TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**TRẦN TRIỀU HUY - 52000449**

**TRẦN THỊ NGỌC ÁNH - 52000008**

**HỆ THỐNG NHẬN DIỆN**

**TIẾNG NÓI DÂN TỘC**

**DỰ ÁN CÔNG NGHỆ THÔNG TIN**

**KỸ THUẬT PHẦN MỀM**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**TRẦN TRIỀU HUY - 52000449**

**TRẦN THỊ NGỌC ÁNH - 52000008**

**HỆ THỐNG NHẬN DIỆN**

**TIẾNG NÓI DÂN TỘC**

**DỰ ÁN CÔNG NGHỆ THÔNG TIN**

**KỸ THUẬT PHẦN MỀM**

Người hướng dẫn

**PGS. TS. Lê Anh Cường**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

**LỜI CẢM ƠN**

Lời đầu tiên, nhóm chúng tôi muốn bày tỏ lòng biết ơn sâu sắc đến người hướng dẫn trực tiếp của chúng tôi, Thầy Lê Anh Cường, vì sự giám sát, hướng dẫn và hỗ trợ của Thầy trong môn học Dự án Công nghệ Thông tin. Những kiến thức tiếp thu được từ Thầy không chỉ là tiền đề cho báo cáo này mà còn là hành trang quý báu cho chúng tôi trong tương lai. Tuy nhiên, kiến thức là vô hạn mà thời gian chỉ có hạn, nhóm không thể tránh khỏi những thiếu sót. Rất mong nhận được ý kiến đóng góp từ Thầy để chúng em hoàn thiện hơn kiến thức cho bản thân.

Chúng em xin chân thành cảm ơn!

*TP. Hồ Chí Minh, ngày 01 tháng 01 năm 2024.*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*Trần Triều Huy*

*Trần Thị Ngọc Ánh*

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của PGS. TS. Lê Anh Cường. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Dự án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Dự án của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 01 tháng 01 năm 2024*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*Trần Triều Huy*

*Trần Thị Ngọc Ánh*

**HỆ THỐNG NHẬN DIỆN TIẾNG NÓI DÂN TỘC**

**TÓM TẮT**

(Time New Romans – 13)

**TITLE**

**ABSTRACT**

(Time New Romans – 13)

**MỤC LỤC**

[DANH MỤC HÌNH VẼ vii](#_Toc13527)

[DANH MỤC BẢNG BIỂU viii](#_Toc32579)

[DANH MỤC CÁC CHỮ VIẾT TẮT ix](#_Toc3781)

[CHƯƠNG 1. GIỚI THIỆU 1](#_Toc25250)

[1.1 Bài toán 1](#_Toc13263)

[1.2 Các nghiên cứu liên quan 1](#_Toc7727)

[1.3 Mục tiêu nghiên cứu 3](#_Toc14540)

[1.4 Phương pháp nghiên cứu 4](#_Toc16747)

[1.5 Nội dung nghiên cứu 5](#_Toc11545)

[CHƯƠNG 2. KIẾN THỨC NỀN TẢNG 6](#_Toc27805)

[2.1 Tổng quan về âm học và tiếng nói 6](#_Toc11476)

[2.1.1 Âm học 6](#_Toc7317)

[2.1.2 Tiếng nói 8](#_Toc30226)

[2.2 Hệ thống ngữ âm tiếng M’Nông 9](#_Toc399)

[2.3 Khái quát về công nghệ nhận diện tiếng nói 10](#_Toc8365)

[2.3.1 Khái niệm 10](#_Toc8772)

[2.3.2 Cách thức hoạt động 10](#_Toc21916)

[2.4 Mô hình Whisper 10](#_Toc21608)

[2.4.1 Khái niệm 10](#_Toc2551)

[2.4.2 Cách thức hoạt động 11](#_Toc7715)

[2.4.3 Các thuật toán liên quan 11](#_Toc8969)

[2.4.4 Ý nghĩa của mô hình Whisper trong nhận diện tiếng nói 12](#_Toc28742)

[CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT 14](#_Toc21548)

[3.1 Xử lý dữ liệu 14](#_Toc26453)

[3.2 Mô hình 14](#_Toc19516)

[CHƯƠNG 4. THỰC NGHIỆM 16](#_Toc29074)

[4.1 Dữ liệu thực nghiệm 16](#_Toc14305)

[4.2 Cài đặt thực nghiệm 16](#_Toc27311)

[CHƯƠNG 5. KẾT LUẬN 17](#_Toc1802)

[5.1 Kết luận 17](#_Toc470)

[5.2 Hướng phát triển 17](#_Toc5960)

[TÀI LIỆU THAM KHẢO 18](#_Toc9508)

# DANH MỤC HÌNH VẼ

[Hình 1 . Biểu diễn tín hiệu âm thanh 6](#_Toc3160)

