|  |  |  |  |
| --- | --- | --- | --- |
| LOGO DHCNTT -hinh.jpg | ĐẠI HỌC QUỐC GIA TP. HCM  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN** | Ngày nhận hồ sơ |  |
| *(Do CQ quản lý ghi)* | |

**THUYẾT MINH**

ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ CẤP SINH VIÊN 2025

# THÔNG TIN CHUNG

## A1. Tên đề tài

* **Tiếng Việt:** XÂY DỰNG MÔ HÌNH NHẬN DIỆN TIẾNG NÓI TỰ ĐỘNG VÀ NHẬN DIỆN CẢM XÚC TỪ GIỌNG NÓI CHO TIẾNG VIỆT
* **Tiếng Anh:** BUILDING AUTOMATIC SPEECH RECOGNITION AND SPEECH EMOTION RECOGNITION MODELS FOR VIETNAMESE

## A2. Loại hình nghiên cứu

(Tham khảo tiêu chuẩn đề tài đối với từng loại hình NC, chọn 01 trong 03 loại hình)

* 🗹Nghiên cứu cơ bản
* ◻Nghiên cứu ứng dụng
* ◻Nghiên cứu triển khai

## A3. Thời gian thực hiện

**06** tháng (kể từ khi được duyệt).

## A3. Tổng kinh phí

*(Lưu ý tính nhất quán giữa mục này và mục B8. Tổng hợp kinh phí đề nghị cấp)*

Tổng kinh phí: **06** triệu đồng,gồm

* Kinh phí từ Trường Đại học Công nghệ Thông tin: **06** triệu đồng

## A4. Chủ nhiệm

## A5. Thành viên đề tài

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | **Khoa** |
|  |  |  | Khoa học Máy tính |
|  |  |  | Khoa học Máy tính |

## B. MÔ TẢ NGHIÊN CỨU

## B1. Giới thiệu về đề tài

## B1.1. Công nghệ giọng nói và trí tuệ nhân tạo

Công nghệ giọng nói kết hợp với trí tuệ nhân tạo (AI) là lĩnh vực nghiên cứu và ứng dụng nhằm giúp máy tính có khả năng nhận diện, hiểu và phản hồi con người theo cách thông minh và tự nhiên nhất. Trong những năm gần đây, công nghệ này đã phát triển nhanh chóng nhờ nhu cầu gia tăng trong các lĩnh vực như giao tiếp từ xa, giáo dục trực tuyến, chăm sóc sức khỏe từ xa và dịch vụ chăm sóc khách hàng. Sự xuất hiện của các trợ lý ảo tiêu biểu như Google Assistant, Amazon Alexa, Apple Siri, cùng với các nền tảng tổng đài chăm sóc khách hàng thông minh, cho thấy vai trò ngày càng quan trọng của công nghệ giọng nói và AI trong đời sống hiện đại.

Theo báo cáo của Grand View Research (2023) [1], thị trường nhận diện giọng nói toàn cầu đạt quy mô khoảng 20,25 tỷ USD và được dự báo tăng trưởng trung bình 14,6% mỗi năm trong giai đoạn 2024–2030, với quy mô ước tính chạm 53,57 tỷ USD vào năm 2030. Đặc biệt, trong các lĩnh vực thương mại và chăm sóc khách hàng, hơn 40% doanh nghiệp trên toàn cầu đã hoặc đang có kế hoạch triển khai công nghệ này nhằm nâng cao trải nghiệm người dùng. Tại Việt Nam, các sản phẩm như trợ lý ảo Kiki của Zalo AI, ViVi trên xe VinFast, tổng đài tự động của Viettel AI và nền tảng FPT.AI đang được ứng dụng ngày càng rộng rãi, minh chứng cho xu hướng phát triển mạnh mẽ của công nghệ giọng nói phủ khắp các lĩnh vực từ giải trí, giáo dục, y tế cho đến thương mại.

Không chỉ dừng lại ở khả năng nhận diện ngôn ngữ, các hệ thống hiện đại còn hướng đến việc phân tích trạng thái cảm xúc của người nói, giúp tăng cường tính tự nhiên và thông minh trong tương tác giữa con người và máy tính. Trong bối cảnh đó, Câu lạc bộ Xử lý ngôn ngữ và tiếng nói tiếng Việt (VLSP) 2025 [2] đã tổ chức cuộc thi Nhận diện tiếng nói tự động (Automatic Speech Recognition – ASR) và Nhận diện cảm xúc từ giọng nói (Speech Emotion Recognition – SER) nhằm thúc đẩy phong trào nghiên cứu trong nước đối với lĩnh vực đầy tiềm năng này. Nhóm đã tham gia cuộc thi và đạt kết quả đáng khích lệ (top 6), từ đó tiếp tục phát triển thành một đề tài nghiên cứu khoa học sinh viên.

## B1.2. Tình hình nghiên cứu

Trên thế giới:

Nghiên cứu của Ahlawat và cộng sự [3] trình bày tổng quan các kỹ thuật học sâu áp dụng cho ASR từ năm 2010 đến nay, nhấn mạnh tính hiệu quả của các mô hình end-to-end với các kiến trúc RNN, CNN và Transformer cho bài toán này. Cụ thể, công trình nghiên cứu phân tích ưu nhược điểm giữa các mô hình học sâu hiện đại và mô hình truyền thống, đánh giá về độ chính xác (Word Error Rate - WER), ảnh hưởng của dữ liệu, tính khái quát và khả năng mở rộng, đồng thời đề xuất xu hướng tương lai như ASR đa ngôn ngữ, mô hình pre-training và các hệ thống thích ứng cho môi trường thực tế.

Yang và các cộng sự [4] đã giới thiệu một phương pháp cá nhân hóa mô hình ASR mà không cần dữ liệu giọng nói thực. Thay vào đó, các tác giả sử dụng văn bản và mô hình tổng hợp giọng nói có thể điều khiển (Controllable TTS) để tạo dữ liệu huấn luyện cho từng người nói. Kết quả cho thấy phương pháp này giúp giảm đáng kể WER và đạt hiệu quả tương đương hoặc vượt trội so với mô hình được huấn luyện bằng giọng thật, mở ra hướng phát triển ASR cá nhân hóa, bảo mật và linh hoạt hơn trong các ứng dụng thực tế.

