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**

**TÌM HIỂU MÔ HÌNH**

**NHẬN DẠNG TIẾNG NÓI**

**CHO TIẾNG DÂN TỘC M’NÔNG VÀ XÂY DỰNG**

**ỨNG DỤNG THỬ NGHIỆM**

**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**

**TÌM HIỂU MÔ HÌNH**

**NHẬN DẠNG TIẾNG NÓI**

**CHO TIẾNG DÂN TỘC M’NÔNG**

**VÀ XÂY DỰNG**

**ỨNG DỤNG THỬ NGHIỆM**

**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*

**TÌM HIỂU MÔ HÌNH NHẬN DẠNG TIẾNG NÓI CHO TIẾNG DÂN TỘC M’NÔNG VÀ XÂY DỰNG ỨNG DỤNG THỬ NGHIỆM**

**TÓM TẮT**

(Time New Romans – 13)

**STUDY SPEECH RECOGNITION MODELS FOR M’NONG ETHNIC LANGUAGE AND BUILD AN EXPERIMENT**

**ABSTRACT**

(Time New Romans – 13)

**MỤC LỤC**

[DANH MỤC HÌNH VẼ viii](#_Toc3055)

[DANH MỤC CÁC CHỮ VIẾT TẮT x](#_Toc30408)

[CHƯƠNG 1. GIỚI THIỆU 1](#_Toc13035)

[1.1 Bài toán 1](#_Toc21165)

[1.2 Các nghiên cứu liên quan 1](#_Toc22617)

[1.3 Mục tiêu nghiên cứu 3](#_Toc13356)

[1.4 Phương pháp nghiên cứu 3](#_Toc18208)

[1.5 Nội dung nghiên cứu 4](#_Toc5165)

[CHƯƠNG 2. KIẾN THỨC NỀN TẢNG 6](#_Toc28599)

[2.1 Tổng quan về âm học và tiếng nói 6](#_Toc22456)

[2.1.1 Âm học 6](#_Toc10187)

[2.1.2 Tiếng nói 7](#_Toc30585)

[2.2 Hệ thống ngữ âm tiếng M’Nông 9](#_Toc43)

[2.3 Khái quát về công nghệ nhận diện tiếng nói 9](#_Toc6292)

[2.3.1 Khái niệm 9](#_Toc10953)

[2.3.2 Cách thức hoạt động 10](#_Toc18699)

[2.4 Mô hình Whisper 10](#_Toc31253)

[2.4.1 Khái niệm 10](#_Toc25653)

[2.4.2 Cách thức hoạt động 11](#_Toc3082)

[2.4.3 Các thuật toán liên quan 11](#_Toc2079)

[2.4.4 Ý nghĩa của mô hình Whisper trong nhận diện tiếng nói 12](#_Toc1683)

[CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT 13](#_Toc2769)

[3.1 Xử lý dữ liệu 13](#_Toc14394)

[3.2 Mô hình 13](#_Toc5829)

[3.3 Multitask Format 14](#_Toc18244)

[3.4 Hiệu suất 15](#_Toc4118)

[3.5 Hướng tiếp cận 17](#_Toc10912)

[CHƯƠNG 4. THỰC NGHIỆM 18](#_Toc12340)

[4.1 Dữ liệu thực nghiệm 18](#_Toc17234)

[4.2 Cài đặt thực nghiệm 19](#_Toc29723)

[4.2.1 Load Dataset 19](#_Toc25158)

[4.2.2 Chuẩn bị Tokenizer, Feature Extractor và xủ lý data 20](#_Toc6497)

[4.2.3 Load WhisperFeatureExtractor 20](#_Toc9647)

[4.2.4 Load WhisperTokenizer 21](#_Toc2199)

[4.2.5 Kết hợp để tạo ra WhisperProcessor 21](#_Toc12890)

[4.2.6 Chuẩn bị dữ liệu 22](#_Toc16336)

[4.2.7 Huấn luyện và đánh giá 23](#_Toc12446)

[4.2.8 Định nghĩa Data Collator 23](#_Toc9809)

[4.2.9 Số liệu đánh giá 25](#_Toc29140)

[4.2.10 Tải pre-trained checkpoint 26](#_Toc15018)

[4.2.11 Định nghĩa Traing Arguments 26](#_Toc2678)

[4.2.12 Huấn luyện mô hình 28](#_Toc22231)

[4.3 Đánh giá 29](#_Toc20057)

[4.4 Thực nghiệm và kết quả 30](#_Toc25105)

[CHƯƠNG 5. KẾT LUẬN 33](#_Toc18338)

[TÀI LIỆU THAM KHẢO 34](#_Toc18035)

# DANH MỤC HÌNH VẼ

[Hình 1 . Biểu diễn tín hiệu âm thanh 6](#_Toc29507)

[Hình 2 . WER trên LibriSpeech dev-clean (%) 15](#_Toc14731)

[Hình 3 . So sánh tính hiệu quả trên nhiều bộ dữ liệu 16](#_Toc9297)

[Hình 4 . Bộ dữ liệu tiếng M’Nông được đẩy lên Hugging Face 18](#_Toc15285)

[Hình 5 . Phân bố độ dài của các file âm thanh trong bộ dữ liệu 19](#_Toc32465)

[Hình 6 . Load dataset 19](#_Toc27272)

[Hình 7 . Kết quả của load dataset 20](#_Toc12279)

[Hình 8 . Load WhisperFeatureExtractor 21](#_Toc24428)

[Hình 9 . Load WhisperTokenizer 21](#_Toc13320)

[Hình 10 . Tạo WhisperProcessor 22](#_Toc19830)

[Hình 11 . In data đầu tiên 22](#_Toc20577)

[Hình 12 . Điều chỉnh tốc độ lấy mẫu 22](#_Toc27533)

[Hình 13 . Resample mẫu âm thanh về tốc độ chuẩn 22](#_Toc28347)

[Hình 14 . Hàm chuẩn bị dữ liệu 23](#_Toc18891)

[Hình 15 . Áp dụng hàm chuẩn bị dữ liệu 23](#_Toc15789)

[Hình 16 . Định nghĩa và khởi tạo Data Collator 25](#_Toc5669)

[Hình 17 . Xác định WER 25](#_Toc14962)

[Hình 18 . Hàm trả về WER 26](#_Toc5199)

[Hình 19 . Tải một pre-trained checkpoint 26](#_Toc2029)

[Hình 20 . Cấu hình lại mô hình 26](#_Toc19674)

[Hình 21 . Ý nghĩa các tham số liên quan 27](#_Toc14058)

[Hình 22 . Chuyển các đối số huấn luyện cho Trainer 28](#_Toc14240)

[Hình 23 . Huấn luyện mô hình 28](#_Toc25511)

[Hình 24 . Kết quả huấn luyện 29](#_Toc25953)

[Hình 25 . Đưa kết quả lên Hugging Face Hub 29](#_Toc19603)

[Hình 26 . Công thức tính WER 30](#_Toc27822)

[Hình 27 . Ví dụ đơn giản sử dụng mô hình đã qua huấn luyện 31](#_Toc18005)

[Hình 28 . Kết quả khi thực hiện kiểm tra trên bộ dữ liệu sử dụng để kiểm định 31](#_Toc7686)

[Hình 29 . Kết quả khi thực hiện kiểm tra trên bộ dữ liệu sử dụng để kiểm định 32](#_Toc12560)

# DANH MỤC CÁC CHỮ VIẾT TẮT

ASR Automatic Speech Reconigtion

ML Machine Learning

WER Word Error Rate

# GIỚI THIỆU

## Bài toán

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

## Các nghiên cứu liên quan

Một lĩnh vực nghiên cứu lớn và khó khăn nhưng ứng dụng vào thực tiễn khá nhiều đó chính là khả năng giao tiếp giữa người - máy. Trong đó, tiếng nói có thể xem như phương thức giao tiếp tự nhiên nhất của con người và vì thế, quá trình tìm tòi, nghiên cứu để máy tính có thể tự động nhận dạng được tiếng nói đã trải qua hơn 70 năm.

