

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Đỗ Nhật Toàn - Đinh Nhật Tường

NGHIÊN CỨU MÔ HÌNH NHẬN DẠNG
NGƯỜI NÓI TRÊN TẬP DỮ LIỆU NHỎ
TIẾNG VIỆT

KHÓA LUẬN TỐT NGHIỆP CỦ NHÂN
CHƯƠNG TRÌNH CHÍNH QUY

Tp. Hồ Chí Minh, tháng 07/2023

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Đỗ Nhật Toàn - 19120688
Đinh Nhật Tường - 19120709

NGHIÊN CỨU MÔ HÌNH NHẬN DẠNG
NGƯỜI NÓI TRÊN TẬP DỮ LIỆU NHỎ
TIẾNG VIỆT

KHÓA LUẬN TỐT NGHIỆP CỦ NHÂN
CHƯƠNG TRÌNH CHÍNH QUY

GIÁO VIÊN HƯỚNG DẪN
TS. Châu Thành Đức

Tp. Hồ Chí Minh, tháng 07/2023

Nhận xét hướng dẫn

Theo bản nhận xét của giảng viên hướng dẫn (có chữ ký) do giáo vụ cung cấp.

Nhận xét phản biện

Theo bản nhận xét của giảng viên phản biện (có chữ ký) do giáo vụ cung cấp.

Lời cảm ơn

Quá trình thực hiện luận văn tốt nghiệp là giai đoạn quan trọng nhất trong quãng đời mỗi sinh viên. Đây là tiền đề nhằm trang bị cho chúng em những kỹ năng nghiên cứu, những kiến thức quý báu trước khi lập nghiệp.

Trước hết, chúng em xin chân thành cảm ơn quý Thầy, Cô khoa Công nghệ thông tin, đặc biệt là các Thầy, Cô trong chuyên ngành *Công nghệ tri thức* đã tận tình chỉ dạy và trang bị cho chúng em những kiến thức cần thiết trong suốt thời gian ngồi trên ghế giảng đường, điều đó là nền tảng cho em có thể hoàn thành được bài luận văn này.

Chúng em xin trân trọng cảm ơn thầy **Châu Thành Đức** đã tận tình giúp đỡ, định hướng cách tư duy và cách làm việc khoa học. Đó là những góp ý hết sức quý báu không chỉ trong quá trình thực hiện luận văn này mà còn là hành trang tiếp bước cho chúng em trong quá trình học tập và lập nghiệp sau này.

Chúng em xin bày tỏ lòng biết ơn sâu sắc với những ý kiến mà các anh, chị và các bạn trong lab đã góp ý cho luận văn này. Các nhận xét và gợi ý của mọi người đã làm gia tăng sự sâu sắc và phong phú cho nghiên cứu của chúng em. Chúng em rất biết ơn sự hỗ trợ và sự hướng dẫn quý báu của mọi người trong suốt quá trình này.

Do chưa có nhiều kinh nghiệm làm để tài cũng như những hạn chế về kiến thức, trong báo cáo khóa luận chắc chắn sẽ không tránh khỏi những thiếu sót. Rất mong nhận được sự nhận xét, ý kiến đóng góp, phê bình từ phía Thầy, Cô để luận văn này được hoàn thiện hơn.

Xin chúc những điều tốt đẹp nhất sẽ luôn đồng hành cùng mọi người.

Lời cam đoan

Chúng em xin cam đoan đây là công trình nghiên cứu của riêng chúng em. Các số liệu và kết quả nghiên cứu trong luận văn này là trung thực và không trùng lặp với các đề tài khác.

Đề cương chi tiết

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP
NGHIÊN CỨU MÔ HÌNH NHẬN DẠNG
NGƯỜI NÓI TRÊN TẬP DỮ LIỆU NHỎ
TIẾNG VIỆT

(*Research on Speaker Recognition Model for Small Vietnamese Dataset*)

1 THÔNG TIN CHUNG

Người hướng dẫn:

– TS. Châu Thành Đức (Khoa Công nghệ Thông tin)

Nhóm viên thực hiện:

1. Đỗ Nhật Toàn (MSSV: 19120688)
2. Đinh Nhật Tường (MSSV: 19120709)

Loại đề tài: Nghiên cứu

Thời gian thực hiện: Từ tháng 1/năm 2023 đến tháng 7/năm 2023

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

Hiện nay, việc nhận dạng người nói vẫn đang nắm giữ vai trò khá quan trọng trong đời sống của chúng ta. Một số ứng dụng có thể kể đến như là xác minh danh tính của một người, xác thực thông tin cá nhân,... Cho đến hiện tại, đã và đang có rất nhiều mô hình và phương pháp có thể phục vụ các mục đích vừa nêu trên, nhưng các mô hình hiện đại trên đều được huấn luyện trên bộ dữ liệu của các ngôn ngữ giàu tài nguyên. Từ đó nhóm quyết định tìm hiểu về những phương pháp nhận dạng người nói tốt nhất hiện nay (state-of-the-art, gọi tắt là SOTA) và thử nghiệm chúng dựa trên bộ dữ liệu nhỏ hơn (dưới 1000 người) như tiếng Việt để khảo sát tính nhất quán với các ngôn ngữ ban đầu.

2.2 Mục tiêu đề tài

Trong đề tài này, nhóm sẽ tổng hợp các mô hình SOTA và cho đánh giá khả năng nhận dạng của các mô hình với nhau trên các bộ dữ liệu bằng tiếng Anh và tiếng Việt. Sau đó nhóm sẽ khảo sát và phân tích khả năng nhận dạng người nói của các mô hình sẽ thay đổi như thế nào nếu dữ liệu bị ảnh hưởng bởi các yếu tố khác nhau như: cắt bỏ một phần băng tần, trộn nhiễu vào dữ liệu. Từ đó nhận xét rằng vai trò của các băng tần là như thế nào trong việc nhận dạng cũng như đánh giá mức độ ảnh hưởng của các loại nhiễu có ảnh hưởng ra sao đến việc nhận dạng người nói.

Ngoài ra, nhóm cũng khảo sát các kết quả thu được khi đánh giá trên bộ dữ liệu tiếng Việt vì dữ liệu bằng tiếng Anh hiện nay khá phổ biến và dồi dào nhưng dữ liệu tiếng Việt lại khá ít ỏi. Với lượng dữ liệu tiếng Việt hạn chế như vậy, nhóm khảo sát hiệu quả của các mô hình nhận dạng người nói có nhất quán với dữ liệu ngôn ngữ tiếng Anh hay không.

Cuối cùng, để có thể nhận dạng người nói tốt hơn trên dữ liệu nhỏ tiếng Việt giữa

lượng dữ liệu dồi dào như tiếng Anh, nhóm tiến hành một số cách tiếp cận như: huấn luyện các mô hình chỉ trên tập dữ liệu tiếng Việt, tinh chỉnh các mô hình (fine-tuning) trên dữ liệu tiếng Việt và huấn luyện các mô hình với tập dữ liệu kết hợp cả tiếng Anh lẫn tiếng Việt. Từ đó nhóm tiến hành khảo sát xem mô hình nào sẽ có hiệu suất vượt trội hơn và cách tiếp cận nào sẽ hiệu quả đối với tập dữ liệu nhỏ như tiếng Việt.

2.3 Phạm vi của đề tài

Nội dung nghiên cứu chính của đề tài là tổng hợp các mô hình nhận dạng người nói tốt nhất hiện nay, tổng hợp bộ dữ liệu tiếng Anh và tiếng Việt, thử nghiệm và so sánh kết quả của chúng đối với trường hợp bình thường và các trường hợp đặc biệt (có ảnh hưởng bởi các yếu tố đã nêu trên) cũng như tìm ra đâu là cách tiếp cận tốt nhất.

Các mô hình sẽ được đề cập: RawNet3 [1], mô hình WavLM [2].

Các tập dữ liệu sử dụng trong đề tài:

- Tiếng Anh: VoxCeleb1 [3], LibriSpeech [4]
- Tiếng Việt: VIVOS [5], Zalo Speaker Verification Dataset.

2.4 Cách tiếp cận dự kiến

Mô hình RawNet3 là mô hình nhận dạng người nói dữ liệu đầu vào là một đoạn lời nói thô (raw waveform) và dữ liệu đầu ra sẽ là một vector biểu diễn người nói tương ứng (speaker embedding vector).

Mô hình WavLM là mô hình tổng hợp đối với các tác vụ xử lý tiếng nói, như là tách kênh người nói (speaker diarization), nhận dạng tiếng nói (speech recognition), phân tách tiếng nói (speech separation), ... và kể cả xác thực người nói (speaker verification).

Nhóm sẽ đánh giá các mô hình trên dựa theo các tập dữ liệu tiếng Anh: VoxCeleb1,

LibriSpeech; các tập dữ liệu tiếng Việt: VIVOS, Zalo Speaker Verification Dataset. Và nhóm sẽ tiến hành dựa theo các mục tiêu đã được nêu trên:

- Đầu tiên là cắt bỏ tần số bằng cách sử dụng bộ lọc thông thấp (low-pass filter) và bộ lọc thông dải (band-pass filter). Gọi x là một giá trị tần số với đơn vị Hz, khi đó bộ lọc thông thấp sẽ giữ lại các mẫu tần số nhỏ hơn x và lọc các mẫu tín hiệu có tần số lớn hơn x . Gọi y là một giá trị tần số khác và lớn hơn x , bộ lọc thông dải sẽ giữ lại các giá trị tần số trong khoảng x và y và lọc các tín hiệu có giá trị tần số còn lại. Mục đích sử dụng hai bộ lọc trên là để tìm hiểu sự thay đổi về mặt hiệu quả của các phương pháp nhận dạng người nói trên nhiều chuẩn lấy mẫu (sampling rate) khác nhau như chuẩn các thiết bị bộ đàm (8,000 Hz), chuẩn điện thoại băng hẹp (16,000 Hz)¹ và xem xét từng khoảng băng tần nào là quan trọng trong nhận dạng người nói.
- Ngoài ra, nhóm cũng tiến hành thêm các loại nhiễu từ nhiễu thấp, nhiễu vừa đến nhiễu cao vào bộ dữ liệu. Cụ thể là nhiễu trắng (white noise) và nhiễu vang (reverb noise) với một lượng tỷ lệ tín hiệu trên nhiễu (Signal to Noise Ratio) được xác định từ 0 dB đến 10 dB để tìm hiểu sự thay đổi hiệu năng của các mô hình trong môi trường thực tế (có nhiễu), sau đó đánh giá và đề xuất mô hình nào tốt trong môi trường có tiếng nhiễu dựa trên kết quả thu được.
- Tiếp theo, nhóm sẽ đánh giá hiệu năng của các mô hình nhận dạng người nói trên bộ dữ liệu của ngôn ngữ chưa được huấn luyện để kiểm tra tính nhất quán của các mô hình khi được sử dụng bởi ngôn ngữ khác. Cụ thể là áp dụng mô hình nhận dạng người nói từ ngôn ngữ giàu tài nguyên (tiếng Anh) sang ngôn ngữ ít tài nguyên hơn (tiếng Việt).
- Cuối cùng, nhóm sẽ thử nghiệm nhiều cách tiếp cận khác nhau cho các mô hình nhận dạng người nói, chẳng hạn như chỉ dùng tập dữ liệu tiếng Việt để

¹<https://www.npmjs.com/package/sample-rate>

huấn luyện, fine-tuning trên dữ liệu tiếng Việt và huấn luyện bằng sự kết hợp của dữ liệu tiếng Anh và tiếng Việt.

Từ những kết quả trên, nhóm đánh giá được với ngôn ngữ có lượng tài nguyên hạn chế thì mô hình nào sẽ cho ra hiệu quả tốt nhất, cũng như phương pháp tiếp cận nào sẽ đạt hiệu quả cao nhất trên bộ dữ liệu tiếng Việt.

Ngoài ra, nhóm cũng khảo sát và so sánh hiệu năng của các mô hình nhận dạng người nói trên tiếng Việt hiện nay. Cụ thể, nhóm sẽ so sánh kết quả nghiên cứu của mình với hai nghiên cứu trên tiếng Việt [6, 7] mà nhóm thu thập được:

- Nghiên cứu thứ nhất [6] đề xuất một mô hình xác thực người nói cho tiếng Việt. Mô hình sử dụng log Mel-filterbank làm đầu vào và trích xuất đặc trưng từng frame bằng mô hình ResNet-34. Kỹ thuật Attentive Statistics Pooling [8] được sử dụng để tổng hợp đặc trưng theo frame thành một đặc trưng toàn cục, và một lớp kết nối đầy đủ (fully connected) để trích xuất embedding. Hàm mất mát là Angular Prototypical và thuật toán tối ưu là Adam.

Dữ liệu được nhóm tác giả thu thập, hợp nhất với danh tính người nói, tiền xử lý bằng loại bỏ dữ liệu không hợp lệ và thống nhất thông tin người nói.

Mô hình được cải tiến bằng việc sử dụng mô hình tiền huấn luyện trên tiếng Anh và hàm mất mát Angular Margin Prototypical Loss, với thuật toán tối ưu là SGD.

Kết quả thể hiện của mô hình khá hiệu quả, với EER của mô hình tiền huấn luyện trên tập dữ liệu tác giả là 14.954%, và sau khi học chuyển đổi là 3.115%.

- Nghiên cứu thứ hai [7] chủ yếu thử nghiệm các mô hình học sâu vào bài toán nhận dạng người nói trên tập dữ liệu tiếng Việt. Nhóm tác giả tiến hành kiểm thử trên 12 sự kết hợp, cũng như khảo sát xem kết hợp nào cho ra hiệu năng cao nhất trên bộ dữ liệu tiếng Việt.

Nhóm tác giả xây dựng bộ dữ liệu tên là VietCeleb, được trích xuất từ các video YouTube, tổng cộng gồm 5 bước được xây dựng như bộ VoxCeleb1.

Các mô hình được sử dụng: ResNetSE34V2, VGG-Vox, ResNetSE34L và ResnetSE34Half. Các hướng tiếp cận đối với các mô hình: (1) Xây dựng mô hình chỉ bằng bộ dữ liệu tiếng Việt, (2) Xây dựng mô hình bằng bộ dữ liệu tiếng Anh và kiểm thử trên tiếng Việt, (3) Xây dựng mô hình kết hợp bộ dữ liệu tiếng Anh và tiếng Việt, mô hình được huấn luyện trên bộ VoxCeleb1 và được tiếp tục trên bộ VietCeleb.

Với kết quả thu được thì phương pháp (3) là tốt nhất và tệ nhất ở phương pháp (2). Ngoài ra, mô hình ResNetSE34V2 kết hợp phương pháp (3) là mô hình tốt nhất so với ba mô hình còn lại với EER là 3%.

2.5 Kết quả dự kiến của đề tài

Các kết quả dự kiến sẽ đạt được:

- Kết quả phân tích các mô hình nhận dạng người nói trên các băng tần và mức độ ảnh hưởng của chúng trong việc nhận dạng người nói.
- Kết quả so sánh các mô hình trên dữ liệu đã được trộn nhiều và mức độ ảnh hưởng của chúng trong việc nhận dạng người nói.
- Kết quả về hiệu quả các mô hình khi được chạy trên bộ dữ liệu tiếng Việt và khi so sánh với bộ dữ liệu của ngôn ngữ tiếng Anh.
- Kết quả việc tinh chỉnh mô hình để đạt hiệu quả cao nhất trên bộ dữ liệu tiếng Việt.
- Kết quả so sánh với các mô hình nhận dạng người nói trên bộ dữ liệu tiếng Việt.
- Bài báo khoa học về việc nghiên cứu nhận dạng người nói trên tiếng Việt.

2.6 Kế hoạch thực hiện

Thời gian	Công việc
Tháng 1/2023	Tìm hiểu về bài toán nhận dạng người nói. Thu thập dữ liệu cho bài toán (các tập dữ liệu VoxCeleb1, LibriSpeech, VIVOS, Zalo Speaker Verification Dataset).
Tháng 2/2023	Tìm hiểu các mô hình đã và đang giải quyết bài toán (các mô hình i-vector, x-vector, SincNet, ...). Tìm hiểu các nghiên cứu đã và đang giải quyết bài toán trên dữ liệu tiếng Việt. Tổng hợp các mô hình SOTA có hiệu quả tốt nhất (RawNet3, WavLM).
Tháng 3/2023	Thực hiện khảo sát và phân tích khả năng nhận dạng người nói của các mô hình sẽ thay đổi như thế nào nếu dữ liệu bị ảnh hưởng bởi các yếu tố khác bằng cách cắt bỏ từng phần tần số và thêm các loại nhiễu. Nhận xét vai trò của băng tần và độ nhiễu trong việc nhận dạng người nói.
Tháng 4/2023	Đánh giá hiệu suất của các mô hình nhận dạng người nói trên ngôn ngữ tiếng Việt và kiểm tra tính nhất quán so với ngôn ngữ tiếng Anh. Thử nghiệm các phương pháp học chuyển giao, tinh chỉnh các mô hình và tiến hành kiểm chuẩn trên bộ dữ liệu đã thu thập được.
Tháng 5/2023	Viết báo cáo khóa luận.
Tháng 6/2023	Đánh giá phương pháp đã sử dụng. Hoàn thiện báo cáo khóa luận.
Tháng 7/2023	Bảo vệ khóa luận tốt nghiệp trước hội đồng.

Tài liệu

- [1] J. weon Jung, Y. J. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, “Pushing the limits of raw waveform speaker recognition,” 2022.
- [2] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, oct 2022.
- [3] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset,” in *INTERSPEECH*, 2017.
- [4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.
- [5] H.-T. Luong and H.-Q. 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)*, (Osaka, Japan), pp. 51–55, The COLING 2016 Organizing Committee, Dec. 2016.
- [6] D. V. Thanh, T. P. Viet, and T. N. T. Thu, “Deep speaker verification model for low-resource languages and Vietnamese dataset,” in *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, (Shanghai, China), pp. 442–451, Association for Computational Linguistics, 11 2021.

- [7] C. T. Tran, D. T. Nguyen, and H. T. Hoang, “Deep representation learning for vietnamese speaker recognition,” in *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 1–4, 2021.
- [8] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive statistics pooling for deep speaker embedding,” in *Interspeech 2018*, ISCA, sep 2018.
- [9] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, “Clova baseline system for the voxceleb speaker recognition challenge 2020,” 2020.

XÁC NHẬN
CỦA NGƯỜI HƯỚNG DẪN
(Ký và ghi rõ họ tên)

Xu
Châu Thành Anh

TP. Hồ Chí Minh, ngày 20 tháng 6 năm 2023
NHÓM SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)

Tuân
Đinh Nhật Trường

Quan
Đỗ Nhật Toàn

Mục lục

Nhận xét của GV hướng dẫn	i
Nhận xét của GV phản biện	ii
Lời cảm ơn	iii
Lời cam đoan	iv
Đề cương	v
Mục lục	xv
1 Giới thiệu	1
1.1 Lý do chọn đề tài	1
1.2 Mục đích nghiên cứu	2
1.3 Đối tượng nghiên cứu	2
1.4 Phạm vi nghiên cứu	3
1.5 Phương pháp nghiên cứu	3
1.6 Đóng góp của đề tài	4
1.7 Nội dung báo cáo	4
2 Các công trình liên quan	6
2.1 Cơ sở lý thuyết	6
2.1.1 Giới thiệu về nhận dạng người nói	6
2.1.2 Phương pháp trích chọn đặc trưng	7

2.1.3	Các nghiên cứu dựa trên học máy	10
2.1.4	Các nghiên cứu dựa trên học sâu	12
2.1.5	Sơ lược về việc khảo sát miền tần số với bộ lọc thông thấp và bộ lọc cấm dải	17
2.1.6	Sơ lược về việc thêm nhiễu	17
2.2	Các nghiên cứu liên quan	21
2.2.1	Nhận dạng người nói tiếng Việt bằng học biểu diễn sâu	21
2.2.2	Mô hình xác thực người nói cho ngôn ngữ ít tài nguyên với bộ dữ liệu tiếng Việt	22
3	Phương pháp đề xuất	25
3.1	Cơ sở lý thuyết	25
3.1.1	Mô hình ECAPA-TDNN	25
3.1.2	Mô hình RawNet3	28
3.1.3	Mô hình WavLM	32
3.2	Câu hỏi nghiên cứu	36
3.3	Phương pháp nghiên cứu	38
3.3.1	Huấn luyện mô hình từ đầu chỉ bằng tập dữ liệu tiếng Việt	38
3.3.2	Sử dụng mô hình tiền huấn luyện trên tập dữ liệu tiếng Anh và đánh giá trên dữ liệu tiếng Việt	39
3.3.3	Huấn luyện mô hình bằng phương pháp transfer learning cho tiếng Việt	39
3.3.4	Khảo sát ảnh hưởng của các miền tần số và nhiễu đến mô hình nhận dạng người nói	41
3.4	Mô tả dữ liệu	41
3.4.1	Bộ dữ liệu VoxCeleb1	41
3.4.2	Bộ dữ liệu VoxCeleb2	42
3.4.3	Bộ dữ liệu Zalo Voice Verification	42
3.5	Độ đo để đánh giá	43

3.5.1	Tỷ lệ lỗi bình đẳng - Equal Error Rate	43
3.5.2	Hàm chi phí phát hiện tối thiểu - Minimum Detection Cost Function	44
4	Thử nghiệm và kết quả	47
4.1	Quy trình thử nghiệm và kết quả chi tiết	47
4.1.1	Huấn luyện mô hình từ đầu bằng tập dữ liệu tiếng Việt	47
4.1.2	Sử dụng mô hình tiền huấn luyện bằng tập dữ liệu tiếng Anh cho tập dữ liệu tiếng Việt	51
4.1.3	Huấn luyện mô hình bằng fine-tuning cho tiếng Việt	52
4.1.4	Huấn luyện mô hình bằng cách kết hợp thêm mô hình nhỏ và huấn luyện trên tiếng Việt	56
4.1.5	Khảo sát ảnh hưởng của các miền tần số và nhiều đến mô hình nhận dạng người nói	62
4.2	Thảo luận	65
4.2.1	Tổng kết các phương pháp đã làm	65
4.2.2	So sánh với các mô hình nhận dạng người nói trên tiếng Việt	71
5	Kết luận và hướng phát triển	73
Tài liệu tham khảo		75

Danh sách hình

2.1	Ví dụ về hai nhiệm vụ chính trong nhận dạng người nói.	7
2.2	Kết quả sau khi áp dụng bộ lọc thông thấp vào một đoạn âm thanh mẫu với giá trị tần số từ 500 Hz đến 7500 Hz với khoảng cách mỗi bước nhảy 500 Hz	18
2.3	Kết quả sau khi áp dụng bộ lọc cầm dải vào một đoạn âm thanh mẫu trong khoảng tần số từ 4000 Hz đến 7000 Hz .	19
3.1	Kiến trúc SE-Res2Block của mô hình ECAPA-TDNN	27
3.2	Kiến trúc của mô hình ECAPA-TDNN	27
3.3	Kiến trúc của mô hình RawNet3 theo [20]	29
3.4	Kiến trúc của khối AFMS-Res2MP trong RawNet3	30
3.5	Kiến trúc của mô hình WavLM	34
3.6	Ví dụ về tỷ lệ lỗi bình đẳng với EER là điểm màu đỏ	45
4.1	Biểu đồ loss trong quá trình huấn luyện mô hình RawNet3 trên tập dữ liệu Zalo	48
4.2	Biểu đồ EER trong quá trình huấn luyện mô hình RawNet3 trên tập dữ liệu Zalo	48
4.3	Biểu đồ loss trong quá trình huấn luyện mô hình WavLM (ECAPA-TDNN) trên tập dữ liệu Zalo	49
4.4	Biểu đồ EER trong quá trình huấn luyện mô hình WavLM (ECAPA-TDNN) trên tập dữ liệu Zalo	50
4.5	Biểu đồ loss trong quá trình fine-tuning mô hình RawNet3 trên tập dữ liệu Zalo	52

4.6	Biểu đồ EER trong quá trình fine-tuning mô hình RawNet3 trên tập dữ liệu Zalo	53
4.7	Biểu đồ loss trong quá trình fine-tuning mô hình WAVLM (ECAPA-TDNN) trên tập dữ liệu Zalo	54
4.8	Biểu đồ EER trong quá trình fine-tuning mô hình WAVLM (ECAPA-TDNN) trên tập dữ liệu Zalo	55
4.9	Biểu đồ loss trong quá trình huấn luyện trên một lớp kết nối đầy đủ của mô hình RawNet3	56
4.10	Biểu đồ EER trong quá trình huấn luyện trên một lớp kết nối đầy đủ của mô hình RawNet3	57
4.11	Biểu đồ loss trong quá trình huấn luyện trên một lớp kết nối đầy đủ của mô hình WavLM (ECAPA-TDNN)	58
4.12	Biểu đồ EER trong quá trình huấn luyện trên một lớp kết nối đầy đủ của mô hình WavLM (ECAPA-TDNN)	58
4.13	Biểu đồ loss trong quá trình huấn luyện trên ba lớp kết nối đầy đủ của mô hình RawNet3	59
4.14	Biểu đồ EER trong quá trình huấn luyện trên ba lớp kết nối đầy đủ của mô hình RawNet3	60
4.15	Biểu đồ loss trong quá trình huấn luyện trên ba lớp kết nối đầy đủ của mô hình WavLM (ECAPA-TDNN)	60
4.16	Biểu đồ EER trong quá trình huấn luyện trên ba lớp kết nối đầy đủ của mô hình WavLM (ECAPA-TDNN)	61
4.17	Biểu đồ khảo sát EER của mô hình RawNet3 + 1 FC và mô hình WavLM (ECAPA-TDNN) + 3 FC trên miền tần số sau khi sử dụng LPF	63
4.18	Biểu đồ khảo sát EER của mô hình RawNet3 fine-tuning và mô hình WavLM (ECAPA-TDNN) fine-tuning khi thêm nhiễu trắng với SNR từ 0 dB đến 30 dB	65

Danh sách bảng

3.1	Bảng tham số và chiều của đầu ra của mô hình WavLM với dữ liệu mẫu là 3 giây	35
3.2	Bảng phân bố tập dữ liệu VoxCeleb1	42
3.3	Bảng phân bố tập dữ liệu VoxCeleb2	42
3.4	Confusion Matrix	44
4.1	Kết quả tốt nhất khi huấn luyện mô hình nhận dạng người nói từ đầu trên bộ dữ liệu Zalo	50
4.2	Kết quả khi áp dụng mô hình tiền huấn luyện trên tập dữ liệu tiếng Anh cho dữ liệu tiếng Việt	51
4.3	Kết quả tốt nhất trong quá trình fine-tuning mô hình nhận dạng người nói trên bộ dữ liệu Zalo	55
4.4	Kết quả tốt nhất trong quá trình huấn luyện mô hình nhỏ trên mô hình nhận dạng người nói với bộ dữ liệu Zalo . . .	62
4.5	Kết quả tổng hợp	66
4.6	Kết quả so sánh giữa các mô hình nhận dạng người nói trên tiếng Việt	71

Chương 1

Giới thiệu

1.1 Lý do chọn đề tài

Ngày nay, nhận dạng người nói là một lĩnh vực phát triển nhanh chóng với nhiều ứng dụng trong xã hội. Các hệ thống bảo mật điều khiển bằng giọng nói kết hợp với phương pháp xác thực dựa trên giọng nói ngày càng trở nên phổ biến hơn, dẫn đến nhu cầu về công nghệ nhận dạng người nói chính xác và đáng tin cậy ngày càng tăng.