Wani và cộng sự [5] trình bày tổng quan toàn diện về lĩnh vực nhận dạng cảm xúc từ giọng nói (SER), bao gồm quy trình tiền xử lý, trích xuất đặc trưng và các kỹ thuật phân loại cảm xúc. Tác giả so sánh giữa các phương pháp học máy truyền thống (SVM, HMM, GMM) và mô hình học sâu như CNN, RNN, LSTM, đồng thời chỉ ra những thách thức về dữ liệu, độ nhiễu và biến thiên cảm xúc giữa các người nói. Nghiên cứu nhấn mạnh xu hướng hiện nay là mô hình học sâu end-to-end và hệ thống đa phương thức, hướng tới việc cải thiện độ chính xác và khả năng ứng dụng của SER trong các hệ thống tương tác người–máy.

Singh và Goel [6] tiến hành khảo sát có hệ thống 152 công trình trong giai đoạn 2000–2021 về nhận dạng cảm xúc từ giọng nói (SER). Nghiên cứu tổng hợp các cơ sở dữ liệu cảm xúc, đặc trưng âm thanh và các mô hình học sâu như CNN, RNN, LSTM; đồng thời so sánh với các phương pháp học máy truyền thống. Kết quả cho thấy mô hình học sâu cải thiện đáng kể độ chính xác nhận dạng cảm xúc, do đó nhóm tác giả đề xuất hướng phát triển tiếp theo là học tự giám sát và đa phương thức để nâng cao khả năng ứng dụng SER trong môi trường thực tế.

Nghiên cứu [7] tổng hợp các tiến bộ gần đây trong lĩnh vực Nhận dạng cảm xúc từ giọng nói (SER), đặc biệt tập trung vào ảnh hưởng của nhiễu âm thanh đối với hiệu suất mô hình. Nhóm tác giả áp dụng các mô hình học sâu như CNN, LSTM, Transformer cùng với các phương pháp tăng cường dữ liệu (data augmentation), thích ứng miền để cải thiện độ bền của mô hình trong môi trường thực. Ngoài ra, công trình còn nhấn mạnh xu hướng kết hợp ASR và SER nhằm khai thác đồng thời thông tin ngữ âm và cảm xúc, hướng tới các hệ thống nhận dạng cảm xúc chính xác và ổn định hơn trong điều kiện nhiễu.

Tại Việt Nam:

Theo tìm hiểu của nhóm, các nghiên cứu về nhận dạng giọng nói tự động (ASR) và nhận dạng cảm xúc trong giọng nói (SER) chủ yếu tập trung vào từng tác vụ riêng lẻ. Chẳng hạn, mô hình PhoWhisper được Le và cộng sự [8] phát triển nhằm nâng cao hiệu quả nhận dạng tiếng nói tiếng Việt, trong khi Tran và các cộng sự [9] trình bày phương pháp xây dựng bộ dữ liệu quy mô lớn và đánh giá các mô hình ASR hiện đại cho các ngôn ngữ ít tài nguyên như tiếng Việt. Bên cạnh đó, Phung và cộng sự [10] đã giới thiệu bộ dữ liệu giúp phát hiện cảm xúc trong giọng nói cho tiếng Việt cùng các mô hình học sâu phục vụ cho bài toán SER. Tuy nhiên, các hệ thống kết hợp giữa ASR và SER cho tiếng Việt hiện vẫn còn hạn chế và chưa có giải pháp toàn diện nào có thể đồng thời nhận dạng nội dung lời nói và phân tích trạng thái cảm xúc. Điều này mở ra cơ hội nghiên cứu và phát triển các mô hình tích hợp thông minh, có khả năng hỗ trợ trong các lĩnh vực như giao tiếp người – máy, tư vấn tâm lý, giáo dục, và chăm sóc khách hàng.

## B1.3. Các thách thức

1. Thiếu dữ liệu gán nhãn cảm xúc và mất cân bằng dữ liệu

* Đối với SER, dữ liệu được gán nhãn thường rất hạn chế (ít số lượng, ít đa dạng) so với dữ liệu ASR nói chung. Điều này làm cho việc huấn luyện mô hình SER trở nên khó khăn [11].
* Dữ liệu gán nhãn cảm xúc thường không cân bằng - một số cảm xúc (ví dụ “bình thường/neutral”) chiếm phần lớn trong khi các cảm xúc “hiếm” như sợ hãi, kinh ngạc thường ít xuất hiện [12], gây ra vấn đề học lệch khi huấn luyện mô hình.

1. Đặc trưng đang dạng của tiếng Việt

* Mỗi ngôn ngữ có cú pháp, đặc trưng ngữ âm, cách thể hiện cảm xúc riêng. Việc áp dụng trực tiếp các mô hình SER từ tiếng Anh vào tiếng Việt thường gặp khó khăn do không tương thích ngôn ngữ.
* Tiếng Việt với đặc trưng vùng miền, phương ngữ và thanh điệu cũng làm tăng độ khó cho cả ASR và SER.
* Ngoài ra, các yếu tố như ngữ điệu (prosody), tốc độ nói, khoảng ngừng, biến đổi âm sắc cũng làm ảnh hưởng đến cảm xúc khi xác định.

1. Thiếu đồng nhất trong định nghĩa cảm xúc/bộ nhãn cảm xúc khác nhau

* Qua tìm hiểu về nhiều bộ dữ liệu khác nhau, nhóm nhận thấy mỗi bộ dữ liệu dùng các tập nhãn khác nhau (ví dụ: 4 cảm xúc, 6 cảm xúc, 8 cảm xúc, hoặc định dạng dimensional - arousal/valence). Điều này gây khó khăn trong việc so sánh kết quả giữa các nghiên cứu và chuyển giao mô hình giữa các bộ dữ liệu.

## B1.4. Lý do chọn đề tài:

1. Nhóm đã tham gia cuộc thi VLSP 2025 [2] - một sân chơi học thuật uy tín và đầy thử thách trong lĩnh vực xử lý ngôn ngữ và tiếng nói. Từ những kết quả và kinh nghiệm đạt được, nhóm mong muốn phát triển sâu hơn thành đề tài nghiên cứu khoa học nhằm cải thiện hiệu suất cho các mô hình ASR, SER và mô hình tích hợp, đóng góp vào thành tựu chung của cộng đồng nghiên cứu trong nhận dạng và phân tích giọng nói cho tiếng Việt.
2. Đề tài giúp nhóm bước đầu làm quen với nghiên cứu khoa học, từ khâu tìm hiểu tổng quan đề tài, xử lý dữ liệu đến xây dựng và đánh giá mô hình. Đây là cơ hội để nhóm từng bước hoàn thiện nền tảng kiến thức và kỹ năng cần thiết cho nghiên cứu chuyên sâu trong tương lai, đặc biệt là khi thực hiện khóa luận tốt nghiệp.
3. Mặc dù đây là một bài toán phức tạp, đòi hỏi nhiều kiến thức liên quan đến xử lý tín hiệu âm thanh, học sâu và xử lý ngôn ngữ tự nhiên cùng với các thách thức về dữ liệu và mô hình, nhóm nhận thấy đây là một hướng đi hấp dẫn và có tiềm năng lớn trong cả nghiên cứu và ứng dụng. Việc tiếp cận và triển khai hệ thống kết hợp ASR và SER không chỉ giúp hiểu sâu hơn về các công nghệ AI hiện đại mà còn mở ra khả năng ứng dụng trong các lĩnh vực như trợ lý ảo, chăm sóc khách hàng, giáo dục hay hỗ trợ tâm lý.

