fit@hcmus

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

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

XÂY DỰNG MÔ HÌNH NHẬN DẠNG ÂM THANH TIẾNG VIỆT

(Building Vietnamese speech recognition model)

1 THÔNG TIN CHUNG

Người hướng dẫn:

– TS. Ngô Huy Biên (Khoa Công nghệ Thông tin)

Nhóm Sinh viên thực hiện:

1. Trần Ngọc Quang (MSSV:1712706)
2. Nguyễn Hoàng Quyên (MSSV:1712712)

Loại đề tài: Nghiên cứu

Thời gian thực hiện: Từ *9/2020* đến *3/2021*

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

- Xây dựng mô hình nhận dạng âm thanh Tiếng Việt.
- Vai trò sinh viên: Data Scientist, Data Collector, Developer, Tester, Project Manager/Scrum Master.

- Kỹ năng yêu cầu : Python programming, Machine learning algorithms.
- Ngữ cảnh: Tiếng Việt được coi là một ngôn ngữ khó học với người nước ngoài bởi ngữ pháp, thanh điệu và đặc trưng vùng miền. Máy tính cũng giống như người nước ngoài - để nó nghe hiểu và diễn giải được giọng nói tiếng Việt thành dạng văn bản không phải là việc dễ dàng. Nhận dạng tiếng nói đóng vai trò quan trọng trong giao tiếp giữa người và máy. Nó giúp máy móc hiểu và thực hiện các hiệu lệnh của con người. Hiện nay trên thế giới, lĩnh vực nhận dạng tiếng nói đã đạt được nhiều tiến bộ vượt bậc. Đối với ngôn ngữ tiếng Anh, việc nhận dạng có thể đạt độ chính xác tới 99%. Khóa luận này nhằm mục đích nghiên cứu, xây dựng và đào tạo một mô hình nhận dạng giọng nói Tiếng Việt với mục tiêu độ chính xác tối thiểu 75%.

2.2 Mục tiêu đề tài

- Viết 120 trang luận văn theo đúng chuẩn yêu cầu và trích dẫn các tài liệu tham khảo đầy đủ.
- Bản luận văn trình bày lí thuyết nền tảng và giải pháp để xử lý việc nhận một tập tin âm thanh tiếng Việt và xuất ra nội dung văn bản ở dạng Tiếng Việt.
- Xây dựng, thu thập dữ liệu, và đào tạo mô hình để nhận một tập tin âm thanh tiếng Việt và xuất ra nội dung ở dạng văn bản.
- Xây dựng ứng dụng web chuyển đổi từ giọng nói sang văn bản Tiếng Việt.
- Cải tiến độ chính xác của mô hình với mục tiêu là 75%.

2.3 Phạm vi của đề tài

- Mô hình nhận dạng giọng nói chuyển tập tin âm thanh Tiếng Việt sang văn bản Tiếng Việt.
- Sản phẩm demo được xây dựng trên nền tảng web.

2.4 Cách tiếp cận dự kiến

Các nghiên cứu về nhận dạng tiếng nói dựa trên ba nguyên tắc cơ bản:

- Tín hiệu tiếng nói được biểu diễn chính xác bởi các giá trị phổ trong một khung thời gian ngắn.
- Nội dung của tiếng nói được biểu diễn dưới dạng một dãy các ký hiệu ngữ âm.
- Nhận dạng tiếng nói là một quá trình nhận thức. Thông tin về ngữ nghĩa và suy đoán có giá trị trong quá trình nhận dạng tiếng nói, nhất là khi thông tin về âm học là không rõ ràng.

2.4.1 Tiếp cận âm học

Phương pháp này dựa trên lý thuyết về Âm học-Ngữ âm học. Lý thuyết đó cho biết có sự tồn tại của các đơn vị ngữ âm trong ngôn ngữ tiếng nói, các đơn vị ngữ âm này được biểu diễn đặc trưng bởi một tập hợp những thuộc tính thể hiện trong tín hiệu âm thanh hay biểu diễn phổ theo thời gian. Đặc điểm của phương pháp nhận dạng tiếng nói theo hướng tiếp cận Âm học-Ngữ âm học:

- Người thiết kế phải có kiến thức khá sâu rộng về Âm học-Ngữ âm học.
- Phân tích các khối ngữ âm mang tính trực giác, thiếu chính xác.
- Phân loại tiếng nói theo các khối ngữ âm thường không tối ưu do khó sử dụng các công cụ toán học để phân tích.

