DEEP LEARNING-BASED AUTOMATIC CONTROL SYSTEM USING AUDIO SIGNAL ANALYSIS FOR VIETNAMESE RECOGNITION

Pham Van Thanh[1*], Nguyen Xuan Nam[1], Vi Anh Quan[1], Pham Tien Lam[2]
[1]VNU University of Science, [2]Phenikaa University

ARTICLE INFO

Received: 20/3/2025
Revised: 09/7/2025
Published: 09/7/2025

KEYWORDS

Deep learning
Speech recognition
Edge computing
Autonomous control
IoT

ABSTRACT

The development of edge computing and TinyAI has opened up new opportunities for intelligent voice-controlled systems. This study proposes a deep learning-based, real-time voice command recognition system for Vietnamese, optimized for deployment on edge devices with limited computational resources. The system utilizes mel-frequency cepstral coefficients for feature extraction and employs a convolutional neural network combined with a long short-term memory network referred to as the CNN-LSTM model to achieve high-accuracy command recognition. The model is deployed and optimized on a Raspberry Pi, leveraging TinyAI to reduce computational demands while maintaining prompt processing capabilities. Experimental results show that the system achieves an accuracy of 96.25% with low latency, making it suitable for robot control and IoT applications without relying on cloud computing. This research highlights the potential of TinyAI in edge computing, enhancing the efficiency and applicability of Vietnamese voice-controlled systems in real-world environments.

PHÁT TRIỂN HỆ THỐNG ĐIỀU KHIỂN TỰ ĐỘNG DỰA TRÊN HỌC SÂU SỬ DỤNG PHÂN TÍCH TÍN HIỆU ÂM THANH ỨNG DỤNG CHO TIẾNG VIỆT

Phạm Văn Thành[1*], Nguyễn Xuân Nam[1], Vi Anh Quân[1], Phạm Tiến Lâm[2]
[1]Trường Đại học Khoa học Tự nhiên – ĐH Quốc gia Hà Nội, [2]Đại học Phenikaa

THÔNG TIN BÀI BÁO

Ngày nhận bài: 20/3/2025
Ngày hoàn thiện: 09/7/2025
Ngày đăng: 09/7/2025

TỪ KHÓA

Học sâu
Nhận diện giọng nói
Điện toán biên
Điều khiển tự động
IoT

TÓM TẮT

Sự phát triển của điện toán biên và TinyAI đã mở ra nhiều cơ hội mới cho các hệ thống điều khiển thông minh dựa trên âm thanh. Nghiên cứu này đề xuất một phương pháp nhận diện tín hiệu điều khiển bằng giọng nói sử dụng tiếng Việt dựa trên học sâu và triển khai trên các thiết bị biên có tài nguyên hạn chế. Hệ thống sử dụng Mel-Frequency Cepstral Coefficients để trích xuất đặc trưng âm thanh và áp dụng mô hình kết hợp mạng nơ-ron tích chập và mạng nơ-ron hồi quy dài-ngắn hạn, nhằm phát triển mô hình TinyAI nhận diện lệnh điều khiển với độ chính xác cao. Chúng tôi triển khai và tối ưu hóa mô hình trên phần cứng Raspberry Pi 3, tận dụng TinyAI để giảm thiểu yêu cầu tính toán mà vẫn đảm bảo hiệu suất xử lý nhanh chóng. Kết quả thực nghiệm cho thấy hệ thống đạt độ chính xác 96,25%, với độ trễ thấp, phù hợp cho các ứng dụng điều khiển robot và thiết bị IoT mà không cần kết nối đám mây. Nghiên cứu này góp phần khẳng định tiềm năng của TinyAI trong điện toán biên, giúp cải thiện hiệu quả và khả năng ứng dụng của các hệ thống điều khiển bằng giọng nói sử dụng tiếng Việt trong môi trường thực tế.

DOI: https://doi.org/10.34238/tnu-jst.12357

[*] Corresponding author. *Email: phamvanthanh@hus.edu.vn*

1. Giới thiệu

Sự phát triển của điện toán biên (Edge Computing) và TinyAI đang thay đổi cách các hệ thống thông minh hoạt động trong thời đại IoT (Internet of Things) [1]. Trong khi các phương pháp truyền thống dựa trên điện toán đám mây đòi hỏi băng thông cao và độ trễ lớn, thì điện toán biên giúp xử lý dữ liệu ngay tại nguồn, giảm thiểu độ trễ và tăng cường tính khả dụng của hệ thống. TinyAI, một nhánh của trí tuệ nhân tạo (Artificial Intelligence - AI), tập trung vào việc tối ưu hóa các giải pháp AI để triển khai trên các thiết bị biên có tài nguyên hạn chế. Nhờ đó, nó mở ra tiềm năng to lớn trong việc triển khai các hệ thống nhận diện và điều khiển thông minh, hoạt động độc lập mà không cần phụ thuộc vào máy chủ trung tâm [2].