## B2. Mục tiêu, nội dung, kế hoạch nghiên cứu

## B2.1. Mục tiêu

**Mục tiêu tổng quát:**  
 Đề tài hướng đến việc nghiên cứu và phát triển các mô hình nhận dạng giọng nói tự động và nhận diện cảm xúc trong giọng nói cho tiếng Việt, đồng thời khai thác khả năng kết hợp giữa hai mô hình này nhằm nâng cao hiệu suất tổng thể, giúp cho tương tác người-máy trở nên tự nhiên hơn. Nghiên cứu kỳ vọng xây dựng, huấn luyện và đánh giá các mô hình đạt tiêu chuẩn của một nghiên cứu khoa học sinh viên, qua đó đóng góp vào nỗ lực thu hẹp khoảng cách nghiên cứu giữa Việt Nam và thế giới trong lĩnh vực công nghệ giọng nói và trí tuệ nhân tạo.

**Mục tiêu cụ thể:**

1. **Khảo sát học thuật:** Tổng hợp và phân tích các công trình nghiên cứu quốc tế giai đoạn 2022–2025 về ASR, SER và mô hình đa nhiệm, đồng thời chỉ ra khoảng trống trong nghiên cứu tiếng Việt.
2. **Tiền xử lý dữ liệu:** Kế thừa bộ dữ liệu từ cuộc thi VLSP 2025 [2], nhóm tiến hành xử lý và chuẩn hóa bộ dữ liệu giọng nói tiếng Việt có gắn nhãn song song cho cả nội dung và cảm xúc, đáp ứng yêu cầu huấn luyện và đánh giá mô hình.
3. **Phát triển mô hình ASR:** Thiết kế và huấn luyện mô hình ASR hiện đại (transformer-based, pre-trained large models) tối ưu cho tiếng Việt, bảo đảm độ chính xác cao trong nhận dạng lời nói.
4. **Phát triển mô hình SER:** Xây dựng mô hình SER phân loại ba trạng thái cảm xúc là tích cực, tiêu cực và trung tính cho tiếng Việt.
5. **Tích hợp pipeline ASR + SER:** Đề xuất phương pháp tích hợp hai mô hình thành một pipeline thống nhất, có khả năng phân tích toàn diện cả nội dung và cảm xúc trong lời nói.
6. **Thực nghiệm và đánh giá:** Đánh giá các mô hình đã huấn luyện trên bộ dữ liệu kiểm thử, xác định những ưu điểm và hạn chế.

## B2.2. Nội dung và phương pháp nghiên cứu

### Nội dung 1: Khảo sát và phân tích tổng quan các phương pháp, thuật toán trên các tài liệu hiện có

**Nội dung:**  
 Nội dung này tập trung vào việc khảo sát, thu thập và phân tích các công trình nghiên cứu trong và ngoài nước liên quan đến Nhận dạng giọng nói và Nhận dạng cảm xúc từ giọng nói. Trọng tâm là đánh giá các xu hướng nghiên cứu, phân tích những mô hình tiêu biểu như Wav2Vec2 [13], HuBERT [14], Whisper [15], đồng thời chỉ ra những thách thức đặc thù khi áp dụng cho tiếng Việt.

**Phương pháp thực hiện:**

* Tìm kiếm và thu thập các tài liệu học thuật trên các nguồn uy tín như Google Scholar, IEEE Xplore, ACL Anthology, Springer.
* Phân loại và phân tích, so sánh giữa các hướng tiếp cận khác nhau, bao gồm:
  + Dựa trên đặc trưng âm học (MFCC, prosody).
  + Dựa trên học sâu (deep learning), đặc biệt là các mô hình tiền huấn luyện (transformer-based pre-trained speech models).
* Xây dựng bảng so sánh ưu điểm và nhược điểm của từng mô hình trong bối cảnh tiếng Việt, từ đó chỉ ra những thách thức và cơ hội.

**Kết quả dự kiến:**

* Báo cáo tổng quan các nghiên cứu về ASR và SER, từ đó xác định ưu, nhược điểm của từng phương pháp.
* Xác định được khoảng trống nghiên cứu: thiếu các bộ dữ liệu lớn và dữ liệu gán nhãn cảm xúc cho tiếng Việt, còn ít nghiên cứu tích hợp ASR và SER trong cùng một pipeline.
* Từ đó chọn được phương pháp phù hợp cho đề tài và định hướng cho các nghiên cứu tiếp theo.

### Nội dung 2: Xử lý bộ dữ liệu giọng nói tiếng Việt có gán nhãn cảm xúc

**Nộidung:**  
 Nhóm kế thừa bộ dữ liệu từ cuộc thi VLSP 2025 Shared Task 2 [2] mà nhóm đã tham gia cho đề tài này. Bộ dữ liệu được sử dụng đồng thời cho cả hai nhiệm vụ: nhận diện giọng nói tự động và nhận diện cảm xúc từ giọng nói. Cụ thể, dữ liệu được gán nhãn đầy đủ, bao gồm bản phiên âm (transcript) phục vụ cho ASR và nhãn cảm xúc cho SER với ba nhãn: tích cực, tiêu cực và trung tính. Trong đó tích cực được hiểu là hạnh phúc, biết ơn, tự hào; tiêu cực được hiểu là sợ hãi, buồn bã, giận dữ, xấu hổ, thất vọng, ghê tởm. Trong quá trình tham gia, một vài tệp dữ liệu vẫn chưa sử dụng, nhóm sẽ sử dụng một phần để bổ sung vào tập dữ liệu huấn luyện và một phần cho tập kiểm tra nhằm đánh giá khách quan hiệu suất của các mô hình (vì hệ thống nộp kết quả “private test” của VLSP 2025 hiện tại đã đóng). Với dữ liệu có được, nhóm tiến hành phân tích và xử lý dữ liệu.

