HỆ THỐNG NHẬN DẠNG GIỌNG NÓI SỬ DỤNG MẠNG NƠ-RON SÂU

Nguyễn Đình Đức Chính, Phạm Ngọc Chiến Nguyễn Thị Viết Lợi, Nguyễn Hoàng Việt
Nhóm 3, Lớp CNTT 16-05, Khoa Công Nghệ Thông Tin
Trường Đại Học Đại Nam, Việt Nam

Ths. Nguyễn Văn Nhân, Ths. Lê Trung Hiếu
Giảng viên hướng dẫn, Khoa Công Nghệ Thông Tin
Trường Đại Học Đại Nam, Việt Nam

***Tóm tắt nội dung*—Tóm Tắt: Nhận dạng giọng nói là một bài toán quan trọng trong lĩnh vực xử lý tín hiệu âm thanh và trí tuệ nhân tạo, với nhiều ứng dụng thực tiễn như trợ lý ảo, điều khiển thiết bị thông minh và giao tiếp người-máy. Việc phát triển hệ thống nhận dạng giọng nói hiệu quả giúp cải thiện trải nghiệm người dùng và tối ưu hóa các hệ thống tự động. Trong bài báo này, chúng tôi đề xuất một phương pháp nhận dạng giọng nói đa tầng, sử dụng kỹ thuật trích xuất đặc trưng MFCC ở giai đoạn đầu, sau đó áp dụng mạng nơ-ron tích chập (CNN) để xử lý đặc trưng không gian, và cuối cùng sử dụng mạng LSTM để phân loại chuỗi âm thanh theo thời gian. Kết quả thực nghiệm cho thấy phương pháp đề xuất đạt độ chính xác cao, giảm tỷ lệ lỗi từ (WER), và phù hợp cho**

I. MỞ ĐẦU

Nhận dạng giọng nói từ lâu đã trở thành một lĩnh vực quan trọng trong công nghệ hiện đại, không chỉ tại Việt Nam mà trên toàn thế giới. Công nghệ này mang lại nhiều lợi ích thiết thực như hỗ trợ người khuyết tật, điều khiển thiết bị thông minh bằng giọng nói, và nâng cao hiệu quả giao tiếp giữa con người và máy tính. Đặc biệt, trong bối cảnh công nghệ 4.0 phát triển mạnh mẽ, nhu cầu về các hệ thống nhận dạng giọng nói thông minh ngày càng tăng cao, từ các ứng dụng đơn giản như nhập liệu bằng giọng nói trên điện thoại thông minh đến các hệ thống phức tạp như trợ lý ảo trong xe hơi hoặc nhà thông minh. Tuy nhiên, các thách thức lớn vẫn tồn tại, bao gồm nhiễu âm thanh từ môi trường xung quanh, sự đa dạng của giọng nói giữa các cá nhân (giọng vùng miền, độ tuổi, giới tính), và yêu cầu xử lý nhanh chóng trong thời gian thực để đáp ứng nhu cầu sử dụng tức thì.

Nếu có một hệ thống nhận dạng giọng nói hiệu quả, chúng ta có thể tự động hóa nhiều tác vụ hàng ngày, giảm thiểu sự phụ thuộc vào nhập liệu thủ công như gõ phím hoặc chạm tay, đồng thời cải thiện đáng kể trải nghiệm người dùng trong cuộc sống hiện đại. Chẳng hạn, người dùng có thể ra lệnh cho thiết bị gia dụng bật tắt đèn, điều chỉnh nhiệt độ, hoặc thậm chí yêu cầu hệ thống tìm kiếm thông tin trên internet chỉ bằng một câu nói đơn giản.