2.4.2 Tiếp cận nhận dạng mẫu thống kê

Đây là một phương pháp sử dụng trực tiếp các mẫu tiếng nói (chính là đoạn tiếng nói cần nhận dạng) mà không cần xác định thật rõ các đặc trưng và cũng không cần phân đoạn tín hiệu. Phương pháp này cũng có 2 bước:

- Bước 1: thu thập các mẫu tiếng nói: sử dụng tập mẫu tiếng nói (cơ sở dữ liệu mẫu tiếng nói) để đào tạo các mẫu tiếng nói đặc trưng (mẫu tham chiếu) hoặc các tham số hệ thống.
- Bước 2: nhận dạng mẫu: đối sánh mẫu tiếng nói từ ngoài với các mẫu đặc trưng để ra quyết định.

Cơ sở dữ liệu tiếng nói cho đào tạo có đủ các mẫu cần nhận dạng thì quá trình đào tạo có thể xác định chính xác các đặc tính âm học của mẫu, từ đó tăng độ chính xác cho mô hình.

Tiếp cận nhận dạng mẫu thường được lựa chọn cho các ứng dụng nhận dạng tiếng nói bởi các lý do sau:

- Tính dễ sử dụng và dễ hiểu trong thuật toán.
- Tính bất biến và khả năng thích nghi đối với những từ vững, người sử dụng, các tập hợp đặc trưng, các thuật toán so sánh mẫu và các quy tắc quyết định khác nhau.
- Khẳng định tính năng cao trong thực tế.

  Hiện nay, một số kỹ thuật nhận dạng mẫu được áp dụng thành công trong nhận dạng tiếng nói là lượng tử hóa vector, so sánh thời gian động (DTW), mô hình Markov ẩn (HMM), mạng nơron nhân tạo (ANN), sử dụng cơ sở tri thức,...

2.4.3 Tiếp cận trí tuệ nhân tạo

Tiếp cận trí tuệ nhân tạo là phương pháp cố gắng "máy móc hóa" chức năng nhận dạng theo cách mà con người áp dụng trí thông minh của mình trong việc quan sát, phân tích và thực hiện những quyết định trên các đặc trưng âm học của tín hiệu. Cách tiếp cận này kết hợp các cách tiếp cận trên nhằm tận dụng tối đa các ưu điểm của chúng.

Đặc điểm của các hệ thống nhận dạng theo cách tiếp cận trí tuệ nhân tạo:

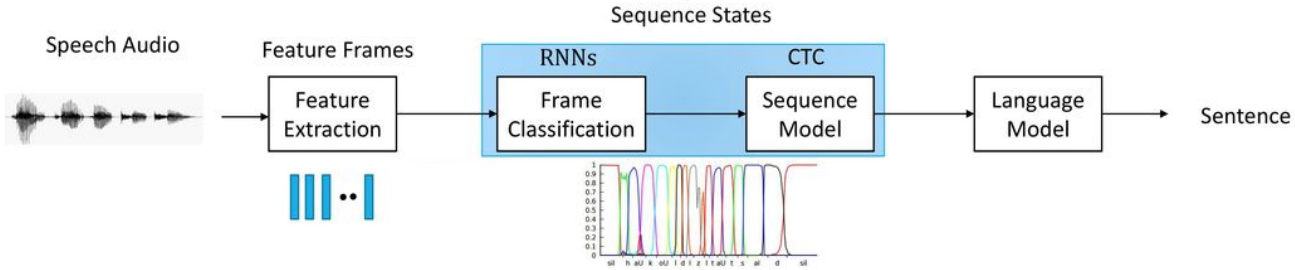
- Sử dụng hệ chuyên gia để phân đoạn, gán nhãn ngữ âm. Điều này làm đơn giản hóa hệ thống so với phương pháp nhận dạng ngữ âm.
- Sử dụng mạng nơron nhân tạo để học mối quan hệ giữa các ngữ âm, sau đó dùng nó để nhận dạng tiếng nói.

Đây sẽ là hướng tiếp cận tương lai của nhận dạng tiếng nói.

2.4.4 Mô hình dự kiến thực hiện trong đề tài

Đề tài dự kiến xây dựng hệ thống nhận dạng giọng nói sử dụng kiến trúc mô hình nhận dạng đầu cuối (End-to-End).

