers and computer scientists, vol. 1, Citeseer, 2009, pp. 18–20.

[12] Anurag Chowdhury and Arun Ross (2019), *Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals*, IEEE transactions on information forensics and security, 15, pp. 1616–1629.

[13] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, *VoxCeleb2: Deep Speaker Recognition*, in: Proc. Interspeech 2018, 2018, pp. 1086–1090, DOI: 10.21437/Interspeech.2018-1929.

[14] Joon Son Chung et al., *VoxSRC 2019: The first VoxCeleb Speaker Recognition Challenge*, 2019, arXiv: 1912.02522 [cs.SD], URL: https://arxiv.org/abs/1912.02522.

[15] Joon Son Chung et al., *In Defence of Metric Learning for Speaker Recognition*, in: Proc. Interspeech 2020, 2020, pp. 2977–2981, DOI: 10.21437/Interspeech.2020-1064.

[16] Sandro Cumani et al., *Comparison of speaker recognition approaches for real applications*, in: Proc. Interspeech 2011, 2011, pp. 2365–2368, DOI: 10.21437/Interspeech.2011-64.

[17] Vi Thanh Dat, Pham Thanh, and Nguyen Thi Thu Trang (2022), *VLSP 2021 - SV challenge: Vietnamese Speaker Verification in Noisy Environments*, VNU Journal of Science: Computer Science and Communication Engineering, 38, ISSN: 2588-1086, DOI: 10.25073/2588-

1086/vnucsce.333, URL: //jcsce.vnu.edu.vn/index.php/jcsce/article/view/333.

[18] Najim Dehak et al. (2011), *Front-End Factor Analysis for Speaker Verification*, IEEE Transactions on Audio, Speech, and Language Processing, 19 (4), pp. 788–798, DOI: 10.1109/TASL.2010.2064307.

[19] Jia Deng et al., *Imagenet: A large-scale hierarchical image database*, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[20] Jiankang Deng et al., *ArcFace: Additive Angular Margin Loss for Deep Face Recognition*, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4685–4694, DOI: 10.1109/CVPR.2019.00482.

[21] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, *ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification*, in: Proc. Interspeech 2020, 2020, pp. 3830–3834, DOI: 10.21437/Interspeech.2020-2650.

[22] Xingping Dong and Jianbing Shen, *Triplet Loss in Siamese Network for Object Tracking*, in: Proceedings of the European Conference on Computer Vision (ECCV), Sept. 2018.

[23] Yue Fan et al., *Cn-celeb: a challenging chinese speaker recognition dataset*, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 7604–7608.

[24] Hume A Feldman, Nick Kaiser, and John A Peacock (1993), *Power spectrum analysis of three-dimensional redshift surveys*, The Astrophysical Journal, 426, pp. 23–37, URL: https://api.semanticscholar.org/CorpusID:15943631.

[25] Qinjian Fu et al., *Research on crane sound clustering of MFCC based on HHT*, in: Journal of Physics: Conference Series, vol. 1693, 1, IOP Publishing, 2020, p. 012134.

[26] Ross Girshick, *Fast r-cnn*, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[27] Emanuel AP Habets (2006), *Room impulse response generator*, Technische Universiteit Eindhoven, Tech. Rep, 2 (2.4), p. 1.

[28] Harish, Hegde Rajesh M Ahmad Waquar, and Karnick, *Cosine Distance Metric Learning for Speaker Verification Using Large Margin Nearest Neighbor Method*, in: Advances in Multimedia Information Processing – PCM 2014, ed. by Cees G M. et al., Springer International Publishing, 2014, pp. 294–303, ISBN: 978-3-319-13168-9.

[29] Md Hasan et al. (2004), *Speaker Identification Using Mel Frequency Cepstral Coefficients*, Proceedings of the 3rd International Conference on Electrical and Computer Engineering (ICECE 2004).

[30] Kaiming He et al., *Deep residual learning for image recognition*, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[31] Kaiming He et al. (2016), *Deep residual learning for image recognition*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem, pp. 770–778, ISSN: 10636919, DOI: 10.1109/CVPR.2016.90.