Nhận thấy được tầm quan trọng của nhận dạng người nói, việc nghiên cứu về lĩnh vực này đã và đang được cải tiến một cách mạnh mẽ. Các kết quả được công bố gần đây chỉ ra rằng độ chính xác và tin cậy trong các mô hình nghiên cứu gần như đạt được con số hoàn hảo. Để thu được kết quả tốt như vậy, các mô hình đã được huấn luyện dựa trên ngôn ngữ có nguồn tài nguyên dữ liệu khá dồi dào. Tuy nhiên, những ngôn ngữ chủ yếu được sử dụng phần lớn là tiếng Anh, còn những ngôn ngữ khác hầu như không hoặc ít được đề cập tới.

Người Việt chúng ta hiện nay đã và đang áp dụng các hệ thống nhận dạng người nói và cũng sẽ áp dụng rộng rãi trong tương lai gần. Tuy nhiên, ngôn ngữ tiếng Việt có các thanh điệu hay các đặc điểm ngữ âm riêng biệt, có một số âm vị khác hoàn toàn so với tiếng Anh. Ngoài ra, lượng tài nguyên dữ liệu của tiếng Việt hiện tại đang khá hạn chế, vì vậy nếu sử

dụng các mô hình tốt nhất ấy cho dữ liệu ít ỏi như tiếng Việt thì liệu có thu được kết quả đáng mong đợi?

Từ câu hỏi trên, nhóm quyết định nghiên cứu và tổng hợp các mô hình nhận dạng người nói tốt nhất hiện nay (state-of-the-art, gọi tắt là SOTA) và thử nghiệm chúng dựa trên tập dữ liệu nhỏ (dưới 1000 người) và để xuất phương pháp tốt nhất trong quá trình huấn luyện các mô hình SOTA trên tập dữ liệu nhỏ.

1.2 Mục đích nghiên cứu

Nghiên cứu này của nhóm nhằm mục đích tổng hợp các mô hình SOTA và tìm hiểu khả năng nhận dạng người nói của các mô hình dựa trên ngôn ngữ tiếng Việt. Ngoài ra, nhóm còn thử nghiệm thêm khả năng nhận dạng của các mô hình sẽ thay đổi như thế nào khi dữ liệu đầu vào bị ảnh hưởng bởi những yếu tố khác, cụ thể như lược bỏ một phần băng tần hoặc trộn nhiễu vào dữ liệu. Từ đó đánh giá được tầm quan trọng của các khoảng băng tần cũng như sức ảnh hưởng của các loại nhiễu tác động đến khả năng nhận dạng người nói. Bên cạnh đó, nhóm cũng tìm giải pháp để có thể cải thiện hơn khả năng nhận dạng của các mô hình trên dữ liệu tiếng Việt.

1.3 Đối tượng nghiên cứu

Nghiên cứu các mô hình SOTA trong việc nhận dạng người nói bằng tập dữ liệu nhỏ trên tiếng Việt và tìm hiểu các cách thử nghiệm về việc cắt bỏ băng tần cũng như thêm các loại nhiễu vào dữ liệu.

Các mô hình sẽ được sử dụng: RawNet3 và WavLM.

Các tập dữ liệu sẽ được sử dụng:

- Tiếng Anh: VoxCeleb1, VoxCeleb2.
- Tiếng Việt: Zalo Speaker Verification Dataset.

1.4 Phạm vi nghiên cứu

- Phạm vi về thời gian: Từ 01/2023 đến 06/2023.
- Phạm vi về không gian: Trường đại học Khoa học Tự nhiên.

1.5 Phương pháp nghiên cứu

Đề tài được thực hiện dựa trên hai phương pháp như sau:

- Phương pháp nghiên cứu về lý thuyết: Tìm hiểu tổng quan bài toán nhận dạng người nói và bài toán xác thực người nói (dựa trên nhận dạng người nói). Sau đó tìm hiểu các mô hình SOTA và tiến hành khảo sát, đánh giá trên các tập dữ liệu tiếng Anh mà các tác giả đã dùng để công bố kết quả. Từ đó đối chiếu kết quả được công bố với kết quả vừa thử nghiệm để xem độ tin cậy của bài báo.
- Phương pháp thử nghiệm với tập dữ liệu mới: Tiến hành đánh giá các mô hình trên các tập dữ liệu khác với dữ liệu mà mô hình đã được huấn luyện từ trước. Các tập dữ liệu ấy gồm tập dữ liệu ngôn ngữ tiếng Anh và các tập dữ liệu ngôn ngữ tiếng Việt. Dữ liệu tiếng Anh dùng để khảo sát xem liệu các mô hình có thể nhận dạng tốt dù cũng một ngôn ngữ tiếng Anh hay không, còn dữ liệu tiếng Việt để khảo sát khả năng nhận dạng trên một ngôn ngữ hoàn toàn khác - ngôn ngữ có thanh điệu, ngữ âm hay âm vị khác biệt với tiếng Anh. Bên cạnh đó còn thử nghiệm thêm các yếu tố ảnh hưởng khác vào dữ liệu để đánh giá khả năng của các mô hình cũng như tiến hành tinh chỉnh mô hình (fine-tuning) để đạt được hiệu năng cao hơn.

1.6 Đóng góp của đề tài

Đề tài này sẽ đóng góp các kết quả thử nghiệm trên các tập dữ liệu thu thập được. Đối với dữ liệu tiếng Việt sẽ có thêm các kết quả về các yếu tố ảnh hưởng khác trên dữ liệu. Các kết quả bao gồm:

- Kết quả về hiệu quả của các mô hình khi được chạy trên tập dữ liệu nhỏ ngôn ngữ tiếng Việt.
- Kết quả việc đề xuất các phương pháp để đạt hiệu quả cao hơn trên bộ dữ liệu nhỏ tiếng Việt.
- Kết quả phân tích các mô hình nhận dạng người nói trên các băng tần và đánh giá mức độ ảnh hưởng của chúng trong việc nhận dạng người nói.
- Kết quả so sánh các mô hình trên dữ liệu đã được trộn nhiều và mức độ ảnh hưởng của chúng đối với khả năng nhận dạng người nói.
- Kết quả so sánh với các mô hình nhận dạng người nói có sử dụng bộ dữ liệu tiếng Việt

1.7 Nội dung báo cáo

Nội dung của đề tài bao gồm 5 chương như sau:

- **Chương 1: Giới thiệu.** Chương này giới thiệu tổng quan lý do tại sao chọn đề tài, nêu rõ các mục đích, đối tượng, phương pháp để tiến hành nghiên cứu và chỉ ra các đóng góp mà đề tài mang lại.
- **Chương 2: Các công trình liên quan.** Chương này giới thiệu chung về nhận dạng người nói, các phương pháp đã và đang được nghiên cứu trên lĩnh vực này.

- **Chương 3: Phương pháp đề xuất.** Chương này trình bày các cơ sở lý thuyết về các mô hình sẽ dùng trong quá trình thực hiện nghiên cứu, các hướng tiếp cận sử dụng cũng như tập dữ liệu để giải quyết bài toán.
- **Chương 4: Kết quả thí nghiệm.** Chương này khảo sát và đánh giá các kết quả thử nghiệm thu được từ các mô hình.
- **Chương 5: Kết luận và hướng phát triển.** Chương này tổng hợp lại quá trình nghiên cứu, chỉ ra được các vấn đề nào đã được giải quyết và những khó khăn gặp phải đồng thời từ đó đưa ra các hướng cải tiến trong tương lai.

Chương 2

Các công trình liên quan

2.1 Cơ sở lý thuyết

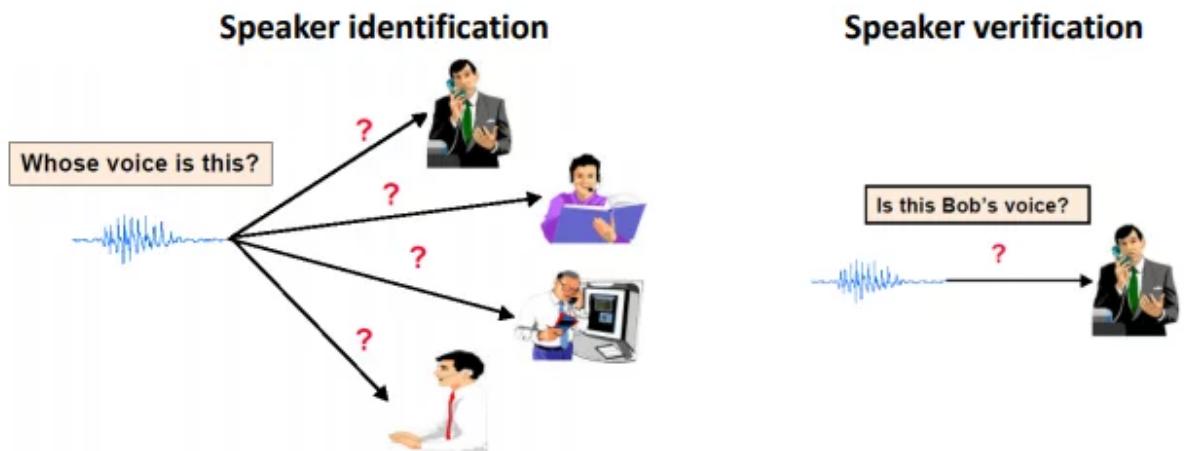
2.1.1 Giới thiệu về nhận dạng người nói

Nhận dạng người nói là quá trình định danh hoặc xác thực danh tính của một người dựa trên những đặc trưng trong tiếng nói của người đó, có thể được ứng dụng trong việc kiểm soát quyền truy cập đối với một số ứng dụng hay dịch vụ bằng tiếng nói. Điều này liên quan đến việc phân tích và so sánh các đặc điểm giọng nói, chẳng hạn như cao độ, âm điệu, nhịp điệu, các kiểu phát âm,... để tạo ra một cấu hình giọng nói riêng biệt của mỗi người. Có hai nhiệm vụ chính trong lĩnh vực nhận dạng người nói là định danh người nói (speaker identification) và xác thực người nói (speaker verification):

- **Định danh người nói:** Đây là nhiệm vụ liên quan đến việc xác định danh tính của một người bằng cách so sánh giọng nói của họ với một cơ sở dữ liệu người nói đã biết trước. Hệ thống, hoặc mô hình, tiến hành phân tích các đặc điểm trong giọng nói của họ rồi cố gắng so khớp chúng với cấu hình giọng nói của từng người có trong cơ sở dữ liệu. Kết quả trả về là **danh tính** của một người cụ thể nếu so khớp thành công, hoặc là **người lạ (unknown speaker)** nếu

không tồn tại cấu hình giọng nói nào đủ phù hợp trong cơ sở dữ liệu.

- **Xác thực người nói:** Nhiệm vụ này nhằm xác thực danh tính được khẳng định của một người nói bằng cách so sánh giọng nói của họ với người đang được khẳng định danh tính. Hệ thống, hoặc mô hình, tiến hành đánh giá sự tương đồng giữa các đặc điểm trong giọng nói của họ với cấu hình giọng nói của người có danh tính được nhắc đến. Kết quả trả về là một trong hai kết quả sau: **cùng một người** nếu độ tương đồng lớn hơn ngưỡng trị cho trước, hoặc **hai người khác nhau** nếu độ tương đồng chưa đủ đạt ngưỡng.



Hình 2.1: Ví dụ về hai nhiệm vụ chính trong nhận dạng người nói.

Hình 2.1 minh họa hai tác vụ chính trong lĩnh vực nhận dạng người nói. Cụ thể, giả sử có một giọng nói x , đối với **Định danh người nói**, ta sẽ trả lời câu hỏi "Giọng nói x là của ai?". Còn đối với **Xác thực người nói**, giả sử giọng nói x được cho là của A thì ta sẽ xác nhận xem "Giọng nói x là của A đúng hay không?"

2.1.2 Phương pháp trích chọn đặc trưng

Nhận dạng người nói, từ ngày xưa, được nghiên cứu bằng cách trích chọn các đặc trưng trong giọng nói một cách thủ công. Điều này đề cập

đến quá trình trích xuất các đặc điểm âm thanh có liên quan từ tín hiệu lời nói để thể hiện các đặc điểm duy nhất trong giọng nói của một người. Đã có rất nhiều phương pháp trích chọn đặc trưng được đề xuất, chẳng hạn một số đặc trưng có thể kể đến như:

- Cao độ (pitch): một đặc trưng âm thanh chủ yếu đại diện cho tần số cơ bản có thể cảm nhận được của giọng nói một người. Tính năng này ghi lại sự thay đổi về cao độ và có thể hữu ích cho việc nhận dạng người nói, đặc biệt là trong các tình huống có những đặc điểm cao độ khác biệt.
- Năng lượng (energy): cường độ tổng thể hoặc sức mạnh của tín hiệu giọng nói. Nó có thể cung cấp thông tin về cường độ hoặc độ to trong giọng nói của người nói. Các đặc trưng dựa trên năng lượng, chẳng hạn như năng lượng ngắn hạn hoặc năng lượng trung bình dài hạn, có thể được sử dụng để mô tả những người nói khác nhau...

Vào năm 1980, một phương pháp trích chọn đặc trưng đã được công bố và trở nên khá phổ biến đó là Mel Frequency Cepstral Coefficients (MFCC) [7]. MFCC cho ra kết quả là các hệ số của cepstral từ Mel filter trên phổ (spectral) lấy được từ các file âm thanh chứa giọng của người nói. Các bước để trích xuất MFCC như sau:

1. **Pre-emphasis:** Tín hiệu giọng nói được tiền xử lý bằng việc nhấn mạnh các tần số cao và cân bằng phổ. Điều này thường được thực hiện bằng cách áp dụng bộ lọc pre-emphasis, giúp khuếch đại các thành phần tần số cao.
2. **Framing:** Tín hiệu giọng nói sau đó được chia thành các khung (frame) ngắn, thường có thời lượng khoảng 20-30 mili giây. Các khung này đảm bảo rằng tín hiệu giọng nói có thể được coi là cố định trong mỗi khung.

3. **Windowing:** Mỗi khung của tín hiệu giọng nói được nhân với một hàm window (ví dụ: Hamming window) để làm mượt phổ tại các ranh giới của khung. Quá trình windowing này làm giảm dần tín hiệu về 0 ở các cạnh của mỗi khung.
4. **Fourier Transform:** Các khung của tín hiệu giọng nói được chuyển đổi thành miền tần số bằng cách sử dụng biến đổi Fourier, chẳng hạn như Discrete Fourier Transform (DFT) hoặc Fast Fourier Transform (FFT). Điều này chuyển đổi tín hiệu từ miền thời gian sang miền tần số, cung cấp thông tin về nội dung phổ của khung.
5. **Mel-scale Filter Bank:** Phổ cường độ thu được từ phép biến đổi Fourier sau đó được truyền qua một dãy các bộ lọc được đặt trong Mel-frequency scale. Bộ lọc Mel-scale hoạt động tương tự như sự nhận thức thính giác phi tuyến tính của con người về tần số. Filter bank được thiết kế để nắm bắt các dải tần số (frequency band) có liên quan để nhận dạng giọng nói và người nói.
6. **Logarithmic Compression:** Đầu ra của filter bank thường được nén bằng hàm logarit. Quá trình nén này thể hiện gần giống như nhận thức của con người về độ ồn và nhấn mạnh các thành phần phổ năng lượng thấp.
7. **Discrete Cosine Transform (DCT):** Sau quá trình nén logarit, tiến hành bước DCT. DCT giải mã năng lượng của filter bank và biểu diễn các đặc trưng phổ một cách cô đọng.
8. **Feature Extraction:** Lấy các hệ số từ bước DCT, các hệ số này được gọi là MFCC và thường được trích xuất làm đặc trưng cuối cùng. Các hệ số này nắm bắt thông tin cần thiết về nội dung phổ của tín hiệu tiếng nói một cách cô đọng và phân biệt.

Nhìn chung, quy trình trích chọn đặc trưng MFCC nhằm mục đích nắm bắt các đặc tính phổ có liên quan của tín hiệu tiếng nói đồng thời giảm

kích thước của biểu diễn đặc trưng trong giọng nói. Điều này giúp cải thiện hiệu suất của giọng nói và hệ thống nhận dạng người nói bằng cách tập trung vào các khía cạnh mang thông tin nhiều nhất của tín hiệu.

2.1.3 Các nghiên cứu dựa trên học máy

So với các phương pháp trích chọn đặc trưng thủ công, các nghiên cứu dựa trên học máy có tiềm năng mang lại nhiều lợi ích hơn. Một số lợi ích có thể kể đến như là khả năng tự động học các đặc trưng tương đồng và riêng biệt từ dữ liệu thô, khả năng nắm bắt các mối quan hệ trong dữ liệu mà phương pháp thủ công không thể hiện đầy đủ, giảm bớt yêu cầu về trình độ chuyên môn cao,... Một số phương pháp học máy đã được nghiên cứu trong lĩnh vực nhận dạng người nói, chẳng hạn như:

- **Gaussian Mixture Models (GMM) [27]:** GMM là một mô hình xác suất được sử dụng trong nhận dạng người nói. GMM biểu diễn sự phân bố các đặc trưng trong giọng nói bằng cách kết hợp nhiều phân phối Gaussian. Trong quá trình huấn luyện, GMM học các thuộc tính thống kê của từng đặc điểm giọng nói của người nói. Trong quá trình nhận dạng người nói, GMM đánh giá khả năng một mẫu giọng nói thuộc về người nói nào bằng cách so sánh các vector đặc trưng của nó với GMM của những người nói đã biết. GMM có thể nắm bắt cả các biến đổi ngắn hạn lẫn dài hạn trong giọng nói, làm cho chúng hiệu quả trong việc xử lý các điều kiện giọng nói khác nhau.
- **Gaussian Mixture Models - Universal Background Model (GMM-UBM) [28]:** Mô hình này liên quan đến việc huấn luyện một Universal Background Model (UBM) nhằm nắm bắt sự phân bố các đặc trưng trong giọng nói từ một tập dữ liệu lớn chứa nhiều người nói. UBM biểu diễn các đặc điểm âm thanh được chia sẻ giữa tất cả người nói. Để mô hình hóa thông tin cụ thể về người nói, các GMM được điều chỉnh từ UBM bằng quy trình gọi là thích ứng mô hình.

Các GMM đã thích ứng sẽ nắm bắt các đặc trưng duy nhất của từng người nói. GMM-UBM rất linh hoạt và có thể được áp dụng cho cả các vụ xác thực lẫn định danh người nói. Cách tiếp cận này đã được chứng minh là hiệu quả trong việc nắm bắt sự đa dạng trong giọng nói của người nói và đã được sử dụng rộng rãi lúc bấy giờ...

Đến năm 2011, một phương pháp học máy trong lĩnh vực nhận dạng người nói được công bố và trở nên phổ biến một cách nhanh chóng là i-vector [8]. Phương pháp này biểu diễn các đặc điểm cụ thể về người nói lẫn những đặc điểm cụ thể về kênh (channel) và i-vector được trích xuất từ tín hiệu giọng nói dưới dạng vector có số chiều thấp (low-dimensional). Các bước thực hiện của i-vector như sau:

1. **Feature Extraction:** Bước đầu tiên xử lý tín hiệu giọng nói và trích xuất các đặc trưng âm thanh. Các đặc trưng này có thể bao gồm MFCC, filter banks hoặc các biểu diễn phổ khác... Tín hiệu giọng nói sau đó thường được chia thành các frame ngắn và các đặc trưng được tính toán cho từng frame.
2. **UBM Training:** UBM được huấn luyện bằng cách sử dụng một lượng lớn dữ liệu giọng nói từ nhiều người nói. UBM biểu diễn sự phân bố của tất cả các vector đặc trưng âm thanh có thể có trong tập huấn luyện. Thông thường GMM được sử dụng làm UBM, trong đó mỗi thành phần Gaussian mô hình hóa một vector đặc trưng âm thanh cụ thể.
3. **Extracting Speaker-Specific Information:** Cho trước một tập hợp các đoạn giọng nói từ một người nói, một mô hình người nói được tạo bằng cách thích ứng UBM với dữ liệu cụ thể về người đó. Điều này được thực hiện bằng cách sử dụng các kỹ thuật như Maximum A Posteriori (MAP) hoặc Joint Factor Analysis (JFA). Quá trình thích ứng này ước tính các tham số của GMM của người này bằng việc kết hợp dữ liệu của người nói trong khi vẫn giữ lại thông tin từ UBM.

4. **Supervector Extraction:** Sau khi mô hình người nói được thích ứng, nó được sử dụng để tạo supervector. Supervector thu được bằng cách ghép nối mean supervector và total variability matrix (ghi lại các biến đổi của một người cụ thể). Mean supervector biểu diễn các giá trị trung bình của các thành phần Gaussian đã thích ứng, ma trận total variability ghi lại các biến đổi của một người cụ thể. Ma trận này có thể được ước tính bằng các kỹ thuật như Factor Analysis (FA) hoặc Linear Discriminant Analysis (LDA).
5. **I-vector Extraction:** Supervector sau đó được rút gọn thành một biểu diễn có số chiều thấp hơn được gọi là i-vector. I-vector nắm bắt thông tin cụ thể của người nói bằng cách chiếu supervector lên một không gian phụ, thường được gọi là không gian i-vector.
6. **Speaker Recognition:** Các i-vector thu được có thể sử dụng cho các tác vụ nhận dạng người nói. Đối với xác thực người nói, i-vector của giọng nói mẫu được so sánh với i-vector của người nói đang được khẳng định danh tính để xác nhận liệu chúng có khớp hay không. Còn đối với định danh người nói thì so sánh với các i-vector của những người nói đã biết để xác định người nào phù hợp nhất. Các độ đo để so sánh thường là cosine similarity.

Phương pháp i-vector đã trở nên phổ biến do tính hiệu quả của nó trong việc nắm bắt thông tin cụ thể về người nói. Ngoài ra, phương pháp này cho phép lưu trữ, xử lý và so sánh hiệu quả các biểu diễn của người nói, làm cho nó phù hợp với các hệ thống nhận dạng người nói quy mô lớn.

2.1.4 Các nghiên cứu dựa trên học sâu

Sau một thời gian, các nghiên cứu dựa trên học sâu ra đời và mang lại hiệu quả vượt trội hẳn so với các phương pháp học máy. Một số khả năng mạnh mẽ của các phương pháp học sâu có thể kể đến như là khả năng huấn luyện end-to-end giúp tối ưu hóa tốt hơn, khả năng thích ứng

tốt với những điều kiện âm thanh khác nhau, khả năng mở rộng giúp xử lý hiệu quả các dữ liệu lớn,... Một số phương pháp học sâu tiêu biểu đã được nghiên cứu như:

D-vector [36], còn được gọi là phương pháp biểu diễn người nói (speaker representation) hay speaker embedding, được sử dụng rộng rãi trong nhận dạng người nói từ khi được công bố. Phương pháp này nhằm mục đích trích xuất các embedding phân biệt và có định nghĩa riêng cho các người nói khác nhau từ các tín hiệu giọng nói của họ. Tổng quan quá trình trích xuất d-vector như sau:

1. **Feature Extraction:** Bước đầu tiên là trích xuất các đặc trưng âm thanh từ tín hiệu giọng nói. Các đặc trưng thường được sử dụng là MFCC, năng lượng filterbank hoặc các biểu diễn phổ khác... Các đặc trưng này nắm bắt các đặc điểm phổ của tín hiệu giọng nói và cung cấp thông tin về giọng nói của người nói.
2. **Neural Network Architecture:** Kiến trúc mạng neural sâu (deep neural network hay DNN) được sử dụng để huấn luyện d-vector. Đầu vào của mạng là chuỗi các đặc trưng âm thanh, sau đó mạng học cách ánh xạ nó thành vector biểu diễn có số chiều cố định, vector này là d-vector. Kiến trúc mạng này được thiết kế để nắm bắt các thông tin phân biệt liên quan đến từng người nói.
3. **Training Objective:** DNN được huấn luyện bằng một lượng lớn dữ liệu người nói được gắn nhãn. Mục tiêu là tối ưu hóa các tham số của mạng để giảm thiểu khoảng cách giữa các d-vector của cùng một người và tối đa hóa khoảng cách giữa các d-vector của những người nói khác nhau. Mục tiêu này đảm bảo rằng mạng học cách mã hóa các đặc điểm cụ thể về người nói trong việc biểu diễn d-vector.
4. **D-vector Extraction:** Sau quá trình huấn luyện, các d-vector có thể được trích xuất bằng cách truyền các đoạn giọng nói qua mạng. Đầu ra của mạng, ở một lớp nhất định (thường là lớp cuối cùng), là

speaker embedding (cũng là d-vector). Embedding này mã hóa các đặc điểm riêng của người nói và được xem như là một biểu diễn thu gọn về đặc trưng giọng nói của họ.

5. **Speaker Recognition:** Các embedding được trích xuất có thể được dùng cho các tác vụ nhận dạng người nói. Các độ đo như cosine similarity hay Euclidean distance thường được áp dụng để so sánh và cách so sánh tương tự như ở phương pháp i-vector.

X-vector [31] cũng là một phương pháp trích xuất embedding mang thông tin cụ thể về người nói, nắm bắt các đặc điểm của người nói tốt hơn. Sau đây là tổng quan các bước của x-vector:

1. **Feature Extraction:** Bước đầu tiên là trích xuất các đặc trưng âm thanh từ tín hiệu giọng nói. Các đặc trưng thường được sử dụng như MFCC, năng lượng filterbank hoặc một số biểu diễn phổ khác... Các đặc trưng này nắm bắt các đặc điểm về thời gian và phổ của tín hiệu giọng nói.
2. **Time-Context Modeling:** Để nắm bắt thông tin về thời gian, kiến trúc mạng time-delay neural network (TDNN) thường được sử dụng. TDNN xử lý bằng cách dùng một cửa sổ (window) trượt trên nhiều frame để nắm bắt được các đặc trưng âm thanh cũng như sự phụ thuộc giữa ngữ cảnh và thời gian trong tín hiệu giọng nói. Đầu ra của TDNN được biểu diễn ở mức độ frame với số chiều cố định.
3. **Context-Aware Pooling:** Thông tin từ nhiều frame được kết hợp bằng một lớp statistical pooling. Lớp này kết hợp các biểu diễn ở mức độ frame theo thời gian, kết hợp cả ngữ cảnh ngắn hạn và dài hạn. Thao tác kết hợp này tổng hợp thông tin ở mức độ frame thành một vector biểu diễn có số chiều cố định, được gọi là x-vector.
4. **Backend Classifier:** Các x-vector được đưa vào một bộ phân loại, thường là feed-forward neural network, để thực hiện các tác vụ nhận

dạng người nói. Bộ phân loại này được huấn luyện trên một tập dữ liệu lớn được gắn nhãn bằng cách sử dụng các kỹ thuật phân loại khác nhau, như softmax regression hoặc Support Vector Machines (SVM), để phân biệt những người nói với nhau.