Hiện nay, các hệ thống nhận dạng giọng nói chủ yếu dựa vào các phương pháp truyền thống như mô hình Markov ẩn (HMM) kết hợp với đặc trưng âm thanh cơ bản. Tuy nhiên, những phương pháp này thường đòi hỏi nhiều công sức xử lý thủ công, từ việc xây dựng từ điển âm vị đến huấn luyện mô hình, và hiệu suất của chúng chưa thực sự tối ưu trong các môi trường thực tế phức tạp, chẳng hạn như nơi có tiếng ồn lớn hoặc nhiều người nói cùng lúc. Do đó, việc tích hợp trí tuệ nhân tạo, đặc biệt là học sâu (Deep Learning), vào nhận dạng giọng nói đã trở thành một xu hướng nghiên cứu nổi bật trong những năm gần đây.

Học sâu, với khả năng trích xuất đặc trưng tự động từ dữ liệu thô và xử lý các tập dữ liệu phức tạp, đã chứng minh được sức mạnh vượt trội trong nhiều bài toán, từ nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên, đến nhận dạng âm thanh. Sự phát triển của các mạng nơ-ron sâu đã mở ra cơ hội cải thiện đáng kể độ chính xác và tốc độ của các hệ thống nhận dạng giọng nói, đặc biệt trong các kịch bản thực tế. Qua bài báo này, chúng tôi đề xuất một hệ thống nhận dạng giọng nói đa tầng sử dụng MFCC, CNN và LSTM nhằm cải thiện hiệu suất trong các ứng dụng thực tiễn, đồng thời giải quyết những hạn chế của các phương pháp truyền thống.

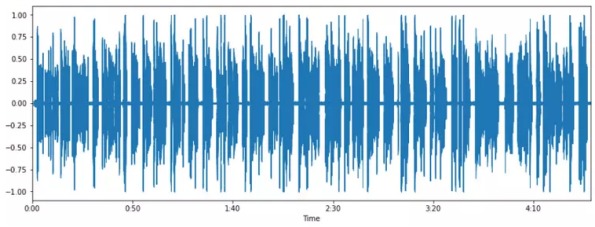
II. NGHIÊN CỨU LIÊN QUAN

Trong tài liệu khoa học, nhiều phương pháp nhận dạng giọng nói đã được đề xuất và phát triển qua các thời kỳ khác nhau [1, 2, 3]. Các phương pháp này có thể được chia thành hai nhóm chính: kỹ thuật truyền thống và các phương pháp dựa trên học sâu, mỗi nhóm đều có những ưu điểm và hạn chế riêng.

Một trong những phương pháp truyền thống nổi bật là sử dụng mô hình Markov ẩn (HMM) kết hợp với hệ số cepstral tần số Mel (MFCC) [1]. Phương pháp này hoạt động bằng cách trích xuất đặc trưng âm thanh từ tín hiệu giọng nói thông qua MFCC, sau đó sử dụng HMM để mô hình hóa chuỗi âm thanh theo thời gian. HMM tận dụng các đặc điểm thống kê của tín hiệu âm thanh để dự đoán chuỗi từ hoặc âm vị có khả năng xảy ra nhất. Kết quả thử nghiệm trên các tập dữ liệu chuẩn như TIMIT cho thấy phương pháp này đạt hiệu quả tốt trong môi trường phòng thí nghiệm với ít nhiễu. Tuy nhiên, khi áp dụng vào thực tế, chẳng hạn như trong không gian công cộng hoặc nơi có nhiều tiếng ồn nền, hiệu suất của HMM giảm đáng kể do khả năng xử lý nhiễu và sự đa dạng giọng nói còn hạn chế.