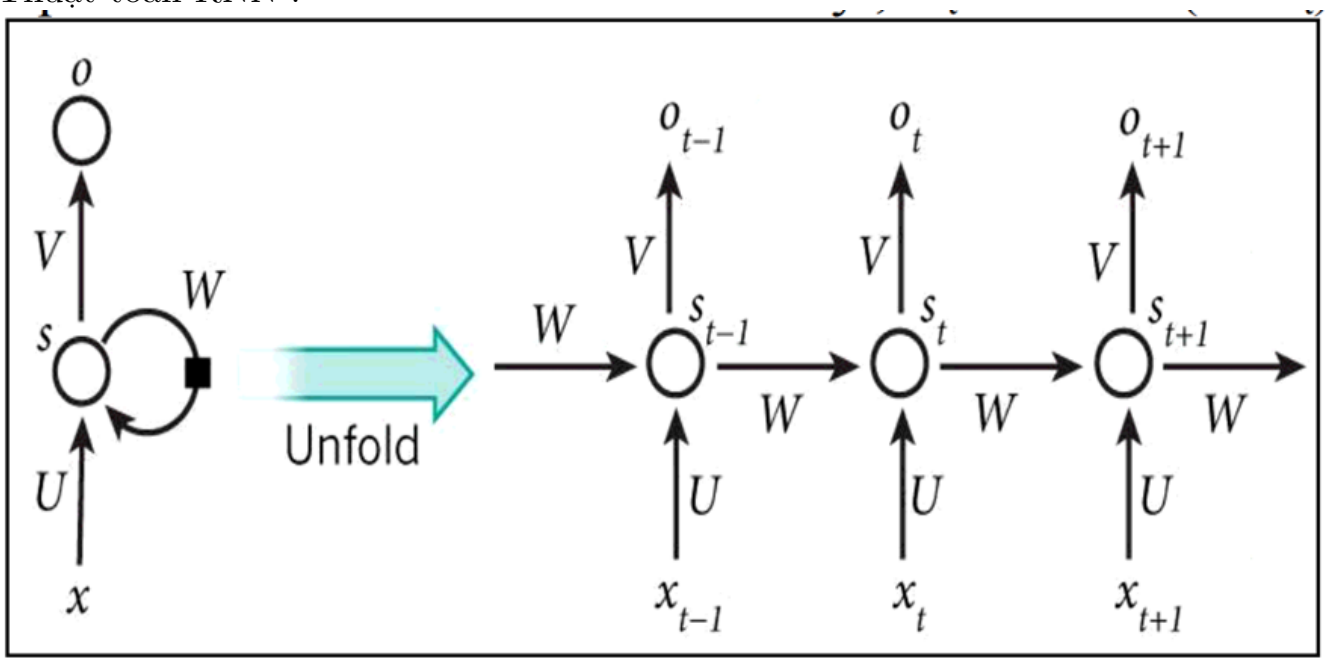
Đề tài này dự kiến sử dụng mô hình Mạng nơ-ron hồi quy (RNN – Recurrent Neural Network) và sử dụng Connectionist temporal classification(CTC) để dự đoán đầu ra:



- Đầu vào của hệ thống là đoạn âm thanh thô chưa được xử lý.
- Feature Extraction là quá trình phân tích các đặc trưng (tham số) tiếng nói bằng cách loại bỏ những thông tin không quan trọng như tiếng ồn của môi trường, nhiễu trên đường truyền, các đặc điểm riêng biệt của người nói...Tiếng nói được phân tích theo các khung thời gian gọi là frame. Kết quả ra của giai đoan này là các vector đặc tính của mỗi khung tín hiệu tiếng nói. Phương pháp chọn đặc trưng tiếng nói dự kiến sử dụng trong đề tài là MFCC (melscale frequency cepstral coefficients).
- Ý tưởng chính của RNN (Recurrent Neural Network) là sử dụng chuỗi các thông tin. Trong các mạng nơ-ron truyền thống tất cả các đầu vào và cả đầu ra là độc lập với nhau.
- Đầu ra của RNN là một câu, nhưng chưa hoàn chỉnh vì có các ký tự lặp lại như "heelllo", "toooo" do giọng nói dài, giọng bị ngắt quãng. CTC được dùng để cho ra được một câu hoàn chỉnh bằng cách căn chỉnh lại đầu ra ấy, loại bỏ các ký tự lặp lại và khoảng trống.

- Mô hình ngôn ngữ (Language Model) là một phân bố xác suất trên các tập văn bản, cho biết xác suất một câu (hoặc cụm từ) thuộc một ngôn ngữ là bao nhiêu.Mô hình ngôn ngữ tốt sẽ đánh giá đúng các câu đúng ngữ pháp, trôi chảy hơn các từ có thứ tự ngẫu nhiên. Mô hình ngôn ngữ dự kiến thực hiện trong đề tài là N-gram.
- Đầu ra của hệ thống là một đoạn văn bản Tiếng Việt hoàn chỉnh.

Thuật toán RNN :



Các bước liên quan đến thuật toán RNN:

- X_t là đầu vào tại thời điểm t, X_{t-1} là đầu vào trước đó và X_{t+1} là đầu vào trong tương lai.

  Ví dụ, X_1 là một vec-tơ one-hot tương ứng với từ thứ 2 của câu
- S_t là trạng thái ẩn tại bước t. S_t được tính là: S_t = f(U*X_t + W * S_{t-1}).
- O_t là đầu ra ở bước t.

  Ví dụ, nếu muốn dự đoán từ tiếp theo trong một câu, nó sẽ là một vectơ xác suất trong từ vựng, O_t = softmax (V * O_t).