Trong thập kỷ 1960s, Vintsyuk - một tác giả người Nga đã đề xuất phương pháp nhận dạng tiếng nói Dynamic Time Warping. Nhưng phải đến những năm 80 thì phương pháp này mới được chú ý đến. Cuối những năm 1960, tác giả Reddy đến từ Đại học CMU (Mỹ) đã đề xuất ý tưởng chưa từng xuất hiện về nhận dạng tiếng nói bằng cách sử dụng kỹ thuật đánh dấu đường đi và truy vết lùi tìm kết quả.

Những năm 1970, hệ thống nhận dạng tiếng nói đã từng bước thu được kết quả khả quan và có thể làm nền tảng cho những mô hình nghiên cứu sau này. Velichko và Zagoruyko, là hai nhà nghiên cứu tiên phong trong cách áp dụng ý tưởng phân lớp mẫu cho hệ thống ASR. Sakoe và Chiba cũng đề xuất các kỹ thuật phương pháp quy hoạch động. Itakura trong lúc ở Bell-lab đã đưa ra phương pháp Linear Predictive Coding (LPC), làm tiền đề cho việc áp dụng các tham số LPC vào hệ thống ASR. Các hệ thống ASR nổi bật của giai đoạn này bao gồm: hệ thống BWIM (BBN), Harpy và Hearsay-II của Đại học CMU (Mỹ),…

Nỗ lực về nghiên cứu hệ thống nhận dạng tiếng nói trong thập kỷ 1980s đánh dấu sự chuyển đổi trong phương pháp luận: từ tiếp cận đối sách mẫu sang tiếp cận sử dụng mô hình thống kê. Phần lớn các hệ thống ASR ngày nay đều dựa trên mô hình thống kê được phát triển vào thời điểm này cùng với những cải tiến mới ở thập kỷ 90. Hidden Markov Model (HMM) là một nghiên cứu quan trọng bậc nhất giai đoạn này và đã được áp dụng thành công tại một vài phòng lab.

Đến thập niên 1990 lại tiếp tục ghi nhận một vài tín hiệu nghiên cứu mới trong lĩnh vực phân lớp mẫu. Cụ thể là 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. Các ứng dụng tiêu biểu được phát triển trong giai đoạn này bao gồm: Air Travel Information Service, Broadcast News Transcription System,…

Những năm đầu của thế kỷ 21, các nghiên cứu tập trung vào việc làm sao có thể nâng cao kết quả nhận dạng tiếng nói thông qua một chương trình được gọi là EARS (Efective Affordable Reusable Speech-to-Text). Lúc này thì tiếng nói đã không còn bị ràng buộc bởi các điều kiện và có thể thu âm trong môi trường bình thường.

Ở Việt Nam hiện nay có 2 nhóm nghiên cứu lớn về nhận dạng tiếng nói với bộ từ vựng lớn. Nhóm đầu thuộc Viện Công nghệ Thông tin do PGS. Lương Chi Mai dẫn dắt với phương pháp ANN và sử dụng công cụ CSLU. Nhóm thứ hai trực thuộc Đại học Khoa học Tự nhiên TPHCM do PGS. Vũ Hải Quân hướng dẫn với phương pháp HMM và dùng bộ công cụ HTK. Các nghiên cứu trên tập trung vào giải quyết bài toán truy vấn thông tin tiếng Việt, nhận dạng tiếng nói,… Ngoài ra, còn có một nghiên cứu nổi bật của Laboratoire Informatique de Grenoble hợp tác cùng với phòng thí nghiệm MICA tại 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 dạng 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 sinh sống tại 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 lần 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 dạng tiếng nói dân tộc M’Nông là một hệ thống có khả năng nhận dạng 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 dạng 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.

## Phương pháp nghiên cứu

Nghiên cứu về hệ thống nhận dạng 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 dạng tiếng nói, đặc điểm ngữ âm của tiếng M’Nông, cũng như các nghiên cứu trước đây về hệ thống nhận dạng tiếng nói dân tộc thiểu số. Mục đích nhằm nắm rõ các nguyên lý, phương pháp và thực tiễn của nhận dạng 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 ba 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 sẽ tương ứng với một câu hoặc một từ trong tiếng M’Nông.

Mô hình nhận dạng tiếng nói là một thuật toán trong học máy và 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 thực hiện 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 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: Mô tả tổng quan: giới thiệu về đề tài, bài toán được đề cập, các nghiên cứu có liên quan đến đề tài nhận dạng 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 về âm học, ngữ âm học, đặc điểm trong tiếng M’Nông, các kiến thức để xây dựng và sử dụng một hệ thống nhận dạng tiếng nói. Các cơ sở lý thuyết về mô hình Whisper, bao gồm các thuật toán có liên quan, ý nghĩa của Whisper trong một hệ thống nhận dạng tiếng nói.

Chương 3: Trình bày phương pháp đề xuất: xây dựng kiến trúc, giải pháp của hệ thống.

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

Chương 5: Nêu kết luận và 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 hướng cải thiện và phát triển dự án.

# KIẾN THỨC NỀN TẢNG

## Tổng quan về âm học và tiếng nói

### Âm học

* Khái niệm

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

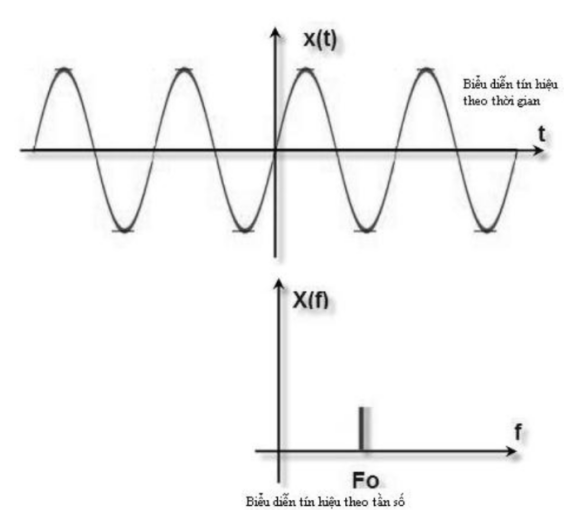
* Biểu diễn tín hiệu âm thanh trong miền thời gian và tần số

Ta có thể dùng hàm toán học x(t) để biểu diễn âm thanh trong miền thời gian. 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 x(t) theo tần số. Với tín hiệu sin đề cập ở trên, đồ thị phổ chính là một vạch có cao độ A tại một điểm có tần số Fο. Khi đó ta nói đó là phổ vạch. Thực tế, với x(t) bất kỳ, biến thiên và không tuần hoàn, ta sẽ dùng Fourier để tính toán phổ tín hiệu. Khi đó ta có được phổ liên tục X(Ω).

* Các loại âm thanh phổ biến

Âm thanh (sound) là những dao động cơ mà tai con người có thể nghe.

Âm thanh có thể được biểu diễn theo thời gian và tần số (chuỗi Fourier, tích phân Fourier). Một âm thanh có thể là tổ hợp từ nhiều đơn âm và có tần số dao động nhất định.

Dải tần số con người nghe được là trong khoảng từ 20 Hz - 20000 Hz. Siêu âm là âm dao động trên 20000 Hz. Hạ âm là âm dao động dưới 20 Hz. Tai người bình thường không nghe được siêu âm và hạ âm.