[32] Yedid Hoshen, Ron J. Weiss, and Kevin W. Wilson, *Speech acoustic modeling from raw multichannel waveforms*, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4624–4628, DOI: 10.1109/ICASSP.2015.7178847.

[33] Danoush Hosseinzadeh and Sridhar Krishnan, *Combining Vocal Source and MFCC Features for Enhanced Speaker Recognition Performance Using GMMs*, in: 2007 IEEE 9th Workshop on Multimedia Signal Processing, 2007, pp. 365–368, DOI: 10.1109/MMSP.2007.4412892.

[34] Andrew G Howard et al. (2017), *Mobilenets: Efficient convolutional neural networks for mobile vision applications*, arXiv preprint arXiv:1704.04861.

[35] Wei-Ning Hsu et al. (2021), *Hubert: Self-supervised speech representation learning by masked prediction of hidden units*, IEEE/ACM transactions on audio, speech, and language processing, 29, pp. 3451–3460.

[36] Jie Hu, Li Shen, and Gang Sun, *Squeeze-and-Excitation Networks*, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141, DOI: 10.1109/CVPR.2018.00745.

[37] Zili Huang, Shuai Wang, and Kai Yu, *Angular Softmax for Short-Duration Text-independent Speaker Verification*, in: Proc. Interspeech 2018, 2018, pp. 3623–3627, DOI: 10.21437/Interspeech.2018-1545.

[38] Jaesung Huh et al., *VoxSRC 2022: The Fourth VoxCeleb Speaker Recognition Challenge*, 2023, arXiv: 2302 . 10248 [cs.SD], URL: https://arxiv.org/abs/2302.10248.

[39] Miquel India, Pooyan Safari, and Javier Hernando, *Self Multi-Head Attention for Speaker Recognition*, 2019, arXiv: 1906.09890 [cs.SD], URL: https://arxiv.org/abs/1906.09890.

[40] Sergey Ioffe, *Probabilistic Linear Discriminant Analysis*, in: Computer Vision – ECCV 2006, ed. by Aleš Leonardis, Horst Bischof, and Axel Pinz, Springer Berlin Heidelberg, 2006, pp. 531–542, ISBN: 978-3-540-33839-0.

[41] Rashid Jahangir et al. (2020), *Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network*, IEEE Access, 8, pp. 32187–32202, DOI: 10.1109/ACCESS.2020.2973541.

[42] H S Jayanna and Mahadeva S (2009), *Analysis, Feature Extraction, Modeling and Testing Techniques for Speaker Recognition*, IETE Technical Review, 26, DOI: 10.4103/0256-4602.50702.

[43] Biswajit Jena, Gopal Nayak, and Sanjay Saxena (2021), *Convolutional neural network and its pretrained models for image classification and object detection: A survey*, Concurrency and Computation: Practice and Experience, 34, DOI: 10.1002/cpe.6767.

[44] Jee weon Jung et al., *RawNet: Advanced End-to-End Deep Neural Network Using Raw Waveforms for Text-Independent Speaker Verification*, in: Proc. Interspeech 2019, 2019, pp. 1268–1272, DOI: 10.21437/Interspeech.2019-1982.

[45] Jee weon Jung et al., *Improved RawNet with Feature Map Scaling for Text-Independent Speaker Verification Using Raw Waveforms*, in: Proc. Interspeech 2020, 2020, pp. 1496–1500, DOI: 10.21437/Interspeech.2020-1011.

[46] Jee weon Jung et al., *Pushing the limits of raw waveform speaker recognition*, in: Proc. Interspeech 2022, 2022, pp. 2228–2232, DOI: 10.21437/Interspeech.2022-126.

[47] Jee weon Jung et al. (2022), *Large-scale learning of generalised representations for speaker recognition*, ArXiv, abs/2210.10985, URL: https://api.semanticscholar.org/CorpusID:253018519.

[48] Jee-weon Jung et al. (2022), *Pushing the limits of raw waveform speaker recognition*, arXiv preprint arXiv:2203.08488.

[49] Zahi N. Karam, William M. Campbell, and Najim Dehak, *Towards reduced false-alarms using cohorts*, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 4512–4515, DOI: 10.1109/ICASSP.2011.5947357.

