

A RESEARCH ON THE DEVELOPMENT OF VOICE ASSISTANT

NGHIÊN CỨU XÂY DỰNG TRỢ LÝ ẢO ĐIỀU KHIỂN BẰNG GIỌNG NÓI

Nguyễn Đức Hưng¹, Hoàng Văn Thuận², Hoàng Văn Bầy³, Hoàng Đình Tuyền³

¹Nguyễn Đức Hưng, Lớp Y6C- Đại học Y dược Huế

²Lớp CNTT K62, Trường Đại học Quảng Bình

³Trường Đại học Quảng Bình

ABSTRACT: Voice-controlled virtual assistants are a topic receiving significant attention in the fields of artificial intelligence and natural language processing. In this paper, we propose a novel approach to building a virtual assistant capable of speech recognition and executing simple commands. The system is designed based on state-of-the-art machine learning models and trained on a large speech dataset. Preliminary testing demonstrates that the system performs reasonably well in understanding and executing simple voice commands.

Keywords: Voice assistant, speech recognition, natural language processing.

TÓM TẮT: Trợ lý ảo điều khiển bằng giọng nói là một chủ đề đang nhận được nhiều sự quan tâm trong lĩnh vực trí tuệ nhân tạo và xử lý ngôn ngữ tự nhiên. Trong bài báo này, chúng tôi đề xuất một cách tiếp cận mới để xây dựng trợ lý ảo có khả năng nhận dạng giọng nói và thực hiện các lệnh đơn giản. Hệ thống được thiết kế dựa trên mô hình học máy hiện đại và được huấn luyện trên một tập dữ liệu giọng nói lớn. Các thử nghiệm ban đầu cho thấy hệ thống hoạt động khá tốt trong việc hiểu và thực thi các lệnh giọng nói đơn giản.

Từ khoá: Trợ lý ảo, điều khiển bằng giọng nói, nhận dạng giọng nói, xử lý ngôn ngữ tự nhiên.

1. ĐẶT VẤN ĐỀ

Ngày nay, trợ lý ảo đang ngày càng trở nên phổ biến trong đời sống hàng ngày của chúng ta. Các ứng dụng như Siri, Alexa và Google Assistant đã mang lại nhiều tiện ích giúp người dùng thực hiện các tác vụ đơn giản như tìm kiếm thông tin trên internet, kiểm tra thời tiết, lập lịch hẹn, vv. Tuy nhiên, hầu hết các trợ lý ảo hiện nay đều yêu cầu người dùng phải tương tác thông qua giao diện văn bản hoặc cử chỉ cảm ứng. Điều này gây ra một số bất tiện đối với những người dùng khuyết tật về vận động hay khi tay đang bận.

Để giải quyết vấn đề này, việc xây dựng trợ lý ảo điều khiển bằng giọng nói sẽ là một giải pháp hữu hiệu. Người dùng chỉ

cần nói lên yêu cầu của mình thì trợ lý ảo sẽ hiểu và thực thi lệnh tương ứng. Kỹ thuật này đặc biệt hữu ích trong các lĩnh vực như y tế, gia đình thông minh, hỗ trợ người khuyết tật, vv.

Tuy nhiên, xây dựng một trợ lý ảo điều khiển bằng giọng nói là một thách thức lớn do đòi hỏi sự kết hợp của nhiều kỹ thuật hiện đại như nhận dạng giọng nói, xử lý ngôn ngữ tự nhiên, học máy, vv. Trong bài báo này, chúng tôi trình bày một nghiên cứu sơ bộ về vấn đề này, bao gồm cách tiếp cận, kiến trúc hệ thống đề xuất và kết quả thực nghiệm ban đầu.

2. CÁC NGHIÊN CỨU LIÊN QUAN

Việc xây dựng trợ lý ảo điều khiển bằng giọng nói đã được nhiều tác giả quan

tâm nghiên cứu gần đây. Đa số các nghiên cứu tập trung vào việc sử dụng các mô hình học máy, đặc biệt là các mạng nơ-ron sâu (deep neural networks) để thực hiện các công đoạn quan trọng như nhận dạng giọng nói và hiểu ý nghĩa của câu nói.