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

Âm nhạc (music) là âm thanh được phát ra từ các loại nhạc cụ. Dải tần số trải từ 20 Hz - 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.

* 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 được phát ra từ miệng của con người. Bộ máy phát âm của con người bao gồm:

- Phổi sẽ đóng vai trò bơm không khí và tạo năng lượng để âm được hình thành.

- Đôi dây thanh là hai cơ thịt nằm trong cuống họng, dao động với tần số cơ bản là Fο. Fο của nam giới nằm trong khoảng từ 100 - 200 Hz, của nữ giới là từ 300 - 400 Hz, của trẻ em là từ 500 - 600 Hz.

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

- Miệng giữ vai trò phát tán âm thanh truyền ra bên ngoài môi trường.

- Lưỡi để tạo ra tần số formant khác nhau.

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

Phân loại tiếng nói dựa theo âm thanh:

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

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

Phân loại tiếng nói:

- Nguyên âm (vowel): âm phát ra có khả năng kéo dài. Nguyên âm là âm hữu thanh, tuần hoàn và khá ổn định (vài chục ms).

- Phụ âm (consonant): âm chỉ phát ra một lần và không có khả năng kéo dài. Có phụ âm hữu thanh và phụ âm vô thanh.

Thanh điệu của tiếng Việt bao gồm các dấu: ngang, huyền, sắc, hỏi, ngã, nặng. Thanh điệu là sự thay đổi Fο và tần số pitch trong quá trình phát âm các nguyên âm.

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

Tiếng bổng hoặc tiếng trầm sẽ tương ứng với dải tần số cao hoặc 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 và được sử dụng bởi 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

Automatic Speech recognition (ASR) là một phần trong Học máy. Để hình dung đơn giản, thay vì chúng ta phải lập trình các quy tắc để chuyển đổi dữ liệu đầu vào từ giọng nói thành đầu ra là văn bản thì mô hình Học máy sẽ được đào tạo bằng cách đưa các tập dữ liệu lớn vào trong một thuật toán tối ưu (chẳng hạn như mô hình Whisper). Trải qua quá trình huấn luyện, mô hình ngày càng cho ra suy luận tốt hơn và có khả năng nhận dạng được tiếng nói của con người.

### Cách thức hoạt động

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

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

Tiếp theo, tín hiệu được chia thành nhiều phần nhỏ hơn (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 noron nhân tạo Transformer.

Mô hình Whisper đã đạt được kết quả ấn tượng trong các thử nghiệm. Độ chính xác mà mô hình mang lại 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 huấn luyện 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 như dịch máy, tóm tắt đoạn 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 đào tạo 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. Các nghiên cứu hiện nay 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 quang phổ Mel với cường độ log 80 channels được tính toán trên cửa sổ 25 mili giây với bước nhảy là 10 mili giây.

Để chuẩn hóa các đặc trưng, đầu vào được tỷ lệ 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 (Hendrycks & Gimpel, 2016), 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.

Năm 2021, Radford và cộng sự đã nhận thấy trong khi tinh chỉnh một mô hình thị giác máy tính trên bộ dữ liệu ImageNet đã tăng 9.2% độ chính xác khi phân loại đối tượng mà không có bất kỳ cải thiện nào về độ chính xác trung bình khi phân loại các đối tượng giống nhau trên 7 bộ dữ liệu khác. Một mô hình có thể đạt hiệu suất vô cùng tốt khi đào tạo trên một bộ dữ liệu vẫn có thể mắc nhiều lỗi cơ bản khi được đánh giá trên bộ dữ liệu khác. Mô hình Whisper nhóm chúng tôi chọn được đánh giá trên speech-to-text thông thường (không phải trên tiếng M’Nông) nhưng nhóm vẫn chọn vì độ chính xác ổn, có thể chấp nhận và thiết kế dễ tiếp cận.

## Multitask Format

Việc dự đoán từ ngữ nào được nói trong một đoạn âm thanh là vấn đề cơ bản của nhận dạng tiếng nói nhưng đó không phải là vấn đề duy nhất. Một hệ thống nhận dạng giọng nói đầy đủ tính năng bao gồm nhiều thành phần bổ sung như phát hiện giọng hoạt động, phân loại người nói và chuẩn hóa văn bản. Những thành phần này thường được xử lý riêng lẻ dẫn đến một hệ thống khá phức tạp. Để giảm độ phức tạp này, mô hình Whisper có thể thực hiện toàn bộ quy trình xử lý giọng nói mà không chỉ là phần nhận diện cơ bản.

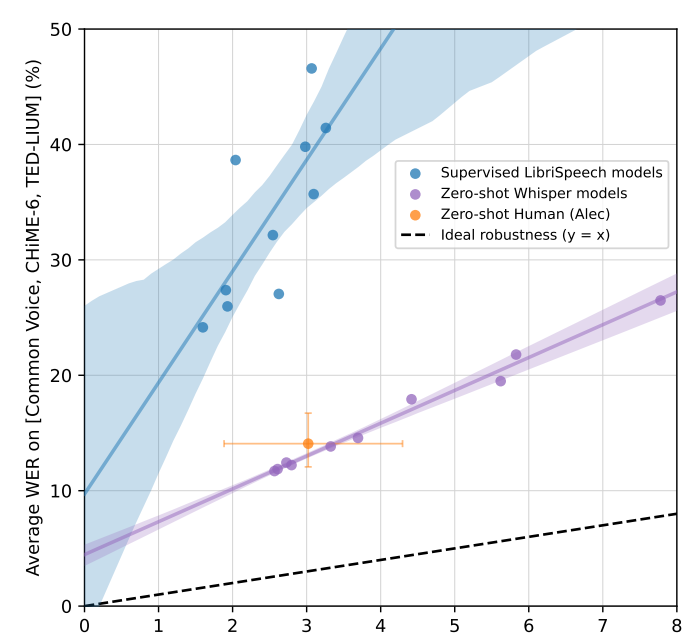
Để làm được điều này, Whisper sử dụng định dạng đơn giản để chỉ định tất cả các nhiệm vụ và thông tin điều kiện dưới dạng một chuỗi các token đầu vào cho bộ giải mã. Whisper chỉ định điểm bắt đầu trước khi dự đoán là một token <start>.

Trước hết, Whisper dự đoán ngôn ngữ đang được nói và biểu diễn bằng một token. Trong trường hợp không có giọng nói được phát hiện, mô hình được huấn luyện để dự đoán token <silence>. Token tiếp theo chỉ định nhiệm vụ “transcription” hoặc “translation” qua token <transcription> hoặc <translation>. Sau đó, mô hình dự đoán liệu có timestamp hay không bằng cách thêm một token <timestamp>. Tại đây, mọi thứ đã được chỉ định đầy đủ và bắt đầu cho ra output.