5. **Training Objective:** Mô hình được huấn luyện để tối ưu hóa hàm mất mát (loss function) sao cho các x-vector của cùng một người nói gần nhau trong không gian embedding, trong khi đẩy các x-vector của những người khác nhau ra xa hơn. Mục đích là để đảm bảo rằng các x-vector nắm bắt thông tin cụ thể về người nói một cách phân biệt.
6. **X-vector Extraction:** Tương tự như d-vector, x-vector có thể được trích xuất bằng việc truyền các đoạn giọng nói qua mô hình, đầu ra là speaker embedding. Embedding này là một biểu diễn thu nhỏ của các đặc trưng giọng nói của người nói cụ thể.
7. **Speaker Recognition:** Các embedding sau khi trích xuất có thể được sử dụng cho các tác vụ nhận dạng người nói. Độ đo chủ yếu thường được dùng là cosine similarity hoặc PLDA. Cách so sánh trong các tác vụ định danh hay xác thực danh tính tương tự như ở phương pháp i-vector.

SincNet [25] là mô hình được thiết kế chủ yếu để nhận dạng người nói trực tiếp từ âm thanh dạng sóng thô (raw waveform). Mục đích của mô hình là trích xuất các đặc trưng cụ thể về người nói trực tiếp từ dạng sóng bằng cách sử dụng tập hợp các bộ lọc thông dải (band-pass filter). Các bộ lọc này có thể huấn luyện được và còn được gọi là bộ lọc Sinc. Các bước chính liên quan đến mô hình SincNet như sau:

1. **Input Processing:** Đầu vào của SincNet là âm thanh dạng sóng thô, thường được biểu thị dưới dạng tín hiệu miền thời gian một chiều, sau đó được chia thành các cửa sổ hoặc phân đoạn có độ dài cố định.

2. **Sinc Filters:** Bộ lọc Sinc dùng để trích xuất thông tin phổ từ dạng sóng đầu vào. Các bộ lọc này được thiết kế để nắm bắt các dải tần số khác nhau. Mỗi bộ lọc Sinc được xác định bởi tần số trung tâm (center frequency) và băng thông (bandwidth) của nó, những tham số này có thể học được trong quá trình huấn luyện. Bộ lọc Sinc được cài đặt tương tự như bộ lọc tích chập (convolutional filter) với kernel shape dựa trên tần số trung tâm và băng thông mong muốn.
3. **Filtering:** Từng phân đoạn đầu vào được truyền qua các bộ lọc Sinc với tần số trung tâm và băng thông khác nhau. Các bộ lọc Sinc thực hiện phép tích chập để lọc các đoạn sóng đầu vào này và đầu ra của bước lọc là một tập hợp các feature map, mỗi feature map biểu diễn một dải tần số khác nhau.
4. **Non-linear Operations:** Kế đến, các feature map được kích hoạt bằng leaky-ReLU và truyền qua lớp pooling. Leaky-ReLU để giảm thiểu độ lớn của các giá trị âm, đồng thời nắm bắt các thành phần dương của tín hiệu sau khi lọc. Lớp pooling làm giảm kích thước của các feature map, giữ lại các mẫu phổ cục bộ quan trọng.
5. **Fully Connected Layers:** Các feature map sau khi gộp được làm phẳng và đưa vào các lớp kết nối đầy đủ (Fully Connected hay FC). Các lớp FC thực hiện thêm các phép biến đổi phi tuyến tính và ánh xạ các đặc trưng sang không gian có số chiều thấp hơn.
6. **Training:** SincNet được huấn luyện bằng các kỹ thuật backpropagation và gradient-based optimization. Hàm mất mát thường được dùng là cross-entropy và dữ liệu người nói thì được gắn nhãn, khi đó mô hình học cách phân biệt những người nói khác nhau dựa trên đặc điểm âm thanh của họ.
7. **Classification:** Đầu ra của lớp FC cuối có thể được trích xuất và xem như là một embedding. Embedding này được dùng trong tác vụ

xác thực người nói. Đối với định danh người nói, embedding được đưa tiếp vào lớp softmax để tiến hành phân loại.

RawNet3 và WavLM là hai mô hình SOTA trong lĩnh vực nhận dạng người nói. Đây là hai mô hình chủ yếu của đề tài nên sẽ được phân tích kỹ hơn ở Chương 3.

2.1.5 Sơ lược về việc khảo sát miền tần số với bộ lọc thông thấp và bộ lọc cầm dải

Bộ lọc thông thấp là bộ lọc mà chỉ cho phép các tín hiệu có giá trị tần số nhỏ hơn một giá trị tần số nào đó đi qua và lược bỏ các tín hiệu mang các giá trị tần số còn lại. Ví dụ minh họa được thể hiện ở hình 2.2, hình này mô tả kết quả sau khi áp dụng bộ lọc thông thấp vào một đoạn âm thanh mẫu với giá trị tần số cắt từ 500 Hz đến 7500 Hz, bước nhảy là 500 Hz.

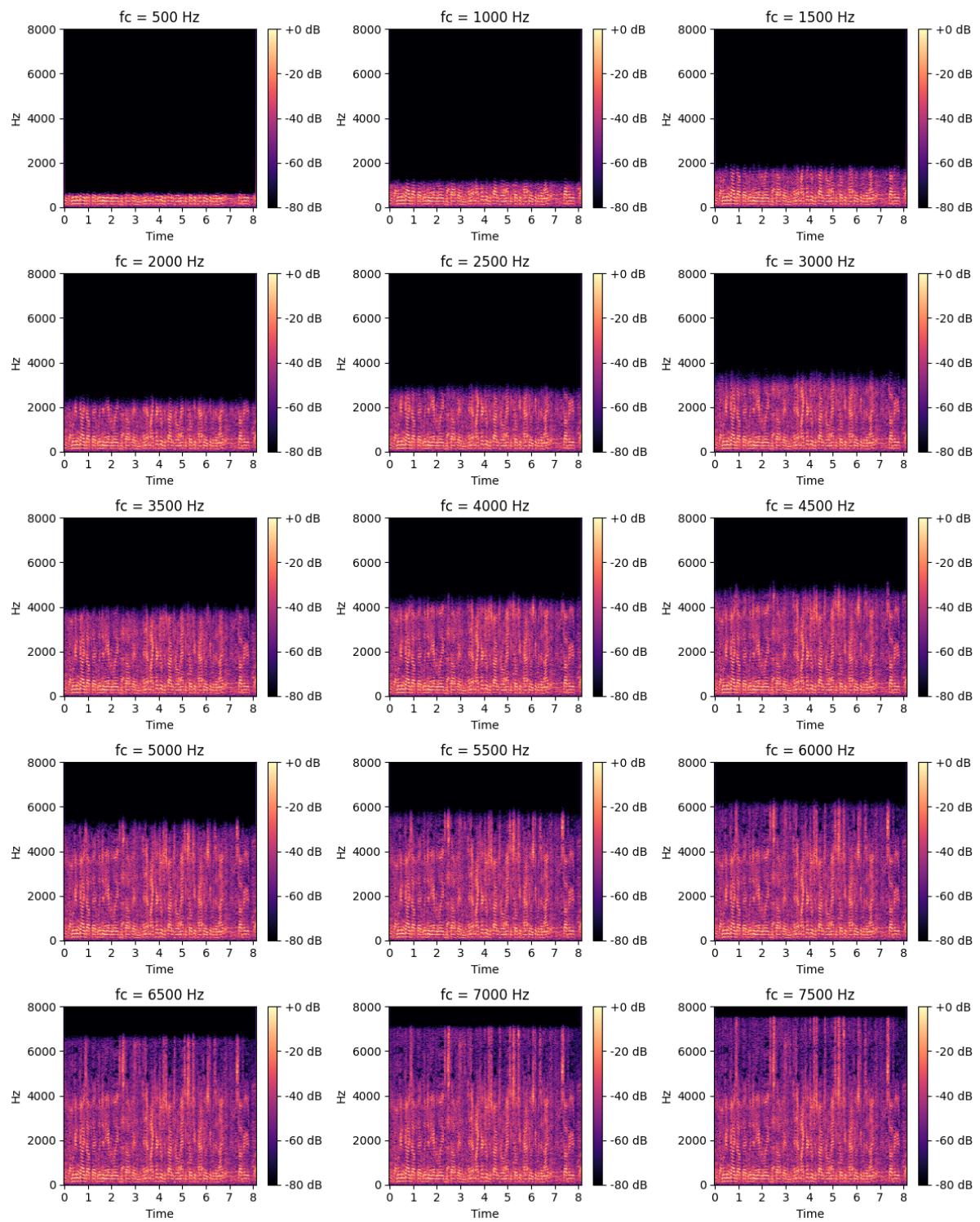
Bộ lọc cầm dải là bộ lọc mà chỉ cho phép các tín hiệu có giá trị tần số ngoài khoảng giá trị tần số nào đó đi qua và lược bỏ các tín hiệu mang các giá trị tần số còn lại. Ví dụ được minh họa ở hình 2.3, mô tả kết quả sau khi áp dụng bộ lọc cầm dải vào một đoạn âm thanh mẫu trong khoảng tần số từ 4000 Hz đến 7000 Hz.

2.1.6 Sơ lược về việc thêm nhiễu

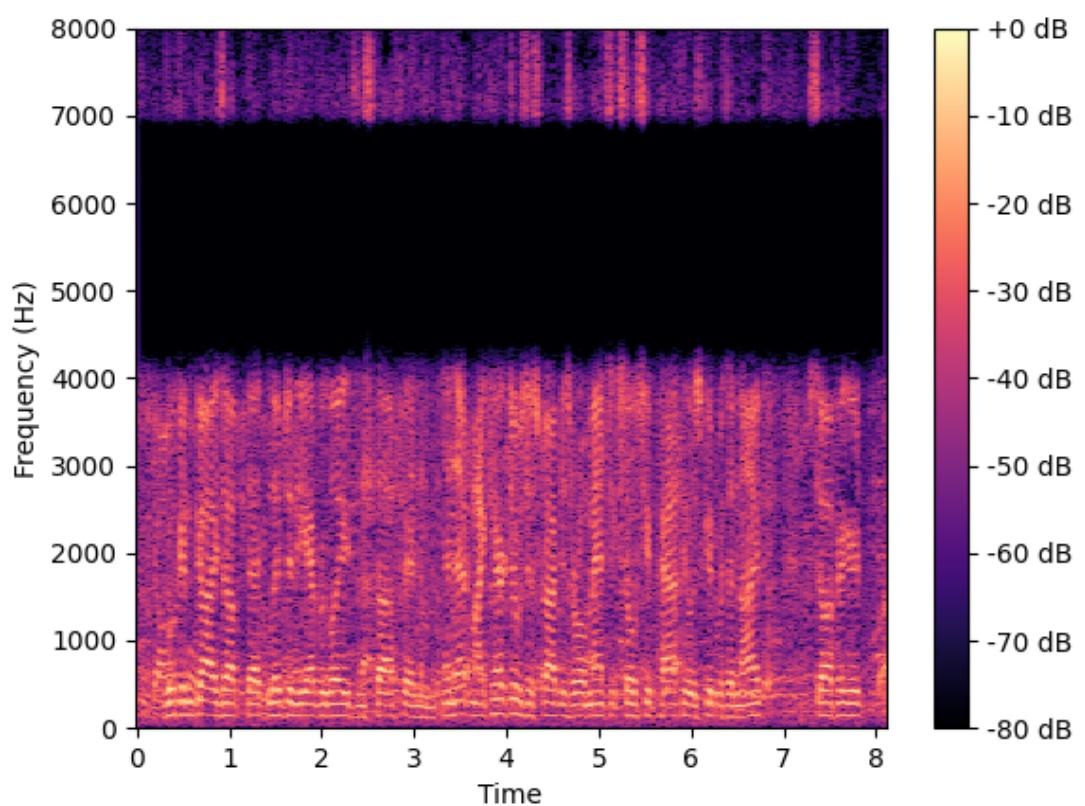
Bên cạnh về việc khảo sát trên miền tần số, nhóm cũng thử nghiệm hiệu suất của các mô hình đối với dữ liệu có nhiễu. Lượng nhiễu thêm vào đoạn lời nói được tính bằng tỉ lệ tín hiệu trên nhiễu (Signal to Noise Ratio, gọi tắt là SNR). Công thức tính SNR¹ 2.1 như sau:

$$\text{SNR} = \frac{\text{Năng lượng của tín hiệu sạch}}{\text{Năng lượng của tín hiệu nhiễu}} = \frac{P_s}{P_n} \quad (2.1)$$

¹https://sites.ualberta.ca/~msacchi/SNR_Def.pdf



Hình 2.2: Kết quả sau khi áp dụng bộ lọc thông thấp vào một đoạn âm thanh mẫu với giá trị tần số từ 500 Hz đến 7500 Hz với khoảng cách mỗi bước nhảy 500 Hz



Hình 2.3: Kết quả sau khi áp dụng bộ lọc cấm dải vào một đoạn âm thanh mẫu trong khoảng tần số từ 4000 Hz đến 7000 Hz

Khi đó, với d_k là tín hiệu đầu ra (d), s_k là tín hiệu sạch (s), n_k là tín hiệu nhiễu (n) với $k = 1..N$, N là số lượng mẫu, ta có 2.2:

$$d_k = s_k + \alpha n_k \Leftrightarrow d = s + \alpha n \quad (2.2)$$

Với α là hệ số sẽ được xác định trước, công thức tính SNR 2.3 như sau:

$$\text{SNR} = \frac{\sum_{k=1}^N s_k^2}{\alpha^2 \sum_{k=1}^N n_k^2} \quad (2.3)$$

Từ đó, ta rút ra được 2.4:

$$\begin{aligned} \alpha^2 &= \frac{\sum_{k=1}^N s_k^2}{\text{SNR} \sum_{k=1}^N n_k^2} \\ &= \frac{\|\mathbf{s}\|_2^2}{\text{SNR} \|n\|_2^2} \end{aligned} \quad (2.4)$$

SNR thu được ở công thức 2.3 mang đơn vị B (bel). Trong trường hợp ta muốn có đơn vị dB (decibels) thì chuyển đổi bằng công thức $\text{SNR}_{\text{dB}} = 10 \log_{10}(\text{SNR})$

Từ đó, nhóm đã xây dựng mã nguồn để thêm một lượng nhiễu SNR (dB) vào tín hiệu s:

```

1 def add_noise_dB(s, SNR_dB):
2     SNR = 10.0 ** (SNR_dB / 10.0)
3     n = np.random.randn(s.size)
4     Es = np.sum(s**2)
5     En = np.sum(n**2)
6     alpha = np.sqrt(Es/(SNR*En))
7     d = s + alpha*n
8     return d, alpha*n

```

2.2 Các nghiên cứu liên quan

2.2.1 Nhận dạng người nói tiếng Việt bằng học biểu diễn sâu

Nhận dạng người nói tiếng Việt bằng học biểu diễn sâu [34] là nghiên cứu tập trung vào việc thử nghiệm các mô hình học sâu vào các bài toán nhận dạng người nói trên tập dữ liệu tiếng Việt. Nghiên cứu đã tiến hành xây dựng một bộ dữ liệu tiếng Việt và tiến hành kiểm thử trên 12 sự kết hợp cũng như khảo sát cách kết hợp nào có hiệu năng cao nhất trên bộ dữ liệu tiếng Việt.

Nhóm tác giả của nghiên cứu này đã thực hiện một phương pháp mới để xây dựng nên bộ dữ liệu *VietCeleb*. Bộ dữ liệu này gồm có 5,800 bản ghi được trích xuất từ các video YouTube với 580 người nói. Nhóm tác giả đã xây dựng quá trình thu thập dữ liệu tương tự như hướng tiếp cận của quá trình thu thập tập dữ liệu VoxCeleb1 [22], cụ thể là một quy trình 5 bước gồm có:

1. Xây dựng một danh sách những người nổi tiếng
2. Tải các video từ YouTube
3. Trích xuất giọng nói từ video
4. Xác thực chuyển động khuôn mặt
5. Chuẩn hóa giọng nói

Nhóm tác giả đã chia thành ba hướng khảo sát nghiên cứu về học chuyển đổi cho bài toán nhận dạng người nói tiếng Việt:

- Phương pháp 1: Xây dựng mô hình nhận dạng người nói với phương pháp học biểu diễn sâu chỉ bằng bộ dữ liệu tiếng Việt.

- Phương pháp 2: Xây dựng mô hình nhận dạng người nói với phương pháp học biểu diễn sâu bằng bộ dữ liệu tiếng Anh và tiến hành kiểm thử trên tiếng Việt.
- Phương pháp 3: Xây dựng mô hình nhận dạng người nói với phương pháp học biểu diễn sâu kết hợp bộ dữ liệu tiếng Anh và tiếng Việt. Trong đó, mô hình sẽ được huấn luyện trước với bộ dữ liệu VoxCeleb1 và sẽ tiếp tục huấn luyện trên bộ dữ liệu VietCeleb.

Các mô hình đã được khảo sát gồm có ResNetSE34V2, ResnetSE34Half, ResNetSE34L và VGG-Vox. Các mô hình này đều được dựa vào những mô hình cơ bản [6]. Các mô hình cơ sở sử dụng lời nói có độ dài 2 giây, được huấn luyện với 500 epoch với tần số lấy mẫu là 16000 Hz và hàm mất mát là Prototypical loss. Ngoài ra, kỹ thuật ASP [23] (Attentive Statistics Pooling) cũng được áp dụng để tổng hợp các đặc trưng theo từng frame.

Trong ba phương pháp, phương pháp kết hợp bộ dữ liệu tiếng Anh và tiếng Việt là phương pháp đạt kết quả tốt nhất, sau đó đến phương pháp chỉ huấn luyện bằng bộ dữ liệu tiếng Việt và trường hợp tệ nhất là huấn luyện chỉ bằng bộ dữ liệu tiếng Anh. Ngoài ra, mô hình ResNetSE34V2 là mô hình tốt nhất khi so sánh với ba mô hình còn lại cho nhận dạng người nói tiếng Việt với $EER = 4\%$

2.2.2 Mô hình xác thực người nói cho ngôn ngữ ít tài nguyên với bộ dữ liệu tiếng Việt

Mô hình xác thực người nói cho ngôn ngữ ít tài nguyên với bộ dữ liệu tiếng Việt [32] là nghiên cứu tập trung vào việc đề xuất một mô hình xác thực người nói để khắc phục việc thiếu tài nguyên, cụ thể là tài nguyên dữ liệu tiếng Việt, với mô hình cơ sở dựa trên mô hình ResNet-34 [15].

Mô hình cơ sở sẽ lấy một lượng log Mel-filterbank làm đầu vào, sau đó sẽ truyền qua mô hình ResNet-34 để trích xuất các đặc trưng theo từng frame. Vì đầu vào là một đoạn âm thanh có chiều dài khả biến nên sẽ sử

dụng kỹ thuật ASP [23] để tổng hợp các đặc trưng theo frame thành một đặc trưng theo lời nói trong khi vẫn giữ được các thông tin quan trọng dựa vào cơ chế attention. Sau đó, đặc trưng theo lời nói sẽ đi qua một lớp kết nối đầy đủ (fully connected) để trích xuất embedding. Hàm mất mát được sử dụng là AP (Angular Prototypical) và thuật toán tối ưu là Adam.

Đối với mô hình đã đề xuất, nhóm tác giả đã áp dụng phương pháp học chuyển đổi để tăng tốc quá trình huấn luyện mô hình và giúp mô hình đạt kết quả tốt hơn khi tinh chỉnh trên bộ dữ liệu tiếng Việt, cụ thể là sử dụng mô hình đã được tiền huấn luyện (pretrain) với bộ dữ liệu tiếng Anh rồi tiếp tục huấn luyện trên dữ liệu tiếng Việt. Nhóm tác giả đã đề xuất sử dụng một phiên bản cải tiến của hàm mất mát Angular Prototypical Loss [6] là Angular Margin Prototypical Loss nhằm cải thiện khả năng phân loại người nói và thuật toán tối ưu là SGD nhằm tăng tính tổng quát hóa.

Nhóm tác giả đã đề xuất một quy trình thu thập dữ liệu dành cho dữ liệu Việt Nam nói riêng dữ liệu ít tài nguyên nói chung. Đầu tiên là tiến hành thu thập dữ liệu, sau đó hợp nhất với danh tính người nói thành dữ liệu thô. Tiếp theo là tiến hành tiền xử lý dữ liệu bằng cách loại bỏ những dữ liệu không hợp lệ, dữ liệu có nhiều nhiễu và thống nhất người nói. Sau đó sẽ chia bộ dữ liệu thành tập train và tập test. Tập test sẽ được cân bằng về số lượng giới tính. Ngoài ra, nhóm tác giả cũng đã xây dựng tập test tiếng Việt từ Common Voice² và xem như là tập test ngoài miền dữ liệu được huấn luyện.

Nhóm tác giả đã tiến hành thử nghiệm và kiểm chuẩn các mô hình với phương pháp đã đề xuất. Cụ thể, các mô hình đã thử nghiệm gồm có mô hình ECAPA [33] và mô hình ResNet-34. Các mô hình được huấn luyện với 2000 epoch, mini batch size là 200, tiền learning rate là 0.005 và sẽ giảm 25% mỗi 50 epoch.

Với kết quả mà nhóm tác giả công bố, có thể thấy rằng mô hình được đề xuất hoạt động khá hiệu quả. Cụ thể, kết quả độ đo equal error rate (EER) của mô hình tiền huấn luyện trên tập dữ liệu nhóm tác giả xây dựng

²Common Voice

là 14.954% và kết quả tốt nhất sau khi mô hình học chuyển đổi là 3.115%. Đối với tập test từ Common Voice, kết quả EER lần lượt là 11.468% và 4.789%.

Chương 3

Phương pháp đề xuất

3.1 Cơ sở lý thuyết

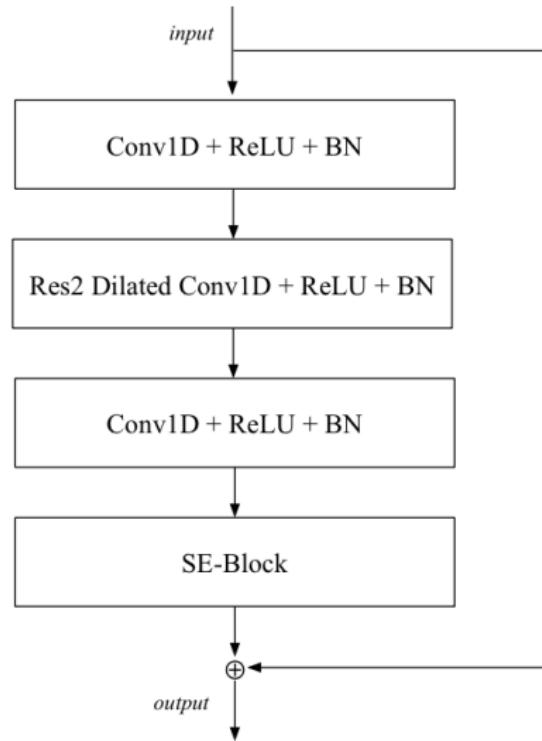
3.1.1 Mô hình ECAPA-TDNN

Đây là mô hình được cả RawNet3 và WavLM sử dụng làm mô hình cơ sở, do đó nhóm sẽ giới thiệu sơ lược về mô hình này một cách cô đọng nhất. ECAPA-TDNN [10] là một mô hình cải tiến đối với các kiến trúc x-vector và các cải tiến của x-vector [30, 13]. Điều này nhằm mục đích nâng cao hiệu suất của kiến trúc TDNN và lớp statistics pooling trong các hệ thống x-vector cho các tác vụ nhận dạng người nói. Nhóm tác giả đã đề xuất các cải tiến trong mô hình như sau:

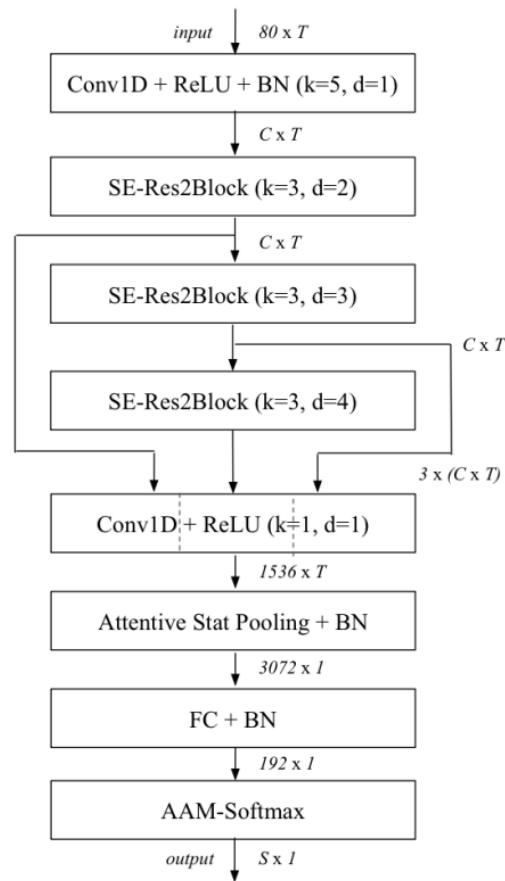
- Kiến trúc ECAPA-TDNN cải thiện sự phụ thuộc ngữ cảnh và channel của statistics pooling. Cải tiến này mở rộng cơ chế soft self-attention được sử dụng trong các kiến trúc x-vector, cho phép mô hình tập trung vào các đặc điểm khác nhau của người nói. Lớp pooling tính toán các vector độ lệch chuẩn và vector trung bình có trọng số (weighted mean) cho mỗi channel dựa trên giá trị channel-dependent self-attention. Ngoài ra, lớp pooling kết hợp các thuộc tính toàn cục của giọng nói (global properties) bằng cách nối ghép đầu vào cục bộ (local input) với giá trị trung bình không trọng số

(non-weighted mean) và độ lệch chuẩn trên miền thời gian. Những điều chỉnh này cải thiện khả năng biểu diễn thông tin của người nói bằng cách nắm bắt các chi tiết cụ thể về channel và xem xét các thuộc tính toàn cục trong giọng nói.