**Phương pháp thực hiện:**

* Kế thừa bộ dữ liệu từ VLSP 2025 Shared Task 2.
* Thực hiện tiền xử lý dữ liệu, gồm các bước:
  + Chuẩn hóa định dạng (.wav).
  + Lọc nhiễu, cắt đoạn, loại bỏ mẫu lỗi hoặc không rõ ràng.
  + Cân bằng dữ liệu giữa các nhãn cảm xúc để tránh hiện tượng lệch phân phối.
* Sử dụng công cụ Praat, Audacity và các thư viện Python như librosa, pyannote để xử lý và phân tích tín hiệu âm thanh.
* Áp dụng kỹ thuật tăng cường dữ liệu như time stretching, pitch shifting, thêm noise nhân tạo để mở rộng và đa dạng hóa tập dữ liệu.
* Chia dữ liệu phục vụ cho việc huấn luyện và đánh giá.

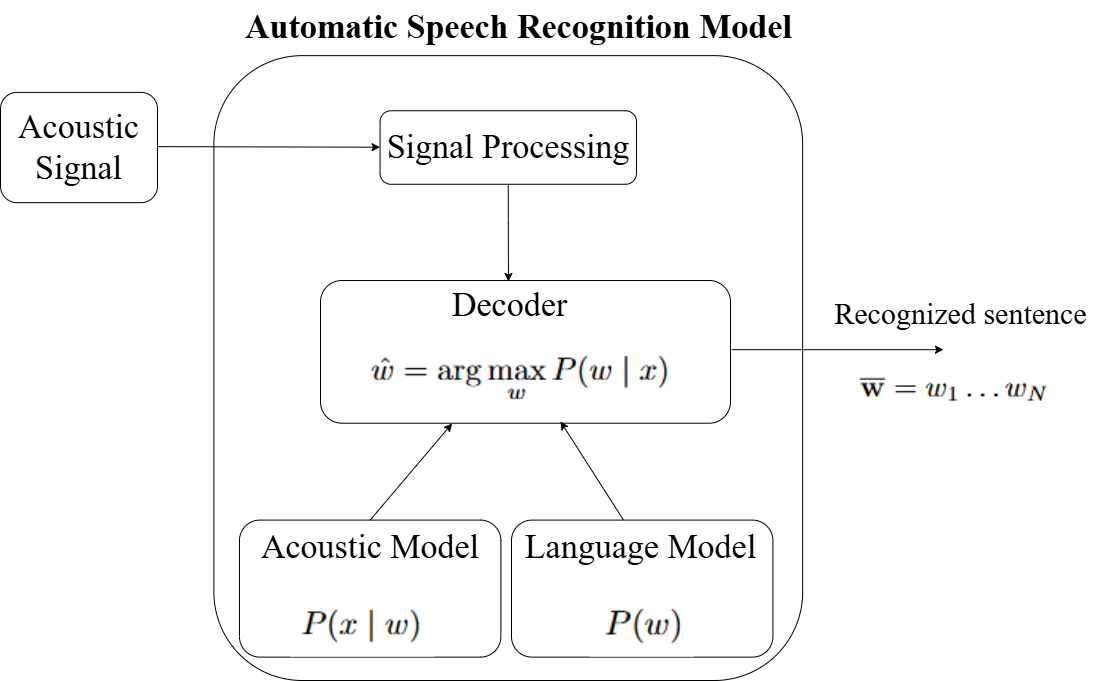
**Dự kiến kết quả:**

* Bộ dữ liệu giọng nói tiếng Việt có gán nhãn cảm xúc, quy mô khoảng 10.000 file audio, phân bố tương đối cân bằng giữa các nhãn cảm xúc.
* Bộ dữ liệu đạt chuẩn sử dụng cho cả hai tác vụ: nhận dạng lời nói (ASR) và nhận dạng cảm xúc (SER).
* Báo cáo thống kê chi tiết về tập dữ liệu, bao gồm: số lượng mẫu, tổng thời lượng, phân phối theo cảm xúc và chất lượng sau tiền xử lý.

### Nội dung 3: Tìm hiểu và triển khai mô hình

### Nội dung 3a: Tìm hiểu và triển khai mô hình ASR tiếng Việt

Nhận dạng tiếng nói tự động là quá trình chuyển đổi tín hiệu giọng nói thành văn bản bằng máy tính. Công nghệ này được xem là một trong những hướng nghiên cứu quan trọng nhất trong lĩnh vực xử lý ngôn ngữ tự nhiên và xử lý tín hiệu số, đóng vai trò nền tảng cho nhiều ứng dụng hiện đại.

****

*Hình 1: Các thành phần cơ bản của một hệ thống ASR.*

Như thể hiện ở Hình 1, một hệ thống ASR cơ bản gồm các giai đoạn chính: xử lý tín hiệu âm thanh, trích xuất và mô hình hóa đặc trưng âm học, và giải mã để nhận diện văn bản. Ở giai đoạn xử lý tín hiệu, tín hiệu giọng nói đầu vào được chuyển đổi thành dạng số và xử lý nhằm loại bỏ nhiễu, chuẩn hóa cường độ và tách khung tín hiệu để dễ dàng phân tích. Tiếp theo, trong giai đoạn mô hình hóa âm học, hệ thống học mối quan hệ giữa các đặc trưng âm thanh và các đơn vị ngôn ngữ như âm vị hoặc ký tự, thông qua các mô hình học sâu. Song song đó, mô hình ngôn ngữ được sử dụng để đánh giá tính hợp lý của chuỗi từ trong ngữ cảnh, giúp xác định câu có xác suất cao nhất. Cuối cùng, bộ giải mã kết hợp thông tin từ hai mô hình này để tạo ra chuỗi văn bản được nhận dạng, thể hiện nội dung lời nói của người nói.

**Nội dung:**

Nhóm lựa chọn mô hình Wav2Vec2-Base-Vietnamese-250h [16] (một biến thể của Wav2vec2 [13] được huấn luyện sẵn trên 250 giờ dữ liệu tiếng Việt) và tích hợp với bộ giải mã ngôn ngữ KenLM 4-gram [17] để cải thiện chất lượng giải mã và khả năng hiểu ngữ cảnh trong tác vụ ASR.