Trong bối cảnh hiện tại, điều khiển bằng giọng nói đang trở thành một phương pháp tương tác quan trọng, đặc biệt trong các ứng dụng robot tự hành [3], thiết bị IoT và nhà thông minh [4]. Tuy nhiên, việc xử lý tín hiệu âm thanh trực tiếp trên thiết bị biên đặt ra nhiều thách thức, bao gồm hạn chế tài nguyên tính toán, bộ nhớ và hiệu suất tiêu thụ năng lượng [5]. Để giải quyết vấn đề này, các mô hình học sâu đã được tối ưu hóa để có thể nhận diện và phân loại tín hiệu âm thanh một cách hiệu quả ngay trên thiết bị biên [6].

Nhận dạng giọng nói tiếng Việt đã ghi nhận bước tiến ban đầu từ năm 2005 tại các phòng thí nghiệm trong nước, tập trung vào việc xây dựng các mô hình Hidden Markov Models (HMM) cùng các biến thể để phân tích âm học và giải mã ngôn ngữ [7], [8]. Tuy nhiên, các nghiên cứu trong giai đoạn này phải đối mặt với nhiều thách thức trong đặc trưng của tiếng Việt, chẳng hạn như hệ thống thanh điệu phức tạp và sự đa dạng trong cách phát âm giữa các vùng miền. Những năm tiếp theo, đã có nhiều nghiên cứu cải thiện được những hạn chế của HMM, chẳng hạn như dựa trên thông tin ngữ điệu để cải thiện sự tự nhiên trong tiếng Việt [9], hay áp dụng những cách trích xuất đặc trưng mới kết hợp với mô hình HMM [10]. Trong thời gian gần đây, sự phát triển mạnh mẽ của các mô hình học sâu (Deep Learning) đã tạo ra những bước đột phá trong lĩnh vực nhận dạng giọng nói. Các mô hình nhận dạng giọng nói tiếng Việt đã nhanh chóng áp dụng những tiến bộ này, mang lại hiệu suất vượt trội so với các phương pháp truyền thống. Đặc biệt, kiến trúc mạng nơ-ron hồi quy hai chiều kết hợp với các lớp LSTM (Bidirectional LSTM) đã chứng minh khả năng nắm bắt tốt các đặc trưng ngữ cảnh dài trong chuỗi âm thanh tiếng Việt [11]. Tuy nhiên, việc triển khai các mô hình nhận diện giọng nói tiếng Việt trên thiết bị biên chưa có được nhiều sự chú ý. Mặc dù công nghệ nhận dạng giọng nói đã và đang phát triển nhanh chóng trên các hệ thống đám mây, việc nghiên cứu và triển khai các mô hình nhẹ, tối ưu cho thiết bị có tài nguyên hạn chế như thiết bị IoT, và các hệ thống nhúng vẫn còn hạn chế. Các mô hình nhận dạng giọng nói hiện đại dựa trên kiến trúc Transformer [12] thường có kích thước lên đến hàng trăm triệu tham số, đòi hỏi thiết bị cần có tài nguyên tính toán và bộ nhớ lớn để hoạt động ổn định. Những yêu cầu này vượt xa khả năng của các thiết bị biên, tạo ra rào cản đáng kể cho việc triển khai ứng dụng nhận dạng giọng nói tiếng Việt trong các tình huống cần phản hồi thời gian thực và hoạt động liên tục. Đặc biệt, tại những môi trường như hệ thống nhà thông minh, thiết bị IoT công nghiệp hay các ứng dụng di động cần hoạt động ngoại tuyến, giới hạn về tài nguyên của thiết bị biên đã trở thành thách thức lớn cho việc ứng dụng các mô hình tiên tiến này.

Việc áp dụng TinyAI không chỉ mở ra tiềm năng to lớn cho các ứng dụng thực tế mà còn giải quyết nhiều thách thức cốt lõi trong việc triển khai công nghệ nhận dạng giọng nói tiếng Việt trên các thiết bị biên. Bằng cách tối ưu hóa mô hình giảm kích thước và tăng hiệu suất, TinyAI cho phép xử lý dữ liệu tại chỗ, giảm độ trễ và đảm bảo quyền riêng tư cho người dùng khi không cần truyền dữ liệu lên đám mây. Điều này đặc biệt quan trọng trong các trường hợp yêu cầu tính bảo mật cao hoặc khi kết nối mạng không ổn định.