- 1-Dimensional Squeeze-Excitation (SE) Res2Blocks được thêm vào để cải thiện bối cảnh về thời gian (temporal context) của các lớp frame và kết hợp các thuộc tính toàn cục của giọng nói. Các khối SE [17] mô hình hóa sự phụ thuộc lẫn nhau của channel toàn cục bằng cách tính toán các channel descriptor và tạo trọng số cho từng channel. SE-Res2Blocks là sự kết hợp của các khối SE với lợi ích của các residual connection [14]. Chúng bao gồm các lớp tích chập giãn nở (dilated convolution), các lớp dense và các khối SE để chia tỷ lệ cho từng channel. Việc tích hợp các ResBlocks truyền thống giúp cải thiện hiệu suất và làm giảm số lượng tham số của mô hình bằng việc xử lý các đặc trưng multi-scale. Kiến trúc SE-Res2Block được mô tả trong hình 3.1.
- Cuối cùng, để tăng khả năng biểu diễn của các embedding, kỹ thuật multi-layer feature aggregation and summation được áp dụng trong ECAPA-TDNN. Mô hình tập hợp các đầu ra của các SE-Res2Block thông qua Multi-layer Feature Aggregation (MFA) giúp tăng cường việc biểu diễn các đặc trưng người nói. Ngoài ra, các feature map từ các khối trước đó được tổng hợp và sử dụng lớp tích chập ban đầu để tận dụng tính phân cấp (hierarchical) của TDNN. Những cải tiến này giúp mô hình tạo nên các embedding biểu diễn mạnh mẽ hơn. Kiến trúc tổng quát của ECAPA-TDNN được mô tả trong hình 3.2.



Hình 3.1: Kiến trúc SE-Res2Block của mô hình ECAPA-TDNN



Hình 3.2: Kiến trúc của mô hình ECAPA-TDNN

Trong đó:

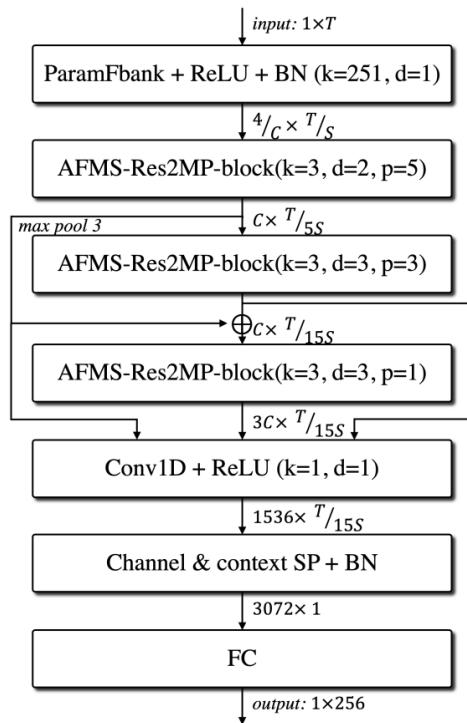
- k : độ dài kernel
- d : độ giãn nở
- C : kích thước channel
- T : kích thước temporal
- S : số lượng người nói

3.1.2 Mô hình RawNet3

RawNet3 [20] là một mô hình nhận dạng người nói end-to-end không phụ thuộc văn bản, tức là mô hình sẽ nhận sóng thô mà không quan tâm đến nội dung người nói làm đầu vào thay vì đi qua các bước trích chọn đặc trưng trước như MFCC hay filterbank. Tuy nhiên, những mô hình end-to-end [21] lại có hiệu suất thấp hơn khi so sánh với các mô hình có sử dụng các phương pháp trích chọn đặc trưng [10, 26]. Do đó, nhóm tác giả đã đề xuất mô hình RawNet3 bằng cách kết hợp kiến trúc của ECAPA-TDNN [10] và RawNet2 [19] để vượt qua thử thách này. Ngoài ra, học tự giám sát đang nổi lên và được xem như là một phương pháp học thay thế học có giám sát và cũng đã có nhiều nghiên cứu về nhận dạng người nói bằng phương pháp học tự giám sát. Tuy nhiên, vẫn chưa có mô hình nhận dạng người nói học tự giám sát nào có kiến trúc end-to-end xử lý trực tiếp sóng thô từ đầu vào và RawNet3 được tạo ra để giải quyết vấn đề này.

Mô hình RawNet3 là sự kết hợp giữa kiến trúc ECAPA-TDNN và mô hình RawNet2. Đầu tiên, dữ liệu đầu vào sẽ được thông qua bước pre-emphasis và truyền vào lớp Instance Normalisation [35]. Sau đó, dữ liệu sẽ được biểu diễn theo tần số-thời gian bằng phép phân tích filterbank đã được tham số hóa [24] mà tại đó, mô hình sẽ học những filterbank có các tham số phức tạp nhưng có giá trị. Tại bước này, khoảng rộng của chuỗi được nén sẽ phụ thuộc vào kích thước stride S , nếu giá trị kích thước stride càng nhỏ thì quá trình sẽ lâu hơn nhưng hiệu suất sẽ mạnh hơn.

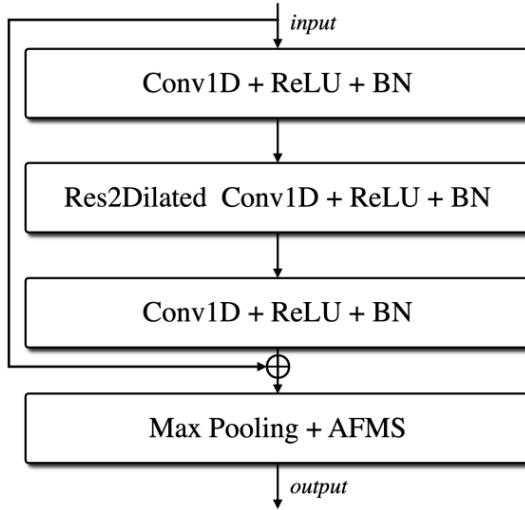
Ngược lại, kích thước stride lớn hơn thì mô hình sẽ xử lý nhanh hơn nhưng hiệu suất sẽ bị giảm. Với kích thước kernel mặc định là 251 và kích thước stride mặc định là 48 thì mỗi window sẽ là xấp xỉ 15 mili giây và dịch sang 3 miligiây cho mỗi frame. Kiến trúc của mô hình RawNet3 được mô tả trong hình 3.3.



Hình 3.3: Kiến trúc của mô hình RawNet3 theo [20]

Trong đó:

- k : kích thước kernel
- d : độ giãn nở
- p : kích thước max pooling
- C : số lượng channel
- S : kích thước stride
- \oplus : cộng ma trận



Hình 3.4: Kiến trúc của khối AFMS-Res2MP trong RawNet3

Sau đó, các filterbank đã tham số hóa sẽ được đưa vào ba khối backbone AFMS-Res2MP (được mô tả trong hình 3.4). Đầu ra của ba khối này sẽ được nối lại tương tự như kiến trúc của ECAPA-TDNN. Ngoài ra, đầu ra của khối thứ nhất và đầu ra của khối thứ hai sẽ được cộng ma trận và sẽ làm đầu vào của khối thứ ba. Theo mặc định, kích thước max pooling là 5 cho block đầu tiên, là 3 cho block thứ hai và 1 cho block cuối. Việc sử dụng max pooling này dựa trên kiến trúc RawNet2. Bên cạnh đó, việc giảm kích thước chuỗi là bắt buộc cho các mô hình nhận dạng người nói sử dụng sóng thô vì kích thước chuỗi của các mô hình này thường dài hơn khi so sánh với mô hình sử dụng các đặc trưng thủ công. Ngoài ra, việc sử dụng max pooling sẽ giúp mô hình giảm khả năng bị overfitting cho các mô hình nhận dạng người nói end-to-end. Những backbone này dựa trên Res2Net [12] và có kiến trúc tương đồng với ECAPA-TDNN nhưng có hai điểm khác biệt chính: trong khi ECAPA-TDNN sử dụng khối squeeze-excitation thì RawNet3 sẽ áp dụng AFMS từ RawNet2 và sẽ áp dụng max pooling trước AFMS.

Sau khi qua ba khối backbone, chuỗi sẽ đi qua một lớp tích chập với Batch Normalization [18] và sẽ tùy thuộc vào hai phương thức học để xây

dựng kiến trúc tiếp theo.

Với mô hình RawNet3, nhóm tác giả đã thực hiện hai phương thức học:

- Học có giám sát: nhóm tác giả sử dụng hàm mục tiêu là AAM-softmax 3.1 [9], còn được gọi là ArcFace, để huấn luyện mô hình cho việc phân loại. Mô hình được huấn luyện bằng dữ liệu có gắn nhãn và đầu phân loại (classification head) có số chiều bằng với số lượng người nói trong tập dữ liệu. Hàm AAM-softmax có thể ép cho khoảng cách giữa những người nói gần nhất lớn hơn dựa vào cosine similarity giữa embedding của người nói và ma trận trọng số của đầu phân loại. Một số tham số còn lại có thể xem chi tiết ở bài báo về ArcFace [9].

$$\mathcal{L}_A = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(\theta_{y_i,i})+m)}}{e^{s(\cos(\theta_{y_i,i})+m)} + \sum_{j=1, j \neq y_i}^N e^{s \cos(\theta_{j,i})}} \quad (3.1)$$

Trong đó:

- \mathbf{x}_i : embedding vector đại diện cho người nói
 - \mathbf{W} : trọng số ma trận của lớp phân loại
 - y_i : nhãn của người nói
 - i : chỉ số của lời nói trong một mini-batch kích thước N với $0 < i < N$
 - $\cos(\theta_{y_i,i})$: tích vô hướng giữa \mathbf{x}_i và \mathbf{W}_j
 - s : hệ số mở rộng
 - m : hệ số lè (margin)
- Học tự giám sát: nhóm tác giả sử dụng DINO framework [2]. DINO bao gồm một mô hình lớn (teacher network) và một mô hình nhỏ (student network), chúng có kiến trúc giống nhau nhưng khác các tham số. Các tham số của student network được cập nhật thông

qua hàm mất mát cross-entropy bằng phương pháp tự chắt lọc (self-distillation), còn các tham số của teacher network được cập nhật thông qua giá trị trung bình động hàm mũ (exponential moving average) của student network. Sau đó, một số kỹ thuật như mài dũa (sharpening) và tập trung hóa (centring) cũng được áp dụng để tránh việc mất biểu diễn trong quá trình huấn luyện. Hàm mất mát của DINO framework được mô tả tại công thức 3.2.

$$\mathcal{L}_D = \sum_{a \in \{a_1^g, a_2^g\}} \sum_{a' \in V, a' \neq a} H(P_t(a), P_s(a')) \quad (3.2)$$

Trong đó:

- V : một tập hợp góc nhìn
- a_1^g và a_2^g : góc nhìn toàn cục
- a^l : góc nhìn cục bộ
- P_t : đầu ra của mô hình lớn (teacher network)
- P_s : đầu ra của mô hình nhỏ (student network)
- $H(\cdot)$: cross entropy

3.1.3 Mô hình WavLM

WavLM [3] là mô hình học tự giám sát, một mô hình lớn tổng hợp nhiều tác vụ xử lý liên quan đến tiếng nói, chẳng hạn như nhận dạng tiếng nói (speech recognition), phân tách tiếng nói (speech separation), tách kênh người nói (speaker diarization), nhận dạng cảm xúc (emotion recognition), chuyển đổi giọng nói (voice conversion),... và kể cả định danh lẩn xác thực người nói. Học tự giám sát đang trở nên rất hiệu quả trong lĩnh vực xử lý ngôn ngữ tự nhiên [11, 38] hoặc xử lý tiếng nói [1, 16]. Tuy nhiên, các mô hình tiền huấn luyện hiện nay vẫn có những hạn chế, ví dụ như đối với các tác vụ liên quan đến nhiều người nói (tách kênh người nói, phân tách tiếng

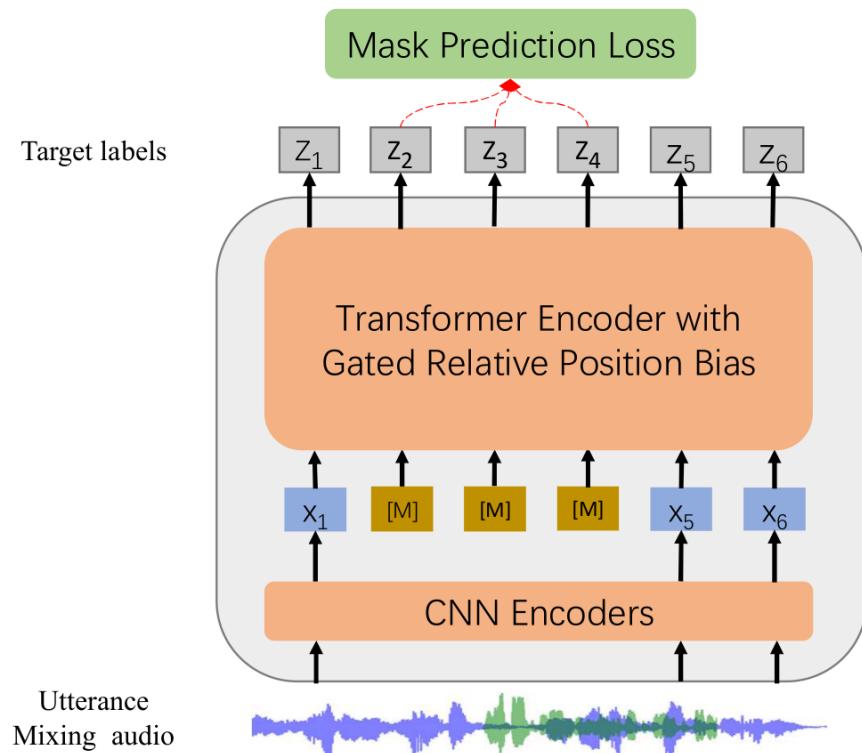
nói,...) hay việc còn phụ thuộc nhiều vào dữ liệu sách nói (audiobook). Việc phát triển một mô hình tiền huấn luyện tổng quát cho tổng hợp các tác vụ về tiếng nói (full-stack speech tasks) là điều cần thiết nhưng lại đầy thách thức.

Để giải quyết các vấn đề trên, nhóm tác giả đã đề xuất mô hình WavLM - một mô hình học các biểu diễn tiếng nói một cách tổng quát từ một lượng dữ liệu không dán nhãn khổng lồ nhưng vẫn thích nghi được với hàng loạt các tác vụ có liên quan đến tiếng nói. Nhóm tác giả đã đề xuất một framework dùng để khử nhiễu và dự đoán tiếng nói đã bị che (masked) trong khi dữ liệu đầu vào đã được giả lập để chèn thêm tiếng nói bị nhiễu hoặc bị lặp lại với mong muốn là dự đoán được nhãn giả (pseudo-label) của tiếng nói ban đầu trên những vùng đã bị che tương tự với mô hình HuBERT [16]. Framework này sẽ kết hợp quá trình khử nhiễu và dự đoán tiếng nói khi tiền huấn luyện. Do đó, mô hình WavLM không những sẽ học được những thông tin cho tác vụ nhận dạng tiếng nói mà còn học được những thông tin phục vụ cho nhiều tác vụ khác bằng mô hình khử nhiễu, ví dụ như các thông tin đặc trưng của người nói sẽ được mô hình hóa bằng nhãn giả.

Ngoài ra, nhóm tác giả cũng đã đề xuất một số cải tiến về kiến trúc mô hình và dữ liệu huấn luyện so với HuBERT và wav2vec2.0 [1]. Cụ thể, nhóm tác giả đã mở rộng bộ dữ liệu dùng để tiền huấn luyện lên tới 94 nghìn giờ. Ngoài ra, nhóm tác giả cải tiến kiến trúc Transformer bằng cách thêm "gated relative position bias" và xem nó như là backbone của mô hình, do đó có thể cải thiện hiệu năng cho tác vụ nhận dạng tiếng nói.

Cụ thể, mô hình WavLM (hình 3.5) sử dụng mô hình Transformer làm mô hình backbone. Lớp tích chập đầu tiên gồm có bảy khối tích chập theo thời gian đi kèm với Layer Normalization và hàm kích hoạt GELU. Các khối tích chập có 512 channel với kích thước stride tương ứng là (5,2,2,2,2,2,2), cũng như kích thước kernel là (10,3,3,3,3,2,2). Nhờ vậy, mỗi đầu ra sẽ đại diện cho một đoạn âm thanh 25 miligiây với khoảng dịch 20 miligiây. Đầu ra biểu diễn x được xem như là đầu vào của mô hình Transformer. Tại

đây, nhóm tác giả đã đề xuất "gated relative position bias" [4] được mã hóa dựa trên offset giữa "key" và "value" trong kiến trúc self-attention của Transformer. Khi so sánh với wav2vec2.0 hay HuBERT, việc có thêm cổng (gate) sẽ giúp mô hình xem xét nhiều hơn đến nội dung lời nói cũng như sẽ giúp thích ứng với nội dung bằng cách áp điều kiện. Bảng 3.1 mô tả kiến trúc của mô hình WavLM cùng với chiều của đầu ra khi ta đưa vào một đoạn âm thanh 3 giây (tương ứng với 48000 mẫu với tần số lấy mẫu là 16 kHZ) và số lượng tham số tương ứng cho mỗi lớp.



Hình 3.5: Kiến trúc của mô hình WavLM

Ngoài ra, nhóm tác giả đã đề xuất framework khử nhiễu và dự đoán tiếng nói để cải thiện sự mạnh mẽ của mô hình trong môi trường nhiều âm thanh mà vẫn bảo toàn được đặc trưng thiên về người nói. Cụ thể, nhóm tác giả thực hiện giả lập nhiều và âm chồng chéo lên nhau bằng cách chèn thêm âm thanh của nhiều người nói ngẫu nhiên trong mỗi batch cũng như trộn với một số nhiễu ngẫu nhiên được chọn từ các vùng khác và cũng sẽ được cắt ngẫu nhiên với một lượng năng lượng ngẫu nhiên sao cho vùng

Lớp	Chiều của đầu ra	Số lượng tham số #
UpstreamExpert: 1-1	—	—
WavLM: 2-1	—	1,024
ConvFeatureExtractionModel: 3-1	[1, 512, 149]	4,206,592
LayerNorm: 3-2	[1, 149, 512]	1,024
Linear: 3-3	[1, 149, 1024]	525,312
Dropout: 3-4	[1, 149, 1024]	—
TransformerEncoder: 3-5	[1, 149, 1024]	310,719,168

Bảng 3.1: Bảng tham số và chiều của đầu ra của mô hình WavLM với dữ liệu mẫu là 3 giây

bị chồng chéo không vượt quá 50% và chọn người nói đầu tiên của lời nói làm người nói chính. Do đó, mô hình có thể được huấn luyện để định danh người nói chính từ môi trường nhiều hoặc có tiếng nói chồng chéo, cũng như dự đoán được nội dung của tiếng nói tương ứng cho người nói chính dựa trên những vùng bị che.

Nhóm tác giả đã sử dụng hàm măt măt "Mask Prediction Loss" để tối ưu WavLM tương tự như HuBERT. Giả sử ta có một lời nói \mathbf{u} và phiên bản giả lập của nó \mathbf{u}' , khi đó ta luôn có thể tạo ra nhãn giả \mathbf{z} bằng cách đưa \mathbf{u} vào mô hình lặp cuối (last iteration network) và ta cũng sẽ có trạng thái ẩn \mathbf{h}_t^L sau khi đưa \mathbf{u}' vào mô hình. Nhóm tác giả sử dụng phép gom nhóm k-means vào các đặc trưng MFCC hoặc các biểu diễn ngầm dưới dạng nhãn giả. Công thức 3.3 mô tả hàm măt măt của mô hình WavLM.

$$\mathcal{L} = - \sum_{l \in K} \sum_{t \in M} \log p(z_t | \mathbf{h}_t^L) \quad (3.3)$$

Trong đó:

- M : tập hợp các chỉ số đã bị che trong miền thời gian
- \mathbf{h}_t^L : đầu ra của Transformer cho L lớp với bước t

Do đó, khi so sánh với các nghiên cứu trước đây, WavLM sẽ có lợi hơn cho những tác vụ không thiêng về nhận dạng tiếng nói do những thông tin

không liên quan đến tiếng nói (ví dụ: đặc trưng người nói) đã được mô hình hóa khi tiền huấn luyện.

Dối với tác vụ xác thực người nói, nhóm tác giả đã sử dụng mô hình ECAPA-TDNN làm mô hình downstream, đồng thời thử nghiệm nhiều cấu hình biểu diễn của đầu vào như sử dụng các đặc trưng thủ công hay các đặc trưng của mô hình tiền huấn luyện. Mô hình sẽ có một bộ mã hóa theo frame để trích xuất đặc trưng của người nói từ chuỗi đầu vào và một lớp pooling thống kê để biến đổi đầu vào thành một biểu diễn có chiều bất biến và một lớp kết nối đầy đủ để trích xuất embedding người nói. Khi huấn luyện, dữ liệu được cắt thành một mẫu 3 giây trong mỗi batch. Nhóm tác giả sử dụng hàm mất mát AAM-Softmax (công thức 3.1) để tối ưu mô hình.

Tóm lại, mô hình WavLM đã tạo tiền đề để phát triển các mô hình tiền huấn luyện **tổng quát** cho các tác vụ liên quan đến xử lý tiếng nói thay vì những mô hình chỉ tập trung vào các tác vụ có liên quan với nhau (chỉ liên quan đến người nói hoặc chỉ liên quan đến lời nói). Ngoài ra, nhóm tác giả đã đề xuất các thay đổi đơn giản nhưng hiệu quả cho các mô hình tiền huấn luyện trước đây, từ đó làm tăng tính tổng quát và tính nhất quán cho các tác vụ downstream (nhận dạng tiếng nói, xác thực người nói, phân tách tiếng nói, tách kênh người nói, ...).

3.2 Câu hỏi nghiên cứu

Trong đề tài này, nhóm sẽ trả lời các câu hỏi chính:

1. **Với tập dữ liệu nhỏ, liệu có thể xây dựng một mô hình nhận dạng người nói tốt hay không?**

Có hai cách hiểu về khái niệm "tập dữ liệu nhỏ", nhưng câu trả lời cho cả hai cách là hoàn toàn có thể. Cụ thể như sau:

- Trường hợp bộ dữ liệu nhỏ đến nỗi chỉ có một hoặc một vài mẫu cho mỗi lớp. Một số phương pháp đã được đề xuất để giải quyết

trường hợp này, là one-shot learning [37] hoặc few-shot learning [29]... Các phương pháp này khắc phục rất tốt các tình huống về lượng dữ liệu cực nhỏ như thế này và kết quả mang lại rất đáng mong đợi.

- Trường hợp bộ dữ liệu nhỏ so với mặt bằng chung, tức là số lượng mẫu ở mỗi lớp là vài chục hoặc hơn, nhưng về tổng thể thì nó vẫn nhỏ (dưới 1000 lớp). Trường hợp này được gọi là bộ dữ liệu hạn chế và đây là một chi tiết chủ yếu trong đề tài của nhóm. Cụ thể hơn, bộ dữ liệu tiếng Việt được công khai hiện nay khá là ít, trong khi đó nói đến bộ dữ liệu tiếng Anh thì có thể kể ngay những cái nên nổi bật như VoxCeleb1, VoxCeleb2, GigaSpeech,... Với tình huống này, phương pháp học chuyển đổi (được trình bày ở mục 3.3), sẽ góp phần mang lại hiệu quả tốt hơn.

2. Tại sao chọn các mô hình nhận dạng người nói trên (RawNet3 và WavLM) và liệu chúng có đáng tin cậy hay không?

- Lý do nhóm chọn hai mô hình RawNet3 và WavLM để nghiên cứu nhận dạng người nói là vì đây là các mô hình SOTA trong lĩnh vực này. Cả hai đều được huấn luyện trên bộ dữ liệu tiếng Anh dồi dào, cụ thể là mô hình RawNet3 được huấn luyện trên tập dữ liệu gồm VoxCeleb1 và VoxCeleb2 (hơn 7000 người nói và hơn 2000 giờ, cụ thể hơn được trình bày ở mục 3.4) và mô hình WavLM được huấn luyện trên tập dữ liệu khổng lồ lên tới 94000 giờ (với mô hình downstream thì được huấn luyện thêm trên các tập VoxCeleb). Bên cạnh đó, nhóm tác giả có công khai mô hình tiền huấn luyện cho mọi người sử dụng. Điều này khuyến khích việc sử dụng phương pháp học chuyển đổi để có thể đạt được hiệu suất cao trên tập dữ liệu hạn chế như tiếng Việt.

- Bên cạnh việc huấn luyện trên tập dữ liệu lớn như tiếng Anh, kết quả thu được từ các mô hình cũng là một yếu tố để xác định liệu các mô hình có đáng tin cậy để sử dụng hay không. Kết quả sử dụng để đánh giá là các độ đo EER và minDCF (được trình bày ở mục 3.5). Với mô hình RawNet3, độ đo EER và minDCF tốt nhất đạt được lần lượt là 0.87% và 0.0593 với phương thức học có giám sát, còn phương thức học tự giám sát thì kết quả lần lượt là 5.40%, 0.3396. Với mô hình WavLM (kết hợp ECAPA-TDNN), nhóm tác giả đã đạt kết quả về độ đo EER lần lượt là 0.383%, 0.48%, 0.986% cho tập Vox1-O, Vox1-E và Vox1-H. Có thể thấy rằng các kết quả này đã chứng minh được đây là các mô hình đang rất mạnh mẽ trong việc nhận dạng người nói.

3.3 Phương pháp nghiên cứu

3.3.1 Huấn luyện mô hình từ đầu chỉ bằng tập dữ liệu tiếng Việt

Như đã nêu trên, hai mô hình mới nhất và hiệu quả nhất được nghiên cứu trong lĩnh vực nhận dạng người nói là RawNet3 và WavLM (kết hợp ECAPA-TDNN, gọi ngắn gọn là WavLM (ECAPA-TDNN)). Các nhóm tác giả đề xuất mô hình với mục đích là nhận dạng người nói cho ngôn ngữ tiếng Anh, do đó các mô hình được huấn luyện trên tập dữ liệu tiếng Anh dồi dào. Tương tự như vậy, để nhận dạng người trên tiếng Việt thì các mô hình cần phải được học qua dữ liệu ngôn ngữ tiếng Việt. Tuy nhiên, tập dữ liệu tiếng Việt lại hạn chế, liệu kết quả có khả quan hay không? Và nhóm sẽ trả lời câu hỏi này bằng cách huấn luyện hai mô hình này từ đầu trên tập dữ liệu tiếng Việt để kiểm chứng sự hiệu quả của mô hình cũng như độ hiệu quả của hướng tiếp cận này. (Kết quả chi tiết được trình bày ở chương 4)

Sau khi thu được kết quả, một câu hỏi mới được đặt ra là liệu có cách

nào để cải thiện kết quả hơn hay không? Quá trình tìm ra câu trả lời cho câu hỏi này sẽ được thể hiện qua các hướng tiếp cận kế tiếp.