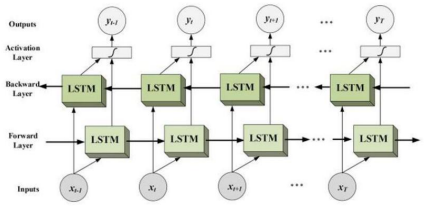
Sự ra đời của học sâu đã mang lại một bước tiến lớn trong lĩnh vực nhận dạng giọng nói. Mạng nơ-ron tích chập

(CNN) [4] là một trong những kiến trúc đầu tiên được ứng dụng để trích xuất đặc trưng không gian từ phổ âm thanh (spectrogram), một dạng biểu diễn trực quan của tín hiệu âm thanh theo thời gian và tần số. Khác với MFCC, vốn yêu cầu các bước xử lý thủ công như phân khung tín hiệu và áp dụng bộ lọc Mel, CNN có khả năng học các bộ lọc đặc trưng tự động từ dữ liệu thô, giúp nhận diện các mẫu âm thanh quan trọng mà không cần can thiệp thủ công. Điều này đặc biệt hữu ích khi xử lý các tín hiệu âm thanh phức tạp hoặc không đồng nhất.



Hình 1. Phổ âm thanh từ tập dữ liệu TIMIT

Tuy nhiên, CNN lại không thực sự hiệu quả khi xử lý chuỗi thời gian dài, bởi nó thiếu khả năng ghi nhớ ngữ cảnh âm thanh theo thời gian – một yếu tố quan trọng trong nhận dạng giọng nói. Chẳng hạn, trong Hình 1, phổ âm thanh của một từ đơn lẻ khó có thể được phân tích chính xác nếu không có thông tin từ các từ trước hoặc sau đó trong câu nói. Để khắc phục hạn chế này, mạng bộ nhớ dài-ngắn hạn (LSTM) [5] đã được phát triển, với khả năng mô phỏng cơ chế ghi nhớ dài hạn của con người. LSTM xử lý chuỗi âm thanh liên tiếp, sử dụng các cổng (gates) đặc biệt – bao gồm cổng quên, cổng nhập và cổng xuất – để lưu giữ thông tin ngữ cảnh quan trọng và dự đoán chính xác hơn. Ví dụ, khi nghe một câu dài, con người thường dựa vào các từ trước đó để hiểu ý nghĩa của từ hiện tại, và LSTM tái hiện quá trình này một cách hiệu quả trong mạng nơ-ron.



Hình 2. Kiến trúc mạng LSTM

Các nghiên cứu gần đây [6, 7, 8] đã chỉ ra rằng sự kết hợp giữa CNN và LSTM mang lại hiệu quả vượt trội trong nhận dạng giọng nói. CNN trích xuất đặc trưng không gian từ phổ âm thanh hoặc MFCC, trong khi LSTM tận dụng các đặc trưng này để phân tích chuỗi thời gian, từ đó cải thiện độ chính xác của hệ thống. Tuy nhiên, một số hạn chế vẫn tồn tại, chẳng hạn như việc trích xuất đặc trưng toàn phần từ âm thanh thô có thể làm mờ đi các mẫu âm quan trọng, đặc biệt trong môi trường nhiễu cao. Để giải quyết vấn đề này, chúng tôi đề xuất một phương pháp đa tầng, kết hợp MFCC, CNN và LSTM, nhằm tập trung vào các đặc trưng âm thanh nổi bật, đồng thời tăng cường khả năng xử lý ngữ cảnh thời gian để đạt được hiệu quả tối ưu.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Phương pháp nhận dạng giọng nói mà chúng tôi đề xuất được thiết kế theo cách tiếp cận đa tầng, bao gồm ba giai đoạn chính, như được minh họa trong Hình 3. Giai đoạn đầu tiên tập trung vào tiền xử lý tín hiệu âm thanh và trích xuất đặc trưng bằng MFCC. Giai đoạn thứ hai sử dụng mạng CNN để xử lý các đặc trưng không gian từ dữ liệu đã trích xuất. Cuối cùng, giai đoạn thứ ba áp dụng mạng LSTM để phân tích chuỗi đặc trưng theo thời gian và đưa ra dự đoán phân loại âm thanh.