# DANH MỤC BẢNG BIỂU

**Error! No table of figures entries found.**

# DANH MỤC CÁC CHỮ VIẾT TẮT

# GIỚI THIỆU

## Bài toán

Đề tài của nhóm chúng tôi mang tên “Hệ thống nhận diện tiếng nói dân tộc”. Để thực hiện được việc nhận diện tiếng nói dân tộc, trước tiên phải xây dựng một hệ thống nhận dạng tiếng nói tự động (Automatic Speech Recognition - ASR), đây là hệ thống có khả năng chuyển đổi chuỗi âm thanh thành chuỗi từ ngữ. Việc xây dựng một hệ thống nhận dạng tiếng nói đòi hỏi nhóm phát triển phải am hiểu các kỹ thuật, lý thuyết từ nhiều kiến thức khác nhau như: học máy, trí tuệ nhân tạo, lý thuyết xác suất thống kê, âm học - vật lý,… Trên thế giới đã có nhiều nhóm nghiên cứu phát triển thành công hệ nhận dạng tiếng nói cho các ngôn ngữ lớn như: tiếng Anh, tiếng Trung Quốc, tiếng Nhật,… nhưng giải pháp nhận diện cho tiếng Việt đặc biệt là tiếng nói dân tộc thiểu số tại Việt Nam vẫn còn rất nhiều hạn chế.

## Các nghiên cứu liên quan

Giao tiếp người - máy là một lĩnh vực nghiên cứu lớn và khó khăn nhưng lại có nhiều ứng dụng thực tiễn. Tiếng nói là một phương thức giao tiếp tự nhiên nhất của con người và vì vậy, nghiên cứu để máy tính có thể nhận dạng tiếng nói tự động (ASR) đã trải qua quá trình phát triển hơn 70 năm. Những cố gắng nghiên cứu đầu tiên về ASR đã được thực hiện trong thập niên 50 với ý tưởng chính là dựa trên ngữ âm. Do kỹ thuật xử lý tín hiệu số cũng như khả năng của máy tính còn giới hạn nên các hệ thống nhận dạng khi đó chỉ tập trung khai thác đặc trưng phổ cộng hưởng đối với các nguyên âm của tín hiệu sau khi đi qua các bộ lọc. Trong giai đoạn này có một số hệ thống nổi bật như: hệ thống nhận dạng ký số rời rạc của Bell-lab (1952), bộ nhận dạng 13 âm vị của Đại học College - Anh (1958),…

Trong tập kỷ 60 của thế kỷ trước, một ý tưởng đáng chú ý của tác giả người Nga - Vintsyuk - khi ông đề xuất phương pháp nhận dạng tiếng nói dựa trên quy hoạch động theo thời gian (Dynamic Time Warping - DTW). Đáng tiếc, phải đến những năm 80 thì phương pháp này mới được thế giới biết đến. Cuối những năm 1960, Reddy (Đại học CMU - Mỹ) đã đề xuất ý tưởng đầu tiên về nhận dạng tiếng nói liên tục bằng kỹ thuật đánh dấu đường đi và truy vết lùi tìm kết quả.

Đến những năm 70, nghiên cứu về nhận dạng tiếng nói đã bước đầu thu được những kết quả mong đợi, làm nền tảng cho những nghiên cứu sau này. Trước hết là bài toán nhận dạng từ rời rạc được hiện thực hóa bởi ý tưởng của các nhà khoa học Nga và Nhật. Velichko và Zagoruyko là hai nhà nghiên cứu tiên phong trong việc áp dụng ý tưởng phân lớp mẫu cho ASR. Sakoe và Chiba đề xuất các kỹ thuật sử dụng phương pháp quy hoạch động. Và Itakura trong khi ở Bell-lab đã đưa ra phương pháp mã hóa dự báo tuyến tính (Linear Predictive Coding - LPC) làm tiền đề cho việc áp dụng các tham số LPC vào ASR. Các hệ thống ASR đáng chú ý của giai đoạn này gồm: Harpy và Hearsay-II của Đại học CMU (Mỹ), hệ thống BWIM (BBN),…

Nỗ lực về ASR trong thập kỷ 80 đánh dấu sự dịch chuyển trong phương pháp luận: từ cách tiếp cận đối sách mẫu sang cách tiếp cận sử dụng mô hình thống kê. Ngày nay, hầu hết các hệ thống ASR đều dựa trên mô hình thống kê được phát triển ở thập kỷ này cùng với những cải tiến ở thập kỷ 90. Một trong những phát minh quan trọng nhất giai đoạn này là mô hình Markov ẩn (Hidden Markov Model - HMM). Dù HMM được áp dụng thành công ở một số phòng lab nhưng phải đợi đến vài năm sau đó, mô hình này mới trở nên phổ biến.