**Phương pháp thực hiện:**

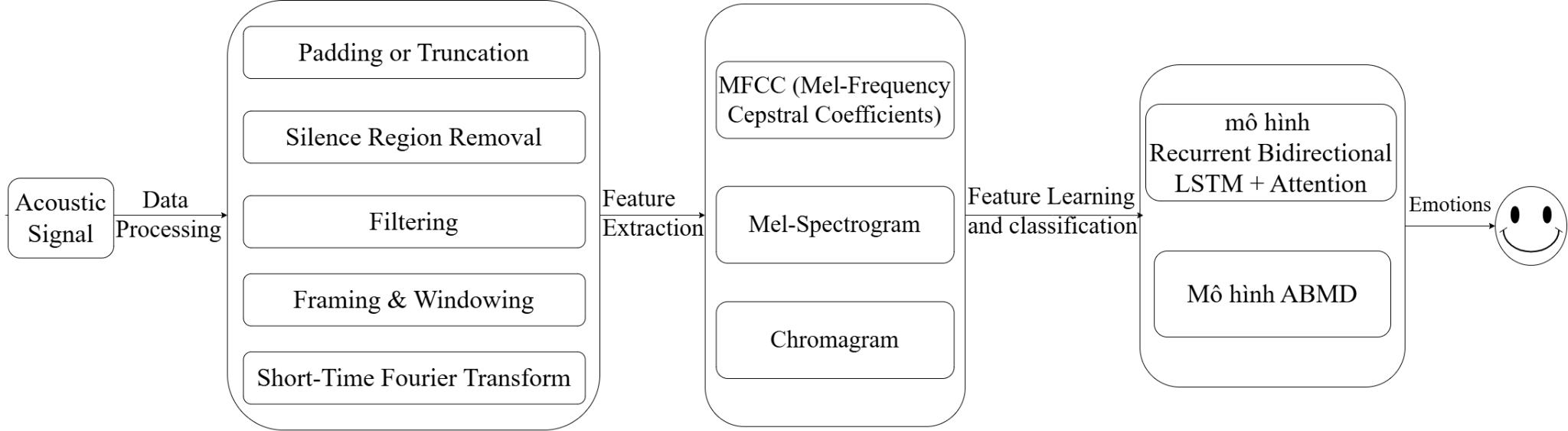
* Lựa chọn mô hình tiền huấn luyện Wav2vec2 phù hợp với tiếng Việt.
* Tiến hành fine-tuning mô hình trên tập dữ liệu huấn luyện đã chuẩn hóa.
* Triển khai huấn luyện và thử nghiệm trên GPU.
* Thử nghiệm các phương pháp giải mã:
  + Greedy search.
  + Beam search kết hợp Language Model là KenLM 4-gram.
* Đánh giá hiệu quả mô hình dựa trên chỉ số WER (Word Error Rate) và CER (Character Error Rate).

**Dự kiến kết quả:**

* Mô hình ASR tiếng Việt đạt độ chính xác cao trên tập dữ liệu kiểm thử.
* Các bản phiên âm đầu ra được chuẩn hóa dưới dạng tệp .tsv, phục vụ cho việc phân tích, tích hợp với mô hình SER và đánh giá kết quả.
* Báo cáo đánh giá chi tiết về hiệu năng mô hình theo từng phương pháp giải mã.

**Nội dung 3b: Xây dựng và huấn luyện mô hình SER tiếng Việt**

Nếu ASR giúp máy tính hiểu nội dung lời nói, thì SER có vai trò xác định trạng thái cảm xúc của người nói qua các tín hiệu âm thanh. Đây là quá trình phân tích tín hiệu âm thanh để xác định cảm xúc, từ đó mang lại khả năng tương tác người-máy tự nhiên và nhân văn hơn.

****

*Hình 2: Các thành phần cơ bản của một hệ thống SER.*

Như thể hiện ở Hình 2, một hệ thống SER cơ bản gồm các giai đoạn chính: xử lý dữ liệu âm thanh, trích xuất đặc trưng từ dữ liệu đã xử lý, huấn luyện mô hình phân loại cảm xúc. Giai đoạn xử lý dữ liệu bao gồm các bước như loại bỏ vùng im lặng, lọc nhiễu, chia khung và biến đổi Fourier ngắn hạn để chuẩn bị cho trích xuất đặc trưng. Tiếp theo, giai đoạn trích xuất đặc trưng tạo ra các đặc trưng đại diện cho cảm xúc. Sau đó, các đặc trưng này được đưa vào mô hình học sâu để học mối quan hệ thời gian và xác định trạng thái cảm xúc. Kết quả cuối cùng là cảm xúc được nhận dạng theo ba nhãn của bộ dữ liệu là tích cực, tiêu cực và trung tính.

**Nội dung:**

Xây dựng mô hình phân loại cảm xúc tiếng nói theo ba nhóm trạng thái cảm xúc cơ bản là: tích cực (hạnh phúc, biết ơn, tự hào), tiêu cực (sợ hãi, buồn bã, giận dữ, xấu hổ, thất vọng, ghê tởm) và trung tính. Cụ thể, nhóm lựa chọn sử dụng mạng nơ-ron tích chập (CNN) kết hợp đặc trưng Mel-frequency Cepstral Coefficients (MFCC) làm cơ sở. Đây là phương pháp kinh điển, đơn giản nhưng hiệu quả, thường được dùng trong nghiên cứu SER để làm mốc so sánh với các mô hình tiên tiến hơn.

**Phương pháp thực hiện:**

* Tiền xử lý dữ liệu giọng nói và chuẩn hóa nhãn cảm xúc.
* Trích xuất đặc trưng âm học từ tín hiệu audio:
  + MFCC.
  + Spectrogram/Mel-Spectrogram.
  + Prosodic features (cao độ, cường độ, trường độ).
* Xây dựng và thử nghiệm các mô hình phân loại:
  + Mô hình CNN kết hợp MFCC.
* Huấn luyện mô hình.
* Đánh giá hiệu quả bằng các chỉ số: Accuracy, F1-score, Confusion Matrix.