Trong nghiên cứu này, chúng tôi đề xuất một mô hình nhận diện lệnh giọng nói sử dụng tiếng Việt dựa trên học sâu, sử dụng mô hình kết hợp mạng nơ-ron tích chập (Convolutional Neural Network - CNN) và mạng nơ-ron hồi quy dài-ngắn hạn (Long Short-Term Memory - LSTM) để phân tích tín hiệu âm thanh và được triển khai trên phần cứng Raspberry Pi. Hệ thống áp dụng

Mel-Frequency Cepstral Coefficients (MFCC) để trích xuất đặc trưng âm thanh [13], sau đó sử dụng mạng CNN để trích xuất đặc trưng không gian và LSTM để xử lý thông tin theo chuỗi thời gian, giúp nhận diện lệnh điều khiển với độ chính xác cao. Hệ thống dựa trên TinyAI hỗ trợ tối ưu hóa toàn bộ giải pháp AI, bao gồm kích thước mô hình và hiệu suất xử lý, từ đó giảm độ trễ và đảm bảo hoạt động hiệu quả trên các thiết bị biên có tài nguyên hạn chế.

Chúng tôi đã xây dựng tập dữ liệu với các giọng đọc đa dạng trong các điều kiện môi trường khác nhau, từ đó phát triển mô hình học sâu nhỏ gọn có thể phân biệt được các hiệu lệnh tiếng Việt: "lên", "xuống", "trái", "phải". Mô hình nhận diện lệnh giọng nói này được nhóm nghiên cứu thiết kế và triển khai để hoạt động hiệu quả trên các thiết bị biên. Chúng tôi đã tiến hành thực nghiệm bằng cách triển khai mô hình lên hệ thống điều khiển xe, sử dụng Raspberry Pi 3 làm phần điều khiển. Kết quả thực nghiệm cho thấy tính khả thi của giải pháp trong điều kiện thực tế, mở ra tiềm năng ứng dụng cho các hệ thống IoT và tự động hóa.

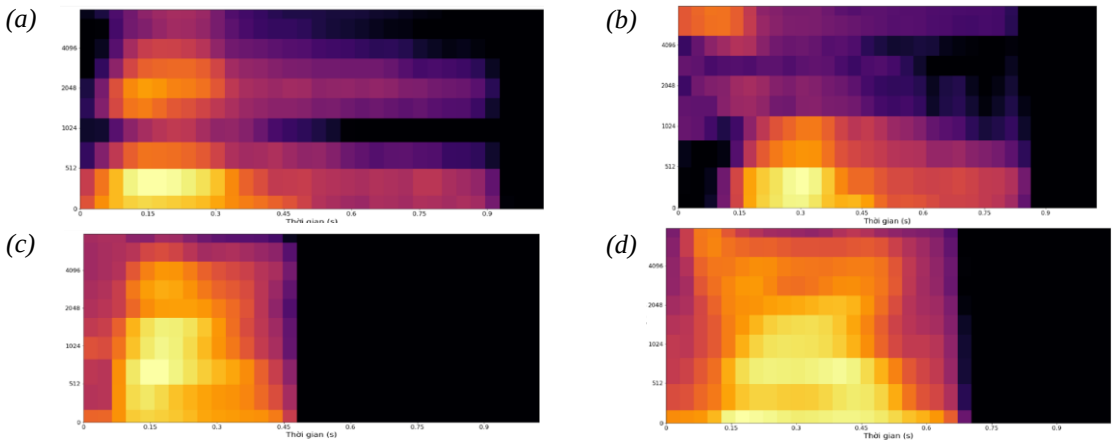
2. Phương pháp nghiên cứu

2.1. Thu thập và tiền xử lý dữ liệu

Việc thu thập dữ liệu là một bước quan trọng trong quá trình xây dựng hệ thống nhận diện lệnh giọng nói, mục tiêu được đưa ra là mô hình có thể học được các đặc trưng âm thanh chính xác và hoạt động ổn định trong điều kiện thực tế cho tiếng Việt. Trong nghiên cứu này, chúng tôi thu thập dữ liệu âm thanh từ giọng nói của nhiều người với các môi trường khác nhau, đảm bảo tính đa dạng và khả năng tổng quát của mô hình. Bộ dữ liệu được xây dựng bao gồm bốn lệnh giọng nói chính để điều khiển xe robot: "Lên" – Di chuyển xe tiến lên; "Xuống" – Di chuyển xe lùi xuống; "Trái" – Rẽ trái; "Phải" – Rẽ phải.