Thập niên 90 tiếp tục ghi nhận một số kết quả nghiên cứu mới trong lĩnh vực phân lớp mẫu. Cụ thể hơn, bài toán phân lớp theo mô hình thống kê (dựa trên luật quyết định Bayes), đòi hỏi phương pháp ước lượng các phân bố cho dữ liệu, được chuyển thành bài toán tối ưu, bao gồm phép cực tiểu lỗi phân lớp bằng thực nghiệm. Cuối cùng, các ứng dụng được phát triển trong giai đoạn này gồm: hệ thống trả lời thông tin tự động cho các chuyến bay (Air Travel Information Service - ATIS), hệ thống ghi lại các bản tin phát thanh (Broadcast News Transcription System),…

Đến những năm đầu thế kỷ 21, các nghiên cứu tập trung vào việc nâng cao kết quả nhận diện tiếng nói thông qua một chương trình có tên gọi EARS (Efective Affordable Reusable Speech-to-Text). Tại thời điểm này, tiếng nói đã không còn bị ràng buộc các điều kiện và có thể thu âm trong môi trường bình thường.

Tại Việt Nam có 2 nhóm nghiên cứu chính về nhận diện tiếng nói liên tục với bộ từ vựng lớn. Nhóm đầu tiên thuộc Viện Công nghệ Thông tin do PGS. Lương Chi Mai đứng đầu, với phương pháp ANN và công cụ CSLU được sử dụng. Nhóm thứ hai thuộc Đại học Khoa học Tự nhiên Tp. HCM do PGS. Vũ Hải Quân đứng đầu với phương pháp HMM và công cụ HTK được sử dụng. Các nghiên cứu tập trung giải quyết bài toán truy vấn thông tin tiếng Việt, nhận dạng tiếng nói, hệ thống giao tiếp người - máy,… Ngoài ra, còn có nghiên cứu của LIG (Laboratoire Informatique de Grenoble) hợp tác với phòng thí nghiệm MICA (Hà Nội) về sự khả chuyển của các mô hình ngữ âm.

## Mục tiêu nghiên cứu

Có thể thấy, lịch sử phát triển của các hệ thống nhận diện tiếng nói từ trong đến ngoài nước là vô cùng phong phú và lâu đời. Tuy nhiên, Việt Nam là một quốc gia với 54 dân tộc anh em cùng sinh sống, bên cạnh tiếng Kinh phổ biến nhất trong cộng đồng người Việt Nam, chúng ta còn rất nhiều tiếng nói khác.

Dân tộc M’Nông là một dân tộc thiểu số ở Việt Nam, với dân số khoảng 127.334 người (theo Tổng điều tra dân số và nhà ở năm 2019), cư trú tại 51 trên tổng số 63 tỉnh, thành phố. Tiếng M’Nông là một ngôn ngữ thuộc ngữ tộc Môn-Khmer của ngữ hệ Nam Á. Tiếng M’Nông được nghiên cứu đầu bởi nhà ngôn ngữ học Richard Phillips vào đầu những năm 1970 và đây là một ngôn ngữ có một hệ thống ngữ âm phức tạp, với nhiều thanh điệu và phụ âm.

Hệ thống nhận diện tiếng nói dân tộc M’Nông là một hệ thống có khả năng nhận diện tiếng nói của người M’Nông thông qua ứng dụng công nghệ thông tin.

Mục tiêu khi chúng tôi thực hiện đề tài là xây dựng một hệ thống nhận diện tiếng nói dân tộc M’Nông có độ chính xác cao, có thể áp dụng vào thực tế và đáp ứng được nhiều yêu cầu của nhiều mục đích sử dụng khác nhau.

Mục tiêu chi tiết:

- Tìm hiểu các khái niệm có liên quan đến hệ thống nhận dạng tiếng nói, làm rõ hơn các yếu tố quan trọng trong việc sử dụng công cụ hỗ trợ.

- Tìm hiểu phương pháp cài đặt công cụ hỗ trợ xây dựng hệ thống nhận diện tiếng nói dân tộc.

- Tìm hiểu, xây dựng mô hình nhận dạng thích hợp cho tiếng M’Nông.

- Xây dựng mô hình, thực nghiệm, từ đó rút ra kết luận và nhận xét.

## Phương pháp nghiên cứu

Nghiên cứu về hệ thống nhận diện tiếng nói dân tộc M’Nông đòi hỏi cần một mô hình kỹ lưỡng và toàn diện. Chính vì thế, cần nghiên cứu và phân tích các tài liệu về lý thuyết nhận diện tiếng nói, đặc điểm ngữ âm của tiếng M’Nông, cũng như các nghiên cứu trước đây về hệ thống nhận diện tiếng nói dân tộc thiểu số. Mục đích nhằm nắm rõ các nguyên lý, phương pháp và thực tiễn của nhận diện tiếng nói, cũng như các đặc điểm của tiếng M’Nông.

Đối tượng được chúng tôi chọn để nghiên cứu là tiếng M’Nông của những người nói giọng huyện Lắk thuộc tỉnh Đắk Lắk. Tất cả các nội dung nghiên cứu trong dự án chỉ giới hạn trong phạm vi điều kiện người nói trong trạng thái bình thường, các vấn đề người nói cố tình cải trang hay giả giọng nói đều nằm ngoài phạm vi nghiên cứu của dự án.