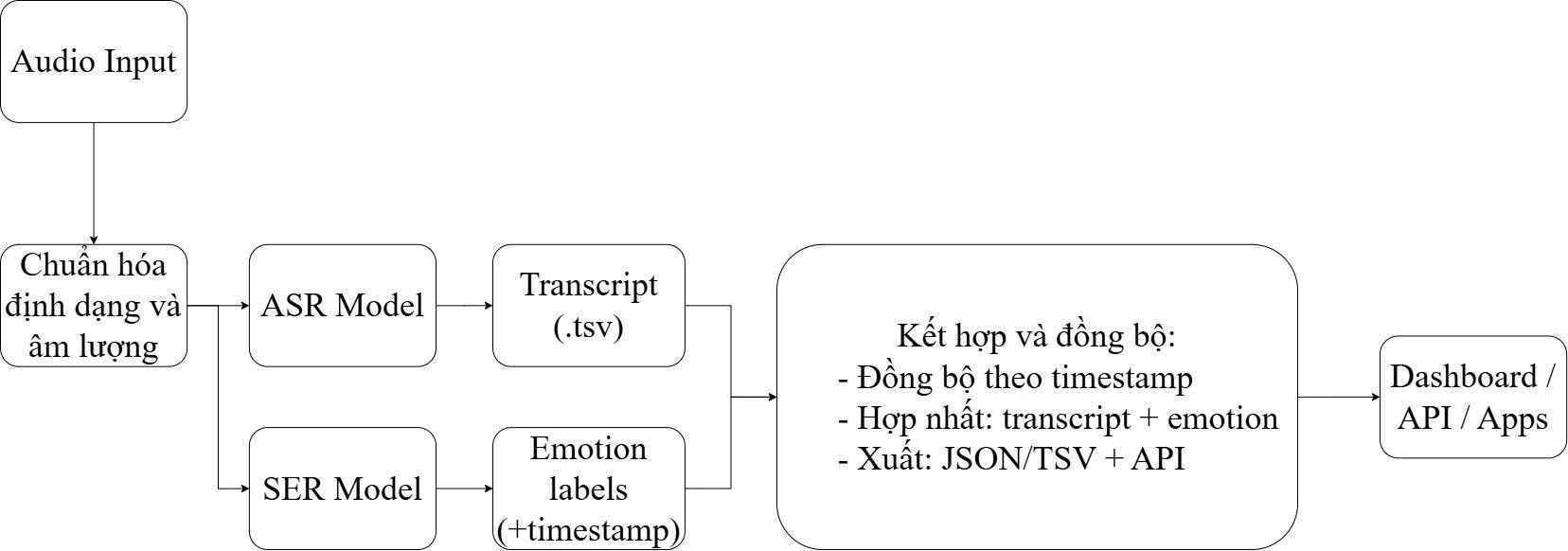
**Dự kiến kết quả:**

* Bảng so sánh hiệu suất của các mô hình.
* Báo cáo chi tiết độ chính xác theo từng lớp cảm xúc.
* Tệp kết quả phân loại cảm xúc được xuất ra kèm mốc thời gian, phục vụ tích hợp với pipeline ASR.
* Đề xuất được kiến trúc mô hình SER tối ưu nhất cho dữ liệu giọng nói tiếng Việt.

### Nội dung 3c: Tích hợp ASR và SER thành hệ thống thống nhất

**Nội dung:**

Nội dung này nhóm tập trung vào việc thiết kế và xây dựng một hệ thống hoàn chỉnh có khả năng vừa nhận dạng nội dung từ lời nói, vừa phân tích cảm xúc của người nói.



*Hình 3: Tổng quan hệ thống kết hợp ASR và SER cơ bản.*

Hình 3 minh họa một pipeline cơ bản tích hợp ASR và SER, vừa có khả năng nhận dạng nội dung lời nói, vừa cho biết trạng thái cảm xúc hiện tại của người nói lời nói đó. Dữ liệu giọng nói đầu vào khi đưa vào mô-đun ASR chuyển đổi thành văn bản; dữ liệu giọng nói đầu vào lại tiếp tục truyền vào mô hình SER để phân tích cảm xúc đồng bộ và xuất kết quả hợp nhất gồm bản phiên âm lời nói và nhãn cảm xúc tương ứng với từng mốc thời gian. Ngoài việc đánh giá độ chính xác của từng mô-đun, nghiên cứu còn đặc biệt chú trọng đến khả năng hoạt động trong môi trường gần thời gian thực. Hai tiêu chí chính được sử dụng để đo lường hiệu năng hệ thống gồm:

* **Latency (độ trễ):** Thời gian từ khi nhận tín hiệu âm thanh đầu vào đến khi trả về kết quả cuối cùng (ASR + SER).
* **Throughput (tốc độ xử lý):** Số giây dữ liệu âm thanh mà hệ thống có thể xử lý trong một giây thời gian thực.

Mục tiêu của hệ thống là đạt độ trễ trung bình thấp và tốc độ xử lý cao, đảm bảo khả năng ứng dụng trong môi trường thực tế.

**Phương pháp thực hiện:**

* Xây dựng pipeline xử lý:
  + Module 1: Nhận dạng giọng nói bằng ASR (sử dụng mô hình huấn luyện ở Nội dung 3a).
  + Module 2: Phân tích cảm xúc bằng SER (sử dụng mô hình huấn luyện ở Nội dung 3b).
  + Module 3: Đồng bộ và hợp nhất kết quả.
* Thiết kế API/Interface để hệ thống có thể tích hợp với các ứng dụng khác.
* Tối ưu hiệu năng: giảm độ trễ xử lý, đảm bảo khả năng hoạt động gần thời gian thực.
* Kiểm thử toàn diện trên các tập dữ liệu thử nghiệm, đánh giá độ chính xác và hiệu quả tích hợp.

**Dự kiến kết quả:**

* Tích hợp thành công hai mô hình ASR và SER.
* Báo cáo so sánh hiệu quả mô hình tích hợp với các giải pháp riêng lẻ tức ASR hoặc SER độc lập.

**Nội dung 4: Đánh giá, phân tích và báo báo kết quả**

**Nội dung:**

Nhóm tiến hành đánh giá các mô hình trên tập dữ liệu kiểm thử được tách ra từ bộ dữ liệu của cuộc thi VLSP 2025 [2]. Do bộ dữ liệu Ban Tổ chức cung cấp có quy mô lớn và đa dạng, nhóm chỉ sử dụng một phần thích hợp cho quá trình huấn luyện, vì vậy phần còn lại sẽ được dùng để đánh giá mô hình. Cách chia này nhằm đảm bảo tính khách quan, đồng thời phản ánh đúng năng lực tổng quát hóa của mô hình trên dữ liệu chưa từng xuất hiện trong quá trình huấn luyện. Việc sử dụng tập dữ liệu này giúp nhóm tập trung hơn vào việc phân tích, tối ưu mô hình và đánh giá hiệu quả hệ thống. Các tiêu chí đánh giá bao gồm độ chính xác nhận dạng lời nói, độ chính xác phân loại cảm xúc, và khả năng hoạt động tổng thể của hệ thống tích hợp trong tình huống thực tế.

**Phương pháp thực hiện:**

* Xây dựng kịch bản thí nghiệm:
  + Đánh giá riêng lẻ mô-đun ASR (theo chỉ số WER hay CER).
  + Đánh giá riêng lẻ mô-đun SER (theo Accuracy, F1-score, Confusion Matrix).
  + Đánh giá hệ thống tích hợp (độ trễ xử lý, tính ổn định, mức độ phù hợp ứng dụng thực tiễn).
* Thực hiện phân tích lỗi để chỉ ra các điểm hạn chế của từng mô-đun và toàn bộ hệ thống.
* Tổng hợp và trình bày kết quả bằng bảng thống kê, biểu đồ minh họa.