3.3.2 Sử dụng mô hình tiền huấn luyện trên tập dữ liệu tiếng Anh và đánh giá trên dữ liệu tiếng Việt

Các kết quả tốt mà ta thấy khi các nhóm tác giả đề xuất là do các mô hình được đánh giá dựa trên tập dữ liệu dồi dào có cùng ngôn ngữ với tập dữ liệu khi huấn luyện, cụ thể là tiếng Anh. Bên cạnh đó, các nhóm tác giả có công khai các mô hình tiền huấn luyện cho mọi người có thể sử dụng. Dựa vào đây, nhóm sẽ sử dụng các mô hình tiền huấn luyện để thực hiện nhận dạng người nói trên tiếng Việt và đánh giá xem độ hiệu quả của nó. Với hướng tiếp cận này, mong muốn chủ yếu của nhóm ở hướng tiếp cận này là đánh giá khả năng nhận dạng người nói của các mô hình trên tập dữ liệu nhỏ tiếng Việt khi đã được huấn luyện bằng lượng lớn dữ liệu tiếng Anh.

3.3.3 Huấn luyện mô hình bằng phương pháp transfer learning cho tiếng Việt

Để các mô hình đạt hiệu suất cao trên tập dữ liệu hạn chế như tiếng Việt thì việc dựa vào các mô hình tiền huấn luyện là điều cần thiết. Vì các mô hình tiền huấn luyện đã được học trên lượng lớn dữ liệu tiếng Anh và đang hoạt động khá hiệu quả nên việc kế thừa nó vừa tốn ít chi phí vừa có khả năng nâng cao hiệu năng trên dữ liệu tiếng Việt do nó đã có sẵn kiến thức về nhận dạng người nói trên tập dữ liệu tiếng Anh. Khi đó, nhóm sử dụng các mô hình tiền huấn luyện như là điểm bắt đầu huấn luyện (thay vì huấn luyện từ đầu) trên tập dữ liệu tiếng Việt. Cách tiếp cận như vậy được gọi là phương pháp học chuyển đổi (transfer learning) và đây cũng chính là cách tiếp cận chính trong đề tài của nhóm. Nhóm sẽ tiến hành áp

dụng hai phương pháp transfer learning sau:

Huấn luyện mô hình bằng kỹ thuật fine-tuning

Với phương pháp này, nhóm tiến hành nạp trọng số vào các kiến trúc mô hình RawNet3 và WavLM (ECAPA-TDNN) bằng mô hình tiền huấn luyện, sau đó bắt đầu huấn luyện trên tập dữ liệu tiếng Việt. Điều này mang lại nhiều lợi ích như là tiết kiệm thời gian huấn luyện mô hình mà vẫn giữ được kiến thức đã được học từ trước, tiết kiệm bộ nhớ vì không cần phải trộn dữ liệu để mô hình học toàn bộ một lượt như đã trình bày ở các cách tiếp cận trước.

Huấn luyện mô hình bằng cách kết hợp thêm mô hình nhỏ

Còn ở phương pháp này, với mong muốn tăng hiệu quả hơn nữa, nhóm tiến hành kết hợp thêm một mô hình nhỏ vào các mô hình ban đầu. Với cách tiếp cận này, mục tiêu là phần mô hình mới này sẽ được huấn luyện với tập dữ liệu tiếng Việt, phần mô hình cũ thì vẫn giữ được tri thức nhận dạng người từ mô hình tiền huấn luyện. Nhóm thực hiện thử nghiệm với hai mô hình nhỏ như sau:

- Mô hình nhỏ chỉ gồm một lớp kết nối đầy đủ: mục đích của lớp này là để ánh xạ embedding (của người nói trên ngôn ngữ tiếng Anh) của hai mô hình RawNet3 và WavLM (ECAPA-TDNN) sang embedding mới trên ngôn ngữ tiếng Việt.
- Mô hình nhỏ gồm ba lớp kết nối đầy đủ: mục đích của mô hình nhỏ này là ngoài việc ánh xạ embedding từ tiếng Anh sang tiếng Việt còn giúp học thêm thông tin về của ngôn ngữ tiếng Việt. Cụ thể, hai lớp kết nối đầy đủ đầu tiên sẽ đảm nhiệm việc học thêm thông tin đặc trưng trong giọng nói của ngôn ngữ tiếng Việt, lớp còn lại để tạo ra embedding mới. Embedding mới này sẽ giúp nhận dạng người nói trên tiếng Việt tốt hơn.

Kết quả thực tế của hướng tiếp cận này, cũng như các hướng tiếp cận trước đó, sẽ được trình bày cụ thể ở chương 4. Các kết quả có thể hiệu quả hơn đúng như ý tưởng đã trình bày, tuy nhiên vẫn có khả năng không đạt được như mong muốn. Nhưng một điều chắc chắn là với các hướng tiếp cận đã nêu sẽ giúp tiết kiệm chi phí nghiên cứu hơn.

3.3.4 Khảo sát ảnh hưởng của các miền tần số và nhiễu đến mô hình nhận dạng người nói

Bên cạnh việc huấn luyện để các mô hình đạt hiệu suất tốt hơn trong việc nhận dạng người nói trên bộ dữ liệu tiếng Việt hạn chế, nhóm còn khảo sát thêm các yếu tố về miền tần số và nhiễu sẽ ảnh hưởng ra sao đối với khả năng nhận dạng của các mô hình:

- Đối với yếu tố tần số, nhóm sử dụng low-pass filter (bộ lọc thông thấp) và band-pass filter (bộ lọc thông dải) để phục vụ cho việc khảo sát. Từ đây, nhóm đánh giá được đâu là băng tần quan trọng cần thiết để nhận dạng người nói.
- Đối với yếu tố nhiễu, nhóm sử dụng white noise (nhiễu trắng). Lượng nhiễu được đo bằng tỷ lệ tín hiệu trên nhiễu, từ đó nhóm đánh giá xem lượng nhiễu các mô hình vẫn có thể chịu được là khoảng bao nhiêu.

3.4 Mô tả dữ liệu

3.4.1 Bộ dữ liệu VoxCeleb1

Bộ dữ liệu VoxCeleb1 [22] là bộ dữ liệu gồm hơn 100,000 bản ghi từ 1251 người nói được trích xuất từ các video đã được tải lên YouTube. Bảng 3.2 mô tả phân bố của tập dữ liệu VoxCeleb1.

Bảng 3.2: Bảng phân bố tập dữ liệu VoxCeleb1

	Train	Test
Số người nói	1211	40
Số lượng video	21819	677
Số bản ghi	122813	30703

3.4.2 Bộ dữ liệu VoxCeleb2

Bộ dữ liệu VoxCeleb2 [5] là bộ dữ liệu gồm hơn 1 triệu bản ghi từ 6112 người nói (chủ yếu là người nổi tiếng) được trích xuất từ các video đã được tải lên YouTube. Bộ dữ liệu VoxCeleb2 không có bất kỳ người nói nào tương đồng với bộ dữ liệu VoxCeleb1. Bảng 3.3 mô tả phân bố của bộ dữ liệu VoxCeleb1.

Bảng 3.3: Bảng phân bố tập dữ liệu VoxCeleb2

	Train	Test
Số người nói	5,994	118
Số lượng video	145,569	4,911
Số bản ghi	1,092,009	36,237

3.4.3 Bộ dữ liệu Zalo Voice Verification

Bộ dữ liệu Zalo Voice Verification là bộ dữ liệu công khai không phụ thuộc vào văn bản được cung cấp bởi Zalo Group và đã được sử dụng trong cuộc thi Zalo AI Challenge 2020 cho bài toán Xác thực giọng nói (Voice Verification). Bộ dữ liệu gồm có 400 người nói với 201 nữ và 199 nam và trung bình có 26.4 bản ghi cho mỗi người nói với độ dài mỗi bản ghi có khoảng từ 0.8 giây đến 11 giây.

Bộ dữ liệu sẽ được chia theo tỷ lệ 8:1:1 tương ứng với ba tập train/dev/test tương ứng, trong đó tập train sẽ gồm 320 người, tập dev và tập test mỗi tập gồm 40 người.

Bộ dữ liệu xác thực người nói được tạo với 20000 cặp lời nói theo từng người nói tương ứng. Không có cặp nào trùng trong bộ dữ liệu xác thực này để tăng độ chính xác khi thử nghiệm.

3.5 Độ đo để đánh giá

3.5.1 Tỷ lệ lỗi bình đẳng - Equal Error Rate

Để đánh giá hiệu quả của mô hình sẽ cần phân tích của hai loại lỗi: tỷ lệ chấp nhận sai (FAR - False Acceptance Rate) và tỷ lệ từ chối sai (FRR - False Rejection Rate). Tỷ lệ chấp nhận sai cho ta biết tỷ lệ những bản ghi khác người nói nhưng vẫn được chấp nhận là cùng một người nói bởi mô hình và tỷ lệ từ chối sai cho ta biết tỷ lệ những bản ghi cùng người nói nhưng bị xác nhận là khác người nói bởi mô hình.

Ta có công thức tính tỷ lệ chấp nhận sai (3.4) và tỷ lệ từ chối sai (3.5) tương ứng như sau:

$$FAR = \frac{FP}{TP + FP} \quad (3.4)$$

$$FRR = \frac{FN}{TN + FN} \quad (3.5)$$

Trong đó:

- **FP (False Positive)** là dương tính giả, tức là số lượng mẫu khác người nói nhưng hệ thống chấp nhận là cùng người nói.
- **TP (True Positive)** là dương tính thực, tức là số lượng mẫu cùng người nói và hệ thống xác nhận là cùng người nói.

- **FN (False Negative)** là âm tính giả, tức là số lượng mẫu cùng người nói nhưng bị hệ thống xác nhận là khác người nói.
- **TN (True Negative)** là âm tính thực, tức là số lượng mẫu khác người nói và hệ thống xác nhận là khác người nói.

Chúng ta có thể biểu diễn các đại lượng trên dưới dạng ma trận 3.4 (hay được gọi là Confusion Matrix)

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Bảng 3.4: Confusion Matrix

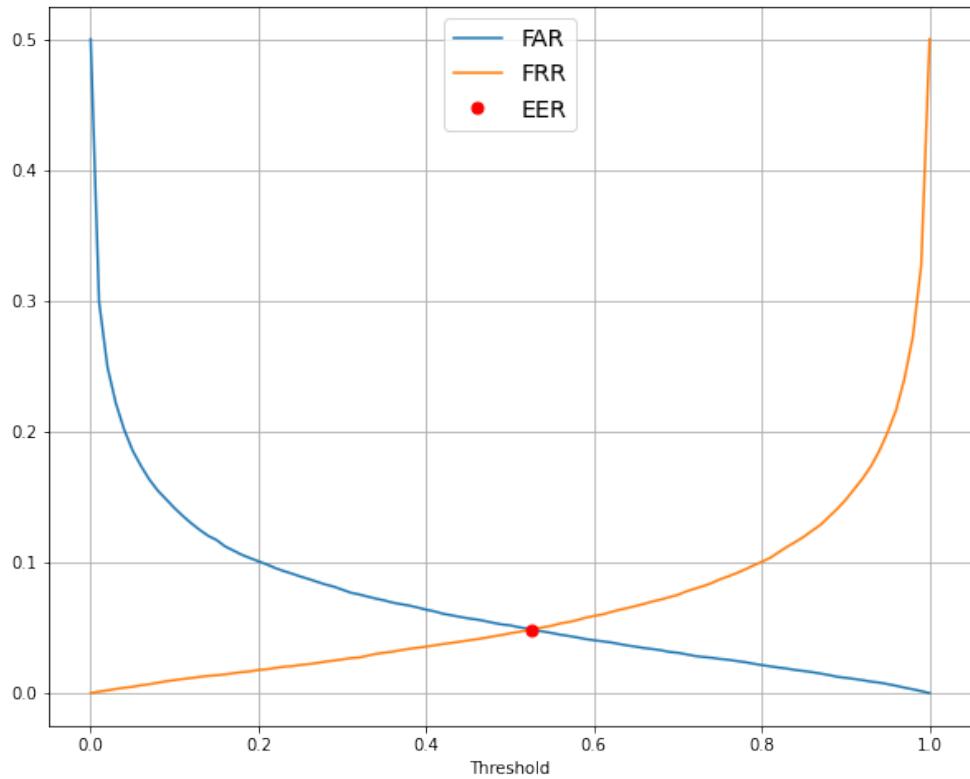
Sau khi tính được tỷ lệ chấp nhận sai và tỷ lệ từ chối sai, ta sẽ tính được tỷ lệ lỗi bình đẳng (Equal Error Rate). Tỷ lệ lỗi bình đẳng là tỷ lệ mà tại đó, $FAR = FRR$. Hình 3.6 là ví dụ về tỷ lệ lỗi bình đẳng tại giao điểm của FAR và FRR.

Giả sử, nếu một mô hình có $EER = 1.5\%$ thì cứ 1000 cặp mẫu được khảo sát thì sẽ có 15 cặp mẫu khác người nói nhưng mô hình xác nhận là cùng một người nói và cũng sẽ có 15 cặp mẫu cùng một người nói nhưng hệ thống xác nhận là khác người nói. Có thể thấy, giá trị tỷ lệ lỗi bình đẳng càng thấp thì mô hình càng chính xác và đáng tin cậy.

3.5.2 Hàm chi phí phát hiện tối thiểu - Minimum Detection Cost Function

Hàm chi phí phát hiện tối thiểu ¹, hay còn được gọi là minDCF, là một độ đo có thể sử dụng kèm với độ đo EER để đánh giá mô hình nhận dạng

¹NIST 2016 Speaker Recognition Evaluation Plan



Hình 3.6: Ví dụ về tỷ lệ lỗi bình đẳng với EER là điểm màu đỏ

người nói. Mục đích của minDCF là tìm sự cân bằng tối ưu giữa những lần phát hiện bị bỏ lỡ (missed detections, hay false negatives) và những lần báo động sai (false alarms, hay false positives) trong quá trình ra quyết định của mô hình.

Độ đo minDCF kết hợp các chi phí liên quan đến phát hiện bị bỏ lỡ và báo động sai, cùng với xác suất tương ứng xảy ra các lỗi này. Công thức để tính minDCF (3.6) như sau:

$$\text{minDCF} = C_{\text{miss}} * P_{\text{miss}} * P_{\text{target}} + C_{\text{fa}} * P_{\text{fa}} * (1 - P_{\text{target}}) \quad (3.6)$$

Trong đó:

- C_{miss} : Chi phí liên quan đến phát hiện bị bỏ lỡ.
- P_{miss} : Xác suất phát hiện bị bỏ lỡ (tỉ lệ từ chối sai).
- P_{target} : Xác suất tiên nghiệm về sự xuất hiện của người nói mục tiêu.

- C_{fa} : Chi phí liên quan đến báo động sai.
- P_{fa} : Xác suất báo động sai (tỉ lệ chấp nhận sai).

Mục tiêu lớn nhất là giảm thiểu chi phí phát hiện này, được thực hiện bằng cách tìm các giá trị tham số phù hợp với các yêu cầu cũng như mức độ ưu tiên cụ thể mà ta quan tâm đến. Việc tối thiểu hóa giá trị minDCF giúp hiệu suất của mô hình được nâng cao về độ chính xác lẫn độ tin cậy.

Giả sử chúng ta mong muốn mô hình nhạy bén hơn trong việc phát hiện ra người lạ, chúng ta có thể tùy chỉnh các yếu tố, chẳng hạn như tăng chi phí liên quan đến báo động sai cao hơn. Điều này giúp mô hình ưu tiên làm giảm số lượng dương tính giả. Bên cạnh đó, chúng ta cũng có thể giảm xác suất tiên nghiệm về sự xuất hiện của người nói mục tiêu, nghĩa là xác suất xuất hiện người lạ cao hơn, giúp mô hình cẩn trọng hơn trong việc nhận dạng người lạ. Do đó, nếu giá trị minDCF đủ nhỏ, tức là mô hình phát hiện ra người lạ rất tốt và minDCF rất phù hợp cho môi trường quan tâm tới từ chối sai và chấp nhận sai nhiều hơn so với từ chối đúng và chấp nhận đúng trong môi trường đề cao tính bảo mật.

Chương 4

Thử nghiệm và kết quả

4.1 Quy trình thử nghiệm và kết quả chi tiết

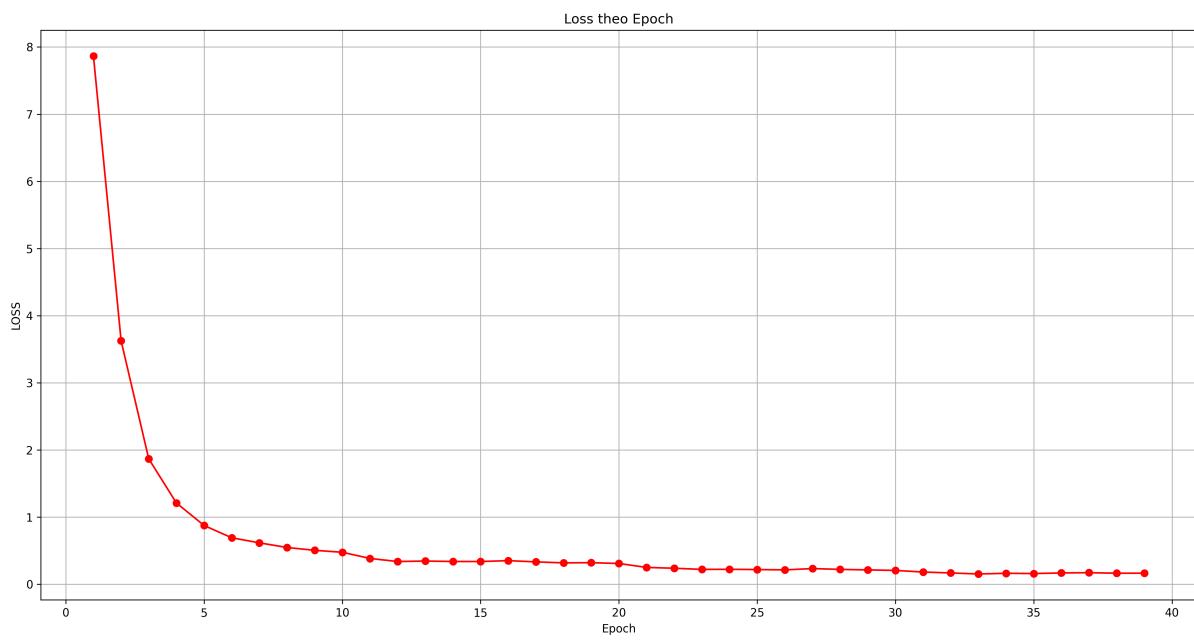
4.1.1 Huấn luyện mô hình từ đầu bằng tập dữ liệu tiếng Việt

Hình 4.1 là biểu đồ loss trong quá trình huấn luyện mô hình RawNet3 trên tập dữ liệu Zalo. Có thể thấy mô hình RawNet3 có khả năng học rất nhanh, cụ thể chỉ trong vòng 5 epoch đầu hàm loss hội tụ rất nhanh và từ epoch 10 trở đi chỉ giảm được thêm đôi chút.

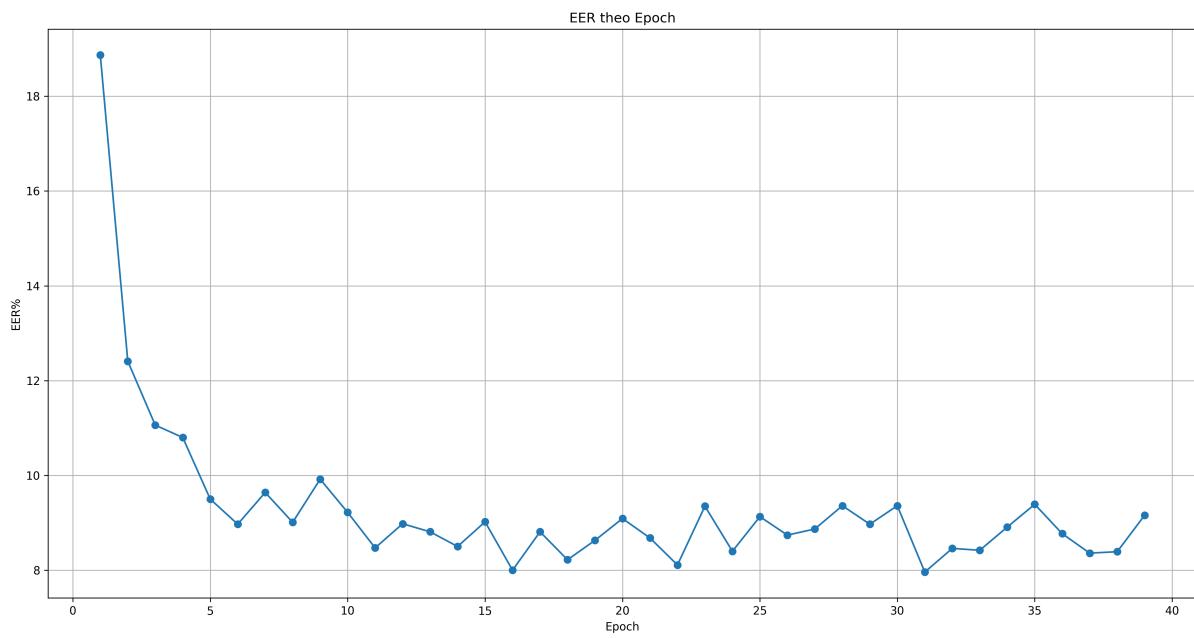
Hình 4.2 là biểu đồ Equal Error Rate trên tập test trong quá trình huấn luyện mô hình RawNet3 trên tập dữ liệu Zalo. Độ đo EER cũng giảm rất nhanh chỉ trong 5 epoch đầu, điều này chứng minh khả năng học khá tốt của RawNet3. Tuy nhiên từ epoch 6 trở đi, cho tới epoch cuối là 39, khả năng nhận dạng người nói của RawNet3 lúc thì tốt lúc thì không.

Hình 4.3 là biểu đồ loss trong quá trình huấn luyện mô hình WavLM (ECAPA-TDNN) trên tập dữ liệu Zalo. Khác một chút với RawNet3, WavLM (ECAPA-TDNN) có tốc độ học chậm hơn và hội tụ trong vòng 15 epoch đầu tiên, sau đó qua mỗi epoch cũng giảm thêm đôi chút.

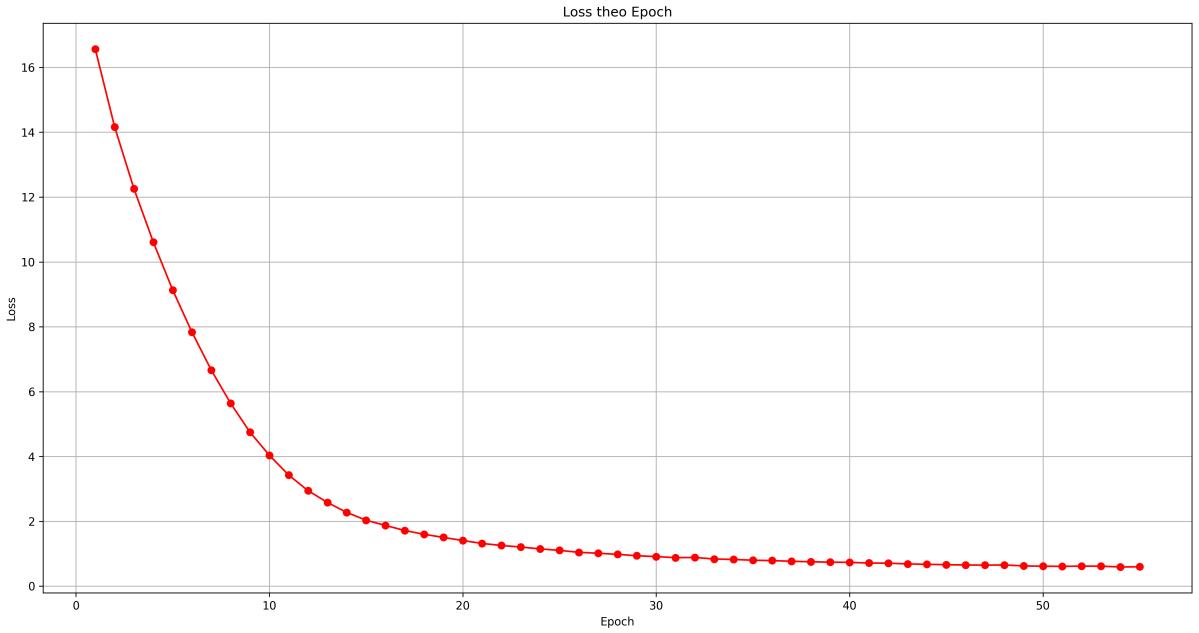
Hình 4.4 là biểu đồ Equal Error Rate trên tập test trong quá trình huấn luyện mô hình WavLM (ECAPA-TDNN) trên tập dữ liệu Zalo. Mặc dù



Hình 4.1: Biểu đồ loss trong quá trình huấn luyện mô hình RawNet3 trên tập dữ liệu Zalo



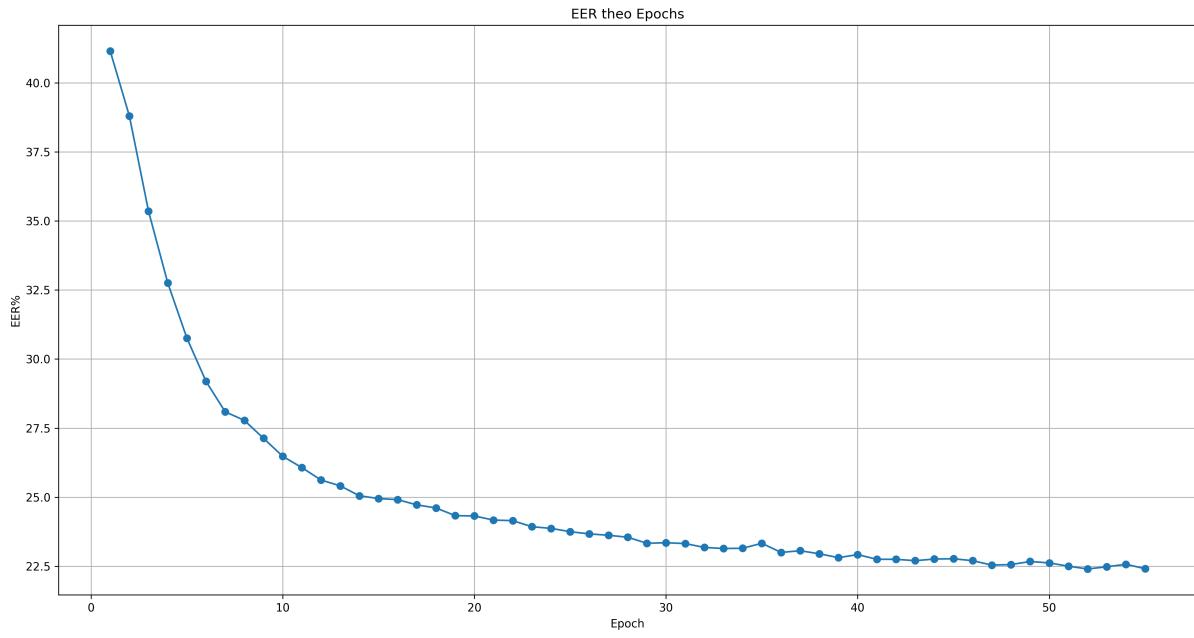
Hình 4.2: Biểu đồ EER trong quá trình huấn luyện mô hình RawNet3 trên tập dữ liệu Zalo



Hình 4.3: Biểu đồ loss trong quá trình huấn luyện mô hình WavLM (ECAPA-TDNN) trên tập dữ liệu Zalo

biểu đồ thể hiện khá tốt nhưng độ đo EER lại quá cao. Nhóm có thực hiện huấn luyện thêm 16 epoch so với RawNet3, nhưng WavLM (ECAPA-TDNN) cũng không cải thiện hơn là bao. Với EER tệ như vậy có thể được lý giải là do mô hình WavLM trích xuất embedding mang quá nhiều thông tin đặc trưng (gồm cả nội dung tiếng nói lẫn đặc trưng người nói), do đó lượng dữ liệu cần thiết để huấn luyện cũng phải lớn để mô hình downstream có thể tập trung vào phần đặc trưng người nói và lọc đi các thông tin về nội dung trong tiếng nói.