[50] Patrick Kenny et al. (2007), *Joint factor analysis versus eigenchannels in speaker recognition*, IEEE Transactions on Audio, Speech, and Language Processing, 15 (4), pp. 1435–1447.

[51] Awais Khan et al. (2022), *Voice spoofing countermeasures: Taxonomy, state-of-the-art, experimental analysis of generalizability, open challenges, and the way forward*, arXiv preprint arXiv:2210.00417.

[52] Umair Khan and Francisco Javier Hernando Pericás, *Unsupervised training of siamese networks for speaker verification*, in: Interspeech 2020: the 20th Annual Conference of the International Speech Communication Association: 25-29 October 2020: Shanghai, China, International Speech Communication Association (ISCA), 2020, pp. 3002–3006.

[53] Prannay Khosla et al., *Supervised Contrastive Learning*, in: Advances in Neural Information Processing Systems, ed. by H. Larochelle et al., vol. 33, Curran Associates, Inc., 2020, pp. 18661–18673, URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

[54] Heekyu Kim et al. (2023), *Non-invasive way to diagnose dysphagia by training deep learning model with voice spectrograms*, Biomedical Signal Processing and Control, 86, p. 105259, ISSN: 1746-8094, DOI: https://doi.org/10.1016/j.bspc.2023.105259, URL: https://www.sciencedirect.com/science/article/pii/S1746809423006924.

[55] Ju-Ho Kim et al., *RawNeXt: Speaker Verification System For Variable-Duration Utterances With Deep Layer Aggregation And Extended Dynamic Scaling Policies*, in: ICASSP 2022 - 2022 IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 7647–7651, DOI: 10.1109/ICASSP43922.2022.9747594.

[56] Diederik Kingma and Jimmy Ba, *Adam: A Method for Stochastic Optimization*, in: International Conference on Learning Representations (ICLR), 2015.

[57] Tom Ko et al., *A study on data augmentation of reverberant speech for robust speech recognition*, in: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2017, pp. 5220–5224.

[58] Alexander Kolesnikov et al. (2019), *Large Scale Learning of General Visual Representations for Transfer*, ArXiv, abs/1912.11370, URL: https://api.semanticscholar.org/CorpusID:209460680.

[59] Nikita Kuzmin, Igor Fedorov, and Alexey Sholokhov, *Magnitude-Aware Probabilistic Speaker Embeddings*, in: Proc. The Speaker and Language Recognition Workshop (Odyssey 2022), 2022, pp. 1–8, DOI: 10.21437/Odyssey.2022-1.

[60] Chao Li et al. (2017), *Deep Speaker: an End-to-End Neural Speaker Embedding System*, CoRR, abs/1705.02304, arXiv: 1705.02304, URL: http://arxiv.org/abs/1705.02304.

[61] Lantian Li et al., *Deep Speaker Feature Learning for Text-Independent Speaker Verification*, in: Interspeech, 2017, URL: https://api.semanticscholar.org/CorpusID:31006202.

[62] Lantian Li et al., *CN-Celeb-AV: A Multi-Genre Audio-Visual Dataset for Person Recognition*, in: Proc. INTERSPEECH 2023, 2023, pp. 2118–2122, DOI: 10.21437/Interspeech.2023-1674.

[63] Weiwei Lin and Man-Wai Mak, *Wav2Spk: A Simple DNN Architecture for Learning Speaker Embeddings from Waveforms*, in: Proc. Interspeech 2020, 2020, pp. 3211–3215, DOI: 10.21437/Interspeech.2020-1287.

[64] Weiwei Lin et al., *Multi-Level Deep Neural Network Adaptation for Speaker Verification Using MMD and Consistency Regularization*, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6839–6843, DOI: 10.1109/ICASSP40776.2020.9054134.

[65] Hui Liu and Longlian Zhao (2019), *A Speaker Verification Method Based on TDNN–LSTMP*, Circuits, Systems, and Signal Processing, 38, pp. 4840–4854, ISSN: 15315878, DOI: 10.1007/s00034-019-01092-3, URL: https://doi.org/10.1007/s00034-019-01092-3.