Chúng tôi sử dụng micro chất lượng cao với tần số lấy mẫu (sample rate) 16 kHz và độ sâu 16-bit (bit depth) để đảm bảo độ chính xác trong quá trình trích xuất đặc trưng. Các thiết bị thu âm bao gồm: Microphone gắn ngoài kết nối với máy tính/laptop thông qua cổng USB; Microphone tích hợp trên Raspberry Pi; và Điện thoại di động. Trước khi đưa vào mô hình học sâu, dữ liệu được xử lý như sau: (1) loại bỏ nhiễu nền bằng bộ lọc số để tăng chất lượng tín hiệu; (2) chuẩn hóa biên độ để đưa tất cả tín hiệu về cùng một mức cường độ; (3) cắt phần âm thanh tĩnh lặng ở đầu và cuối của tệp âm thanh nhằm loại bỏ khoảng lặng không cần thiết. Nhóm nghiên cứu đã thu thập được 400 mẫu âm thanh từ 40 người tham gia, với các bản ghi được thực hiện trong nhiều điều kiện môi trường khác nhau. Sau đó, bộ dữ liệu được chia ngẫu nhiên theo tỷ lệ: 80% dùng cho huấn luyện mô hình và 20% còn lại cho việc kiểm thử và đánh giá hiệu suất.

2.2. Trích xuất đặc trưng chuỗi âm thanh sử dụng MFCC

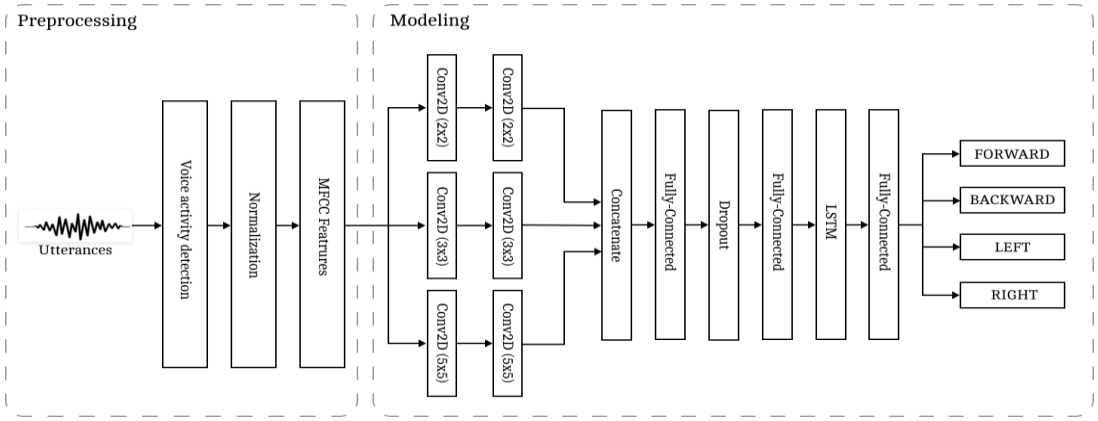


Hình 1. *Mẫu phổ MFCC tiêu biểu cho các hiệu lệnh: (a) tín hiệu "lên", (b) tín hiệu "xuống", (c) tín hiệu "phải" và (d) tín hiệu "trái"*

Sau khi hoàn thành quá trình thu thập và tiền xử lý dữ liệu âm thanh, chúng tôi thực hiện trích xuất đặc trưng MFCC (Mel-Frequency Cepstral Coefficients) để chuyển đổi tín hiệu âm thanh thành một dạng đặc trưng phù hợp cho mô hình học sâu. MFCC là một trong những phương pháp phổ biến nhất để trích xuất đặc trưng từ tín hiệu âm thanh, đặc biệt là trong nhận diện giọng nói, nhờ khả năng mô phỏng cách con người nghe và nhận diện âm thanh theo phổ tần Mel. Tín hiệu âm thanh được chia thành các khung thời gian (frames) có độ dài từ 128 ms, với bước trượt (stride) khoảng 32 ms. Để tránh hiện tượng biến dạng phổ (spectral leakage), mỗi khung tín hiệu được nhân với cửa sổ Hamming [14]. Cửa sổ Hamming giúp làm mượt các cạnh của khung tín hiệu, giảm thiểu nhiễu khi thực hiện biến đổi Fourier. Chuyển đổi từng khung tín hiệu từ miền thời gian sang miền tần số bằng cách áp dụng Biến đổi Fourier nhanh (FFT - Fast Fourier Transform). Để mô phỏng cơ chế cảm nhận âm thanh của tai người, chúng tôi áp dụng bộ lọc Mel, tập trung vào các dải tần quan trọng mà tai người nhạy cảm nhất, đồng thời giảm nhiễu và thông tin dư thừa trong quá trình xử lý tín hiệu âm thanh [15]. Trong nghiên cứu này, chúng tôi sử dụng 13 hệ số MFCC chính cho mỗi khung, tạo thành biểu diễn đặc trưng phổ MFCC. Hình 1 minh hoạ phổ MFCC tiêu biểu cho các tín hiệu âm thanh: "lên", "xuống", "phải", và "trái". Khi thu thập dữ liệu, các mẫu âm thanh nhận được có độ dài không đồng đều, gây khó khăn cho quá trình huấn luyện mô hình. Để khắc phục vấn đề này, chúng tôi đã sử dụng kỹ thuật zero-padding, bổ sung thêm giá trị 0 vào cuối các chuỗi ngắn hơn, giúp chuẩn hóa dữ liệu và đảm bảo tất cả các mẫu có cùng độ dài.