**Dự kiến kết quả:**

* Báo cáo định lượng:
  + ASR đạt WER ≤ 10% trên tập kiểm thử.
  + SER đạt độ chính xác trung bình ≥ 85%.
  + Hệ thống tích hợp vận hành với độ trễ thấp, có khả năng chạy gần thời gian thực.
* Báo cáo phân tích lỗi: các trường hợp nhận dạng sai, nhầm lẫn cảm xúc, nguyên nhân và đề xuất giải pháp.
* Viết báo cáo cho đề tài.

## B2.3. Kế hoạch nghiên cứu

Đề tài đề xuất có kế hoạch thực hiện như sau:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tham chiếu** | **Tháng thứ 1** | **Tháng thứ 2** | **Tháng thứ 3** | **Tháng thứ 4** | **Tháng thứ 5** | **Tháng thứ 6** |
| Nội dung 1 |  |  |  |  |  |  |
| Nội dung 2 |  |  |  |  |  |  |
| Nội dung 3a |  |  |  |  |  |  |
| Nội dung 3b |  |  |  |  |  |  |
| Nội dung 3c |  |  |  |  |  |  |
| Nội dung 4 |  |  |  |  |  |  |

## B3. Kết quả dự kiến

* Các mô hình được huấn luyện có hiệu suất cao hơn kết quả đã đạt được ở cuộc thi nếu hệ thống của VLSP 2025 cho tập kiểm tra riêng mở lại, hoặc đạt hiệu suất cao hơn trên tập kiểm tra công khai.
* Các kiến thức và kỹ năng nghiên cứu khoa học có được trong quá trình thực hiện đề tài.
* Báo cáo tổng kết đề tài.
* Demo với đầu vào là một đoạn âm thanh, hệ thống xuất ra văn bản cùng nhãn cảm xúc theo từng mốc thời gian, đồng thời được thiết kế với API/Interface dễ tích hợp và tối ưu hiệu năng nhằm giảm độ trễ, đáp ứng yêu cầu gần thời gian thực.

# B4. Tài liệu tham khảo

[1] Grand View Research, “Voice Recognition Market Size, Share & Trends Analysis Report By Deployment Mode, By Technology, By Vertical, By Region, And Segment Forecasts, 2024–2030,” Grand View Research, San Francisco, CA, USA, 2023. [Online]. Available:<https://www.grandviewresearch.com/industry-analysis/voice-recognition-market>

[2] VLSP, “Vietnamese Language and Speech Processing Workshop (VLSP 2025),” *Vietnamese Language and Speech Processing (VLSP)*, 2025. [Online]. Available:<https://vlsp.org.vn/vi/node>

[3] H. Ahlawat, N. Aggarwal, and D. Gupta, “Automatic speech recognition: A survey of deep learning techniques and approaches,” *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 201–237, Dec. 2025. [Online]. Available:<https://www.sciencedirect.com/science/article/pii/S2666307424000573>

[4] K. Yang, T.-Y. Hu, J.-H. R. Chang, H. S. Koppula, and O. Tuzel, “Text is all you need: Personalizing ASR models using controllable speech synthesis,” *Apple Machine Learning Research*, Apr. 2023. [Online]. Available:<https://machinelearning.apple.com/research/personalizing-asr-models>

[5] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, “A comprehensive review of speech emotion recognition systems,” IEEE Access, vol. 9, pp. 77 598–77 620, 2021. doi: 10.1109/ACCESS.2021.3082309. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9383000>

[6] Y. B. Singh and S. Goel, “A systematic literature review of speech emotion recognition approaches,” *Neurocomputing*, vol. 492, pp. 245–263, Jul. 2022. [Online]. Available:<https://doi.org/10.1016/j.neucom.2022.04.028>

[7] S. M. George and P. M. Ilyas, “A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise,” Neurocomputing, vol. 568, p. 127015, Feb. 2024. doi: 10.1016/j.neucom.2023.127015. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0925231223011384>

[8] T.-T. Le, L. T. Nguyen, and D. Q. Nguyen, “PhoWhisper: Automatic speech recognition for Vietnamese,” arXiv preprint arXiv:2406.02555, 2024. doi: 10.48550/arXiv.2406.02555. [Online]. Available: <https://arxiv.org/abs/2406.02555>

[9] L. T. T. Tran, H.-G. Kim, H. M. La, and S. V. Pham, “Automatic speech recognition of Vietnamese for a new large-scale corpus,” Electronics, vol. 13, no. 5, p. 977, Mar. 2024. doi: 10.3390/electronics13050977. [Online]. Available: <https://www.mdpi.com/2079-9292/13/5/977>

[10] T.-N. Phung, D.-B. Nguyen, and N.-P. Pham, “A review on speech recognition for under-resourced languages: A case study of Vietnamese,” International Journal of Knowledge and Systems Science, vol. 15, no. 1, pp. 1–16, Jan. 2024. doi: 10.4018/IJKSS.332869. [Online]. Available: <https://www.igi-global.com/article/a-review-on-speech-recognition-for-under-resourced-languages/332869>

[11] S. Wang, J. Guðnason, and D. Borth, “Learning Emotional Representations from Imbalanced Speech Data for Speech Emotion Recognition and Emotional Text-to-Speech,” in *Proc. Interspeech 2023*, Dublin, Ireland, 2023, pp. 351–355. [Online].Available:<https://arxiv.org/abs/2306.05709>

[12] G. Alhussein, I. Ziogas, S. Saleem, and L. J. Hadjileontiadis, “Speech emotion recognition in conversations using artificial intelligence: A systematic review and meta-analysis,” Artificial Intelligence Review, vol. 58, article no. 198, Apr. 2025. doi: 10.1007/s10462-025-11197-8. [Online]. Available: <https://link.springer.com/article/10.1007/s10462-025-11197-8>

[13] PyTorch Team, “Speech Recognition with Wav2Vec2 — Torchaudio tutorial,” *PyTorch Documentation*, 2023. [Online]. Available:<https://docs.pytorch.org/audio/stable/tutorials/speech_recognition_pipeline_tutorial.html>