## Hiệu suất

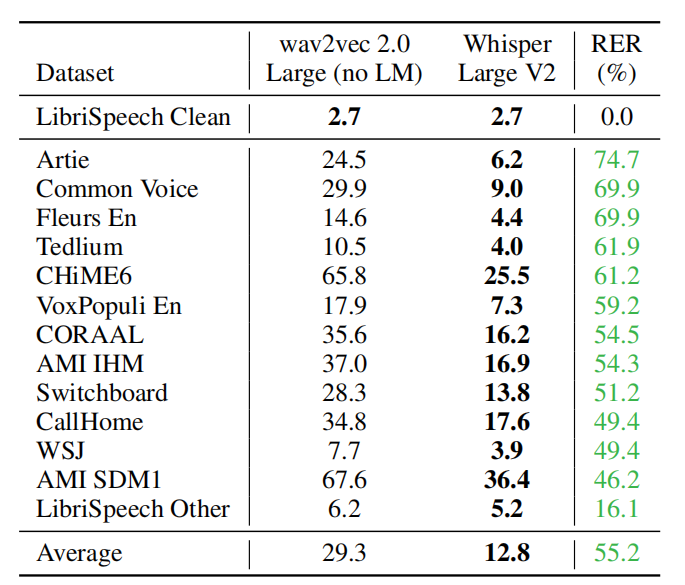
Các mô hình Whisper được huấn luyện trên dữ liệu âm thanh rộng và đa dạng và được đánh giá trong cài đặt zero-shot có thể cho hiệu suất tốt hơn các hệ thống đang tồn tại hiện nay. Có thể so sánh các mô hình Whisper với hiệu suất con người và các mô hình học máy tiêu biểu hiện nay và kiểm tra hiệu suất của chúng.

Các mô hình Whisper zero-shot mạnh mẽ gần với con người. Dù có thể sánh ngang hoặc vượt qua hiệu suất của con người trên tập dev-clean của LibriSpeech, các mô hình LibriSpeech lại mắc khoảng gấp đôi số lỗi so với con người trên các bộ dữ liệu khác. Tuy nhiên trong trường hợp này, độ ổn định ước lượng của các mô hình Whisper zero-shot bao gồm khoảng 95% khoảng tin cậy.



Hình 2. WER trên LibriSpeech dev-clean (%)

Chúng ta xem xét tổng thể hiệu suất trung bình trên nhiều bộ dữ liệu và tính hiệu quả để đo lường sự khác biệt trong hiệu suất giữa một tập dữ liệu tham chiếu và một hoặc nhiều tập dữ liệu bên ngoài. Nhóm nghiên cứu Whisper sử dụng LibriSpeech làm tập dữ liệu tham chiếu do vai trò quan trọng của nó trong nghiên cứu nhận diện giọng nói ngày nay và có nhiều mô hình được phát hành dựa trên LibriSpeech.



Hình 3. So sánh tính hiệu quả trên nhiều bộ dữ liệu

Cả hai mô hình đều cho hiệu suất trong khoảng 0.1% trên LibriSpeech nhưng mô hình Whisper thực hiện tốt hơn nhiều trên các bộ dữ liệu khác và gây ra ít lỗi hơn trung bình 55.2%. Kết quả được đánh giá dưới dạng tỉ lệ lỗi từ (WER) cho cả hai mô hình.

Mặc dù mô hình Whisper cho một tỉ lệ lỗi WER trên LibriSpeech clean-test khá bình thường là 2.5, tương đương với hiệu suất của một mô hình cơ bản vào giữa năm 2019, nhưng mô hình Whisper có các đặc tính mạnh mẽ khác biệt so với các mô hình LibriSpeech được giám sát và vượt qua tất cả các mô hình LibriSpeech trên các bộ dữ liệu khác. Ngay cả mô hình Whisper nhỏ nhất, có chỉ 39 triệu tham số và tỷ lệ lỗi WER trên LibriSpeech test-clean là 6.7, cũng vượt trội hơn so với mô hình LibriSpeech tốt nhất khi được đánh giá trên các bộ dữ liệu khác. Khi so sánh với một con người, những mô hình Whisper cho thấy độ chính xác và tính mạnh mẽ của chúng khá tương đồng. Mặc dù hiệu suất rất gần nhau nhưng mô hình Whisper đạt được độ giảm lỗi trung bình là 55.2% khi được đánh giá trên các bộ dữ liệu nhận diện giọng nói khác.

## Hướng tiếp cận

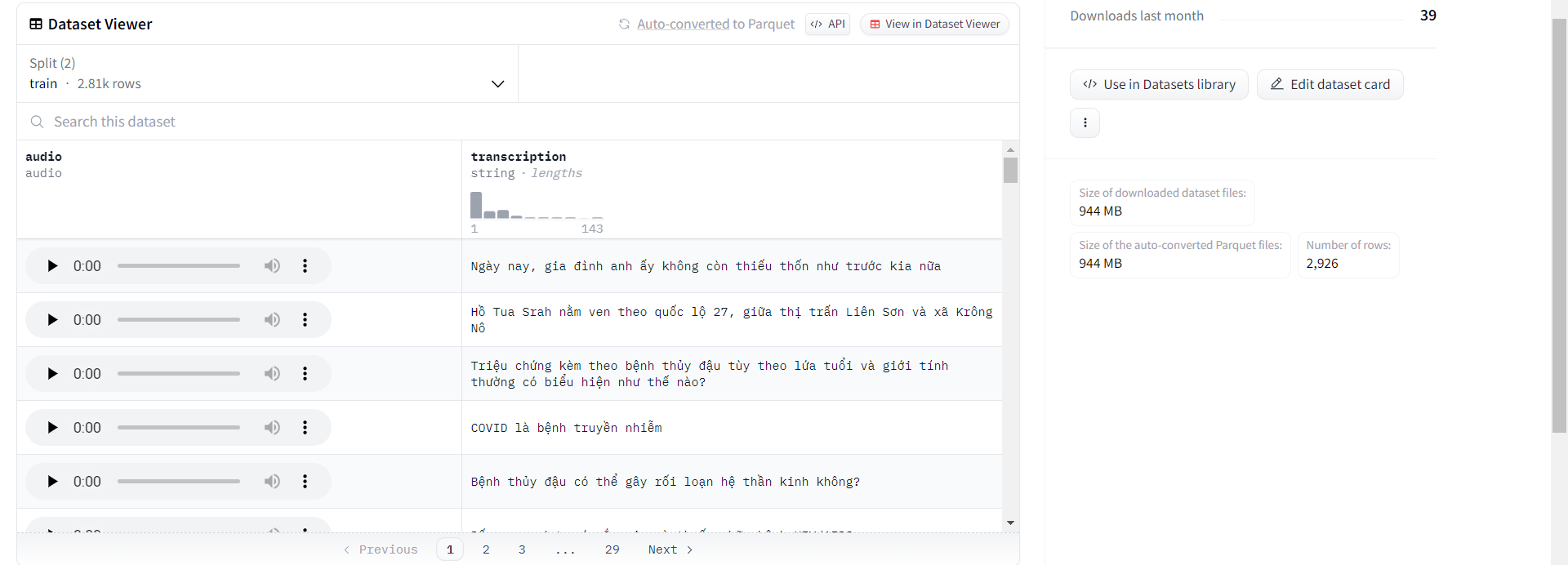
Bài toán được đặt ra là biên dịch tiếng M’Nông ở dạng âm thanh thành văn bản tiếng Việt. Trong ngôn ngữ nói hàng ngày của người dân tộc thiểu số nói chung và dân tộc M’Nông nói riêng, họ sử dụng rất nhiều từ mượn của tiếng nói phổ thông. Nên ta sẽ muốn trong quá trình phiên dịch, mô hình phải vừa phát hiện được tiếng M’Nông lẫn tiếng Việt thông thường. Tiếng Việt và tiếng M’Nông có cùng cấu trúc ngữ pháp giờ đây trở thành một đặc điểm ta có thể tận dụng được để tìm phương pháp xử lý bài toán: tinh chỉnh (fine-tune) lại một mô hình phát hiện tiếng Việt phổ thông.