Chúng tôi sử dụng bộ dữ liệu bao gồm các mẫu âm thanh tiếng M’Nông do nhóm thu thập. Bao gồm các câu, từ thường dùng trong đời sống hằng ngày và được thu âm bởi chính người dân địa phương nói tiếng M’Nông. Dữ liệu âm thanh được thu từ hai người nói và với các giọng nói khác nhau (nam, nữ) để hệ thống có thể nhận diện được nhiều giọng nói.

Dữ liệu thu thập cần được xử lý để chuẩn hóa trước khi tiến hành huấn luyện. Trong quá trình xử lý dữ liệu, chúng tôi phân đoạn âm thanh thành các đoạn âm thanh nhỏ hơn, mỗi đoạn tương ứng với một câu hoặc một từ.

Mô hình nhận diện tiếng nói là một thuật toán học máy được sử dụng để phân loại các mẫu âm thanh. Trong đề tài này, chúng tôi sử dụng mô hình Whisper của OpenAI vì đây là một hệ thống nhận dạng giọng nói tự động mã nguồn mở, thiết kế dễ sử dụng và cho độ chính xác cao.

Trong quá trình huấn luyện mô hình, chúng tôi cung cấp cho mô hình các mẫu âm thanh và nhãn của các mẫu âm thanh đó. Mô hình sẽ học cách phân loại các mẫu âm thanh dựa trên các nhãn đã cung cấp và cho ra sản phẩm là văn bản.

Sau khi huấn luyện, mô hình sẽ được thử nghiệm trên dữ liệu mới để đánh giá độ chính xác và có các điều chỉnh phù hợp.

## Nội dung nghiên cứu

Chương 1: Trình bày tổng quan: giới thiệu về đề tài, bài toán được đề cập, các nghiên cứu có liên quan đến đề tài nhận diện tiếng nói trong và ngoài nước, các mục tiêu nghiên cứu đề ra cũng như phương pháp để tiến hành nghiên cứu.

Chương 2: Trình bày kiến thức nền tảng bao gồm: lý thuyết cơ bản về âm học, ngữ âm học, đặc điểm trong tiếng M’Nông, các kiến thức cơ bản để xây dựng và sử dụng một hệ thống nhận dạng tiếng nói. Cơ sở lý thuyết về mô hình Whisper, bao gồm các thuật toán liên quan, ý nghĩa của Whisper trong một hệ thống nhận diện tiếng nói.

Chương 3: Trình bày phương pháp đề xuất: xây dựng kiến trúc, giải pháp của hệ thống và giải thích mô tả hệ thống. Xây dựng về thiết kế và cơ sở dữ liệu.

Chương 4: Trình bày chi tiết thực nghiệm, thống kê về dữ liệu, các mô hình thực nghiệm là gì, xây dựng bộ huấn luyện, tiến hành huấn luyện, giải thích kết quả huấn luyện và xây dựng ứng dụng demo.

Chương 5: Nêu lên kết luận, trình bày những kết quả đạt được, những điểm còn hạn chế, cũng như kinh nghiệm sau quá trình thực hiện dự án, từ đó nêu lên các hướng cải thiện và phát triển nghiên cứu.

# KIẾN THỨC NỀN TẢNG

## Tổng quan về âm học và tiếng nói

### Âm học

* Khái niệm

Khi có nguồn phát ra âm thanh (tiếng nói, tiếng nhạc cụ, tiếng động vật kêu,…) ta sẽ nghe và cảm nhận được âm thanh phát ra. Vật tạo ra âm thanh được gọi là nguồn phát âm. Âm thanh chính là sự dao động cơ của các thành phần vật chất trong môi trường lan truyền và khi nó đến tai ta, ta cảm nhận được âm thanh. Trong môi trường không có vật chất tồn tại như chân không sẽ không có âm thanh bởi không có dao động cơ. Trong đời sống xã hội, âm thanh là phương tiện giao tiếp, tuyền đạt thông tin xuất hiện lâu đời nhất trong lịch sử nhân loại. Khi nghiên cứu về âm thanh, người ta thường quan tâm đến 2 đặc điểm: đặc trưng vật lý và đặc trưng sinh học.

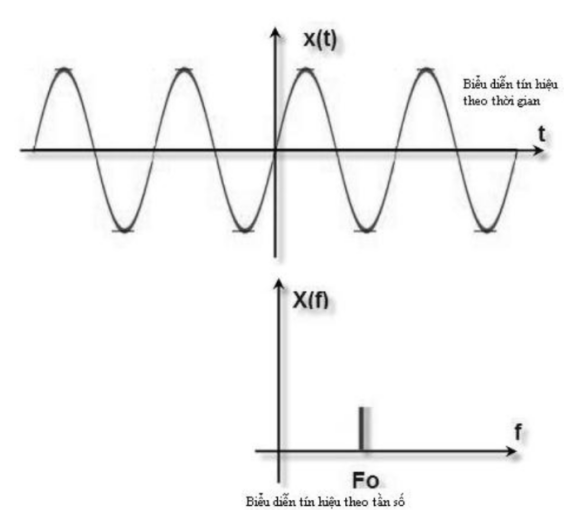
* Biểu diễn tín hiệu âm thanh trong miền thời gian và tần số