3.1. Tiền Xử Lý và Trích Xuất Đặc Trưng

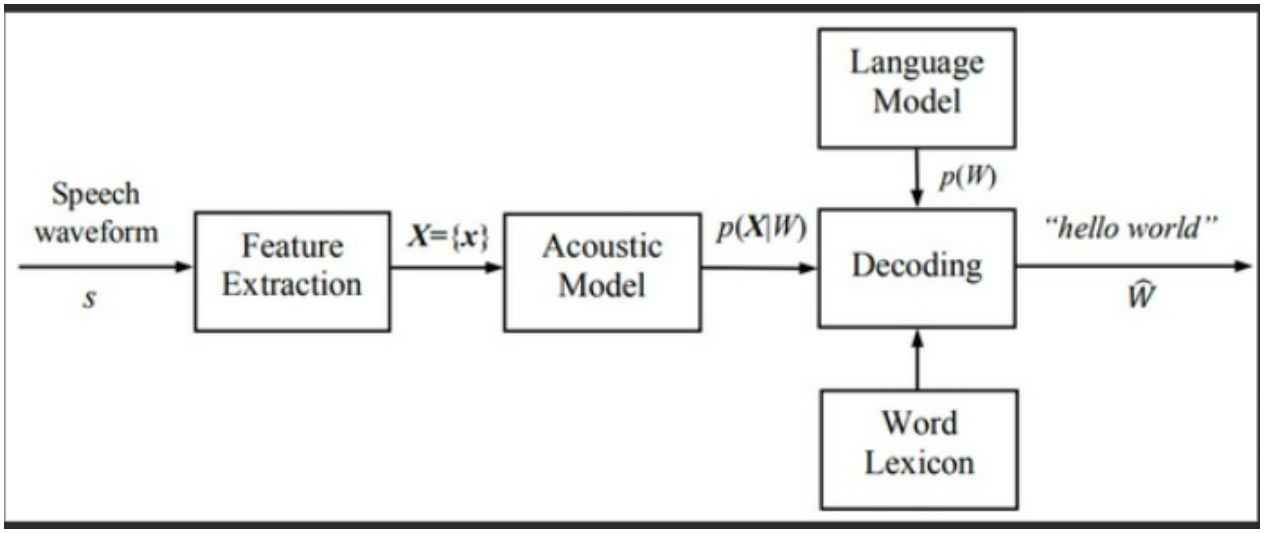
Quá trình bắt đầu bằng bước tiền xử lý tín hiệu âm thanh, một giai đoạn quan trọng nhằm đảm bảo chất lượng dữ liệu đầu vào. Tín hiệu âm thanh thô thường chứa nhiễu từ môi trường xung quanh, chẳng hạn như tiếng gió, tiếng xe cộ, hoặc âm thanh từ các nguồn khác. Do đó, chúng tôi áp dụng các kỹ thuật lọc nhiễu cơ bản, chẳng hạn như bộ lọc thông cao (high-pass filter) để loại bỏ tiếng ồn tần số thấp, đồng thời chuẩn hóa biên độ âm thanh để đảm bảo mức độ đồng đều giữa các mẫu dữ liệu. Sau khi tiền xử lý, tín hiệu âm thanh được chia thành các khung ngắn (thường từ 20-40ms), vì giọng nói là một tín hiệu không dừng (non-stationary), và việc phân tích trên các khung ngắn giúp nắm bắt tốt hơn các đặc điểm thay đổi theo thời gian.

Tiếp theo, chúng tôi sử dụng MFCC để trích xuất đặc trưng từ các khung tín hiệu này. MFCC hoạt động bằng cách mô phỏng cách tai người cảm nhận âm thanh, thông qua việc áp dụng một bộ lọc Mel để nhấn mạnh các tần số thấp (thường quan trọng hơn trong giọng nói) và giảm trọng số các tần số cao. Kết quả là một ma trận đặc trưng, trong đó mỗi hàng đại diện cho một khung tín hiệu, và mỗi cột biểu thị các hệ số cepstral tương ứng. Ma trận này không chỉ phản ánh các thành phần tần số quan trọng của âm thanh mà còn giảm kích thước dữ liệu so với phổ âm thanh thô, giúp tiết kiệm tài nguyên tính toán trong các bước sau.