Bảng 4.1 cho ta kết quả tốt nhất trong quá trình huấn luyện mô hình nhận dạng người nói từ đầu trên bộ dữ liệu tiếng Việt nhỏ. EER tốt nhất của RawNet3 là 7.96% trong 39 epoch, trong khi đó WavLM (ECAPA-TDNN) được huấn luyện đến 55 epoch nhưng hiệu quả lại tệ hơn gần gấp 3 lần và EER tốt nhất chỉ 22.4%. Mô hình WavLM (ECAPA-TDNN) có kết quả khá tệ so với mô hình RawNet3 (kém hơn 14.44% EER) do mô hình WavLM (ECAPA-TDNN) quá lớn, xấp xỉ 325 triệu tham số, trong



Hình 4.4: Biểu đồ EER trong quá trình huấn luyện mô hình WavLM (ECAPA-TDNN) trên tập dữ liệu Zalo

khi mô hình RawNet3 chỉ có 16 triệu tham số nên mô hình càng lớn phải đi đôi với dữ liệu lớn. Tuy nhiên, bộ dữ liệu nhỏ được sử dụng để huấn luyện lại nhỏ (dưới 1000 người). Do đó, ta kết luận rằng việc sử dụng mô hình WavLM (ECAPA-TDNN) để huấn luyện bộ dữ liệu nhỏ là không khả thi, nên sử dụng mô hình có kích thước nhỏ hơn để huấn luyện sẽ tốt hơn (RawNet3).

Mô hình	EER%	minDCF
RawNet3	7.96	0.4826
WavLM (ECAPA-TDNN)	22.4	0.9403

Bảng 4.1: Kết quả tốt nhất khi huấn luyện mô hình nhận dạng người nói từ đầu trên bộ dữ liệu Zalo

4.1.2 Sử dụng mô hình tiền huấn luyện bằng tập dữ liệu tiếng Anh cho tập dữ liệu tiếng Việt

Bảng 4.2 cho ta kết quả khi áp dụng mô hình nhận dạng người nói đã được tiền huấn luyện trên dữ liệu tiếng Anh (đã được tiền huấn luyện trên tập dữ liệu VoxCeleb1 và VoxCeleb2) cho tập dữ liệu nhỏ tiếng Việt.

Mô hình	EER%	minDCF
RawNet3	9.74	0.4431
WavLM (ECAPA-TDNN)	11.26	0.5410

Bảng 4.2: Kết quả khi áp dụng mô hình tiền huấn luyện trên tập dữ liệu tiếng Anh cho dữ liệu tiếng Việt

Có thể thấy đối với mô hình RawNet3, việc áp dụng mô hình đã được huấn luyện trên một lượng lớn thông tin đặc trưng người nói bằng lượng lớn dữ liệu ngôn ngữ tiếng Anh vào ngôn ngữ tiếng Việt lại mang hiệu quả kém hơn so với khi mô hình được huấn luyện chỉ trên dữ liệu tiếng Việt, cụ thể EER là 9.74% so với 7.96% trong bảng 4.1.

Ngược lại, mô hình tiền huấn luyện của WavLM (ECAPA-TDNN) khi nhận dạng người nói trên tiếng Việt lại hoạt động tốt hơn so với khi nó được huấn luyện chỉ bằng dữ liệu tiếng Việt, cụ thể kết quả EER là 11.26% so với 22.4% trong bảng 4.1. Điều này có thể được lý giải rằng do lượng dữ liệu tiếng Anh được dùng để huấn luyện mô hình tiền huấn luyện đủ lớn để mô hình có thể học một cách hiệu quả, từ đó rút trích được nhiều đặc trưng người nói và loại bỏ các thông tin không liên quan đến người nói.

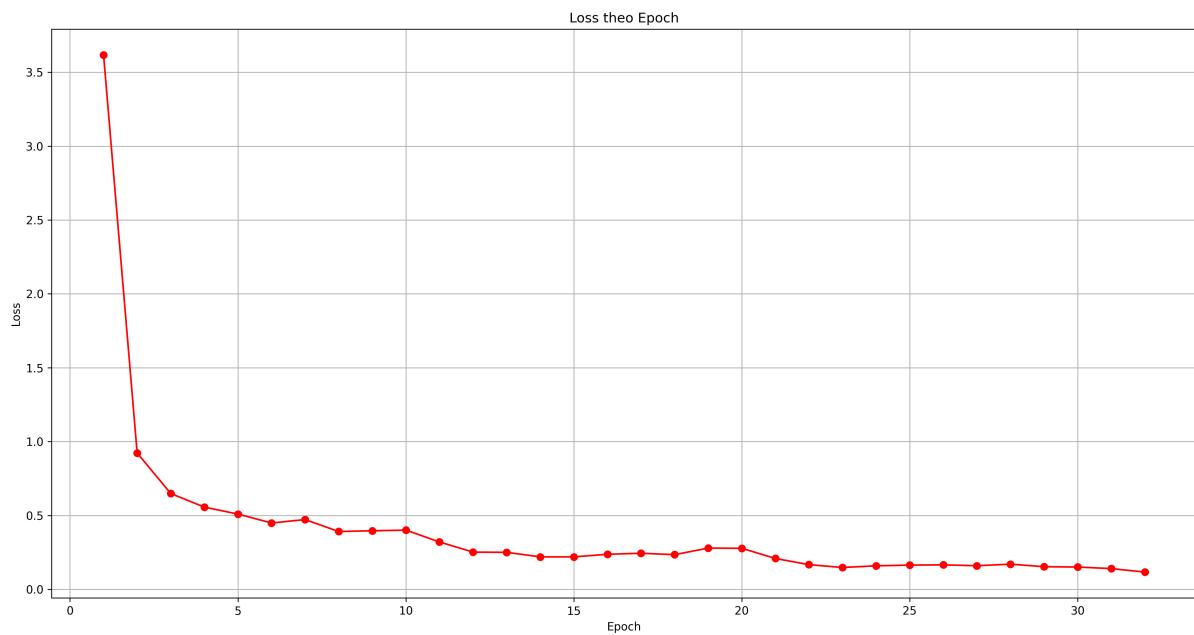
Thực tế, mô hình RawNet3 [20] đạt được $EER = 0.89\%$ trong khi mô hình WavLM [3] đạt được $EER = 0.383\%$ khi thử nghiệm trên tập Vox1-O. Tuy nhiên, khi thử nghiệm trên tập dữ liệu nhỏ tiếng Việt thì mô hình RawNet3 lại nhỉnh hơn so với mô hình WavLM (ECAPA-TDNN) một khoảng 1.52% EER do mô hình WavLM là mô hình phụ thuộc vào ngôn ngữ¹, tức WavLM sẽ hoạt động tốt hơn trên ngôn ngữ đã được tiền huấn

¹<https://github.com/microsoft/unilm/issues/1001>

luyện (cụ thể là tiếng Anh) so với khi sử dụng trên ngôn ngữ khác và có thể hoặc thậm chí tệ hơn khi sử dụng ngôn ngữ ngoài miền huấn luyện.

4.1.3 Huấn luyện mô hình bằng fine-tuning cho tiếng Việt

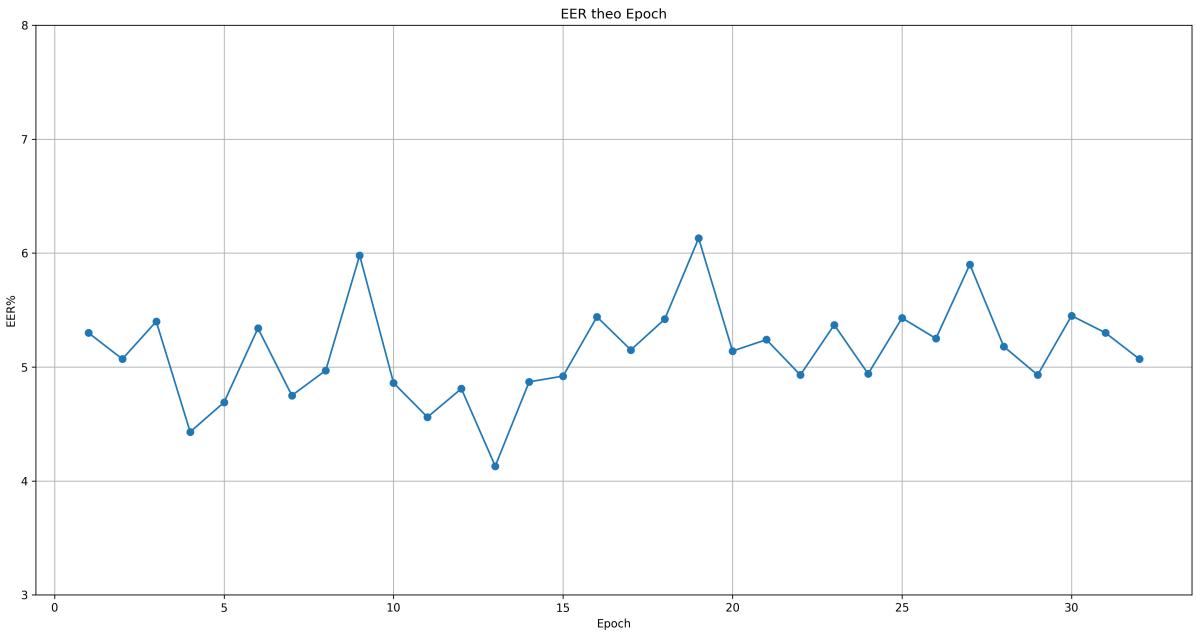
Hình 4.5 là biểu đồ loss trong quá trình fine-tuning mô hình RawNet3 trên tập dữ liệu Zalo. Tốc độ học của RawNet3 rất nhanh như thử nghiệm ở phần 4.1.1. Ở hướng tiếp cận này, do kế thừa những kiến thức về đặc trưng người nói từ mô hình tiền huấn luyện, mô hình RawNet3 chỉ cần học thêm một số đặc trưng của người nói tiếng Việt là có thể có kết quả đáng mong đợi.



Hình 4.5: Biểu đồ loss trong quá trình fine-tuning mô hình RawNet3 trên tập dữ liệu Zalo

Hình 4.6 là biểu đồ EER trong quá trình fine-tuning mô hình RawNet3 trên tập dữ liệu Zalo. Có thể thấy rằng việc fine-tuning mô hình RawNet3 đã hoạt động tốt khi huấn luyện chỉ trong vài epoch đầu, tốt hơn nhiều so

với chỉ sử dụng mô hình tiền huấn luyện (so với kết quả trong bảng 4.2 thì độ đo EER giảm lên đến 5.61%). Tuy nhiên, việc fine-tuning trên nhiều epoch là không có ý nghĩa đối với mô hình RawNet3 và kết quả tốt nhất đạt được tại epoch thứ 13 và kết quả đạt tốt thứ nhì tại epoch 4 dựa vào biểu đồ.

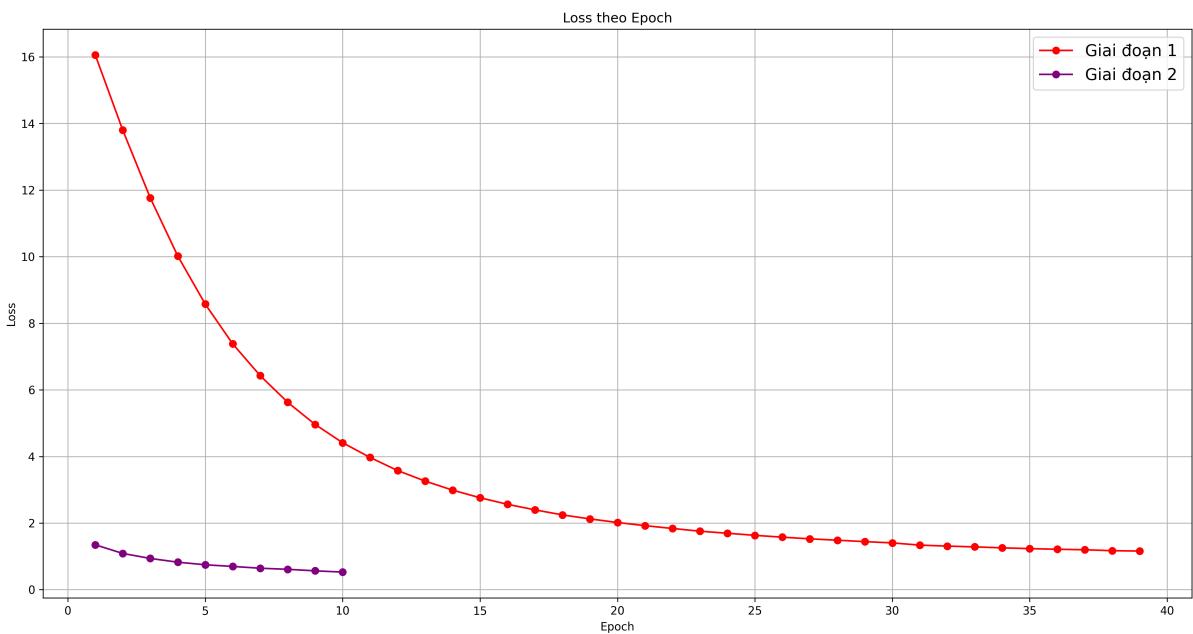


Hình 4.6: Biểu đồ EER trong quá trình fine-tuning mô hình RawNet3 trên tập dữ liệu Zalo

Hình 4.7 là biểu đồ loss trong quá trình fine-tuning mô hình WavLM (ECAPA-TDNN) trên tập dữ liệu Zalo. Nhóm đã tiến hành finetuning mô hình WavLM (ECAPA-TDNN) trong hai giai đoạn:

- Giai đoạn 1: Đóng băng (freeze) mô hình WavLM, fine-tuning mô hình downstream ECAPA-TDNN cho 39 epoch. Có thể thấy tốc độ học của mô hình WavLM (ECAPA-TDNN) chậm hơn so với mô hình RawNet3 và trong khoảng 15 epoch đầu, hàm loss của mô hình hội tụ khá nhanh nhưng từ epoch 16 trở về sau thì khả năng hội tụ bị chậm lại.

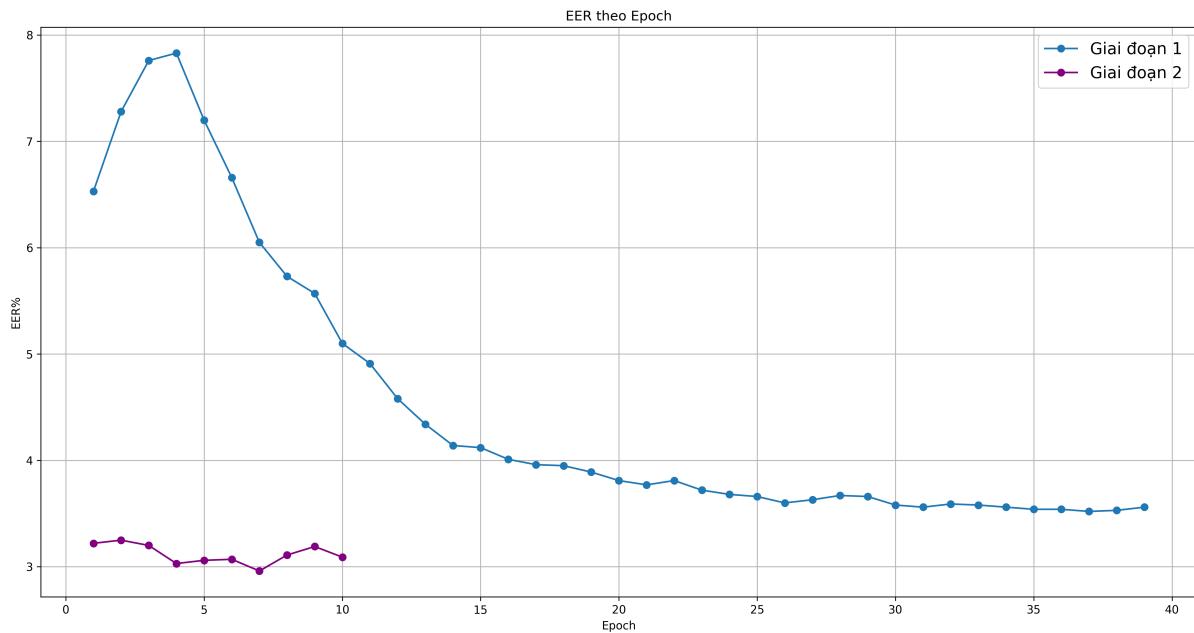
- Giai đoạn 2: Mở đóng băng (unfreeze) mô hình WavLM và fine-tuning cả upstream WavLM và downstream ECAPA-TDNN cho 10 epoch. Từ đó giúp cho mô hình upstream có thể học được thêm những đặc trưng của người nói tiếng Việt.



Hình 4.7: Biểu đồ loss trong quá trình fine-tuning mô hình WAVLM (ECAPA-TDNN) trên tập dữ liệu Zalo

Hình 4.8 là biểu đồ EER trong quá trình fine-tuning mô hình WavLM (ECAPA-TDNN) trên tập dữ liệu Zalo. Trong giai đoạn đầu, mô hình đạt EER tốt nhất là 3.52% ở epoch 37. Việc mở đóng băng đã cải thiện EER xuống còn 2.96%.

Bảng 4.3 cho ta kết quả tốt nhất trong quá trình fine-tuning mô hình nhận dạng người nói trên tập dữ liệu Zalo. Có thể thấy, việc thừa hưởng khả năng nhận dạng người nói từ mô hình tiền huấn luyện với lượng lớn dữ liệu trên ngôn ngữ tiếng Anh đã giúp giảm thiểu gánh nặng việc các mô hình phải học nhận dạng từ đầu với tập dữ liệu nhỏ. Từ đó, khi được đưa tập dữ liệu nhỏ tiếng Việt, các mô hình tiền huấn luyện chỉ cần học thêm các đặc trưng của người nói tiếng Việt nữa là có thể cải thiện hiệu suất. Kết



Hình 4.8: Biểu đồ EER trong quá trình fine-tuning mô hình WAVLM (ECAPA-TDNN) trên tập dữ liệu Zalo

quả EER tốt nhất mà RawNet3 đạt được là 4.13% và WavLM (ECAPA-TDNN) là 3.52%. Ở WavLM (ECAPA-TDNN), với việc học thêm các đặc trưng trong tiếng nói ngôn ngữ tiếng Việt của mô hình upstream, độ đo EER được cải thiện hơn nữa và đạt 2.96%, vượt trội hoàn toàn mô hình RawNet3.

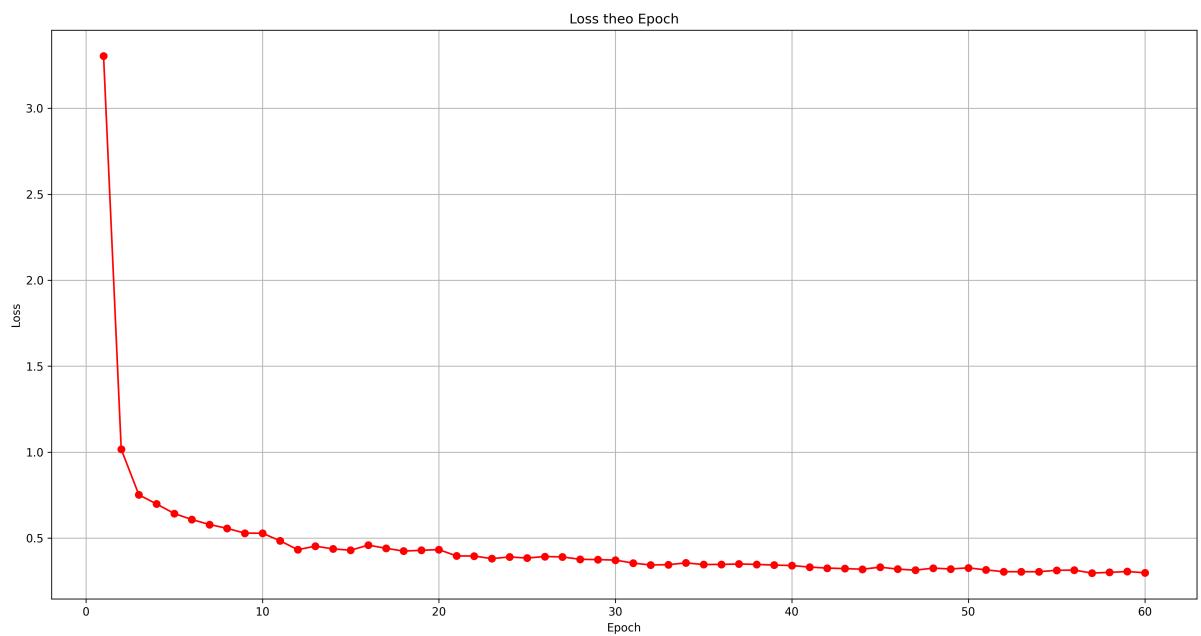
Mô hình	EER%	minDCF
RawNet3	4.13	0.268
WavLM (ECAPA-TDNN) giai đoạn 1	3.52	0.2056
WavLM (ECAPA-TDNN) giai đoạn 2	2.96	0.1779

Bảng 4.3: Kết quả tốt nhất trong quá trình fine-tuning mô hình nhận dạng người nói trên bộ dữ liệu Zalo

4.1.4 Huấn luyện mô hình bằng cách kết hợp thêm mô hình nhỏ và huấn luyện trên tiếng Việt

Gắn thêm một lớp kết nối đầy đủ

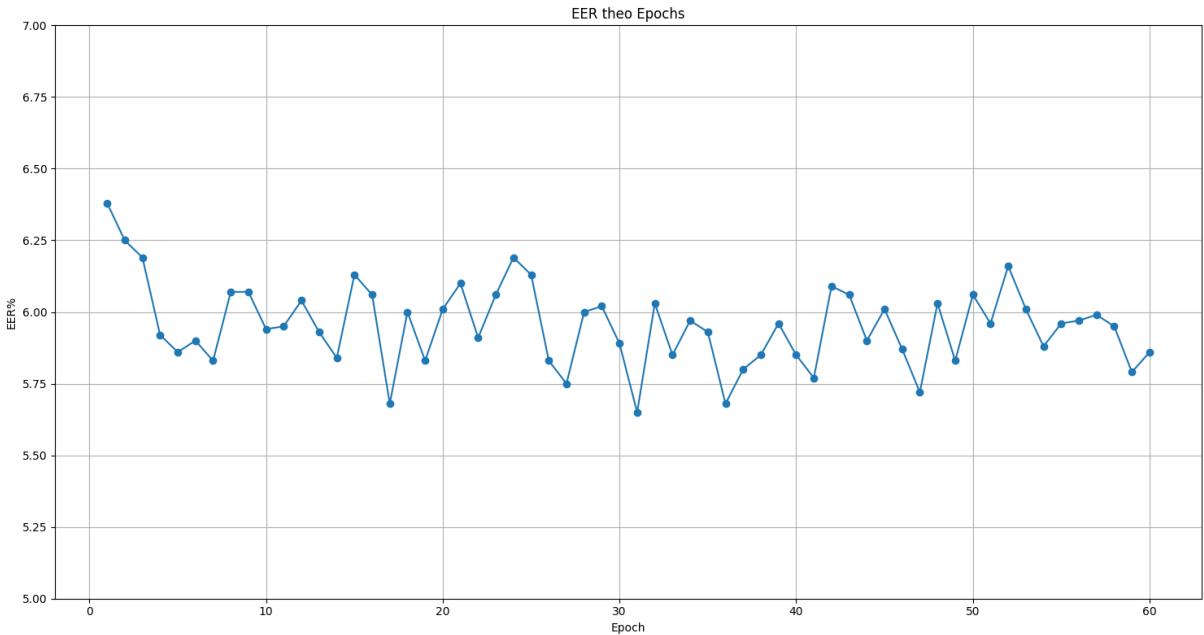
Hình 4.9 là biểu đồ loss trong quá trình huấn luyện trên một lớp kết nối đầy đủ của mô hình RawNet3. Với khả năng học nhanh của RawNet3 đã được thể hiện từ kết quả của các hướng tiếp cận trước đó, thì việc gắn thêm vào mô hình một lớp kết nối đầy đủ vẫn không ảnh hưởng đến tốc độ học của mô hình. Biểu đồ cho thấy cũng chỉ trong 10 epoch đầu tiên, hàm loss đã hội tụ rất nhanh và từ epoch 10 trở đi giá trị hàm loss giảm từng chút một.