Để biểu diễn âm thanh trong miền thời gian, ta có thể dùng hàm toán học x(t). Trong đó:

- t: thời gian

- x: biên độ biến thiên (ly độ)

Biểu diễn x(t) bằng đồ thị thời gian. Đặt x(t) = A.sin οt = A.sin 2πFοt:



Hình 1. Biểu diễn tín hiệu âm thanh

Phổ tính hiệu: là cách biểu diễn các thành phần cấu tạo nên x(t) theo tần số. Với tín hiệu sin nói trên, đồ thị phổ là một vạch có cao độ A tại điểm có tần số Fο. Ta nói đó là phổ vạch. Trong thực tế, với x(t) bất kỳ, biến thiên, không tuần hoàn, người ta sẽ dùng phân tích Fourier để tính toán phổ tín hiệu. Khi đó ta có phổ liên tục X(Ω).

* Các loại âm thanh

Những dao động cơ mà con người nghe được gọi là âm thanh (sound).

Âm thanh có thể biểu diễn theo thời gian và tần số do có thể phân tích một tín hiệu âm thanh thành tổ hợp các thành phần tần số khác nhau (chuỗi Fourier, tích phân Fourier). Đơn giản hơn, một âm thanh có thể là tổ hợp từ nhiều đơn âm có tần số dao động nhất định.

Dải tần số nghe được là từ 20 Hz - 20000 Hz. Siêu âm là âm dao động ngoài 20000 Hz. Hạ âm là âm dao động dưới 20 Hz. Tai người không nghe được siêu âm và hạ âm.

Tiếng nói (voice, speech) là âm thanh phát ra từ miệng con người, được truyền đi trong không khí đến tai người nghe. Dải tần số đủ nghe rõ là từ 300 Hz đến 3500 Hz (dải tần tiêu chuẩn). Dải tần số có chất lượng cao có thể từ 200 Hz đến 7000 Hz (ampli hội trường).

Âm nhạc (music) là âm thanh phát ra từ các nhạc cụ. Dải tần số từ 20 Hz đến 15000 Hz.

Tiếng kêu là âm thanh phát ra từ miệng động vật.

Tiếng động là âm thanh phát ra từ va chạm giữa các đồ vật.

Tiếng ồn (noise) là âm thanh không mong muốn.

Nhìn chung, có 2 loại âm: tuần hoàn (tiếng nói, âm nhạc) và không tuần hoàn (tín hiệu tạp nhiễu).

* Đơn vị đo âm thanh

Cảm nhận độ to của âm thanh dựa vào hàm số mũ:

Bel = 10 lg P2/P1

Deccibel = 20 lg I2/I1

### Tiếng nói

Là âm thanh phát ra từ miệng người. Nghiên cứu tiếng nói gồm: Bộ máy phát âm của con người. Thụ cảm âm thanh của tai người. Phân loại tiếng nói.

Bộ máy phát âm của con người gồm:

- Phổi đóng vai trò bơm không khí, tạo năng lượng hình thành âm.

- Đôi dây thanh (vocal fold) là hai cơ thịt trong cuống họng, có hai đầu tính vào nhau, còn hai đầu dao động với tần số cơ bản là Fο. Fο của nam giới nằm trong khoảng 100 - 200 Hz, của nữ giới là 300 - 400 Hz, của trẻ em là 500 - 600 Hz.

- Thanh quản và vòm miệng đóng vai trò như hốc và cộng hưởng, tạo ra phân biệt tần số khi tín hiệu dao động từ đôi dây thanh phát ra.

- Miệng đóng vai trò phát tán âm thanh ra bên ngoài.

- Lưỡi thay đổi để tạo ra tần số formant khác nhau.

- Các âm khác nhau là do vị trí tương đối của formants.

Phân loại tiếng nói theo thanh:

- Âm hữu thanh (voiced sounds): là âm khi phát ra có sự dao động của đôi dây thanh, nên nó tuần hoàn với Fο. Vì vậy phổ của nguyên âm là phổ vạch, khoảng cách giữa các vạch bằng chính Fο.

- Âm vô thanh (unvoiced sounds): phát ra khi đôi dây thanh không dao động. Phổ tín hiệu có dạng nhiễu trắng và phổ phân bố đều.

Phân loại tiếng nói:

- Nguyên âm (vowel): là âm phát ra có thể kéo dài. Nguyên âm là âm hữu thanh, tuần hoàn và khá ổn định trong một đoạn thời gian vài chục ms.

- Phụ âm (consonant): là âm chỉ phát ra một nhát, không thể kéo dài. Có phụ âm hữu thanh và phụ âm vô thanh.