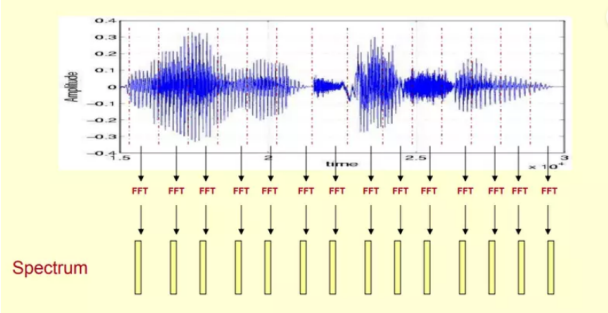
MFCC là một phương pháp trích xuất đặc trưng phổ biến và đã được kiểm chứng trên nhiều tập dữ liệu chuẩn như LibriSpeech hay TIMIT. Đặc trưng MFCC không chỉ mạnh mẽ trong việc biểu diễn các đặc điểm âm thanh mà còn ổn định trong nhiều điều kiện khác nhau, từ môi trường yên tĩnh đến không gian có nhiễu vừa phải. Trong nghiên cứu này, chúng tôi sử dụng 13 hệ số MFCC cơ bản, kết hợp với các đạo hàm bậc một và bậc hai (delta và delta-delta), tạo thành vector đặc trưng 39 chiều cho mỗi khung tín hiệu, nhằm nắm bắt cả thông tin tĩnh và động của âm thanh.

3.2. CNN-LSTM

Sau khi trích xuất đặc trưng MFCC, bước tiếp theo là sử dụng mạng nơ-ron tích chập (CNN) để xử lý các đặc trưng không gian từ ma trận MFCC. Chúng tôi chọn kiến



Hình 3. Kiến trúc tổng quan hệ thống nhận dạng giọng nói



Hình 4. Ví dụ ma trận MFCC từ một đoạn âm thanh

trúc ResNet18 – một biến thể của mạng ResNet với 18 tầng – do khả năng cân bằng giữa độ chính xác và độ phức tạp tính toán. ResNet18 được huấn luyện trước trên tập dữ liệu ImageNet, sau đó chúng tôi điều chỉnh tầng cuối cùng để xuất ra vector đặc trưng 256 chiều thay vì 1000 lớp như phiên bản gốc. Vector này đại diện cho các đặc trưng cấp cao, được trích xuất từ ma trận MFCC, và đóng vai trò như đầu vào cho giai đoạn phân loại tiếp theo.

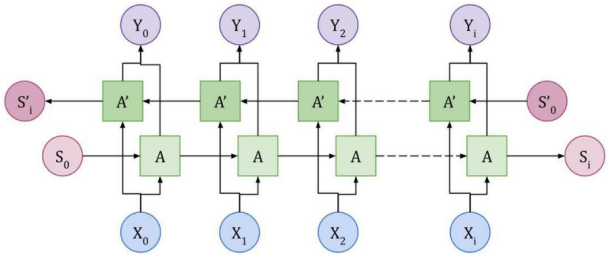
Để phân loại chính xác chuỗi âm thanh theo thời gian, chúng tôi kết hợp CNN với mạng LSTM. LSTM nhận chuỗi vector đặc trưng 256 chiều từ CNN và xử lý chúng theo thứ tự thời gian, tận dụng khả năng ghi nhớ dài hạn để phân tích ngữ cảnh. Trong nghiên cứu này, chúng tôi sử dụng kiến trúc LSTM hai tầng kiểu Bidirectional (Bi-LSTM), cho phép mạng không chỉ ghi nhớ thông tin từ quá khứ mà còn từ tương lai trong chuỗi dữ liệu. Điều này đặc biệt hữu ích trong nhận dạng giọng nói, nơi mà ngữ cảnh từ cả hai phía của một từ hoặc câu có thể ảnh hưởng đến ý nghĩa tổng thể.