[66] Hieu-Thi Luong and Hai-Quan Vu, *A non-expert Kaldi recipe for Vietnamese Speech Recognition System*, in: Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016), ed. by Yohei Murakami et al., The COLING 2016 Organizing Committee, Dec. 2016, pp. 51–55, URL: https://aclanthology.org/W16-5207.

[67] Yufeng Ma et al., *SpeakIn Team for the VoxCeleb Speaker Recognition Challange 2020*, 2020, URL: https://www.robots.ox.ac.uk/~vgg/data/voxceleb/data_workshop_2020/speakin.pdf.

[68] Yufeng Ma et al. (2021), *Rep works in speaker verification*, arXiv preprint arXiv:2110.09720.

[69] N J Metilda Sagaya Mary, S Umesh, and Sandesh V Katta, *S-vectors and TESA: Speaker Embeddings and a Speaker Authenticator Based on Transformer Encoder*, 2021, arXiv: 2008.04659 [eess.AS], URL: https://arxiv.org/abs/2008.04659.

[70] Pavel Matějka et al., *Analysis of Score Normalization in Multilingual Speaker Recognition*, in: Proc. Interspeech 2017, 2017, pp. 1567–1571, DOI: 10.21437/Interspeech.2017-803.

[71] Rafizah Mohd Hanifa, Khalid Isa, and Shamsul Mohamad (2021), *A review on speaker recognition: Technology and challenges*, Computers Electrical Engineering, 90, p. 107005, ISSN: 0045-7906, DOI: https://doi.org/10.1016/j.compeleceng.2021.107005, URL: https://www.sciencedirect.com/science/article/pii/S0045790621000318.

[72] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, *VoxCeleb: A Large-Scale Speaker Identification Dataset*, in: Proc. Interspeech 2017, 2017, pp. 2616–2620, DOI: 10.21437/Interspeech.2017-950.

[73] Arsha Nagrani et al., *VoxSRC 2020: The Second VoxCeleb Speaker Recognition Challenge*, 2020, arXiv: 2012.06867 [cs.SD], URL: https://arxiv.org/abs/2012.06867.

[74] Arsha Nagrani et al. (2020), *Voxceleb: Large-scale speaker verification in the wild*, Computer Speech & Language, 60, p. 101027.

[75] Seiichi Nakagawa, Kouhei Asakawa, and Longbiao Wang, *Speaker recognition by combining MFCC and phase information*, in: Proc. Interspeech 2007, 2007, pp. 2005–2008, DOI: 10.21437/Interspeech.2007-161.

[76] Son T. Nguyen et al., *Vietnamese Speaker Authentication Using Deep Models*, in: Proceedings of the 9th International Symposium on Information and Communication Technology, SoICT '18, Association for Computing Machinery, 2018, 177–184, ISBN: 9781450365390, DOI: 10.1145/3287921.3287954, URL: https://doi.org/10.1145/3287921.3287954.

[77] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, *Attentive Statistics Pooling for Deep Speaker Embedding*, in: Interspeech 2018, 2018, pp. 2252–2256, DOI: 10.21437/Interspeech.2018-993.

[78] Manuel Pariente et al., *Filterbank design for end-to-end speech separation*, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 6364–6368.

[79] Daniel S. Park et al., *SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition*, in: Interspeech, 2019, URL: https://api.semanticscholar.org/CorpusID:121321299.

[80] Heewoong Park et al., *Training Utterance-level Embedding Networks for Speaker Identification and Verification*, in: Proc. Interspeech 2018, 2018, pp. 3563–3567, DOI: 10.21437/Interspeech.2018-1044.

[81] Adam Paszke et al., *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, in: Advances in Neural Information Processing Systems, ed. by H Wallach et al., vol. 32, Curran Associates, Inc., 2019, pp. 8024–8035, URL: https://proceedings.neurips.cc/paper/2019/file/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[82] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, *A time delay neural network architecture for efficient modeling of long temporal contexts.* In: Interspeech, 2015, pp. 3214–3218.

[83] Viet Thanh Pham et al., *Vietnam-Celeb: a large-scale dataset for Vietnamese speaker recognition*, in: Proc. INTERSPEECH 2023, 2023, pp. 1918–1922, DOI: 10.21437/Interspeech.2023-1989.