Nếu ta coi việc biên dịch tiếng M’Nông thành tiếng Việt là đang phiên âm tiếng Việt với các cách phát âm khác nhau của cùng một từ ngữ, hướng đi hiệu quả nhất cho bài toán sẽ là tinh chỉnh lại mô hình phát hiện tiếng nói tự động dựa trên bộ dữ liệu tiếng M’Nông đã thu được. Khi tinh chỉnh, các từ tiếng Việt xuất hiện trong câu sẽ được dịch y nguyên, giúp ta không cần phải huấn luyện lại phần từ vựng tiếng Việt cho mô hình. Với mô hình Whisper trên - đã được huấn luyện cho rất nhiều ngôn ngữ trong đó có tiếng Việt, việc huấn luyện sẽ thuận lợi đáng kể.

# THỰC NGHIỆM

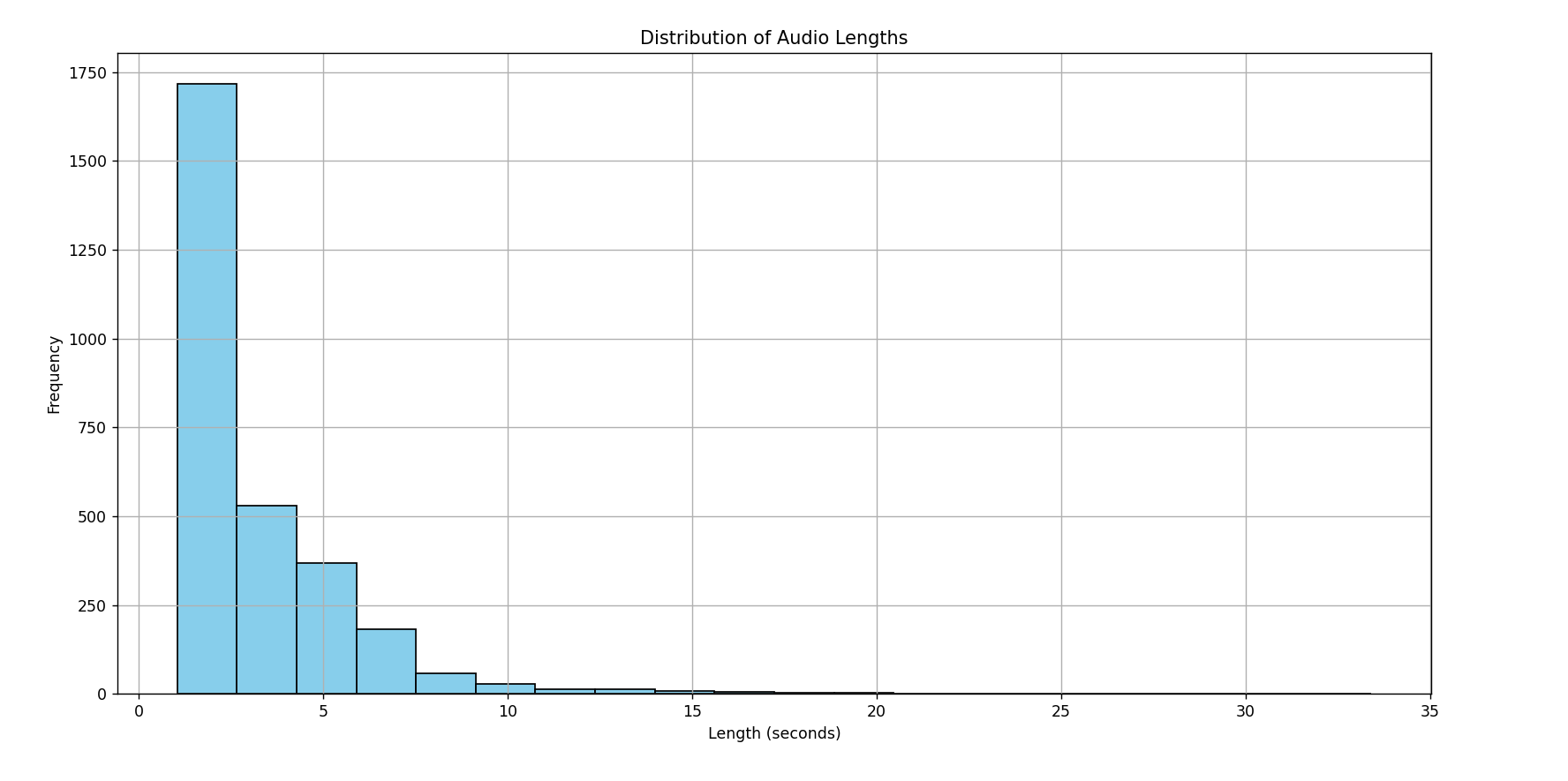
## Dữ liệu thực nghiệm

Dữ liệu được lưu trữ trên nền tảng lưu trữ của Hugging Face. Dữ liệu được chia làm 2 phần: cột “audio” chứa file audio tiếng M’Nông và cột “transcription” chứa bản dịch tiếng Việt của audio tương ứng. Bộ dữ liệu sẽ được chia thành các bộ nhỏ hơn, mang vai trò tương ứng như: “train”, “test”.



Hình 4. Bộ dữ liệu tiếng M’Nông được đẩy lên Hugging Face

Chi tiết bộ dữ liệu: 2926 file ghi âm, bao gồm 1338 file thu âm câu, 1588 file thu âm từ với tổng thời lượng là 2 giờ 34 phút. Bộ dữ liệu này mới chỉ gồm có 3 giọng đọc, thời lượng còn khá thấp, nhưng cũng vừa đủ để ta thử nghiệm phương án được đưa ra.



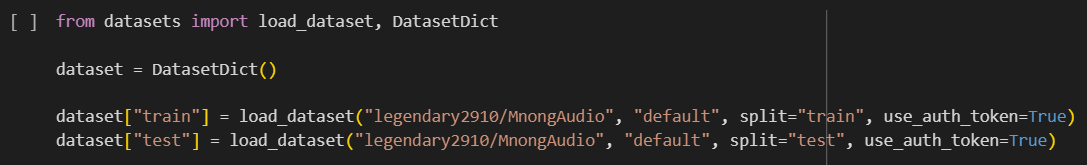
Hình 5. Phân bố độ dài của các file âm thanh trong bộ dữ liệu

## Cài đặt thực nghiệm

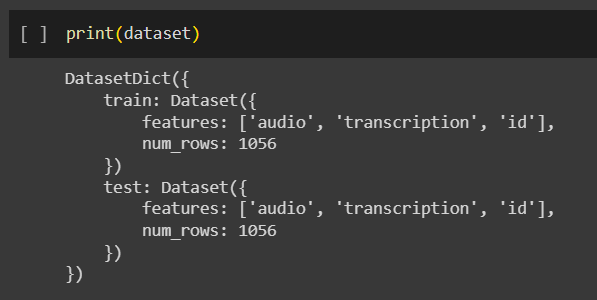
Những bước tiến hành cài đặt thực nghiệm sau đây dựa trên “Step-by-step Guide” được cung cấp bởi chính nhóm phát triển mô hình Whisper.

### Load Dataset

Đầu tiên, ta tải bộ dữ liệu đã được lưu trên Hugging Face Hub. Sau đó lấy phần “train” và “test” của bộ dữ liệu.



Hình 6. Load dataset



Hình 7. Kết quả của load dataset

### Chuẩn bị Tokenizer, Feature Extractor và xủ lý data

Quy trình ASR có thể được phân rã thành ba thành phần:

1. Một Feature Extractor có tác dụng tiền xử lý dữ liệu thô.
2. Một mô hình thực hiện nhiệm vụ sequence-to-sequence mapping.
3. Một Tokenizer thực hiện xử lý kết quả trả về thành văn bản.