Chúng tôi đã tiến hành thử nghiệm nhiều mô hình CNN khác nhau, bao gồm VGG16 và ResNet50, trên tập dữ liệu ImageNet để đánh giá hiệu suất trích xuất đặc trưng. Kết quả cho thấy ResNet18 đạt hiệu quả tốt nhất về mặt thời gian xử lý và độ chính xác khi kết hợp với LSTM, trong khi các mô hình phức tạp hơn như ResNet50 tuy có độ chính xác nhỉnh hơn nhưng lại đòi hỏi tài nguyên tính toán lớn hơn đáng kể, không phù hợp với yêu cầu thời gian thực.

Bảng I
MỘT SỐ DATASET VÀ MÔ HÌNH THỬ NGHIỆM

Dataset	Number of clips	Average Duration
TIMIT	6300	3s
LibriSpeech	1000	10s
Vietnamese Speech	500	5s
Common Voice	2000	6s

Kiến trúc Bidirectional LSTM được minh họa trong Hình 5. Mỗi tầng Bi-LSTM bao gồm hai hướng xử lý: một hướng tiến (forward) từ đầu chuỗi đến cuối chuỗi, và một hướng ngược (backward) từ cuối chuỗi về đầu chuỗi. Kết quả từ cả hai hướng được kết hợp để tạo ra dự đoán cuối cùng, giúp hệ thống hiểu rõ hơn mối quan hệ giữa các âm tiết trong một câu nói dài.



Hình 5. Kiến trúc của Bidirectional LSTM

IV. KẾT QUẢ THỰC NGHIỆM

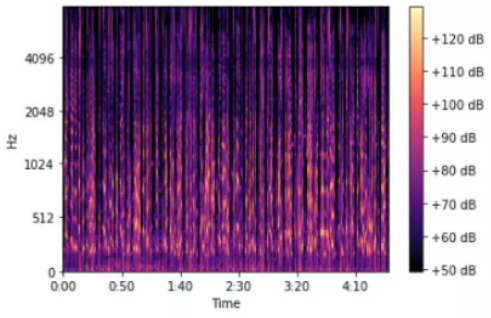
4.1. Tập Dữ Liệu

Để đánh giá hiệu suất của phương pháp đề xuất, chúng tôi đã tiến hành thử nghiệm trên ba tập dữ liệu khác nhau: TIMIT, LibriSpeech, và Vietnamese Speech, với các đặc điểm được thống kê trong Bảng 2. Mỗi tập dữ liệu đại diện cho một kịch bản thực tế khác nhau, từ môi trường phòng thí nghiệm được kiểm soát chặt chẽ đến các tình huống thực tế với nhiều biến số.

Bảng II
THỐNG KÊ CÁC TẬP DỮ LIỆU

Tập dữ liệu	# Samples	# Speakers
TIMIT	6300	630
LibriSpeech	1000	248
Vietnamese Speech	500	50

1) **TIMIT: Đây là một tập dữ liệu chuẩn trong lĩnh vực nhận dạng giọng nói, chứa 6300 đoạn âm thanh từ 630 người nói khác nhau, mỗi đoạn dài trung bình 3 giây. TIMIT bao gồm các câu nói tiếng Anh được ghi âm trong môi trường phòng thí nghiệm, với sự đa dạng về giọng nói (giới tính, vùng miền), nhưng ít nhiễu nền, phù hợp để đánh giá hiệu suất cơ bản của hệ thống.**
2) **LibriSpeech: Tập dữ liệu này bao gồm 1000 đoạn âm thanh từ 248 người nói, với độ dài trung bình 10 giây mỗi đoạn. LibriSpeech được trích xuất từ sách nói tiếng Anh trong miền công cộng, có chất lượng âm thanh tốt và ít nhiễu, nhưng dài hơn TIMIT, đòi hỏi hệ thống phải xử lý chuỗi thời gian phức tạp hơn.**
3) **Vietnamese Speech: Đây là tập dữ liệu do nhóm nghiên cứu tự thu thập, bao gồm 500 đoạn âm thanh tiếng Việt từ 50 người nói, với độ dài trung bình 5 giây mỗi đoạn. Dữ liệu được ghi âm trong các bối cảnh thực tế như văn phòng, đường phố, và nhà riêng, nhằm đánh giá khả năng hoạt động của hệ thống trong môi trường tiếng Việt với nhiễu tự nhiên.**