[84] Tuan Phan, Nam Vu, and Cuong Pham, *Multi-task Learning based Voice Verification with Triplet Loss*, in: 2020 International Conference on Multimedia Analysis and Pattern Recognition, MAPR 2020, Institute of Electrical and Electronics Engineers Inc., Oct. 2020, ISBN: 9781728165554, DOI: 10.1109/MAPR49794.2020.9237767.

[85] Arnab Poddar, Md Sahidullah, and Goutam Saha (2017), *Speaker Verification with Short Utterances: A Review of Challenges, Trends and Opportunities*, IET Biometrics, 7, DOI: 10.1049/iet-bmt.2017.0065.

[86] Gang Qian et al., *Similarity between Euclidean and cosine angle distance for nearest neighbor queries*, in: Proceedings of the 2004 ACM Symposium on Applied Computing, Association for Computing Machinery, 2004, pp. 1232–1237, ISBN: 1581138121, DOI: 10.1145/967900.968151, URL: https://doi.org/10.1145/967900.968151.

[87] L R Rabiner (1989), *A tutorial on hidden Markov models and selected applications in speech recognition*, Proceedings of the IEEE, 77, pp. 257–286, DOI: 10.1109/5.18626.

[88] K. Sreenivasa Rao and K. E. Manjunath (2017), *Speech Recognition Using Articulatory and Excitation Source Features*, 1st, Springer Publishing Company, Incorporated, ISBN: 3319492195.

[89] Mirco Ravanelli and Yoshua Bengio, *Speaker Recognition from Raw Waveform with SincNet*, in: 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 1021–1028, DOI: 10.1109/SLT.2018.8639585.

[90] Mirco Ravanelli et al., *SpeechBrain: A General-Purpose Speech Toolkit*, arXiv:2106.04624, 2021, arXiv: 2106.04624 [eess.AS].

[91] Douglas Reynolds, *Gaussian Mixture Models*, in: Encyclopedia of Biometrics, ed. by Stan Z. Li and Anil Jain, Springer US, 2009, pp. 659–663, ISBN: 978-0-387-73003-5, DOI: 10.1007/978-0-387-73003-5_196, URL: https://doi.org/10.1007/978-0-387-73003-5_196.

[92] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn (2000), *Speaker Verification Using Adapted Gaussian Mixture Models*, Digital Signal Processing, 10, pp. 19–41, ISSN: 1051-2004, DOI: https://doi.org/10.1006/dspr.1999.0361, URL: https://www.sciencedirect.com/science/article/pii/S1051200499903615.

[93] Johan Rohdin et al., *Speaker Verification Using End-to-end Adversarial Language Adaptation*, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6006–6010, DOI: 10.1109/ICASSP.2019.8683616.

[94] Mukund K Roy and Ushaben Keshwala, *Res2Net based Text Independent Speaker recognition system*, in: 2022 12th International Conference on Cloud Computing, Data Science Engineering (Confluence), 2022, pp. 612–616, DOI: 10.1109/Confluence52989.2022.9734175.

[95] Seyed Omid Sadjadi et al. (2019), *The 2018 NIST speaker recognition evaluation*, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2019-Septe, pp. 1483–1487, ISSN: 19909772, DOI: 10.21437/Interspeech.2019-1351.

[96] Pooyan Safari, Miquel India, and Javier Hernando, *Self-Attention Encoding and Pooling for Speaker Recognition*, in: Proc. Interspeech 2020, 2020, pp. 941–945, DOI: 10.21437/Interspeech.2020-1446.

[97] Md Sahidullah and Goutam Saha (2012), *A novel windowing technique for efficient computation of MFCC for speaker recognition*, IEEE signal processing letters, 20 (2), pp. 149–152.

[98] Tara N. Sainath et al., *Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks*, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2015-August, Institute of Electrical and Electronics Engineers Inc., Aug. 2015, pp. 4580–4584, ISBN: 9781467369978, DOI: 10.1109/ICASSP.2015.7178838.