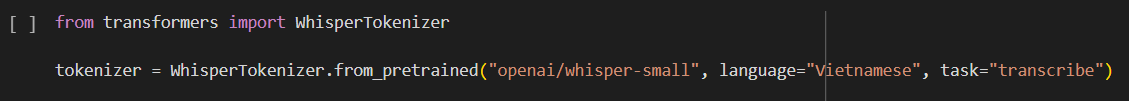
Trong Transformer, mô hình Whisper đi kèm với một bộ feature extractor và một bộ tokenizer lần lượt là WhisperFeatureExtractor và WhisperTokenizer.

### Load WhisperFeatureExtractor

Tốc độ của nguồn âm thanh đầu vào phải phù hợp với tốc độ lấy mẫu mà mô hình của chúng ta lựa chọn. Nếu lấy một mẫu âm thanh với tốc độ lấy mẫu là 16kHz và nghe nó với tốc độ lấy mẫu là 8kHz sẽ khiến âm thanh nghe như đang chạy ở nửa tốc độ. Tương tự, truyền âm thanh với tốc độ lấy mẫu không đúng có thể làm sai lệch mô hình ASR. Bộ feature extractor của Whisper cần đầu vào âm thanh với tốc độ lấy mẫu là 16kHz, vì vậy chúng ta cần tinh chỉnh đầu vào với dữ liệu theo giá trị này.

Bộ feature extractor của Whisper thực hiện hai hoạt động. Trước hết, nó chia độ dài của các mẫu âm thanh sao cho tất cả các mẫu có chiều dài đầu vào là 30 giây. Các mẫu ngắn hơn 30 giây được thêm độ dài bằng cách thêm các số không vào cuối chuỗi (số không trong tín hiệu âm thanh tương ứng với không có tín hiệu hoặc im lặng). Các mẫu dài hơn 30 giây được cắt bớt thành 30 giây.

Hoạt động thứ hai mà bộ feature extractor của Whisper thực hiện là chuyển đổi các mảng âm thanh đã được chia độ dài thành các biểu đồ tần số log-Mel.



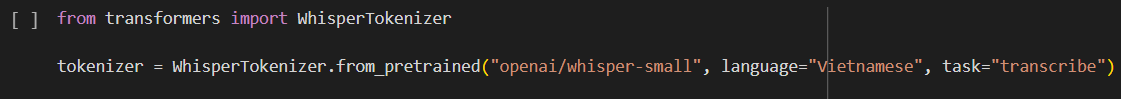
Hình 8. Load WhisperFeatureExtractor

### Load WhisperTokenizer

Mô hình Whisper xuất ra các token văn bản chỉ định chỉ số của văn bản dự đoán trong từ điển các mục từ vựng. Tokenizer ánh xạ một chuỗi các token văn bản thành chuỗi văn bản thực tế (ví dụ: [1169, 3797, 3332] -> "the cat sat").

Trong trước đây, khi sử dụng các mô hình chỉ có trình mã hóa cho ASR, chúng ta giải mã bằng cách sử dụng Connectionist Temporal Classification (CTC). Ở đây, chúng ta cần đào tạo một tokenizer CTC cho mỗi bộ dữ liệu chúng ta sử dụng. Một trong những ưu điểm của việc sử dụng kiến trúc mã hóa-giải mã là chúng ta có thể trực tiếp tận dụng tokenizer từ mô hình được đào tạo trước.

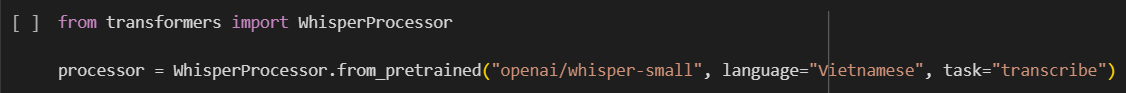
Tokenizer của Whisper được đào tạo trên các bản ghi chép cho 96 ngôn ngữ được đào tạo trước. Do đó, nó có một byte-pair mở rộ rất phù hợp cho hầu hết các ứng dụng ASR đa ngôn ngữ.



Hình 9. Load WhisperTokenizer

### Kết hợp để tạo ra WhisperProcessor

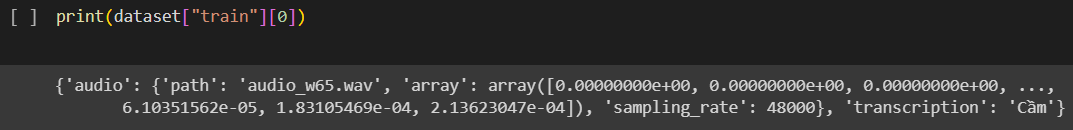
Để đơn giản hóa việc sử dụng bộ feature extractor và tokenizer, chúng ta có thể kết hợp cả hai vào một lớp duy nhất là WhisperProcessor. Đối tượng này kế thừa từ WhisperFeatureExtractor và WhisperProcessor và có thể được sử dụng trên đầu vào âm thanh và dự đoán của mô hình khi cần thiết.



Hình 10. Tạo WhisperProcessor

### Chuẩn bị dữ liệu

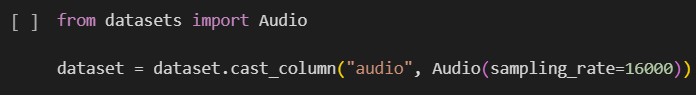
Bây giờ hãy in ra ví dụ đầu tiên trong bộ dữ liệu để xem dữ liệu ở dạng như thế nào.



Hình 11. In data đầu tiên

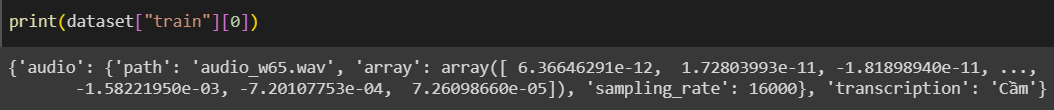
Có thể thấy rằng chúng ta có một mảng âm thanh đầu vào 1 chiều và bản dịch tương ứng. Chúng ta đã nói về tầm quan trọng của tốc độ lấy mẫu và thực tế là chúng ta cần phải điều chỉnh tốc độ lấy mẫu của âm thanh với tốc độ của mô hình Whisper (16kHz). Vì âm thanh đầu vào của chúng ta được lấy mẫu ở 48kHz, cần giảm mẫu xuống 16kHz trước khi truyền vào bộ feature extractor của Whisper.

Chúng ta sẽ đặt tốc độ lấy mẫu của đầu vào âm thanh đúng bằng cách sử dụng phương thức cast\_column của bộ dữ liệu. Thao tác này không thay đổi âm thanh ngay tại chỗ, mà thay vào đó báo hiệu cho bộ dữ liệu để resample các mẫu âm thanh trực tiếp lần đầu tiên chúng được tải.



Hình 12. Điều chỉnh tốc độ lấy mẫu

Việc tải lại mẫu âm thanh sẽ resample nó về tốc độ chúng ta mong muốn.

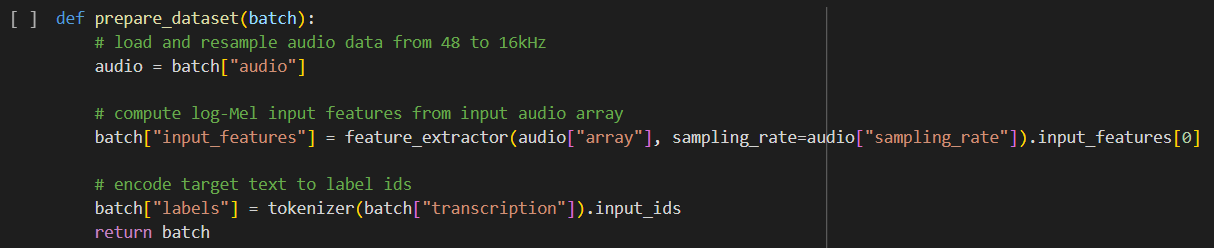


Hình 13. Resample mẫu âm thanh về tốc độ chuẩn

Có thể thấy rằng tốc độ lấy mẫu đã được giảm xuống còn 16kHz, giá trị mảng cũng khác nhau.