3. Kết quả và bàn luận

3.1. Mô hình nhận diện lệnh



Hình 2. *Mô hình Hybrid CNN-LSTM*

Trong quá trình nghiên cứu, chúng tôi đã tiến hành thử nghiệm và đánh giá hiệu suất của ba mô hình học sâu khác nhau nhằm mục đích nhận diện các lệnh điều khiển bằng giọng nói: CNN, LSTM, và Hybrid (CNN-LSTM).

Mô hình CNN được sử dụng trong nghiên cứu này để trích xuất đặc trưng không gian từ biểu diễn MFCC của tín hiệu âm thanh. CNN hoạt động bằng cách áp dụng các bộ lọc tích chập (convolutional filters) để phát hiện các đặc trưng quan trọng, chẳng hạn như mẫu tần số của giọng nói trong các khung thời gian khác nhau. Cấu trúc mô hình bao gồm hai lớp tích chập (Conv2D) với các bộ lọc kích thước (3×3), mỗi lớp đều đi kèm với hàm kích hoạt ReLU để tăng tính phi tuyến và một lớp MaxPooling để giảm chiều dữ liệu mà vẫn giữ được thông tin quan trọng. Dữ liệu đầu vào là biểu diễn MFCC, được coi như một ảnh hai chiều, giúp CNN tận dụng khả năng xử lý không gian của mình để học các đặc trưng giọng nói theo tần số và thời gian. Sau khi qua các lớp tích chập, dữ liệu được làm phẳng (Flatten) và đưa vào các lớp Fully Connected (Dense Layers) để tổng hợp thông tin và thực hiện phân loại. Lớp đầu ra sử dụng Softmax, với

bốn neuron tương ứng với bốn lệnh giọng nói cần nhận diện. Kết quả thực nghiệm cho thấy CNN đạt độ chính xác 87,5% khi sử dụng learning rate = 0,0001.

Mô hình LSTM được sử dụng để xử lý dữ liệu chuỗi thời gian, cho phép hệ thống nhận diện giọng nói hiệu quả bằng cách ghi nhớ các đặc trưng quan trọng trong khoảng thời gian dài. Trong nghiên cứu này, dữ liệu đầu vào là biểu diễn MFCC, được sắp xếp theo thứ tự thời gian, giúp mô hình học được sự thay đổi của đặc trưng âm thanh theo thời gian. Mô hình gồm hai lớp LSTM, trong đó lớp đầu tiên có 128 units và sử dụng return_sequences = True để duy trì thông tin chuỗi, còn lớp thứ hai có 64 units để tổng hợp dữ liệu trước khi đưa vào lớp fully connected. Lớp đầu ra sử dụng Softmax với bốn neuron, tương ứng với bốn lệnh giọng nói. Kết quả thực nghiệm cho thấy LSTM đạt 93,75% độ chính xác khi sử dụng learning rate = 0,0001, cao hơn so với CNN, nhờ khả năng xử lý tốt hơn các mối quan hệ theo thời gian trong tín hiệu âm thanh.