2.5 Kết quả dự kiến của đề tài

- Mô hình nhận dạng giọng nói Tiếng Việt.
- Dịch vụ web (API) sử dụng mô hình nhận dạng giọng nói tiếng Việt cho phép nhận vào một tập tin âm thanh Tiếng Việt sau đó chuyển sang dạng văn bản Tiếng Việt.
- Website mẫu việc sử dụng API của mô hình dịch máy từ tiếng Anh sang tiếng Việt đã xây dựng.

2.6 Kế hoạch thực hiện

Thời gian thực hiện	**Công việc thực hiện**	**Người thực hiện**
7/9/2020-13/9/2020	• Nhận đề tài. • Xây dựng bản kế hoạch sơ bộ cho các công việc cần thực hiện.	Quang, Quyên
14/9/2020-20/9/2020	• Tìm hiểu và phân tích các yêu cầu về kiến thức nền cho đề tài. • Khảo sát và dùng thử các hệ thống cung cấp dịch vụ mẫu có sẵn trên thị trường: fpt.ai, https://vais.vn/,.. • Tạo Trello.	Quang, Quyên
21/9/2020-27/9/2020	• Thống nhất nội dung chính của ứng dụng demo việc sử dụng API. • Biên soạn đề cương cho luận văn (dạng slide).	Quang, Quyên

28/9/2020- 11/10/2020	• Tìm hiểu lý thuyết nền tảng trong máy học. • Biên soạn đề cương cho luận văn (dạng word). • Tìm hiểu lý thuyết nền tảng trong việc nhận dạng giọng nói.	Quang, Quyên
19/10/2020- 1/11/2020	• Tìm hiểu về các thư viện Scikit-Learn, Tensorflow, Keras. • Tìm hiểu các model và kiến trúc, chạy thử các ví dụ để đánh giá.	Quang, Quyên
2/11/2020-8/11/2020	• Chạy thử mô hình nhận dạng âm thanh tiếng Anh sang văn bản.	Quang, Quyên
9/11/2020- 22/11/2020	• Thu thập dữ liệu âm thanh. • Viết chương 1 luận văn. • Chỉnh sửa dữ liệu âm thanh.	Quang, Quyên
23/11/2020- 29/11/2020	• Tìm hiểu và xây dựng mô hình nhận dạng âm thanh Tiếng Việt. • Chỉnh sửa chương 1 luận văn.	Quang, Quyên
30/11/2020- 13/12/2020	• Tiếp tục xây dựng mô hình nhận dạng âm thanh Tiếng Việt.	Quang, Quyên
14/12/2020- 20/12/2020	• Huấn luyện mô hình. • Viết chương 2 luận văn.	Quang, Quyên

21/12/2020-3/1/2020	• Cải tiến mô hình. • Chỉnh sửa chương 2 luận văn.	Quang, Quyên
4/1/2020-17/1/2020	• Viết chương 3 luận văn. • Chỉnh sửa chương 3 luận văn.	Quang, Quyên
18/1/2020-31/1/2020	• Xây dựng và triển khai hệ thống cung cấp dịch vụ web (API). • Viết chương 4 luận văn.	Quang, Quyên
1/2/2020-7/2/2020	• Xây dựng ứng dụng demo việc sử dụng API trên nền tảng web. • Chỉnh sửa chương 4 luận văn.	Quang, Quyên
8/2/2020-14/2/2020	• Viết chương 5 luận văn. • Chỉnh sửa chương 5 luận văn.	Quang, Quyên
15/2/2020-21/2/2020	• Hoàn thành luận văn. • Chỉnh sửa và cải thiện hiệu năng ứng dụng demo. • Nâng cấp mô hình hoàn thiện hơn. • Cải thiện hiệu năng hệ thống cung cấp dịch vụ web(API).	Quang, Quyên
22/2/2020-28/2/2020	• Hoàn chỉnh cuốn luận văn.	Quang, Quyên
1/3/2020-7/3/2020	• Hoàn chỉnh slide trình bày. • Hoàn chỉnh sản phẩm khoá luận.	Quang, Quyên

Tài liệu

[1] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014.

[2] V. H. Nguyen, "An end-to-end model for vietnamese speech recognition," 2019.

[3] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks. proceedings of the 31st international conference on international conference on machine learning," vol. 32, 2014.

[4] I. L. Tom Hope, Yehezkel S. Resheff, *Learning TensorFlow: A Guide to Building Deep Learning Systems 1st Edition.* 2017.

[5] "Chương 3 : Lý thuyết nhận dạng giọng nói." `http://read.pudn.com/downloads443/doc/comm/1868404/Nhan%20dang%20tieng%20noi%20-%20Mo%20phong%20bang%20Matlab/Chuong3%20-%20Ly%20thuyet%20nhan%20dang%20tieng%20noi_completed.pdf`.

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

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