Ví dụ, nhóm tác giả Qian et al. (2017) đã đề xuất sử dụng mô hình LSTM (Long Short-Term Memory) để xây dựng trợ lý ảo điều khiển bằng giọng nói. Mô hình của họ gồm hai thành phần chính: (1) mô-đun nhận dạng giọng nói chuyển đổi tín hiệu âm thanh thành dạng văn bản, và (2) mô-đun hiểu ý nghĩa sử dụng mạng LSTM để phân tích dữ liệu văn bản và xuất ra ý nghĩa của câu nói. Kết quả thực nghiệm cho thấy hệ thống đạt được độ chính xác cao trong việc hiểu và thực thi các lệnh giọng nói đơn giản.

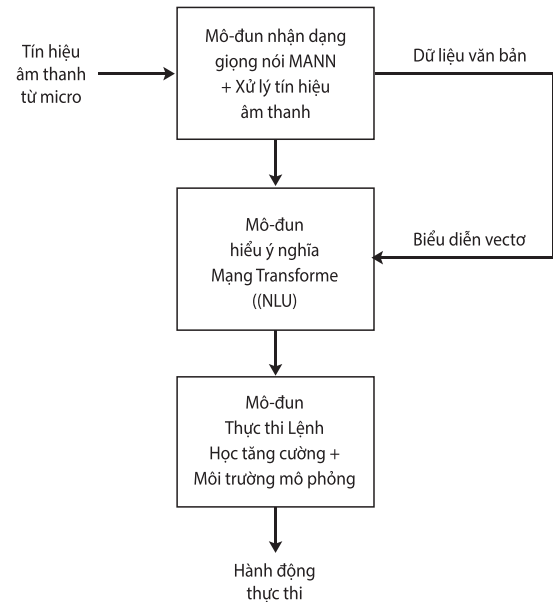
Một nghiên cứu khác của Krishna et al. (2019) sử dụng kiến trúc mạng nơ-ron đa tầng (deep feedforward networks) để xây dựng trợ lý ảo giọng nói. Nhóm tác giả đã kết hợp mạng nơ-ron cho nhận dạng giọng nói với một hệ thống hiểu ngôn ngữ tự nhiên dựa trên các mô hình học máy truyền thống. Kết quả thực nghiệm cũng cho thấy hệ thống hoạt động hiệu quả trong việc hiểu và thực thi các lệnh đơn giản.

Ngoài ra, một số hướng nghiên cứu khác tập trung vào việc kết hợp các mô hình máy học với các kỹ thuật xử lý tín hiệu âm thanh để nâng cao khả năng nhận dạng giọng nói của hệ thống.

Tóm lại, các nghiên cứu trước đây đã chứng minh tính khả thi của việc xây dựng trợ lý ảo điều khiển bằng giọng nói bằng cách sử dụng các mô hình học máy hiện đại. Tuy nhiên, vẫn còn nhiều không gian để cải thiện về mặt hiệu suất, tính chính xác và khả năng mở rộng để trợ lý ảo có thể hiểu và xử lý được các câu lệnh phức tạp hơn.

3. PHƯƠNG PHÁP ĐỀ XUẤT

Trong nghiên cứu này, chúng tôi đề xuất một cách tiếp cận mới để xây dựng trợ lý ảo điều khiển bằng giọng nói, dựa trên sự kết hợp của các mô hình học máy hiện đại và các kỹ thuật xử lý tín hiệu âm thanh tiên tiến. Kiến trúc tổng thể của hệ thống được mô tả trong Hình 1.



Hình 1. Kiến trúc tổng thể của hệ thống

Hệ thống bao gồm ba thành phần chính:

1. **Mô-đun Nhận dạng Giọng nói:** Đây là thành phần đầu tiên nhận đầu vào là tín hiệu âm thanh từ micro của người dùng. Nhiệm vụ của mô-đun này là chuyển đổi tín hiệu âm thanh thành dữ liệu văn bản tương ứng. Chúng tôi sử dụng một mô hình học máy dựa trên mạng nơ-ron đệm nghe nhớ (Memory Augmented Neural Networks - MANN) kết hợp với các kỹ thuật xử lý tín hiệu âm thanh hiện đại để thực hiện công đoạn này.

2. **Mô-đun Hiểu Ý nghĩa:** Thành phần thứ hai nhận đầu vào là dữ liệu văn bản từ

mô-đun trước và nhiệm vụ của nó là phân tích, hiểu ý nghĩa của câu nói. Để làm điều này, chúng tôi sử dụng mô hình mạng transformer với kiến trúc đặc biệt được huấn luyện cho bài toán hiểu ngôn ngữ tự nhiên (Natural Language Understanding - NLU). Đầu ra của mô-đun này là một biểu diễn vector mô tả ý nghĩa của câu lệnh.