Mô hình Hybrid CNN-LSTM kết hợp sức mạnh của CNN trong việc trích xuất đặc trưng không gian từ biểu diễn MFCC và LSTM để xử lý thông tin chuỗi thời gian thể hiện như Hình 2. Cấu trúc mô hình bao gồm ba nhánh CNN song song, mỗi nhánh sử dụng bộ lọc tích chập (2×2), (3×3) và (5×5) để thu thập đặc trưng ở nhiều mức độ khác nhau. Các đặc trưng từ ba nhánh này sau đó được ghép lại bằng lớp Concatenate, giúp mô hình có khả năng học tốt hơn các mẫu âm thanh phức tạp. Sau giai đoạn trích xuất đặc trưng, dữ liệu được đưa vào lớp LSTM với 64 units để học thông tin chuỗi từ các khung thời gian của tín hiệu âm thanh. Tiếp theo, dữ liệu được đưa vào các lớp Dense với 256 và 128 neurons. Cuối cùng, lớp đầu ra Softmax có bốn neuron, tương ứng với bốn lệnh giọng nói. Kết quả thực nghiệm cho thấy Hybrid CNN-LSTM đạt 96,25% độ chính xác, vượt trội hơn so với mô hình CNN (87,5%) và LSTM (93,75%), nhờ vào sự kết hợp tối ưu giữa trích xuất đặc trưng và xử lý chuỗi thời gian. Bảng 1 so sánh các mô hình theo Accuracy, Average Precision và Average Recall.

Bảng 1. *So sánh hiệu suất các mô hình nhận diện lệnh giọng nói*

Mô hình	Accuracy (%)	Average Precision (%)	Average Recall (%)
CNN	87,50	87,50	87,30
LSTM	93,75	92,83	94,65
CNN-LSTM	96,25	96,25	96,24

Kết quả nghiên cứu cho thấy mô hình Hybrid CNN-LSTM là phương pháp hiệu quả nhất trong việc nhận diện lệnh giọng nói để điều khiển xe robot. Mô hình tận dụng sức mạnh của CNN để trích xuất đặc trưng không gian từ biểu diễn MFCC, đồng thời sử dụng LSTM để xử lý mối quan hệ chuỗi thời gian trong tín hiệu giọng nói, giúp phân loại lệnh chính xác ngay cả trong môi trường có nhiễu.

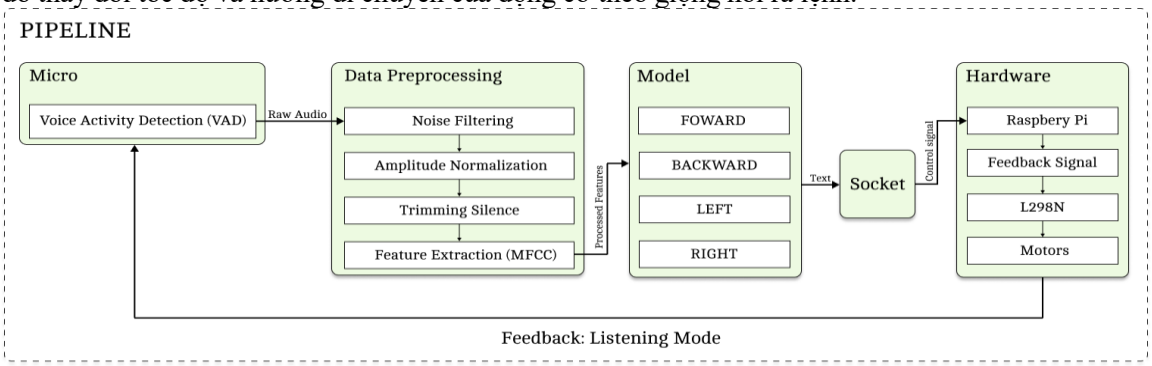
3.2. Tích hợp vào hệ thống điều khiển xe robot

Hệ thống điều khiển xe robot bằng giọng nói trong nghiên cứu này được thiết kế nhằm mục tiêu xây dựng một hệ thống tương tác người-máy thông qua giọng nói mà không cần tiếp xúc vật lý. Hệ thống bao gồm ba thành phần chính: bộ thu âm và xử lý tín hiệu giọng nói, mô hình học sâu nhận diện lệnh và bộ điều khiển động cơ xe robot. Toàn bộ hệ thống được triển khai theo mô hình điện toán biên, giúp giảm độ trễ và tăng tính khả dụng trong môi trường thực tế.