Thanh điệu của tiếng Việt tương ứng với các dấu: ngang, huyền, sắc, hỏi, ngã, nặng khi viết. Phân tích cho thấy thanh điệu là sự thay đổi Fο, tần số cơ bản pitch, trong quá trình phát âm các nguyên âm và tai người cảm nhận được. Tiếng Việt có 6 thanh thể hiện sự phong phú và độc đáo, trong khi tiếng Trung Quốc chỉ có 4 thanh.

Giọng bổng (high voiced pitch) hay giọng trầm (low voiced pitch) là Fο cao hay thấp. Như vậy Fο đóng vai trò rất quan trọng trong cảm nhận, thụ cảm âm thanh của con người.

Tiếng bổng hay tiếng trầm tương ứng dải tần số cao hay thấp.

Khi nói, không khí được đẩy từ phổi qua miệng và khoang mũi. Luồng không khí này bị cản trở và thay đổi do hoạt động của lưỡi và môi. Điều này tạo ra các co giãn của không khí, sóng âm, âm thanh. Những âm thanh tạo thành, tương ứng với nguyên âm và phụ âm, thường được gọi là âm vị. Các âm vị được kết hợp với nhau tạo thành từ. Mỗi âm vị được hình thành trong quá trình nói phụ thuộc vào ngữ cảnh của nó, tức là phụ thuộc vào âm vị đứng ngay trước và âm vị đứng ngay sau nó. Tuy nhiên, tiếng nói không chỉ là chuỗi các âm vị tạo thành từ và câu. Có nhiều thành phần của tiếng nói mang thông tin, ví dụ: phong cách nói thể hiện qua âm điệu, cao độ, âm lượng, ngữ điệu. Thông tin này đôi khi được gọi là thông tin phi từ vựng (paralinguistic) của tiếng nói và cũng được sử dụng trong nhiều phương pháp nhằm nâng cao độ chính xác của hệ thống nhận diện.

## Hệ thống ngữ âm tiếng M’Nông

Tiếng M’Nông là một ngôn ngữ thuộc nhóm ngôn ngữ Môn-Khmer, được sử dụng bởi các dân tộc M’Nông ở Việt Nam, Campuchia và Lào. Tiếng M’Nông có hệ thống ngữ âm tương đối phức tạp, với 6 thanh điệu, 44 âm vị, bao gồm 26 phụ âm và 18 nguyên âm được phân loại dựa trên cách phát âm.

Ngoài ra, tiếng M’Nông còn có các phụ âm và nguyên âm biến thể, được tạo thành bằng cách kết hợp các âm vị cơ bản.

Tính chất ngữ âm của tiếng M’Nông: Tiếng M’Nông là một ngôn ngữ có trật tự âm tiết mở, nghĩa là âm tiết bắt buộc phải có nguyên âm. Tiếng M’Nông cũng là một ngôn ngữ có phân chia âm tiết thành âm tiết chính và âm tiết phụ. Âm tiết chính là âm tiết chứa thanh điệu và âm tiết phụ là âm tiết không chứa thanh điệu.

Thanh điệu tiếng M’Nông có thể được phân loại thành 2 nhóm chính: thanh điệu đơn và thanh điệu kép. Thanh điệu đơn là thanh điệu chỉ có một độ cao nhất định. Thanh điệu kép là thanh điệu có hai độ cao.

## Khái quát về công nghệ nhận diện tiếng nói

### Khái niệm

ASR là một nhánh của Học máy (Machine Learning – ML). Về cơ bản, thay vì lập trình các quy tắc để chuyển đổi dữ liệu đầu vào (giọng nói) thành đầu ra (văn bản), thì mô hình Học máy được đào tạo bằng cách đưa các tập dữ liệu lớn vào một thuật toán, chẳng hạn như mô hình Whisper. Trải qua quá trình đào tạo, mô hình ngày càng suy luận tốt hơn, và có khả năng nhận dạng tiếng nói của con người.

### Cách thức hoạt động

Để chuyển giọng nói sang văn bản, hệ thống phải thực hiện một quá trình gồm nhiều bước phức tạp. Khi nói, ta sẽ tạo ra những rung động trong không khí. Bộ chuyển đổi tín hiệu tương tự sang số (Analog-to-Digital Converter - ADC) chuyển các sóng tương tự (analog) này thành dữ liệu mà máy có thể hiểu được.

Để làm điều này, hệ thống thu thập các mẫu (hoặc số hóa) âm thanh bằng cách đo chính xác sóng âm ở các khoảng thời gian gần nhau, sau đó lọc âm thanh đã được số hoá để loại bỏ tiếng ồn, đôi khi tách chúng thành các dải tần số khác nhau. Nó cũng tinh chỉnh âm thanh đến một mức âm lượng không thay đổi hoặc sắp xếp theo thời gian. Không phải lúc nào con người cũng nói với tốc độ như nhau nên âm thanh phải được điều chỉnh cho phù hợp với tốc độ mà âm thanh mẫu được ghi nhận trong bộ nhớ máy.