Hình 4.9: Biểu đồ loss trong quá trình huấn luyện trên một lớp kết nối đầy đủ của mô hình RawNet3

Hình 4.10 là biểu đồ EER trong quá trình huấn luyện trên một lớp kết nối đầy đủ của mô hình RawNet3. Có thể thấy rằng kết quả này khá tương đồng với hình 4.6 và cũng chứng minh rằng việc gắn một component trên RawNet3 đã hoạt động hiệu quả khi so sánh kết quả với bảng 4.2 nhưng

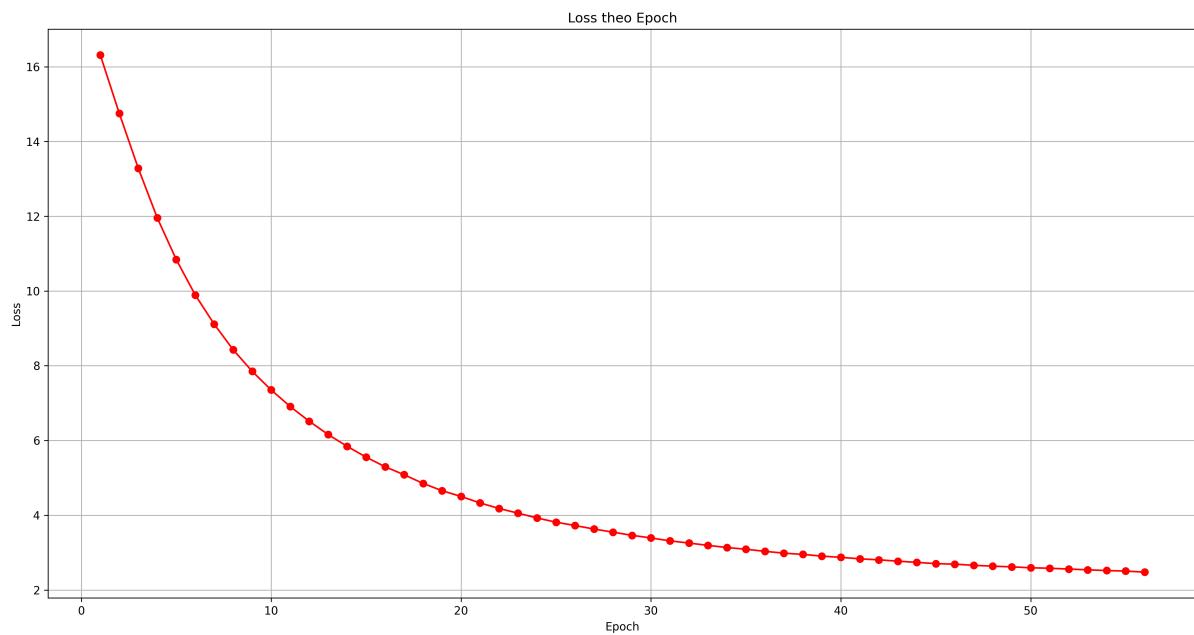
việc huấn luyện nhiều epoch cũng không có ý nghĩa nhiều khi mô hình đã tốt sau khi huấn luyện với lượng epoch nhỏ.



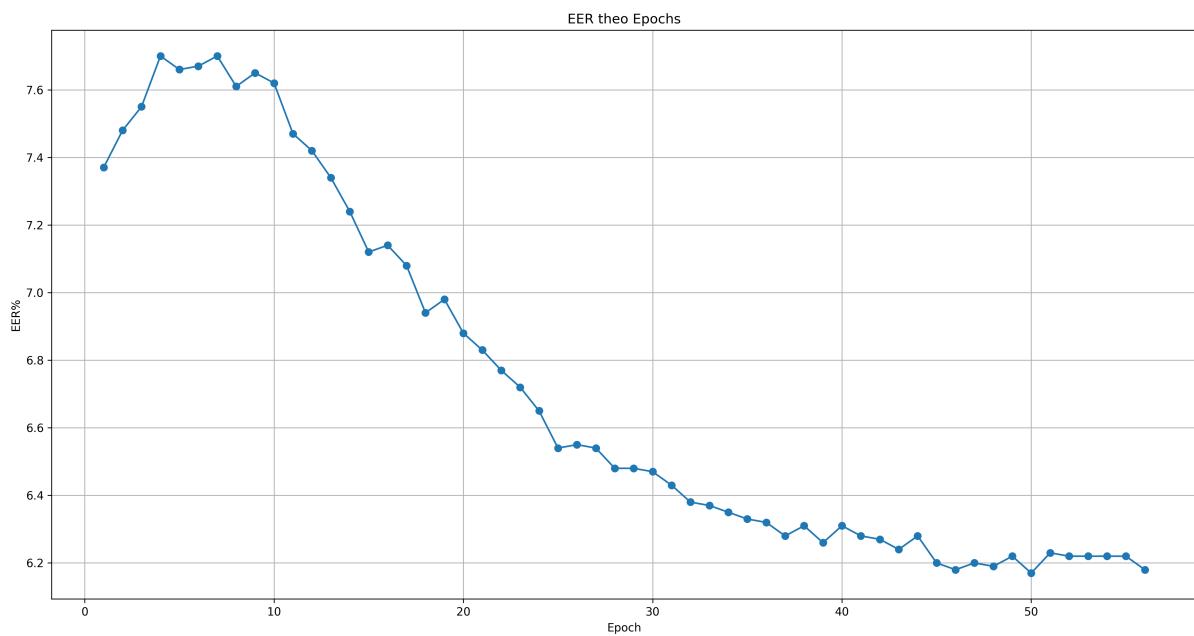
Hình 4.10: Biểu đồ EER trong quá trình huấn luyện trên một lớp kết nối đầy đủ của mô hình RawNet3

Hình 4.11 là biểu đồ loss trong quá trình huấn luyện trên một lớp kết nối đầy đủ của mô hình WavLM (ECAPA-TDNN). Lúc này, hàm loss của mô hình hội tụ khá nhanh trong 10 epoch đầu tiên, tiếp tục hội tụ nhưng chậm hơn trong 10 epoch kế tiếp. Với 10 epoch nữa, hàm loss cũng giảm đôi chút và sau đó chỉ giảm nhẹ.

Hình 4.12 là biểu đồ EER trong quá trình huấn luyện trên một lớp kết nối đầy đủ của mô hình WavLM (ECAPA-TDNN). Có thể thấy rằng kết quả này khá tương đồng với hình 4.8, đó là EER trở nên tệ hơn trong vài epoch đầu nhưng lại dần cải thiện hơn trong các epoch sau đó và độ EER tốt nhất là 6.17% ở epoch 50.



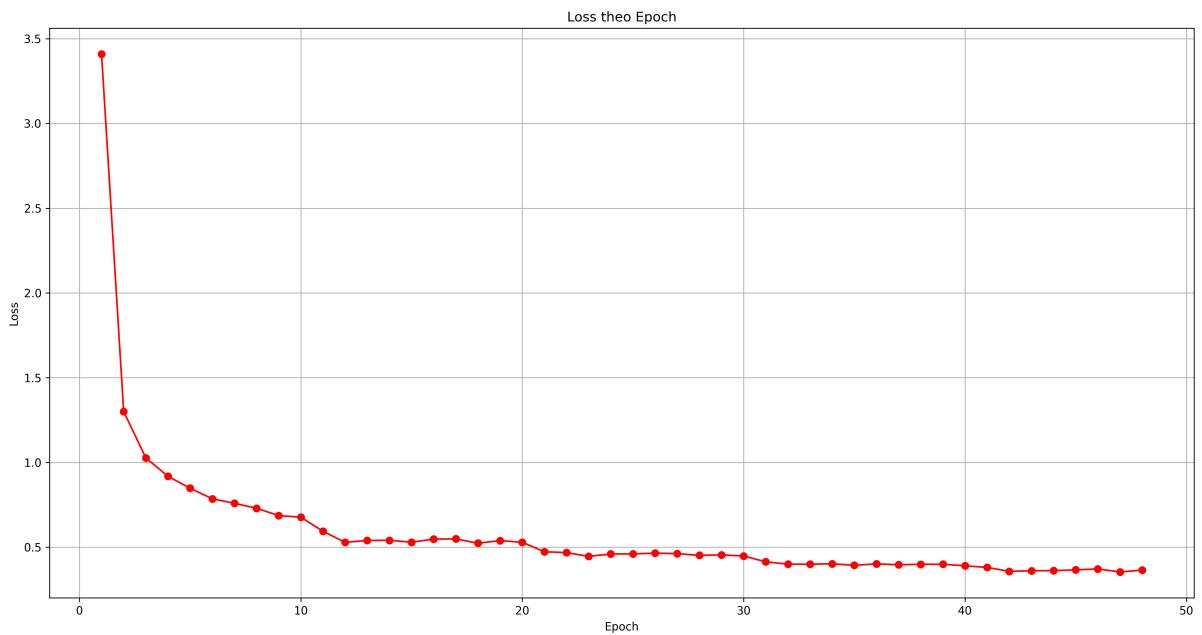
Hình 4.11: Biểu đồ loss trong quá trình huấn luyện trên một lớp kết nối đầy đủ của mô hình WavLM (ECAPA-TDNN)



Hình 4.12: Biểu đồ EER trong quá trình huấn luyện trên một lớp kết nối đầy đủ của mô hình WavLM (ECAPA-TDNN)

Gắn thêm ba lớp kết nối đầy đủ

Hình 4.13 là biểu đồ loss trong quá trình huấn luyện trên ba lớp kết nối đầy đủ của mô hình RawNet3. Tốc độ học của mô hình với việc gắn thêm ba lớp kết nối đầy đủ cũng tương đồng như việc gắn thêm một lớp như hình 4.9 và mô hình đã hội tụ một cách nhanh chóng trong vòng 10 epoch đầu. Sau đó, cứ qua mỗi epoch thì hàm loss cũng chỉ giảm từng chút một.

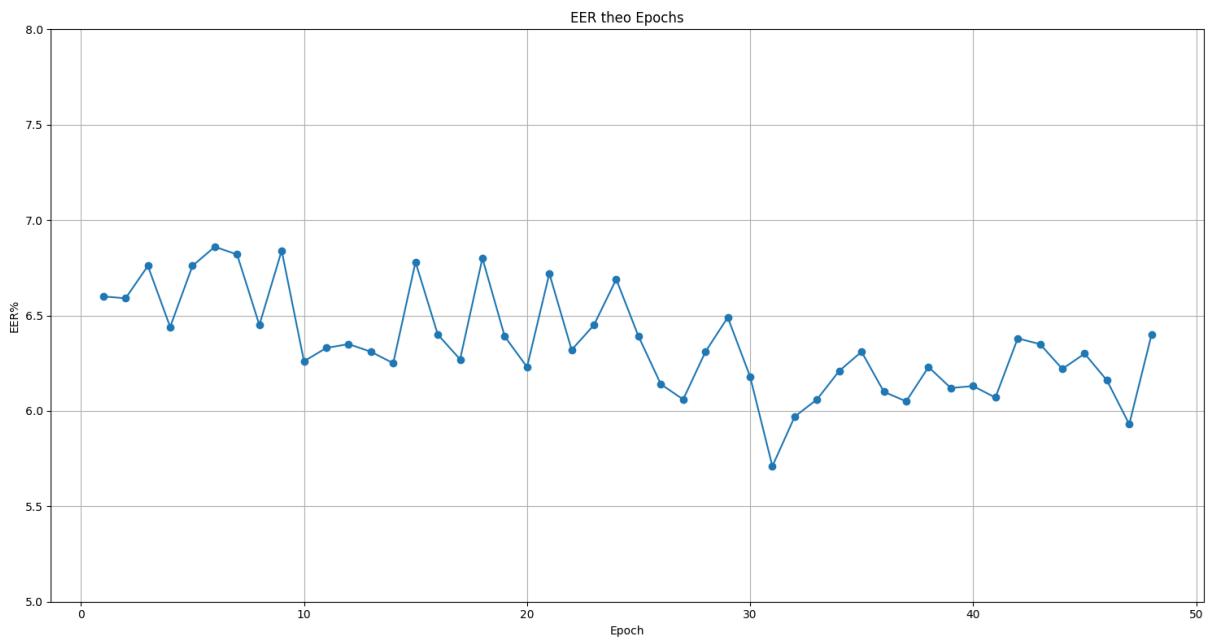


Hình 4.13: Biểu đồ loss trong quá trình huấn luyện trên ba lớp kết nối đầy đủ của mô hình RawNet3

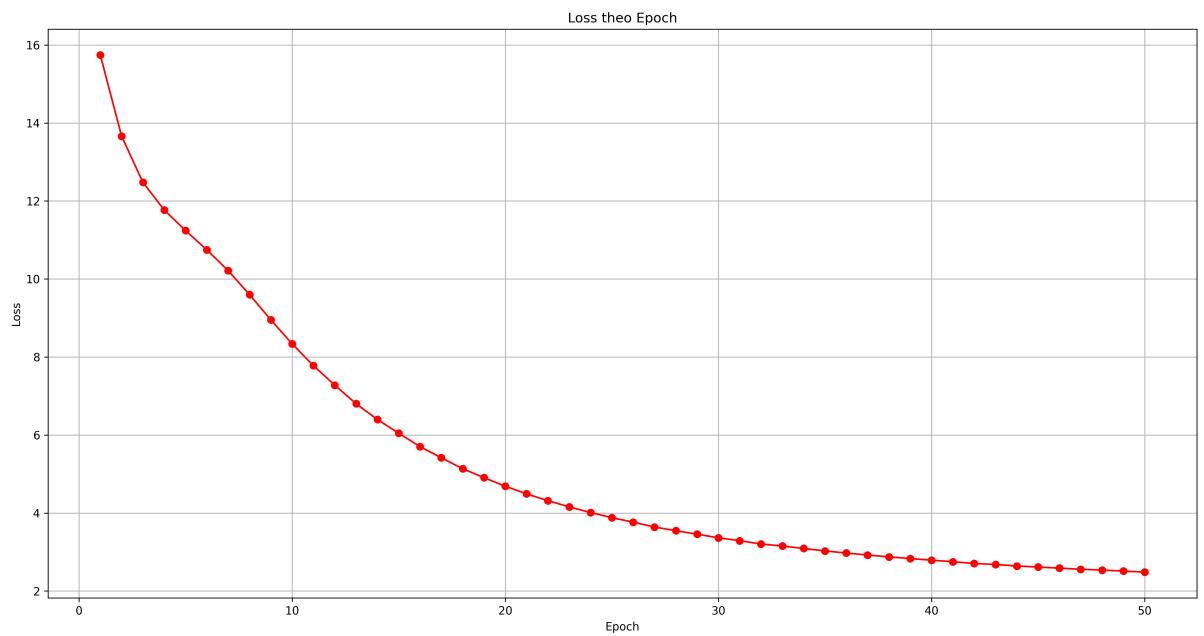
Hình 4.14 là biểu đồ EER trong quá trình huấn luyện trên ba lớp kết nối đầy đủ của mô hình RawNet3. Dựa vào biểu đồ, ta thấy kết quả khá tương đồng với kết quả thử nghiệm tại hình 4.10.

Hình 4.15 là biểu đồ loss trong quá trình huấn luyện trên ba lớp kết nối đầy đủ của mô hình WavLM (ECAPA-TDNN). Cũng tương tự như việc gắn thêm một lớp, hàm loss của mô hình hội tụ trong khoảng 20 epoch, với 10 epoch đầu tiên hội tụ nhanh hơn 10 epoch kế. Mỗi epoch sau đó, hàm loss chỉ giảm dần từng khoảng nhỏ.

Hình 4.16 là biểu đồ EER trong quá trình huấn luyện trên ba lớp kết

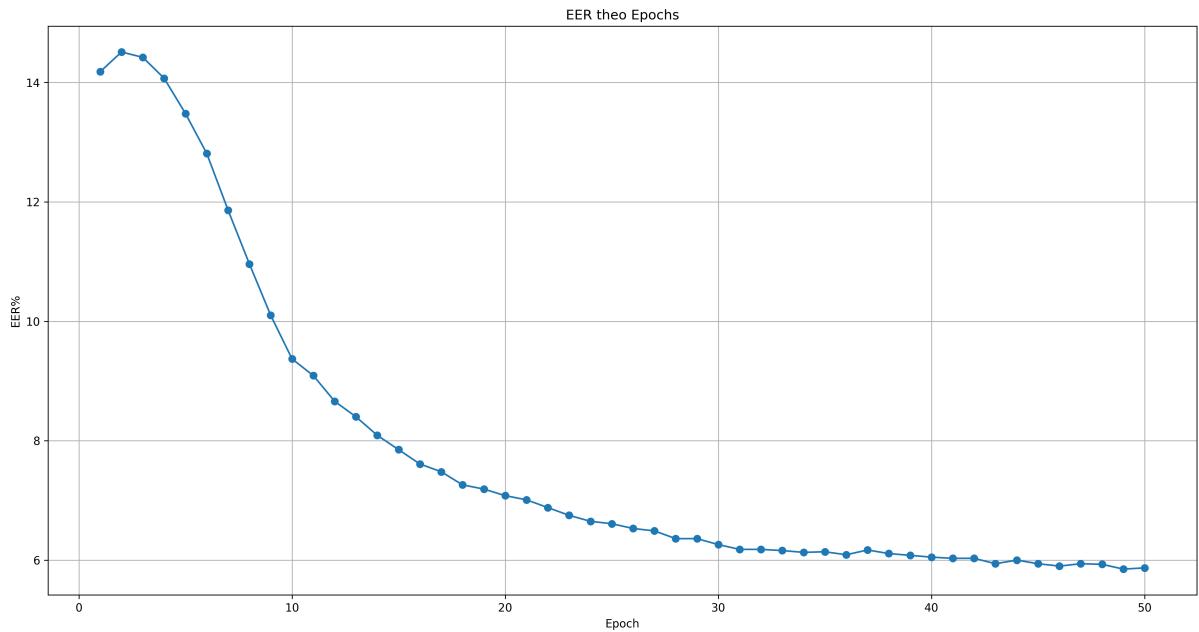


Hình 4.14: Biểu đồ EER trong quá trình huấn luyện trên ba lớp kết nối đầy đủ của mô hình RawNet3



Hình 4.15: Biểu đồ loss trong quá trình huấn luyện trên ba lớp kết nối đầy đủ của mô hình WavLM (ECAPA-TDNN)

nối đầy đủ của mô hình WavLM (ECAPA-TDNN). Việc gắn ba lớp kết nối đầy đủ thay vì một lớp kết nối đầy đủ như thử nghiệm tại hình 4.12 đã giúp mô hình cải thiện nhiều hơn khi so sánh với nhau.



Hình 4.16: Biểu đồ EER trong quá trình huấn luyện trên ba lớp kết nối đầy đủ của mô hình WavLM (ECAPA-TDNN)

Tổng kết về việc gắn mô hình nhỏ

Bảng 4.4 cho ta kết quả tốt nhất trong quá trình huấn luyện mô hình nhỏ trên mô hình nhận dạng người nói với tập dữ liệu tiếng Việt nhỏ. Với mô hình RawNet3, việc gắn thêm một lớp kết nối đầy đủ cho kết quả tốt hơn so với ba lớp, cụ thể kết quả độ đo EER của việc gắn thêm một lớp là 5.65% và gắn thêm ba lớp là 5.71%. Điều này có thể lý giải là do bản thân RawNet3 khi tiến hành ánh xạ embedding từ ngôn ngữ tiếng Anh sang ngôn ngữ tiếng Việt rất tốt. Còn với mô hình WavLM (ECAPA-TDNN), ba lớp kết nối đầy đủ mang lại hiệu quả tốt hơn so với chỉ một lớp. Điều này được giải thích là do việc sử dụng ba lớp kết nối đầy đủ giúp mô hình học được nhiều thông tin về người nói hơn do lượng thông tin từ đầu vào

của ECAPA-TDNN là quá lớn khi được trích xuất bởi WavLM.

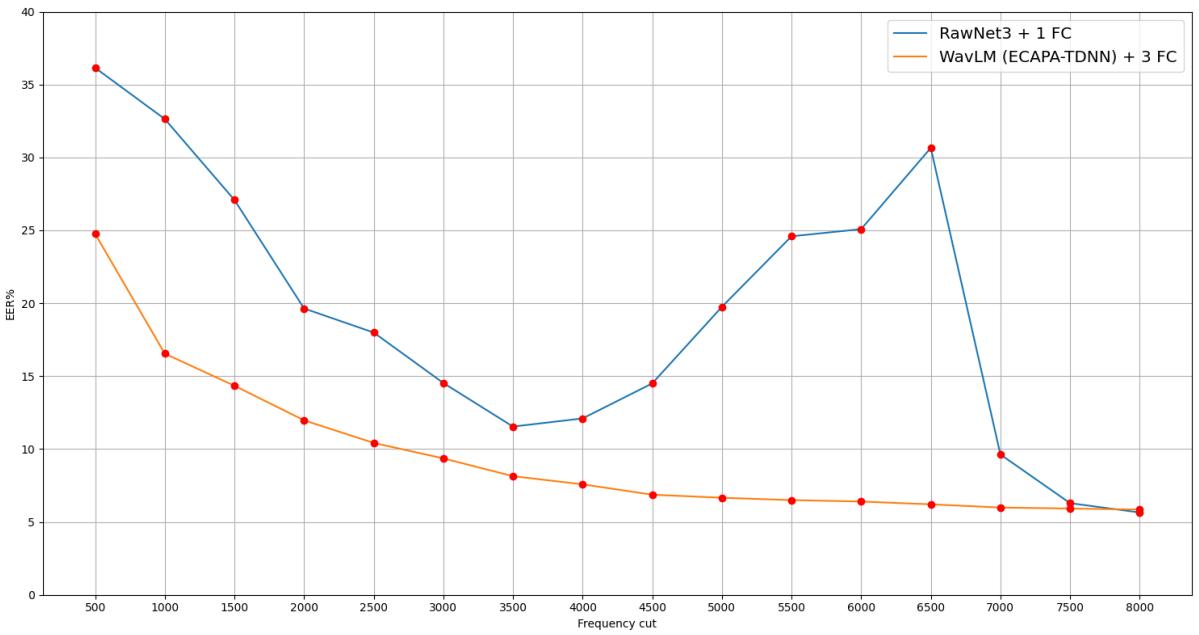
Mô hình	EER%	minDCF
RawNet3 + 1 FC	5.65	0.2864
RawNet3 + 3 FC	5.71	0.3103
WavLM (ECAPA-TDNN) + 1 FC	6.17	0.3475
WavLM (ECAPA-TDNN) + 3 FC	5.85	0.3109

Bảng 4.4: Kết quả tốt nhất trong quá trình huấn luyện mô hình nhỏ trên mô hình nhận dạng người nói với bộ dữ liệu Zalo

4.1.5 Khảo sát ảnh hưởng của các miền tần số và nhiễu đến mô hình nhận dạng người nói

Hình 4.17 là biểu đồ khảo sát EER của mô hình RawNet3 + 1 FC và mô hình WavLM (ECAPA-TDNN) + 3 FC trên miền tần số sau khi sử dụng bộ lọc thông thấp (LPF) lần lượt từ 500 đến 7500 Hz và khoảng cách bước nhảy là 500 Hz. Nhóm chọn hướng tiếp cận kết hợp thêm mô hình nhỏ để khảo sát ảnh hưởng của miền tần số, mục đích là để theo dõi khả năng ánh xạ embedding đại diện cho người nói từ embedding cho người nói tiếng Anh sang embedding cho người nói tiếng Việt của hai mô hình sẽ như thế nào.

Như đã minh họa về LPF ở hình 2.2, LPF với giá trị càng cao sẽ giúp giữ lại thêm những tín hiệu ở tần số cao nhằm mục đích mang nhiều thông tin hơn và mô hình WavLM (ECAPA-TDNN) + 3 FC đã thể hiện được điều đó. Cụ thể, với đoạn âm thanh chỉ còn những giá trị tần số dưới 500 Hz, mô hình này đạt kết quả khá tệ với EER xấp xỉ 25%. Nhưng mỗi khi tăng giá trị của LPF thêm 500 Hz thì mô hình cải thiện được EER tốt hơn trước đó, vì lúc này mô hình dần trích xuất được thêm các thông tin trong giọng nói ở tần số cao hơn trước đó 500 Hz, dần đến khả năng nhận dạng người nói dần tốt hơn. Cho đến cuối cùng, LPF với giá trị tần số đạt 8000 Hz, tức là toàn bộ tín hiệu giọng nói, EER đạt giá trị tốt nhất là 5.85%.



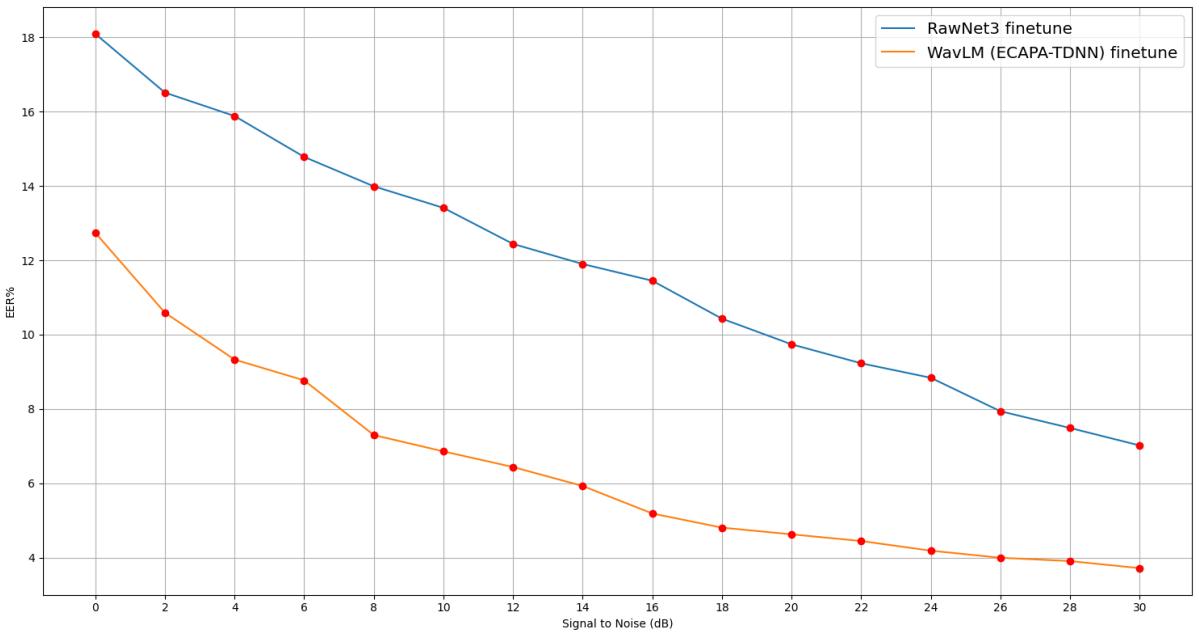
Hình 4.17: Biểu đồ khảo sát EER của mô hình RawNet3 + 1 FC và mô hình WavLM (ECAPA-TDNN) + 3 FC trên miền tần số sau khi sử dụng LPF

Khác với WavLM (ECAPA-TDNN) + 3 FC, biểu đồ của mô hình RawNet3 + 1 FC lại thể hiện kết quả có chút khác biệt so với lý thuyết đã được đề cập. Cho đến LPF ở giá trị 3500 Hz, mô hình này vẫn thể hiện rõ quá trình cải thiện EER vì lúc này thông tin trong đoạn âm thanh là nhiều hơn so với giá trị LPF thấp hơn mỗi 500 Hz trước đó. Nhưng bắt đầu từ giá trị LPF ở mức 4000 Hz, hiệu quả của mô hình bắt đầu có xu hướng tệ dần khi giá trị LPF tăng thêm mỗi 500Hz sau đó. Độ đo EER tệ nhất (cục bộ) lên đến hơn 30% ở giá trị LPF 6500 Hz. Tuy nhiên với LPF có giá trị 7000 Hz thì mô hình đạt hiệu suất tốt hơn đột biến, hơn cả khi LPF với giá trị 3500 Hz và sau đó EER được cải thiện như lý thuyết đã nêu. Với toàn bộ tín hiệu giọng nói, EER tốt nhất là 5.65%.