Bây giờ chúng ta viết một hàm để chuẩn bị dữ liệu sẵn sàng cho mô hình:

1. Tải và resample dữ liệu âm thanh bằng cách gọi batch[“audio”]
2. Sử dụng bộ feature extractor để tính các đặc trưng đầu vào log-Mel spectrogram từ mảng âm thanh 1 chiều.
3. Mã hóa transcriptions thành các nhãn thông qua tokenizer.



Hình 14. Hàm chuẩn bị dữ liệu

Áp dụng hàm chuẩn bị dữ liệu cho tất cả các ví dụ đào tạo bằng cách sử dụng phương thức .map của bộ dữ liệu.



Hình 15. Áp dụng hàm chuẩn bị dữ liệu

Đến đây chúng ta đã chuẩn bị đầy đủ dữ liệu cho quá trình huấn luyện.

### Huấn luyện và đánh giá

Để Trainer thực hiện các công việc liên quan đến huấn luyện, chúng ta cần:

1. Xác định data collator: Lấy dữ liệu đã được xử lý trước và chuẩn bị các tensor PyTorch.
2. Đo độ chính xác: Cần một hàm compute\_metrics để xử lý tính toán các chỉ số lỗi từ vựng (WER) để đánh giá mô hình.
3. Tải một pre-trained chechkpoint và cấu hình nó cho quá trình huấn luyện.
4. Xác định đối số được sử dụng bởi Trainer trong khi lên lịch huấn luyện.

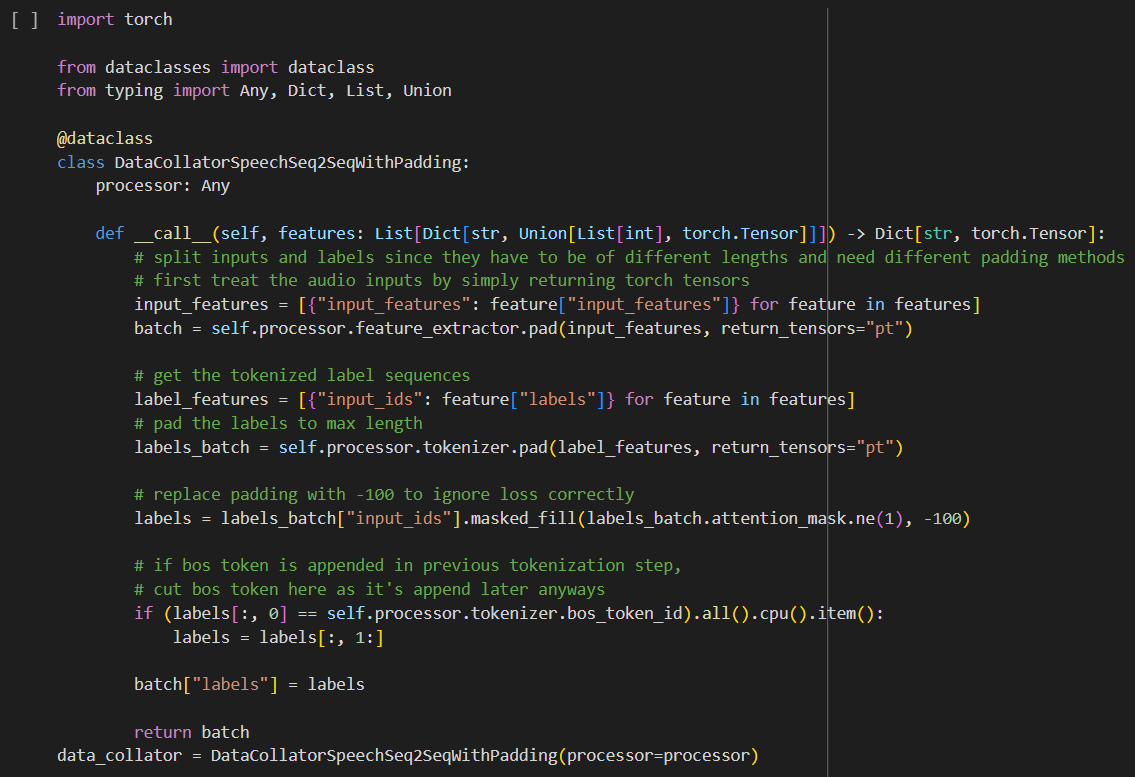
### Định nghĩa Data Collator

Bộ thu thập dữ liệu cho một mô hình nói chung sequence-to-sequence xử lý input\_features và nhãn một cách độc lập: input\_features phải được xử lý bởi bộ feature extractor và gán nhãn bởi tokenizer.

Các input\_features đã được chia độ dài thành 30 giây và chuyển đổi thành log-Mel spectrogram có kích thước cố định, vì thế những gì chúng ta cần phải làm là chuyển đổi chúng thành các tensor PyTorch. Chúng ta thực hiện điều này bằng cách sử dụng phương thức .pad của bộ feature extractor với return\_tensors=pt.

Ngược lại, các nhãn không được chia độ dài. Đầu tiên, chúng ta chia độ dài các chuỗi đến độ dài tối đa trong lô bằng cách sử dụng phương thức .pad của tokenizer. Các token sau đó được thay thế bằng -100 để những token này không được tính vào khi tính toán mất mát. Sau đó, chúng ta cắt bỏ token bắt đầu của bản trích từ đầu của chuỗi nhãn vì chúng ta sẽ thêm nó vào sau này trong quá trình huấn luyện.

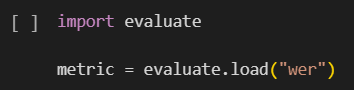
Chúng ta có thể tận dụng WhisperProcessor đã định nghĩa trước đó để thực hiện cả hai thao tác feature extractor và tokenizer. Sau đó khởi tạo bộ thu thập dữ liệu vừa định nghĩa.



Hình 16. Định nghĩa và khởi tạo Data Collator

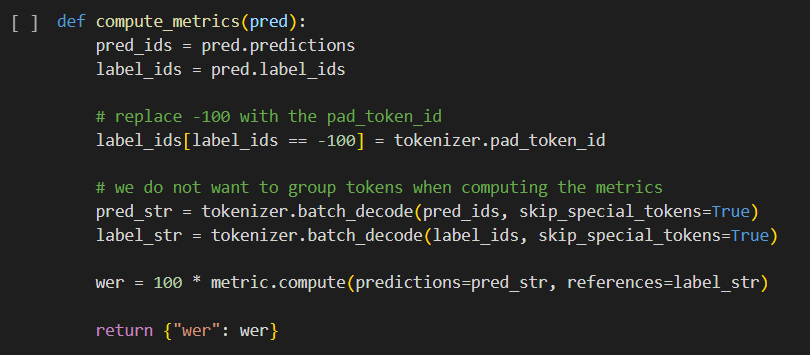
### Số liệu đánh giá

Tiếp theo, cần xác định chỉ số đánh giá mà chúng ta sẽ sử dụng trên tập đánh giá của mình. Ta sẽ sử dụng chỉ số Word Error Rate (WER), chỉ số 'de-facto' để đánh giá các hệ thống ASR.