Tiếp theo, tín hiệu được chia thành nhiều phần nhỏ (thời gian khoảng vài phần trăm giây, thậm chí là phần ngàn giây trong trường hợp có phụ âm cuối khó phân biệt). Chương trình sau đó đặt những phần âm thanh này vào các âm vị có sẵn trong ngôn ngữ thích hợp.

## Mô hình Whisper

### Khái niệm

Mô hình Whisper là một mô hình nhận dạng tiếng nói đa ngôn ngữ và đa nhiệm, được phát triển bởi OpenAI. Mô hình này được huấn luyện trên một tập dữ liệu khổng lồ gồm 680.000 giờ âm thanh tiếng nói, bao gồm 96 ngôn ngữ khác nhau.

Mô hình Whisper hoạt động theo hướng end-to-end, sử dụng kiến trúc mạng nơ-ron nhân tạo Transformer.

Mô hình Whisper đã đạt được kết quả ấn tượng trong các thử nghiệm. Độ chính xác của mô hình trong các tác vụ nhận dạng giọng nói và dịch giọng nói tương đương với các mô hình nhận dạng tiếng nói chuyên biệt cho từng ngôn ngữ.

### Cách thức hoạt động

Whisper là một công cụ được tạo ra bởi OpenAI có khả năng hiểu và chuyển đổi ngôn ngữ nói thành văn bản, tương tự như cách Siri hoặc Alexa hoạt động. Cách mà Whisper hoạt động giống như một thông dịch viên. Nó sử dụng một kiến trúc mã hóa-giải mã Transformer. Để đơn giản hóa, hãy tưởng tượng ai đó dịch từ tiếng Anh sang tiếng Việt. Phần ‘mã hóa’ sẽ hiểu câu tiếng Anh, và phần ‘giải mã’ sau đó sẽ tạo ra câu tiếng Việt. Trong Whisper, phần ‘mã hóa’ hiểu ngôn ngữ nói (âm thanh), và phần ‘giải mã’ tạo ra văn bản.

Để học cách hiểu và chuyển đổi ngôn ngữ nói, Whisper đã được đào tạo bằng cách sử dụng một lượng lớn dữ liệu từ internet - tương đương với việc liên tục lắng nghe suốt hơn 77 năm. Dữ liệu này bao gồm nhiều ngôn ngữ khác nhau và đa nhiệm (bảo gồm nhiều loại nhiệm vụ khác nhau, không chỉ là chuyển đổi). Dữ liệu âm thanh mà Whisper học được xử lý theo một cách cụ thể để làm cho hệ thống dễ hiểu hơn. Điều này giống như khi bạn điều chỉnh cài đặt trên TV để làm cho hình ảnh rõ ràng hơn. Âm thanh được ‘tái tạo’ thành chất lượng chuẩn (16.000 Hz), sau đó nó được biến đổi thành biểu đồ thị trực quan mà hệ thống có thể học. Quá trình biến đổi này được thực hiện trong các phần nhỏ chồng lên nhau để đảm bảo không có phần nào của âm thanh bị bỏ lỡ.

### Các thuật toán liên quan

Mô hình Whisper sử dụng các thuật toán sau:

- Kiến trúc mạng nơ-ron nhân tạo Transformer: Kiến trúc Transformer là một kiến trúc mạng nơ-ron nhân tạo được sử dụng phổ biến trong các tác vụ xử lý ngôn ngữ tự nhiên, chẳng hạn như dịch máy, tóm tắt văn bản, và nhận dạng tiếng nói. Kiến trúc Transformer có khả năng xử lý thông tin theo hướng end-to-end, giúp mô hình có thể học hỏi các mối quan hệ phức tạp giữa các từ và các đoạn văn bản.

- Kỹ thuật chuyển đổi attention: Kỹ thuật chuyển đổi attention là một kỹ thuật được sử dụng trong kiến trúc Transformer để giúp mô hình tập trung vào các thông tin quan trọng trong một đoạn văn bản. Kỹ thuật này hoạt động bằng cách tính toán một trọng số cho mỗi từ trong đoạn văn bản, dựa trên mức độ liên quan của từ đó với từ đang được xử lý.

- Kỹ thuật học có giám sát: Kỹ thuật học có giám sát là một kỹ thuật học máy trong đó mô hình được huấn luyện trên một tập dữ liệu có nhãn. Trong trường hợp của mô hình Whisper, tập dữ liệu huấn luyện bao gồm các mẫu âm thanh tiếng nói và nhãn của các mẫu âm thanh đó. Mô hình sẽ học cách phân loại các mẫu âm thanh dựa trên các nhãn đã cung cấp.

Ngoài các thuật toán trên, mô hình Whisper còn sử dụng một số kỹ thuật khác để cải thiện hiệu quả của mô hình:

- Kỹ thuật giảm kích thước: Kỹ thuật giảm kích thước được sử dụng để giảm kích thước của các mảng dữ liệu, giúp mô hình có thể chạy nhanh hơn và hiệu quả hơn.