[99] S. Shahnawazuddin et al., *In-Domain and Out-of-Domain Data Augmentation to Improve Children's Speaker Verification System in Limited Data Scenario*, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7554–7558, DOI: 10.1109/ICASSP40776.2020.9053891.

[100] Suwon Shon, Hao Tang, and James Glass, *Frame-Level Speaker Embeddings for Text-Independent Speaker Recognition and Analysis of End-to-End Model*, in: 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 1007–1013, DOI: 10.1109/SLT.2018.8639622.

[101] Karen Simonyan and Andrew Zisserman (2014), *Very deep convnets for large-scale image recognition*, Computing Research Repository.

[102] Vrijendra Singh and Narendra Meena, *Engine Fault Diagnosis using DTW, MFCC and FFT*, in: Proceedings of the First International Conference on Intelligent Human Computer Interaction, ed. by U. S. Tiwary et al., Springer India, 2009, pp. 83–94, ISBN: 978-81-8489-203-1.

[103] Jake Snell, Kevin Swersky, and Richard Zemel, *Prototypical networks for few-shot learning*, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., 2017, 4080–4090, ISBN: 9781510860964.

[104] David Snyder, Guoguo Chen, and Daniel Povey (2015), *Musan: A music, speech, and noise corpus*, arXiv preprint arXiv:1510.08484.

[105] David Snyder et al., *Deep neural network-based speaker embeddings for end-to-end speaker verification*, in: 2016 IEEE Spoken Language Technology Workshop (SLT), 2016, pp. 165–170, DOI: 10.1109/SLT.2016.7846260.

[106] David Snyder et al., *Deep Neural Network Embeddings for Text-Independent Speaker Verification*, in: INTERSPEECH, 2017.

[107] David Snyder et al., *Deep Neural Network Embeddings for Text-Independent Speaker Verification*, in: Proc. Interspeech 2017, 2017, pp. 999–1003, DOI: 10.21437/Interspeech.2017-620.

[108] David Snyder et al., *X-vectors: Robust dnn embeddings for speaker recognition*, in: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2018, pp. 5329–5333.

[109] Gilbert Strang (1999), *The Discrete Cosine Transform*, SIAM Rev., 41, pp. 135–147, URL: https://api.semanticscholar.org/CorpusID:6693738.

[110] Christian Szegedy et al., *Inception-v4, inception-resnet and the impact of residual connections on learning*, in: Proceedings of the AAAI conference on artificial intelligence, vol. 31, 1, 2017.

[111] Cao Truong Tran, Dinh Tan Nguyen, and Ho Tan Hoang, *Deep Representation Learning for Vietnamese Speaker Recognition*, in: 2021 13th International Conference on Knowledge and Systems Engineering (KSE), 2021, pp. 1–4, DOI: 10.1109/KSE53942.2021.9648808.

[112] Xavier Valero and Francesc Alias (2012), *Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification*, IEEE transactions on multimedia, 14, pp. 1684–1689.

[113] Dinh Van Hung et al. (2022), *SV-VLSP2021: The Smartcall-ITS's Systems*, VNU Journal of Science: Computer Science and Communication Engineering, 38 (1).

[114] Ehsan Variani et al., *Deep neural networks for small footprint text-dependent speaker verification*, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 4052–4056, DOI: 10.1109/ICASSP.2014.6854363.

[115] Ashish Vaswani et al., *Attention is all you need*, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., 2017, 6000–6010, ISBN: 9781510860964.

[116] Feng Wang and Huaping Liu (2020), *Understanding the Behaviour of Contrastive Loss*, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2495–2504, URL: https://api.semanticscholar.org/CorpusID:229297730.

[117] Feng Wang et al. (2018), *Additive Margin Softmax for Face Verification*, IEEE Signal Processing Letters, 25 (7), 926–930, ISSN: 1558-2361, DOI: 10.1109/lsp.2018.2822810, URL: http://dx.doi.org/10.1109/LSP.2018.2822810.

[118] Jixuan Wang et al. (2019), *Centroid-based deep metric learning for speaker recognition*, CoRR, abs/1902.02375, arXiv: 1902.02375, URL: http://arxiv.org/abs/1902.02375.