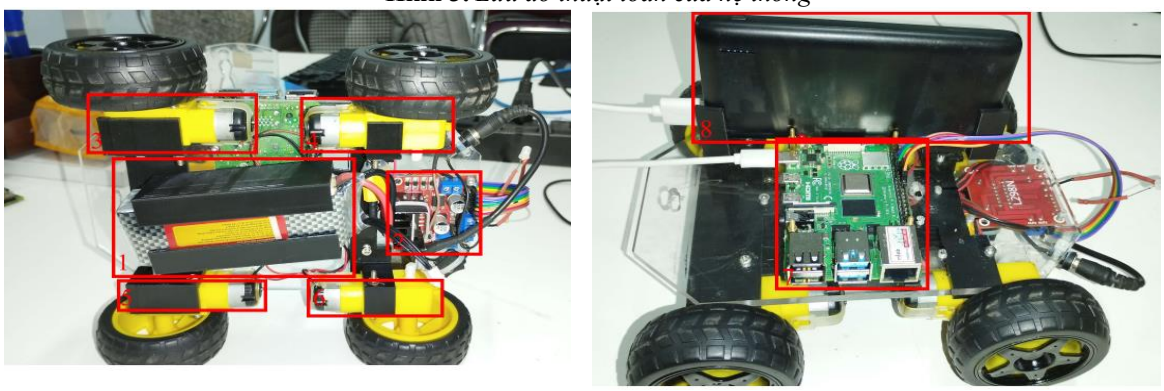
Lưu đồ thuật toán của hệ thống điều khiển robot thể hiện như Hình 3. Hệ thống nhận diện giọng nói tiếng Việt được thu thập sử dụng mic của máy tính xách tay. Tín hiệu giọng nói sau thu thập được xử lý sử dụng mô hình Hybrid CNN-LSTM trực tiếp trên máy tính xách tay. Sau đó tín hiệu xử lý sẽ được gửi tới Bo mạch điều khiển Raspberry Pi để trực tiếp ra lệnh cho bo mạch L298N điều khiển 4 motor của robot.

Hệ thống phần cứng thể hiện như Hình 4 của robot gồm các thành phần sau: (1) Nguồn 12V - 2A được sử dụng để cấp nguồn cho 4 motor DC 6V và module L298N; (2) Module L298N được sử dụng để thay đổi tốc độ và chiều quay của bánh xe; (3,4,5,6) các Motor DC 6V; (7) Bo mạch Raspberry 3 được kết nối với module L298N để điều khiển các động cơ; (8) Nguồn 5V - 2A được

sử dụng để cấp nguồn cho bo mạch. Microphone USB được sử dụng để cắm vào Laptop để thu âm lệnh giọng nói và và xử lý tín hiệu âm thanh thu được sử dụng phương pháp Hybrid CNN-LSTM nhận diện lệnh giọng nói để điều khiển xe robot. Dữ liệu sau xử lý âm thanh được gửi đến Raspberry Pi 3 của robot sử dụng giao tiếp Socket [16] (Socket Communication). Giao tiếp Socket đảm bảo tốc độ truyền tin nhanh, giảm độ trễ khi thực hiện lệnh điều khiển. Raspberry Pi 3 của xe robot nhận lệnh và sử dụng các tín hiệu từ GPIO của Raspberry Pi để điều khiển module L298N, từ đó thay đổi tốc độ và hướng di chuyển của động cơ theo giọng nói ra lệnh.



Hình 3. *Lưu đồ thuật toán của hệ thống*



(a) Mặt dưới robot *(b) Mặt trên robot*

Hình 4. *Phần cứng của robot: (a) Mặt dưới và (b) mặt trên robot*

Hệ thống đã được kiểm thử trên một xe robot thử nghiệm với bốn lệnh điều khiển: "Lên" (di chuyển tiến), "Xuống" (di chuyển lùi), "Trái" (rẽ trái) và "Phải" (rẽ phải). Kết quả cho thấy, xe robot tự hành di chuyển chính xác theo lệnh giọng nói với độ trễ điều khiển trung bình khoảng 01 giây. Hệ thống điều khiển xe robot bằng giọng nói trong nghiên cứu này đã ứng dụng thành công mô hình học sâu trong nhận diện giọng nói, đặc biệt xử lý các lệnh tiếng Việt, đồng thời tận dụng điện toán biên để xử lý câu lệnh nhanh chóng. Kết quả điều khiển xe robot được thể hiện trong video minh họa: https://youtu.be/zJp0PE88iCg.

Việc kết hợp mô hình Hybrid CNN-LSTM, giao tiếp Socket và phần cứng nhúng Raspberry Pi 3 giúp tạo ra một hệ thống nhanh, chính xác, ổn định và giá thành thấp. Hệ thống có tiềm năng mở rộng cho các ứng dụng trong nhà thông minh, robot tự hành và các hệ thống điều khiển không tiếp xúc trong tương lai ra lệnh bằng giọng nói sử dụng tiếng Việt.