Để làm rõ chỗ bất thường ở mô hình RawNet3 + 1 FC, nhóm đã tiến hành áp dụng bộ lọc cấm dải trên miền tần số từ 4000 Hz đến 6500 Hz và đã thu được kết quả **EER = 9.06%**, **minDCF = 0.4377**. Có thể thấy kết quả độ đo EER này xấp xỉ với EER khi LPF có giá trị 7000 Hz. Điều này

chứng tỏ rằng mô hình RawNet3 + 1 FC không học tốt những đặc trưng người nói trong miền tần số này, thậm chí chúng còn làm tệ embedding của người nói khi trích xuất.

Hình 4.18 là biểu đồ khảo sát EER của mô hình RawNet3 fine-tuning và mô hình WavLM (ECAPA-TDNN) fine-tuning khi thêm nhiễu trắng với SNR từ 0 dB đến 30 dB. Với lượng nhiễu lớn nhất, tức là SNR có giá trị là 0 dB, mô hình RawNet3 có kết quả EER khá tệ là hơn 18%, còn EER của mô hình WavLM (ECAPA-TDNN) là gần 13%. Khi lượng nhiễu giảm dần, tức là giá trị SNR tăng dần, hiệu quả của hai mô hình đều được cải thiện. Điều này là hiển nhiên vì lượng nhiễu càng ít thì thông tin trong giọng nói của người nói càng được thể hiện rõ hơn, dẫn đến hiệu suất sẽ cao hơn. Nhìn chung, ở bất kỳ lượng nhiễu nào, mô hình WavLM (ECAPA-TDNN) đều có kết quả tốt hơn mô hình RawNet3, lý do là vì trong kiến trúc của WavLM có sử dụng framework khử nhiễu (đã được trình bày ở mô hình WavLM trong chương 3), còn kiến trúc RawNet3 thì không có sử dụng bất kỳ kỹ thuật khử nhiễu nào.



Hình 4.18: Biểu đồ khảo sát EER của mô hình RawNet3 fine-tuning và mô hình WavLM (ECAPA-TDNN) fine-tuning khi thêm nhiễu trắng với SNR từ 0 dB đến 30 dB

4.2 Thảo luận

4.2.1 Tổng kết các phương pháp đã làm

Trong ngữ cảnh với một tập dữ liệu nhỏ (dưới 1000 người) thì làm sao để có thể xây dựng một mô hình nhận dạng người nói tốt, nhóm đã tiến hành chạy các thử nghiệm trên các mô hình SOTA mà nhóm đã tổng hợp được từ quá trình nghiên cứu và thu được kết quả tổng hợp các quá trình tại bảng 4.5.

Trong **thử nghiệm 1**: nhóm đã tiến hành huấn luyện hai mô hình RawNet3 và WavLM (ECAPA-TDNN) từ đầu với bộ dữ liệu nhỏ tiếng Việt. Với cách tiếp cận truyền thống này, mô hình RawNet3 chiếm ưu thế với EER = 7.96%. Trong khi đó, mô hình WavLM (ECAPA-TDNN) đạt kết quả EER = 22.4% và tệ hơn mô hình RawNet3 rất nhiều. Có thể lý

Thử nghiệm 1: Huấn luyện từ đầu bằng tập dữ liệu tiếng Việt

Mô hình	EER%	minDCF
RawNet3	7.96	0.4826
WavLM (ECAPA-TDNN)	22.4	0.9403

Thử nghiệm 2: Sử dụng mô hình tiền huấn luyện

Mô hình	EER%	minDCF
RawNet3	9.74	0.4431
WavLM (ECAPA-TDNN)	11.26	0.5410

Thử nghiệm 3: Huấn luyện mô hình bằng fine-tuning

Mô hình	EER%	minDCF
RawNet3	4.13	0.268
WavLM (ECAPA-TDNN) giai đoạn 1	3.52	0.2056
WavLM (ECAPA-TDNN) giai đoạn 2	2.96	0.1779

Thử nghiệm 4: Kết hợp thêm mô hình nhỏ và huấn luyện

Mô hình	EER%	minDCF
RawNet3 + 1 FC	5.65	0.2864
RawNet3 + 3 FC	5.71	0.3103
WavLM (ECAPA-TDNN) + 1 FC	6.17	0.3475
WavLM (ECAPA-TDNN) + 3 FC	5.85	0.3109

Bảng 4.5: Kết quả tổng hợp

giải bằng các lý do sau:

- Mô hình WavLM (upstream) là mô hình phụ thuộc vào ngôn ngữ, mô hình WavLM chỉ trích xuất thông tin đặc trưng tốt khi sử dụng ngôn ngữ giống như khi tiền huấn luyện là tiếng Anh. Do đó, việc sử dụng ngôn ngữ khác ngoài tiếng Anh có thể dẫn tới kết quả tệ hơn.
- Số lượng đặc trưng sau khi rút trích từ WavLM là quá nhiều, lên tới [149 x 1024] (bảng 3.1) với dữ liệu mẫu là một đoạn âm thanh 3 giây nên sẽ làm khó khăn cho mô hình ECAPA-TDNN (downstream) để loại bỏ những đặc trưng tiếng nói và học những đặc trưng người nói. Để giúp mô hình downstream thực hiện tốt điều này thì phải đi kèm với tập dữ liệu lớn thay vì nhỏ như thực nghiệm.

Do đó, thử nghiệm 1 về cơ bản đã chứng minh hoạt động của mô hình RawNet3 khi được huấn luyện từ đầu trên bộ dữ liệu nhỏ nhưng đồng thời cũng chứng minh mô hình WavLM (ECAPA-TDNN) **không** phù hợp để huấn luyện từ đầu chỉ với bộ dữ liệu nhỏ tiếng Việt.

Trong **thử nghiệm 2**: nhóm đã áp dụng mô hình đã được tiền huấn luyện trên dữ liệu tiếng Anh vào bộ dữ liệu nhỏ tiếng Việt. Mô hình RawNet3 đạt kết quả tốt hơn với EER = 9.74% trong khi mô hình WavLM (ECAPA-TDNN) đạt kết quả EER = 11.26%. Nguyên nhân WavLM (ECAPA-TDNN) tệ hơn một chút so với RawNet3 như đã đề cập ở thí nghiệm 1 là do WavLM là mô hình phụ thuộc vào ngôn ngữ nên sẽ có kết quả không tốt khi sử dụng ngôn ngữ ngoài tiếng Anh. Do đó, trong thử nghiệm 2, nhóm đề xuất sử dụng mô hình end-to-end là RawNet3 sẽ có kết quả vượt trội hơn.

Trong **thử nghiệm 3**: nhóm đã huấn luyện hai mô hình bằng fine-tuning. Cụ thể, nhóm đã tiến hành nạp trọng số của mô hình đã được tiền huấn luyện từ thử nghiệm 2 và tiếp tục huấn luyện trên bộ dữ liệu nhỏ tiếng Việt.

- Đối với mô hình RawNet3, việc fine-tuning đã thể hiện ý nghĩa khi EER giảm từ trước khi fine-tuning tại thử nghiệm 2 là 9.74% xuống

còn 4.13%. Mô hình RawNet3 đã học rất tốt trong vài epoch đầu và việc finetuning thêm nhiều epoch là không có ý nghĩa với RawNet3 do mô hình này đã đủ những kiến thức về đặc trưng người nói từ mô hình tiền huấn luyện và việc fine-tuning sẽ giúp mô hình thích nghi hơn với những đặc trưng về người nói tiếng Việt.

- Đối với mô hình WavLM, nhóm đã chia việc fine-tuning thành hai giai đoạn:
 - Giai đoạn 1: Nhóm tiến hành fine-tuning trên mô hình ECAPA-TDNN đạt kết quả EER = 3.52%.
 - Giai đoạn 2: Nhóm tiến hành mở đóng bằng mô hình WavLM và fine-tuning trên toàn bộ mô hình và đạt kết quả EER = 2.96%. Việc mở đóng bằng mô hình upstream đã giúp cải thiện EER hơn 0.56%.

Dựa vào kết quả, mô hình WavLM (ECAPA-TDNN) đã thể hiện thế mạnh khi fine-tuning so với RawNet3 trong cả hai giai đoạn vì fine-tuning sẽ giúp mô hình thích nghi với tập dữ liệu nhỏ ngoài miền tiền huấn luyện, từ đó cải thiện khả năng phân biệt người nói và giảm thiểu sai sót. Việc mở đóng bằng mô hình upstream cũng đã cải thiện khả năng đặc trưng hóa ngôn ngữ tiếng Việt và qua các thử nghiệm này, nhóm đã xác minh được hiệu quả của việc sử dụng fine-tuning để cải thiện hiệu suất của các mô hình trong nhận dạng người nói.

Trong **thử nghiệm 4**, nhóm tiếp tục nghiên cứu về khả năng ánh xạ embedding đại diện từ ngôn ngữ tiếng Anh sang tiếng Việt của hai mô hình. Để làm điều này, nhóm đã đóng toàn bộ mô hình của hai kiến trúc RawNet3 và WavLM (ECAPA-TDNN), tiếp đến nhóm gắn thêm một và ba lớp mạng kết nối đầy đủ (FC), sau đó tiến hành huấn luyện trên dữ liệu để đánh giá hiệu suất, cụ thể:

- Đối với mô hình RawNet3, việc gắn thêm một mô hình nhỏ đã hoạt động tốt với kết quả EER là 5.65% cho một lớp FC và 5.71% cho ba lớp FC.
- Đối với mô hình WavLM (ECAPA-TDNN), việc gắn thêm một mô hình nhỏ cũng đã hoạt động tốt với kết quả EER là 6.17% cho một lớp FC và 5.85% cho ba lớp FC.

Kết quả trên cho ta thấy rằng việc thêm các lớp FC đã giúp cải thiện khả năng ánh xạ embedding đại diện từ ngôn ngữ gốc sang ngôn ngữ đích. Việc kết hợp mô hình nhỏ tuy có kết quả không bằng với phương pháp fine-tuning nhưng phương pháp này mới và đã hoạt động dựa trên thực nghiệm. Từ đó cho thấy tiềm năng của phương pháp này trong việc giúp mô hình lớn thích nghi tốt với tập dữ liệu nhỏ và rất hữu ích trong thực tế, nhất là trong viễn cảnh không có một tập dữ liệu lớn để huấn luyện. Ngoài ra, thay vì phải huấn luyện lại toàn bộ mô hình, việc gắn thêm một mô hình nhỏ và chỉ cần huấn luyện trên phần mô hình này sẽ giúp tiết kiệm hơn rất nhiều về thời gian và công sức.

Ngoài ra, nhóm đã tiến hành thêm hai khảo sát gồm khảo sát ảnh hưởng của miền tần số và ảnh hưởng của nhiễu đến mô hình nhận dạng người nói, cụ thể:

- Đối với khảo sát về ảnh hưởng của miền tần số (hình 4.17), nhóm đã sử dụng bộ lọc thông thấp để lọc bỏ tín hiệu với giá trị tần số được cắt lần lượt từ 500 Hz đến 7500 Hz, bước nhảy 500 Hz. Mục đích của việc khảo sát này là để tìm hiểu khả năng nhận dạng người nói trên nhiều băng tần khác nhau cũng như khả năng ánh xạ embedding đại diện cho người nói của hai mô hình, cụ thể là từ embedding đại diện cho người nói tiếng Anh sang embedding đại diện cho người nói tiếng Việt. Đối với mô hình WavLM (ECAPA-TDNN), việc cắt miền tần số từ giá trị thấp đến cao cho thấy sự cải thiện EER rõ rệt khi qua mỗi lần tăng ngưỡng cắt tần số thêm 500 Hz. Điều này giúp mô hình WavLM (ECAPA-TDNN) dần thấy được các tín hiệu thông tin về

người nói từ tần số thấp đến tần số cao nên kết quả cũng dần được cải thiện theo. Còn đối với mô hình RawNet3, sự cải thiện EER chỉ thể hiện rõ cho tới ngưỡng cắt tần số ở 3500 Hz, bắt đầu bước nhảy kế tiếp thì hiệu suất của mô hình lại trở nên tệ hơn và đến ngưỡng 7000 Hz thì mô hình mới hoạt động tốt trở lại. Sau đó nhóm tiến hành sử dụng bộ lọc cấm dải để khảo sát sự bất thường này ở mô hình RawNet3 đạt kết quả **EER = 9.06%, minDCF = 0.4377**. Kết quả thu được đã phản ánh rằng mô hình RawNet3 không học tốt những đặc trưng người nói trong miền tần số này, thậm chí chúng còn làm tệ embedding của người nói khi trích xuất.

- Đối với khảo sát về ảnh hưởng của nhiễu (hình 4.18), nhóm sử dụng lượng nhiễu với giá trị SNR từ 0 dB đến 30 dB. Mục đích của thử nghiệm này là để so sánh khả năng chịu đựng nhiễu của hai mô hình trong nhận dạng người nói. Có thể thấy rằng ở bất kỳ giá trị SNR, mô hình WavLM (ECAPA-TDNN) luôn vượt trội hơn so với mô hình RawNet3. Lý do là vì với mô hình WavLM (ECAPA-TDNN), ở phần kiến trúc của WavLM có sử dụng framework khử nhiễu, cho nên tổng thể WavLM (ECAPA-TDNN) sẽ cho ra kết quả tốt. Ngược lại thì với kiến trúc RawNet3 không có sử dụng bất kỳ kỹ thuật khử nhiễu nào, nên hiệu quả sẽ kém hơn.

Tổng kết lại, với tập dữ liệu nhỏ hiện tại, để đạt được kết quả cao thì việc huấn luyện bằng fine-tuning mô hình lớn là hiệu quả nhất. Nhưng cách nhanh nhất để các mô hình lớn có thể thích nghi với tập dữ liệu nhỏ là cách kết hợp thêm mô hình. Kết quả thu được ở cách này sẽ thấp hơn một chút so với cách fine-tuning nhưng thời gian cần thiết để huấn luyện là nhanh hơn rất nhiều. Nhìn chung, cả hai cách này đều đã cải thiện kết quả tốt hơn nhiều so với việc huấn luyện lại từ đầu và việc dùng mô hình tiền huấn luyện.

4.2.2 So sánh với các mô hình nhận dạng người nói trên tiếng Việt

Nghiên cứu thứ nhất [34]			
Mô hình	Hướng tiếp cận	Tập dữ liệu	EER%
ResNetSE34V2	Only Vietnamese	VietCeleb	5
ResNetSE34V2	Train on VoxCeleb1	VietCeleb	11
ResNetSE34V2	Fine-tuning	VietCeleb	3

Nghiên cứu thứ hai [32]			
Mô hình	Hướng tiếp cận	Tập dữ liệu	EER%
ResNet-34	Pretrained Model	VietSV	14.954
ResNet-34	Fine-tuning	VietSV	3.115
ResNet-34	Pretrained Model	Common Voice	11.468
ResNet-34	Fine-tuning	Common Voice	4.789

Nghiên cứu của nhóm			
Mô hình	Hướng tiếp cận	Tập dữ liệu	EER%
RawNet3	Pretrain Model	Zalo	9.74
RawNet3	Fine-tuning	Zalo	4.13
WavLM (ECAPA-TDNN)	Pretrained Model	Zalo	11.26
WavLM (ECAPA-TDNN)	Fine-tuning stage 1	Zalo	3.52
WavLM (ECAPA-TDNN)	Fine-tuning stage 2	Zalo	2.96

Bảng 4.6: Kết quả so sánh giữa các mô hình nhận dạng người nói trên tiếng Việt

Bảng 4.6 so sánh các mô hình nhận dạng người nói trên tiếng Việt. Việc so sánh này chỉ mang tính chất tương đối do hai nghiên cứu liên quan không công bố mã nguồn và tập dữ liệu để nhóm có thể thu thập và thử nghiệm mô hình đề xuất vào các nghiên cứu liên quan. Nhìn chung, cả hai nghiên cứu đều có hướng tiếp cận tốt nhất là dùng mô hình được tiền huấn luyện bằng dữ liệu ngôn ngữ tiếng Anh để đánh giá trên tập dữ liệu tiếng Việt, sau đó áp dụng kỹ thuật fine-tuning để huấn luyện tiếp mô hình tiền huấn luyện với dữ liệu tiếng Việt. Có thể thấy rằng việc fine-tuning được áp dụng khá rộng rãi vì lợi ích về kiến thức đã học của các mô hình tiền

huấn luyện cũng như khả năng cải thiện tốt kết quả trên tập dữ liệu mới, cụ thể là tập dữ liệu tiếng Việt.

Chương 5

Kết luận và hướng phát triển

Tóm lại, để giải quyết bài toán nhận dạng người nói cho bộ dữ liệu nhỏ (dưới 1000 người), nhóm đã nghiên cứu và tổng hợp các mô hình SOTA cho bài toán nhận dạng người nói là RawNet3 và WavLM (ECAPA-TDNN) đồng thời đã thử nghiệm nhiều phương pháp từ những phương pháp huấn luyện mô hình từ đầu, sử dụng mô hình tiền huấn luyện và huấn luyện dựa trên transfer learning.

Trong tất cả các phương pháp nhóm đã thử nghiệm, ngoài phương pháp truyền thống như huấn luyện từ đầu và sử dụng mô hình tiền huấn luyện, các phương pháp dựa trên transfer learning đều cải thiện EER so với các phương pháp truyền thống. Trong đó, phương pháp fine-tuning cho kết quả tốt nhất với mô hình WavLM (ECAPA-TDNN) cho hai giai đoạn. Ngoài ra, kết quả thực nghiệm của phương pháp kết hợp mô hình cũng đã cho thấy tiềm năng trong việc thích nghi với tập dữ liệu nhỏ, nhất là trong viễn cảnh không có một tập dữ liệu lớn để huấn luyện trong thực tế và cũng như tiết kiệm tài nguyên và chi phí huấn luyện.

Nhóm cũng đã tiến hành khảo sát ảnh hưởng của miền tần số và nhiễu của các mô hình nhận dạng người nói. Thực tế cho thấy đối với mô hình WavLM (ECAPA-TDNN) khi cắt miền tần số từ giá trị thấp đến cao cho thấy một quá trình cải thiện kết quả rõ rệt. Ngược lại, mô hình RawNet3 nhận dạng đặc trưng tại miền tần số từ 4000 Hz - 6500 Hz khá tệ và việc bỏ hoàn toàn miền tần số này lại thu được kết quả tốt hơn khá nhiều.

Do đó cần thận trọng trong việc downsampling hoặc upsampling dữ liệu có ảnh hưởng đến miền tần số này đối với RawNet3. Đối với ảnh hưởng của nhiều, mô hình WavLM (ECAPA-TDNN) cho kết quả vượt trội hơn so với mô hình RawNet3 ở bất kỳ giá trị SNR nào. Vì thế, khi cần nhận dạng người nói trong môi trường có lượng nhiễu nặng thì sử dụng WavLM (ECAPA-TDNN) sẽ có lợi thế hơn.

Trong tương lai, nhóm dự định sẽ nghiên cứu thêm về hướng gắn thêm mô hình nhỏ vì hướng này khá tiềm năng để phát triển. Ngoài ra, nếu có điều kiện sử dụng các tập dữ liệu nhỏ khác trên các ngôn ngữ ít tài nguyên, nhóm cũng sẽ thử nghiệm để tăng tính nhất quán của kết quả thu được.

Tài liệu tham khảo

Tiếng Anh

- [1] Baevski, Alexei et al. *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. 2020. arXiv: 2006.11477 [cs.CL].
- [2] Caron, Mathilde et al. *Emerging Properties in Self-Supervised Vision Transformers*. 2021. arXiv: 2104.14294 [cs.CV].
- [3] Chen, Sanyuan et al. “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing”. In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022), pp. 1505–1518. DOI: 10.1109/jstsp.2022.3188113.
- [4] Chi, Zewen et al. *XLM-E: Cross-lingual Language Model Pre-training via ELECTRA*. 2022. arXiv: 2106.16138 [cs.CL].
- [5] Chung, Joon Son, Nagrani, Arsha, and Zisserman, Andrew. “Vox-Celeb2: Deep Speaker Recognition”. In: *Interspeech 2018*. ISCA, 2018. DOI: 10.21437/interspeech.2018-1929.
- [6] Chung, Joon Son et al. “In Defence of Metric Learning for Speaker Recognition”. In: *Interspeech 2020*. ISCA, 2020. DOI: 10.21437/interspeech.2020-1064.
- [7] Davis, S. and Mermelstein, P. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29.4 (1981), pp. 746–754. DOI: 10.1109/tassp.1981.1163510.

Processing 28.4 (1980), pp. 357–366. DOI: 10.1109/TASSP.1980.1163420.

- [8] Dehak, Najim et al. “Front-End Factor Analysis for Speaker Verification”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011), pp. 788–798. DOI: 10.1109/TASL.2010.2064307.
- [9] Deng, Jiankang et al. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.10 (2022), pp. 5962–5979. DOI: 10.1109/tpami.2021.3087709.
- [10] Desplanques, Brecht, Thienpondt, Jenthe, and Demuynck, Kris. “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification”. In: *Interspeech 2020*. ISCA, 2020. DOI: 10.21437/interspeech.2020-2650.
- [11] Devlin, Jacob et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].
- [12] Gao, Shang-Hua et al. “Res2Net: A New Multi-Scale Backbone Architecture”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.2 (2021), pp. 652–662. DOI: 10.1109/tpami.2019.2938758.
- [13] Garcia-Romero, Daniel et al. “Jhu-HLT COE System for the Voxsrc Speaker Recognition Challenge”. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 7559–7563. DOI: 10.1109/ICASSP40776.2020.9053209.
- [14] He, Kaiming et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].
- [15] Heo, Hee Soo et al. *Clova Baseline System for the VoxCeleb Speaker Recognition Challenge 2020*. 2020. arXiv: 2009.14153 [eess.AS].

- [16] Hsu, Wei-Ning et al. *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units*. 2021. arXiv: 2106.07447 [cs.CL].
- [17] Hu, Jie et al. *Squeeze-and-Excitation Networks*. 2019. arXiv: 1709.01507 [cs.CV].
- [18] Ioffe, Sergey and Szegedy, Christian. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. arXiv: 1502.03167 [cs.LG].
- [19] Jung, Jee weon et al. *Improved RawNet with Feature Map Scaling for Text-independent Speaker Verification using Raw Waveforms*. 2020. arXiv: 2004.00526 [eess.AS].
- [20] Jung, Jee weon et al. *Pushing the limits of raw waveform speaker recognition*. 2022. arXiv: 2203.08488 [eess.AS].
- [21] Kim, Ju ho et al. *RawNeXt: Speaker verification system for variable-duration utterances with deep layer aggregation and extended dynamic scaling policies*. 2022. arXiv: 2112.07935 [eess.AS].
- [22] Nagrani, Arsha, Chung, Joon Son, and Zisserman, Andrew. “Vox-Celeb: A Large-Scale Speaker Identification Dataset”. In: *Interspeech 2017*. ISCA, 2017. DOI: 10.21437/interspeech.2017-950.
- [23] Okabe, Koji, Koshinaka, Takafumi, and Shinoda, Koichi. “Attentive Statistics Pooling for Deep Speaker Embedding”. In: *Interspeech 2018*. ISCA, 2018. DOI: 10.21437/interspeech.2018-993.
- [24] Pariente, Manuel et al. *Filterbank design for end-to-end speech separation*. 2020. arXiv: 1910.10400 [cs.SD].
- [25] Ravanelli, Mirco and Bengio, Yoshua. *Speaker Recognition from Raw Waveform with SincNet*. 2019. arXiv: 1808.00158 [eess.AS].
- [26] Ravanelli, Mirco et al. *SpeechBrain: A General-Purpose Speech Toolkit*. 2021. arXiv: 2106.04624 [eess.AS].

- [27] Reynolds, D.A. and Rose, Richard. “Robust text-independent speaker identification using Gaussian Mixture speaker models”. In: *Speech and Audio Processing, IEEE Transactions on* 3 (Feb. 1995), pp. 72–83. DOI: 10.1109/89.365379.
- [28] Reynolds, Douglas A., Quatieri, Thomas F., and Dunn, Robert B. “Speaker Verification Using Adapted Gaussian Mixture Models”. In: *Digital Signal Processing* 10.1 (2000), pp. 19–41. ISSN: 1051-2004. DOI: <https://doi.org/10.1006/dspr.1999.0361>.
- [29] Snell, Jake, Swersky, Kevin, and Zemel, Richard S. *Prototypical Networks for Few-shot Learning*. 2017. arXiv: 1703.05175 [cs.LG].
- [30] Snyder, David et al. “Speaker Recognition for Multi-speaker Conversations Using X-vectors”. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 5796–5800. DOI: 10.1109/ICASSP.2019.8683760.
- [31] Snyder, David et al. “X-Vectors: Robust DNN Embeddings for Speaker Recognition”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 5329–5333. DOI: 10.1109/ICASSP.2018.8461375.
- [32] Thanh, Dat Vi, Viet, Thanh Pham, and Thu, Trang Nguyen Thi. “Deep Speaker Verification Model for Low-Resource Languages and Vietnamese Dataset”. In: *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*. Shanghai, China: Association for Computational Linguistics, Nov. 2021, pp. 442–451. URL: <https://aclanthology.org/2021.paclic-1.47>.
- [33] Thienpondt, Jenthe, Desplanques, Brecht, and Demuynck, Kris. “Cross-Lingual Speaker Verification with Domain-Balanced Hard Prototype Mining and Language-Dependent Score Normalization”. In: *Interspeech 2020*. ISCA, 2020. DOI: 10.21437/interspeech.2020-2662.

- [34] Tran, Cao Truong, Nguyen, Dinh Tan, and Hoang, Ho Tan. “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](https://doi.org/10.1109/KSE53942.2021.9648808).
- [35] Ulyanov, Dmitry, Vedaldi, Andrea, and Lempitsky, Victor. *Instance Normalization: The Missing Ingredient for Fast Stylization*. 2017. arXiv: [1607.08022 \[cs.CV\]](https://arxiv.org/abs/1607.08022).
- [36] Variani, Ehsan 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](https://doi.org/10.1109/ICASSP.2014.6854363).
- [37] Vinyals, Oriol et al. *Matching Networks for One Shot Learning*. 2017. arXiv: [1606.04080 \[cs.LG\]](https://arxiv.org/abs/1606.04080).
- [38] Yang, Zhilin et al. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. 2020. arXiv: [1906.08237 \[cs.CL\]](https://arxiv.org/abs/1906.08237).