[119] Shuai Wang et al., *Investigation of Specaugment for Deep Speaker Embedding Learning*, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7139–7143, DOI: 10.1109/ICASSP40776.2020.9053481.

[120] Zhiming Wang et al., *AntVoice Neural Speaker Embedding System for FFSVC 2020*, in: Interspeech 2021, 2021, pp. 1069–1073, DOI: 10.21437/Interspeech.2021-966.

[121] Weidi Xie et al. (2019), *Utterance-level Aggregation for Speaker Recognition in the Wild*, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2019-May, pp. 5791–5795, ISSN: 15206149, DOI: 10.1109/ICASSP.2019.8683120.

[122] Samir S Yadav and Shivajirao M Jadhav (2019), *Deep convolutional neural network based medical image classification for disease diagnosis*, Journal of Big data, 6 (1), pp. 1–18.

[123] Shiqing Yang and Min Liu, *Data augmentation for speaker verification*, in: Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering, 2022, pp. 1247–1251.

[124] Hui Yin, Volker Hohmann, and Climent Nadeu (2011), *Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency*, Speech Communication, 53 (5), Perceptual and Statistical Audition, pp. 707–715, ISSN: 0167-6393, DOI: https://doi.org/10.1016/j.specom.2010.04.008, URL: https://www.sciencedirect.com/science/article/pii/S0167639310000919.

[125] Steve J. Young et al. (2006), *The HTK Book Version 3.4*, Cambridge University Press.

[126] Hossein Zeinali et al. (2020), *SdSV Challenge 2020: Large-scale evaluation of short-duration speaker verification*, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2020-Octob, pp. 731–735, ISSN: 19909772, DOI: 10.21437/Interspeech.2020-1485.

[127] Li Zhang, Jian Wu, and Lei Xie, *NPU Speaker Verification System for INTERSPEECH 2020 Far-Field Speaker Verification Challenge*, in: Proc. Interspeech 2020, 2020, pp. 3471–3475.

[128] Li Zhang, Jian Wu, and Lei Xie (2020), *NPU speaker verification system for INTERSPEECH 2020 far-field speaker verification challenge*, arXiv preprint arXiv:2008.03521.

[129] Li Zhang et al., *Multi-Level Transfer Learning from Near-Field to Far-Field Speaker Verification*, in: Interspeech 2021, 2021, pp. 1094–1098, DOI: 10.21437/Interspeech.2021-1980.

[130] Li Zhang et al., *Distance-based weight transfer for fine-tuning from near-field to far-field speaker verification*, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[131] Li Zhang et al. (2024), *Whisper-SV: Adapting Whisper for low-data-resource speaker verification*, Speech Communication, 163, p. 103103, ISSN: 0167-6393, DOI: https://doi.org/10.1016/j.specom.2024.103103, URL: https://www.sciencedirect.com/science/article/pii/S016763932400075X.

[132] Yang Zhang et al., *MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification*, in: Proc. Interspeech 2022, 2022, pp. 306–310, DOI: 10.21437/Interspeech.2022-563.

[133] Miao Zhao et al. (2021), *The SpeakIn System for VoxCeleb Speaker Recognition Challange 2021*, ArXiv, abs/2109.01989.

[134] Xiaojia Zhao and DeLiang Wang, *Analyzing noise robustness of MFCC and GFCC features in speaker identification*, in: 2013 IEEE international conference on acoustics, speech and signal processing, IEEE, 2013, pp. 7204–7208.

[135] Ge Zhu, Fei Jiang, and Zhiyao Duan, *Y-Vector: Multiscale Waveform Encoder for Speaker Embedding*, in: Proc. Interspeech 2021, 2021, pp. 96–100, DOI: 10.21437/Interspeech.2021-1707.

[136] Yingke Zhu, Tom Ko, and Brian Mak, *Mixup Learning Strategies for Text-Independent Speaker Verification*, in: Proc. Interspeech 2019, 2019, pp. 4345–4349, DOI: 10.21437/Interspeech.2019-2250.

[137] And (2005), *Improved MFCC-Based Feature for Robust Speaker Identification*, Tsinghua Science and Technology, 10 (2), pp. 158–161.