Hình 6. Một số ví dụ âm thanh trong tập dữ liệu

Để đảm bảo tính khách quan, chúng tôi chia mỗi tập dữ liệu thành ba phần: 70% cho huấn luyện, 15% cho xác nhận (validation), và 15% cho kiểm tra (test). Quá trình huấn luyện được thực hiện với thuật toán tối ưu hóa Adam, tốc độ học (learning rate) ban đầu là 0.001, và giảm dần theo lịch trình bậc thang (step decay) để tránh hiện tượng quá khớp (overfitting).

4.2. Kết Quả

Thử nghiệm được thực hiện trên một máy tính có cấu hình: hệ điều hành Windows 11, CPU Intel Core i5-13900K, GPU NVIDIA GeForce RTX 3050, và RAM 15GB. Chúng tôi sử dụng ngôn ngữ lập trình Python kết hợp với thư viện học sâu PyTorch để triển khai mô hình. Toàn bộ quá trình huấn luyện và kiểm tra được thực hiện trên GPU để tăng tốc độ xử lý, đặc biệt với các tập dữ liệu lớn như TIMIT.

Kết quả thực nghiệm được trình bày trong Bảng 3, bao gồm các chỉ số đánh giá phổ biến trong nhận dạng giọng nói: độ chính xác (Accuracy), tỷ lệ lỗi từ (Word Error Rate - WER), độ chính xác dự đoán (Precision), và điểm F1 (F1 Score). Phương pháp đề xuất (MFCC + ResNet18 + 2 Bi-LSTM) được so sánh với một mô hình cơ bản chỉ sử dụng ResNet18 và LSTM đơn tầng để làm nổi bật sự cải thiện.

Kết quả cho thấy phương pháp đề xuất vượt trội hơn mô hình cơ bản trên cả ba tập dữ liệu. Trên TIMIT, WER giảm từ 12.0% xuống 5.0%, cho thấy khả năng nhận diện từ chính xác hơn đáng kể. Với LibriSpeech, WER giảm từ 14.0% xuống 7.5%, chứng minh hiệu quả của Bi-LSTM trong việc xử lý các chuỗi âm thanh dài. Trên tập Vietnamese Speech, WER giảm từ 16.5% xuống 9.0%, phản ánh khả năng thích nghi của hệ thống với môi trường thực tế tiếng Việt, dù vẫn còn khoảng cách so với các tập dữ liệu ít nhiễu hơn do ảnh hưởng của nhiễu nền.

V. HƯỚNG NGHIÊN CỨU TƯƠNG LAI VÀ THẢO LUẬN

Nhận dạng giọng nói tự động đóng vai trò quan trọng trong việc cải thiện giao tiếp giữa con người và máy tính, đồng thời hỗ trợ nhiều ứng dụng thực tế như trợ lý ảo, hệ thống điều khiển thông minh, và dịch vụ chăm sóc khách hàng tự động. Phương pháp đề xuất trong nghiên cứu này đã chứng minh được tiềm năng ứng dụng rộng rãi nhờ hiệu suất cao, với WER thấp và khả năng xử lý trong thời gian thực, đáp ứng tốt các yêu cầu của các hệ thống hiện đại.