4. Kết luận

Nghiên cứu này đã đề xuất và triển khai một hệ thống điều khiển xe robot bằng giọng nói dựa trên học sâu và điện toán biên, nhằm giảm sự phụ thuộc vào điện toán đám mây và cải thiện khả năng vận hành trong thời gian ngắn. Hệ thống sử dụng mô hình Hybrid CNN-LSTM, kết hợp giữa CNN để trích xuất đặc trưng không gian từ MFCC và LSTM để xử lý mối quan hệ chuỗi thời gian,

giúp đạt độ chính xác nhận diện lệnh 96,25%, vượt trội so với các mô hình CNN và LSTM riêng lẻ. Khi triển khai mô hình trên phần cứng Raspberry Pi kết hợp với giao thức Socket Communication, hệ thống xử lý lệnh với độ trễ thấp, đảm bảo độ tin cậy trong môi trường thực tế. Kết quả thực nghiệm cho thấy hệ thống có thể nhận diện chính xác các lệnh "Lên", "Xuống", "Trái" và "Phải" với khả năng hoạt động ổn định ngay cả khi có nhiễu âm nhẹ. Xe robot di chuyển chính xác theo lệnh giọng nói với độ trễ điều khiển trung bình khoảng 01 giây. Việc tối ưu hóa mô hình bằng TinyAI giúp giảm tải tính toán, mở ra tiềm năng ứng dụng trong các hệ thống nhúng và IoT có tài nguyên hạn chế, đặc biệt trong việc ra lệnh cho các hệ thống sử dụng tiếng Việt. Trong tương lai, hệ thống sẽ tiếp tục được nghiên cứu, tối ưu thông qua cải tiến thuật toán và phần cứng với mục đích giảm độ trễ của mô hình, hướng tới đáp ứng yêu cầu xử lý thời gian thực.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 854-864, 2016.

[2] A. V. Dastjerdi and R. Buyya, "Fog computing: Helping the Internet of Things realize its potential," *Computer*, vol. 49, no. 8, pp. 112-116, 2016.

[3] M. Prasad *et al.*, "Voice-controlled autonomous navigation system for mobile robots in dynamic environments," *IEEE Transactions on Robotics and Automation*, vol. 39, no. 4, pp. 1852-1867, 2023.

[4] A. Khan and R. Johnson, "Efficient voice control systems for IoT devices and smart homes: A comprehensive review," *Internet of Things Journal*, vol. 11, no. 2, pp. 345-362, 2024.

[5] L. Zhang *et al.*, "Challenges and solutions for on-device audio processing in resource-constrained edge devices," *IEEE Transactions on Edge Computing*, vol. 8, no. 3, pp. 512-528, 2023.

[6] Y. Chen and P. Smith, "Lightweight deep learning architectures for real-time speech recognition on edge devices," *Neural Computing and Applications*, vol. 36, no. 1, pp. 78-93, 2024.

[7] Q. Vu *et al.*, "Vietnamese Automatic Speech Recognition: the FLaVoR Approach," in *Proc. International Symposium on Chinese Spoken Language Processing*, Singapore, December 2006, vol. 4274, pp. 464–474.

[8] T. Le, H. Nguyen, and Q. Vu, "Progress in Transcription of Vietnamese Broadcast News," in *Proc. International Conference on Communications and Electronics (ICCE'06)*, October 2006, pp. 300–304.

[9] T.-S. Phan, T.-C. Duong, A.-T. Dinh, T.-T. Vu, and C.-M. Luong, "Improvement of naturalness for an HMM-based Vietnamese speech synthesis using the prosodic information," *Proceedings of the 2013 RIVF International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, 2013, pp. 276–281.

[10] Q. B. Nguyen, T. T. Vu, and C. M. Luong, "Improving acoustic model for Vietnamese large vocabulary continuous speech recognition system using deep bottleneck features," *Proceedings of the Sixth International Conference on Knowledge and Systems Engineering (KSE 2014)*, 2015, pp. 49–60.

[11] P. Hung, T. Minh, L. Hoang, and M. Phan, "Vietnamese speech command recognition using recurrent neural networks," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, 2019, doi: 10.14569/IJACSA.2019.0100728.

[12] T.-T. Le, L. T. Nguyen, and D. Q. Nguyen, "PhoWhisper: Automatic speech recognition for Vietnamese," *arXiv preprint arXiv:2406.02555*, 2024.

[13] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51-83, Jan. 1978.

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980.

[15] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Communication*, vol. 54, no. 4, pp. 543-565, 2012.

[16] M. Xue and C. Zhu, "The Socket Programming and Software Design for Communication Based on Client/Server," *2009 Pacific-Asia Conference on Circuits, Communications and Systems*, Chengdu, China, 2009, pp. 775-777, doi: 10.1109/PACCS.2009.89.