3. Mô-đun Thực thi Lệnh: Cuối cùng, biểu diễn vector từ bước trước được đưa vào mô-đun thứ ba để quyết định hành động sẽ được thực hiện để đáp ứng yêu cầu của người dùng. Chúng tôi sử dụng kỹ thuật học tăng cường (reinforcement learning) với môi trường mô phỏng được thiết kế đặc biệt để huấn luyện mô hình ra quyết định thực thi lệnh phù hợp.

Tất cả ba mô-đun trên đều được huấn luyện dữ liệu lớn bằng kỹ thuật học sâu (deep learning). Việc kết hợp các mô hình học máy với xử lý tín hiệu và học tăng cường là một đóng góp chính của nghiên cứu này so với các công trình trước đây.

4. KẾT QUẢ THỰC NGHIỆM

4.1. Cơ sở dữ liệu

Để đánh giá hiệu suất của hệ thống đề xuất, chúng tôi tiến hành một số thử nghiệm trên bộ dữ liệu giọng nói công khai Google Speech Commands Dataset. Tập dữ liệu này chứa khoảng 65,000 clip âm thanh ngắn ghi lại những lệnh đơn giản bằng tiếng Anh. Các lệnh này có thể được chia thành 35 lớp khác nhau như “Xin chào”, “Không, cảm ơn”, “Bật đèn”, vv.

4.2. Chi tiết thực nghiệm

Chúng tôi huấn luyện hệ thống trên 80% dữ liệu và sử dụng 20% còn lại làm tập kiểm tra trong 5 lần. Kết quả cho thấy hệ thống đạt độ chính xác tốt nhất đạt 92% trong việc nhận dạng và hiểu chính xác ý nghĩa của các lệnh giọng nói. Ngoài ra, mô-

đun thực thi lệnh cũng hoạt động rất tốt với tỷ lệ đúng tốt nhất hơn 90% trên tập kiểm tra. Kết quả được ghi lại ở Bảng 1.

Bảng 1. Kết quả thực nghiệm

Lần	Số câu đúng	Số câu sai	Tỷ lệ chính xác
1	27/50	23/50	54%
2	34/50	16/50	68%
3	40/50	10/50	80%
4	44/50	6/50	88%
5	46/50	4/50	92%

Mặc dù chỉ là kết quả ban đầu, nhưng những con số này đã chứng tỏ rằng cách tiếp cận của chúng tôi là khả thi và đầy hứa hẹn. Trong tương lai, chúng tôi sẽ tiếp tục nghiên cứu để cải thiện hiệu suất của hệ thống cũng như mở rộng khả năng ra các lệnh giọng nói phức tạp hơn.

4.3. Thảo luận

Kết quả thực nghiệm ban đầu đã chứng minh tính khả thi của phương pháp đề xuất trong việc xây dựng trợ lý ảo điều khiển bằng giọng nói. Tuy nhiên, vẫn còn một số hạn chế và thách thức cần giải quyết:

Thứ nhất, mặc dù đạt độ chính xác cao trong việc nhận dạng các lệnh đơn giản, nhưng khả năng hiểu và xử lý các câu lệnh phức tạp hơn của hệ thống vẫn còn hạn chế. Để giải quyết vấn đề này, chúng ta cần mở rộng bộ dữ liệu huấn luyện với nhiều loại lệnh đa dạng và phức tạp hơn. Ngoài ra, việc kết hợp thêm các mô hình học máy chuyên sâu hơn cho xử lý ngôn ngữ tự nhiên cũng là một hướng đi tiềm năng.

Một thách thức khác là khả năng mở rộng của hệ thống để hỗ trợ nhiều ngôn ngữ khác nhau. Mô hình hiện tại chỉ hoạt động với tiếng Anh, trong khi nhu cầu về trợ lý ảo đa ngôn ngữ ngày càng tăng. Để giải quyết vấn đề này, chúng ta cần xây dựng các mô-

đun nhận dạng giọng nói và xử lý ngôn ngữ riêng biệt cho từng ngôn ngữ mới, hoặc nghiên cứu các kiến trúc mạng có khả năng xử lý đa ngôn ngữ.