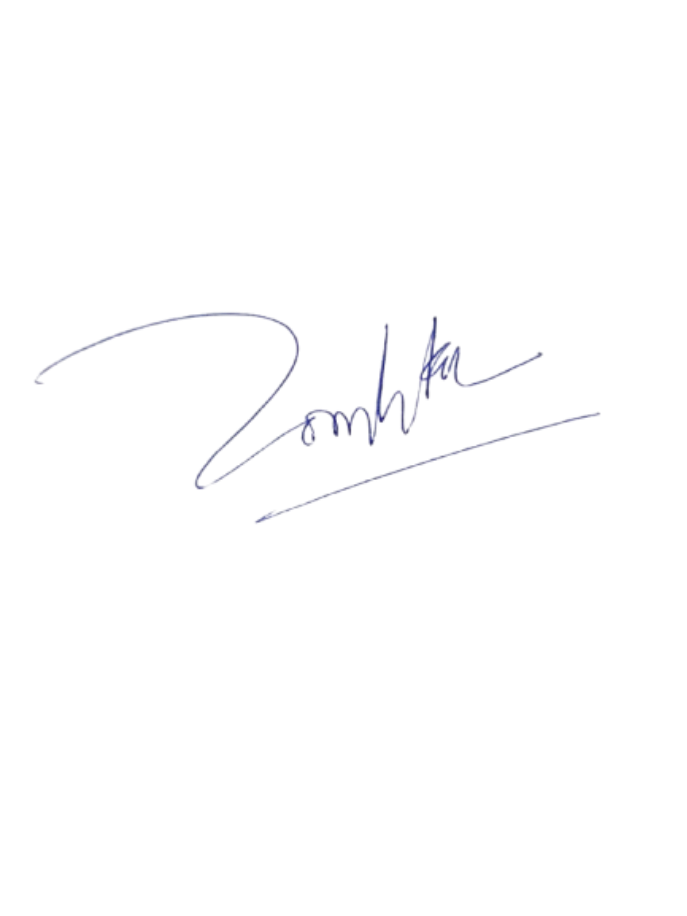
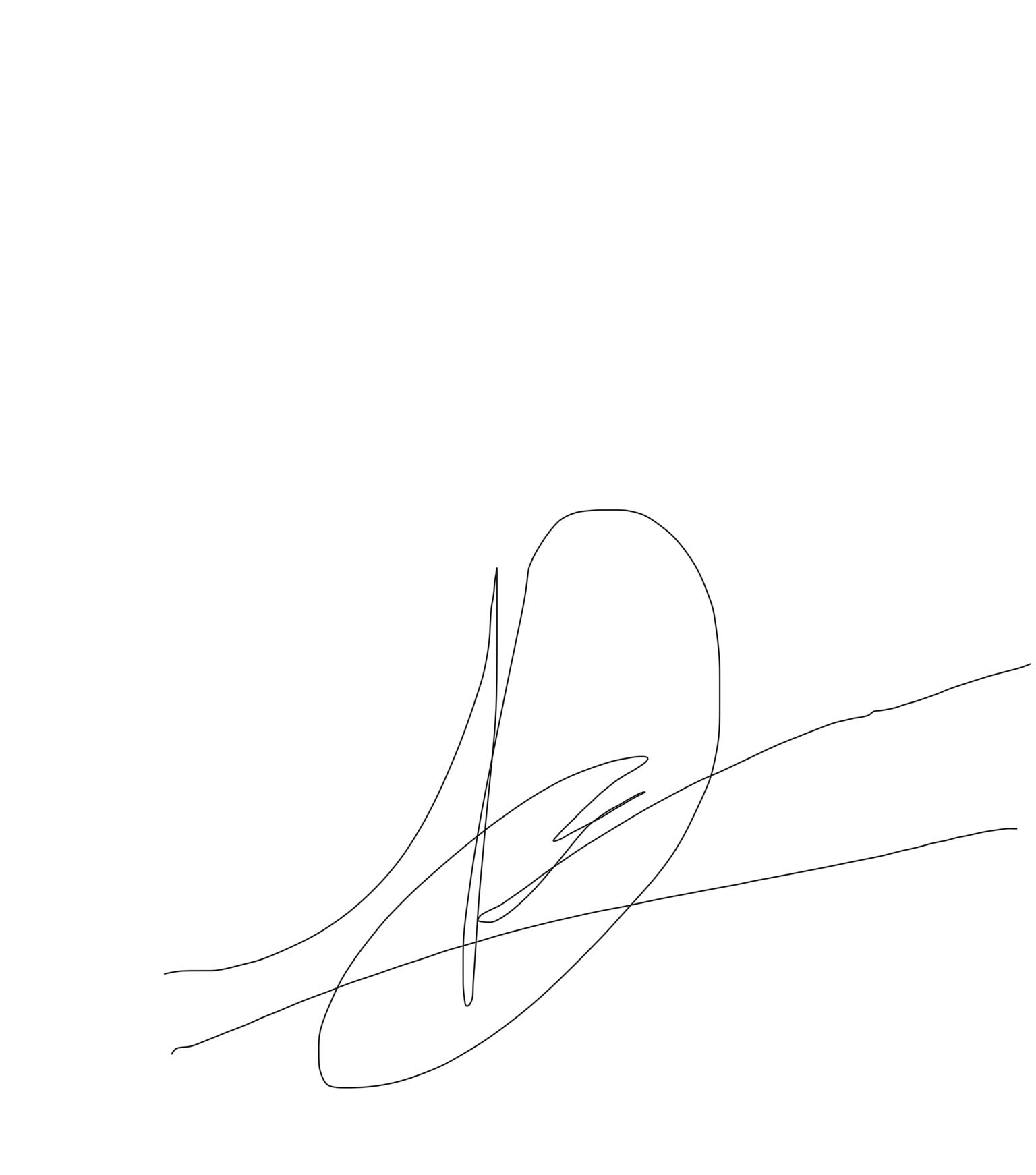
[14] A. Chakhtouna, S. Sekkate, and A. Adib, “Unveiling embedded features in Wav2vec2 and HuBERT models for speech emotion recognition,” Procedia Computer Science, vol. 232, pp. 2560–2569, 2024. doi: 10.1016/j.procs.2024.02.074. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050924002515>

[15] OpenAI, “Whisper: Robust Speech Recognition via Large-Scale Weak Supervision,” *OpenAI Research*, Sept. 2022. [Online]. Available:<https://openai.com/index/whisper/>

[16] T. B. Nguyen, *“nguyenvulebinh/wav2vec2-base-vietnamese-250h,”* Hugging Face Model Hub, 2021. [Online]. Available:<https://huggingface.co/nguyenvulebinh/wav2vec2-base-vietnamese-250h>

[17] Kenneth Heafield, “KenLM Language Model Toolkit,” *KenLM*, [Online]. Available: [https://kheafield.com/code/kenlm/](https://kheafield.com/code/kenlm/?utm_source=chatgpt.com)

|  |  |
| --- | --- |
| *Ngày 15 tháng 10 năm 2025*  **Giảng viên hướng dẫn**  (Ký và ghi rõ họ tên) | *Ngày 15 tháng 10 năm 2025*  **Chủ nhiệm đề tài**  (Ký và ghi rõ họ tên) |



Võ Minh Tâm Nguyễn Cao Duy