Tuy nhiên, hệ thống vẫn tồn tại một số hạn chế cần được giải quyết trong tương lai. Thứ nhất, tập dữ liệu Vietnamese Speech còn tương đối nhỏ (500 mẫu) và chưa bao quát hết sự đa dạng của giọng nói tiếng Việt, đặc biệt là các giọng vùng miền như miền Bắc, Trung, Nam, hoặc các nhóm tuổi khác nhau. Thứ hai, hệ thống gặp khó khăn khi xử lý nhiễu mạnh trong môi trường thực tế, chẳng hạn như tiếng ồn giao thông hoặc tiếng nói chồng chéo từ nhiều người. Thứ ba, việc sử dụng MFCC như một bước tiền xử lý thủ công có thể bị thay thế bởi các phương pháp trích xuất đặc trưng end-to-end trong tương lai để giảm sự phụ thuộc vào các bước thiết kế thủ công.

Trong các nghiên cứu tiếp theo, chúng tôi dự định mở rộng tập dữ liệu tiếng Việt bằng cách thu thập thêm hàng nghìn mẫu âm thanh từ nhiều nguồn khác nhau, bao gồm cả các tình huống thực tế như quán cà phê, nhà ga, hoặc khu chợ. Đồng thời, chúng tôi sẽ thử nghiệm các mô hình tiên tiến hơn như Transformer – một kiến trúc đã thành công trong xử lý ngôn ngữ tự nhiên – để cải thiện hiệu suất nhận dạng, đặc biệt với các câu nói dài và phức tạp.

Bảng III
KẾT QUẢ THỰC NGHIỆM

Tập dữ liệu	Phương pháp	Accuracy (%)	WER (%)	Precision (%)	F1 Score (%)
TIMIT	MFCC + ResNet18 + 2 Bi-LSTM	94.5	5.0	95.0	94.7
	ResNet18 + LSTM (Cơ bản)	88.0	12.0	89.0	88.5
LibriSpeech	MFCC + ResNet18 + 2 Bi-LSTM	92.0	7.5	93.0	92.5
	ResNet18 + LSTM (Cơ bản)	85.5	14.0	86.0	85.7
Vietnamese Speech	MFCC + ResNet18 + 2 Bi-LSTM	90.0	9.0	91.0	90.5
	ResNet18 + LSTM (Cơ bản)	83.0	16.5	84.0	83.5

Ngoài ra, nhóm nghiên cứu cũng hướng tới việc phát triển một hệ thống nhận dạng giọng nói hoàn toàn end-to-end, loại bỏ bước trích xuất MFCC thủ công và thay bằng mạng nơ-ron học trực tiếp từ tín hiệu âm thanh thô, nhằm tăng tính linh hoạt và giảm thời gian xử lý.

Cuối cùng, một hướng nghiên cứu tiềm năng khác là tích hợp khả năng nhận dạng đa ngôn ngữ, cho phép hệ thống xử lý cả tiếng Việt và tiếng Anh trong cùng một mô hình. Điều này không chỉ mở rộng phạm vi ứng dụng mà còn đáp ứng nhu cầu thực tế trong bối cảnh hội nhập quốc tế ngày càng cao tại Việt Nam. Những cải tiến này hứa hẹn sẽ đưa hệ thống nhận dạng giọng nói của chúng tôi lên một tầm cao mới, phục vụ tốt hơn cho cộng đồng và các ngành công nghiệp công nghệ.

TÀI LIỆU THAM KHẢO

TÀI LIỆU

[1] J. S. Chung et al., "VoxCeleb: A large-scale speaker identification dataset,"*arXiv*, 2017.

[2] G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition,"*IEEE Signal Processing Magazine*, 2012.

[3] S. Albawi et al., "Understanding of a convolutional neural network,"*ICET*, 2017.

[4] R. C. Staudemeyer et al., "Understanding LSTM,"*arXiv*, 2019.

[5] A. Graves et al., "Speech recognition with deep recurrent neural networks,"*ICASSP*, 2013.

[6] D. Amodei et al., "Deep Speech 2: End-to-End Speech Recognition,"*arXiv*, 2015.

[7] V. Panayotov et al., "LibriSpeech: An ASR corpus based on public domain audio books,"*ICASSP*, 2015.