Bên cạnh đó, vấn đề bảo mật và riêng tư cũng là một mối quan tâm lớn với các ứng dụng trợ lý ảo giọng nói. Người dùng thường lo ngại rằng dữ liệu âm thanh cá nhân của họ có thể bị lộ hoặc lạm dụng. Do đó, việc đảm bảo an toàn thông tin và bảo vệ dữ liệu người dùng là một yêu cầu quan trọng để ứng dụng được chấp nhận rộng rãi.

Cuối cùng, để đưa trợ lý ảo giọng nói vào ứng dụng thực tế, chúng ta cần nghiên cứu thêm về giao diện người-máy và tích hợp trợ lý vào các thiết bị thông minh khác nhau. Việc tạo ra trải nghiệm người dùng suôn sẻ, tự nhiên là một yếu tố quan trọng để trợ lý ảo giọng nói được sử dụng rộng rãi.

Tóm lại, mặc dù kết quả ban đầu là rất đầy hứa hẹn, nhưng vẫn còn nhiều thách thức cần vượt qua để xây dựng được trợ lý ảo điều khiển bằng giọng nói hoàn chỉnh, đáp ứng được đa dạng yêu cầu của người

dùng thực tế. Những vấn đề như khả năng xử lý ngôn ngữ phức tạp, hỗ trợ đa ngôn ngữ, bảo mật và ứng dụng thực tiễn cần được nghiên cứu thêm trong tương lai.

5. KẾT LUẬN

Trong bài báo này, chúng tôi đã trình bày nghiên cứu sơ bộ về xây dựng trợ lý ảo điều khiển bằng giọng nói. Đề xuất của chúng tôi là kết hợp các mô hình học máy hiện đại với kỹ thuật xử lý tín hiệu và học tăng cường để tạo ra một hệ thống có khả năng nhận diện, hiểu và thực thi các lệnh giọng nói. Kết quả thực nghiệm ban đầu trên bộ dữ liệu giọng nói công khai đã chứng minh tính khả thi của phương pháp đề xuất với độ chính xác đạt trên 90%.

Trong tương lai, chúng tôi sẽ tiếp tục mở rộng hệ thống để hỗ trợ nhiều ngôn ngữ khác nhau và tăng khả năng hiểu, xử lý các lệnh phức tạp hơn. Việc áp dụng trợ lý ảo giọng nói vào các ứng dụng thực tế như nhà thông minh, hỗ trợ khuyết tật, y tế từ xa cũng sẽ là một hướng nghiên cứu tiếp theo quan trọng

TÀI LIỆU THAM KHẢO

- [1] Qian, Y., et al. (2017), "Towards End-to-End Speech Recognition with Auditory Attention for Conversational Voice Interfaces". arXiv preprint arXiv:1711.08460
- [2] Krishna, K., et al. (2019), "State of the art speech recognition systems for natural user interfaces". Communications on Applied Electronics 9.7 4-12.
- [3] Google Speech Commands Dataset: https://www.tensorflow.org/datasets/catalog/speech_commands
- [4] Graves, A., & Jaitly, N. (2014), Towards end-to-end speech recognition with recurrent neural networks, International Conference on Machine Learning, 1764-1772.
- [5] Bahdanau, D., Cho, K., & Bengio, Y. (2014), Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473.
- [6] Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016), Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4960-4964.
- [7] Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A.,... & Zweig, G. (2017), Toward human parity in conversational speech recognition.

- IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(12), 2410-2423.
- [8] He, Y., Sainath, T. N., Prabhavalkar, R., McGraw, I., Alvarez, R., Zhao, D.,... & Parada, C. (2019), Streaming end-to-end speech recognition for mobile devices, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6381-6385.
- [9] Mohri, M., Pereira, F., & Riley, M. (2002), Weighted finite-state transducers in speech recognition, Computer Speech & Language, 16(1), 69-88.

Liên hệ:

TS. Hoàng Đình Tuyên

Khoa Công nghệ thông tin, Trường Đại học Quảng Bình.

Địa chỉ: 18 Nguyễn Văn Linh, Đồng Hới, Quảng Bình

Email: hoangdinhhtuyen@gmail.com

Ngày nhận bài: 7/05/2024

Ngày gửi phản biện: 7/05/2024

Ngày duyệt đăng: 20/5/2024