- Kỹ thuật tăng tốc: Kỹ thuật tăng tốc được sử dụng để tăng tốc độ của mô hình, chẳng hạn như sử dụng các kỹ thuật tính toán song song hoặc sử dụng các phần cứng chuyên dụng.

### Ý nghĩa của mô hình Whisper trong nhận diện tiếng nói

Với 680,000 giờ âm thanh được gán nhãn, bộ dữ liệu của Whisper là một trong những bộ dữ liệu lớn nhất từng được tạo ra trong lĩnh vực nhận diện tiếng nói có giám sát.

Mô hình Whisper trong hệ thống nhận diện giọng nói giúp cải thiện độ chính xác của việc nhận dạng giọng nói trong các môi trường có tiếng ồn hoặc khi người nói nói nhỏ. Điều này có thể được ứng dụng trong nhiều lĩnh vực như: trợ lý ảo, tạo phụ đề, dịch ngôn ngữ, bảo mật,…

# PHƯƠNG PHÁP ĐỀ XUẤT

## Xử lý dữ liệu

Theo xu hướng của các nghiên cứu gần đây, chúng tôi áp dụng một phương pháp tối giản để xử lý dữ liệu. Khác với các hệ thống ASR khác, mô hình Whisper có thể dự đoán bản thô của audio mà không cần bất kỳ tiêu chuẩn hay ràng buộc nào. Chúng tôi phân đoạn âm thanh thành các đoạn âm thanh nhỏ hơn, mỗi đoạn tương ứng với một câu hoặc một từ. Tất cả các đoạn này được huấn luyện và sử dụng làm dữ liệu huấn luyện.

Nhiều bản ghi trên internet thực tế không phải do con người tạo ra mà là kết quả của các hệ thống ASR. Nghiên cứu gần đây đã chỉ ra rằng việc huấn luyện trên các bộ dữ liệu chứa cả dữ liệu được tạo ra bởi con người và máy có thể ảnh hưởng đáng kể đến hiệu suất của các hệ thống dịch. Để tránh việc học "ngôn ngữ bản ghi" không tự nhiên, chúng tôi xây dựng bộ dữ liệu âm thanh từ chính người dân địa phương nói tiếng M’Nông tại huyện Lắk của tỉnh Đắk Lắk, điều này làm tăng cường độ tin cậy cho bộ dữ liệu và chất lượng âm thanh tốt cũng giúp mô hình hoạt động mạnh mẽ hơn.

## Mô hình

Chúng tôi sử dụng một kiến trúc có sẵn là Transformer mã hóa-giải mã vì kiến trúc này đã được xác minh có thể mở rộng và đáng tin cậy. Tất cả âm thanh được lấy mẫu thành 16,000 Hz, và biểu diễn spectrogram Mel log-magnitude với 80 kênh được tính toán trên cửa sổ 25 mili giây với bước nhảy là 10 mili giây.

Để chuẩn hóa các đặc trưng, chúng tôi mở rộng đầu vào từ -1 đến 1 với giá trị trung bình xấp xỉ 0 trên toàn bộ dữ liệu tiền đào tạo. Bộ mã hóa xử lý biểu diễn đầu vào này với một tầng stem nhỏ bao gồm hai tầng tích chập với chiều rộng bộ lọc là 3 và hàm kích hoạt GELU, trong đó tầng tích chập thứ hai có bước nhảy là 2. Phần nhúng có vị trí hình sin sau đó được thêm vào đầu ra của stem, tiếp đó là áp dụng các khối Transformer của bộ mã hóa. Transformer sử dụng các khối pre-activation và áp dụng một lớp chuẩn hóa cuối cho đầu ra của bộ mã hóa. Bộ giải mã sử dụng phần nhúng được học và biểu diễn thông tin đầu vào-đầu ra được ràng buộc. Bộ mã hóa và bộ giải mã có cùng chiều rộng và số lượng khối Transformer.

# THỰC NGHIỆM

## Dữ liệu thực nghiệm

...

## Cài đặt thực nghiệm

…

# KẾT LUẬN

## Kết luận

Whisper cho thấy rằng việc mở rộng quy mô đào tạo có giám sát cho đến nay vẫn chưa được đánh giá cao trong nghiên cứu nhận diện giọng nói. Chúng ta đạt được các kết quả mà không cần các kỹ thuật tự giám sát và tự đào tạo mà đã là điều quan trọng của các hệ thống ASR gần đây. Đồng thời, việc đào tạo trên một tập dữ liệu giám sát lớn, đa dạng và tập trung vào chuyển giao không giám sát có thể cải thiện đáng kể tính ổn định của một hệ thống nhận dạng giọng nói.

Chúng tôi đã bước đầu nhận diện được

## Hướng phát triển

# TÀI LIỆU THAM KHẢO

Tiếng Việt

…

Tiếng Anh

[1] Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.S., and Nguyen, A. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.4845–4854, 2019.

[2] Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509, 2019.

[3] Seide, F., Li, G., Chen, X., and Yu, D. Feature engineering in context dependent deep neural networks for conversational speech transcription. In 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, pp.24–29. IEEE, 2011.