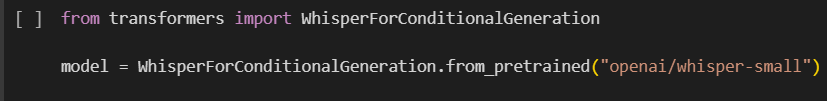
Hình 17. Xác định WER

Sau đó, chúng ta chỉ cần định nghĩa một hàm lấy dự đoán của mô hình và trả về chỉ số WER. Hàm này được gọi là compute\_metrics, đầu tiên thay thế -100 bằng pad\_token\_id trong label\_ids (hoàn ngược bước chúng ta áp dụng trong bộ thu thập dữ liệu để bỏ qua đúng các token). Sau đó, nó giải mã predicted và label ids thành chuỗi. Cuối cùng, nó tính toán chỉ số WER giữa các dự đoán và nhãn tham chiếu.



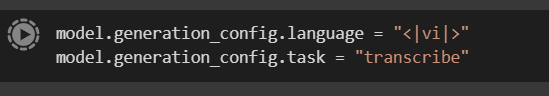
Hình 18. Hàm trả về WER

### Tải pre-trained checkpoint



Hình 19. Tải một pre-trained checkpoint

Mô hình Whisper có khả năng dự đoán những ngôn ngữ khác nhau xuất hiện trong câu được huấn luyện. Điều này khiến cho việc huấn luyện và dự đoán gặp khó khăn khi những từ mới của tiếng M’Nông xuất hiện có thể được mô hình dự đoán thành một ngôn ngữ khác. Nên ra ép mô hình phải sử dụng tiếng Việt và tác vụ phiên âm bằng việc cấu hình lại mô hình.



Hình 20. Cấu hình lại mô hình

### Định nghĩa Traing Arguments

Trong bước cuối cùng, chúng ta định nghĩa tất cả các tham số liên quan đến quá trình huấn luyện. Một phần của các tham số được giải thích dưới đây:

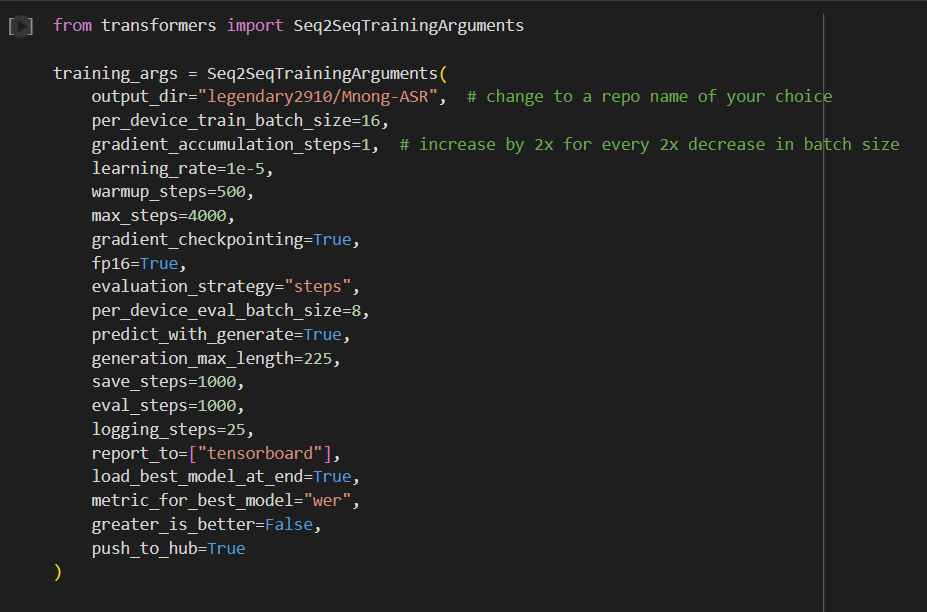
- `output\_dir`: thư mục cục bộ để lưu trữ trọng số của mô hình. Đây cũng sẽ là tên kho chứa trên Hugging Face Hub.

- `generation\_max\_length`: số lượng tối đa các token để tạo ra theo cách phát sinh tự động trong quá trình đánh giá.

- `save\_steps`: trong quá trình huấn luyện, các điểm kiểm tra trung gian sẽ được lưu và tải lên bất đồng bộ lên Hub mỗi save\_steps.

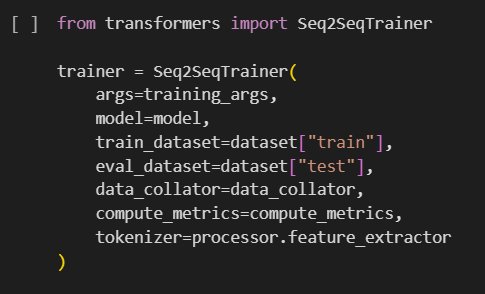
- `eval\_steps`: trong quá trình huấn luyện, đánh giá các điểm kiểm tra trung gian sẽ được thực hiện mỗi eval\_steps.

- `report\_to`: nơi lưu trữ nhật ký huấn luyện. Các nền tảng được hỗ trợ là "azure\_ml", "comet\_ml", "mlflow", "neptune", "tensorboard" và "wandb". Chọn nền tảng ưa thích của bạn hoặc để là "tensorboard" để đăng nhập vào Hub.



Hình 21. Ý nghĩa các tham số liên quan

Chúng ta có thể chuyển các đối số huấn luyện cho Trainer cùng với mô hình, bộ dữ liệu, bộ thu thập dữ liệu và hàm compute\_metrics.



Hình 22. Chuyển các đối số huấn luyện cho Trainer

Và với thao tác vừa rồi, mô hình đã sẵn sàng để huấn luyện.

### Huấn luyện mô hình

Với tổng gần 3000 dữ liệu thu được, nhóm sử dụng dữ liệu trên để huấn luyện và trích ra 8% tổng dữ liệu (khoảng 230 dữ liệu) cho việc đánh giá thông qua phương pháp WER (sẽ đề cập trong phần tiếp theo).

Việc huấn luyện được thực hiện trên Google Colab trong tổng thời gian 3 tiếng, với các thông số kỹ thuật như sau:

- CPU: Intel(R) Xeon(R) CPU @ 2.00GHz.

- GPU: Tesla V100-SXM2-16GB.

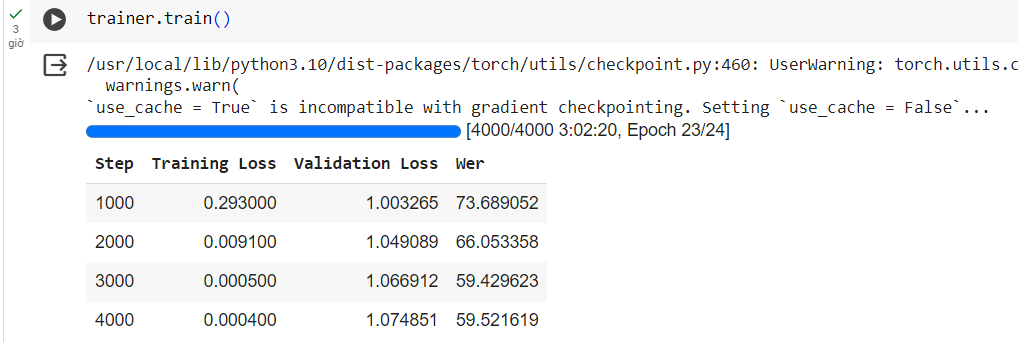
- RAM: 52GB.

Sau khi huấn luyện xong, mô hình đã được huấn luyện sẽ được đẩy lên Huggingface Hub cho việc sử dụng sau này dưới tên “legendary2910/MNong-ASR”.

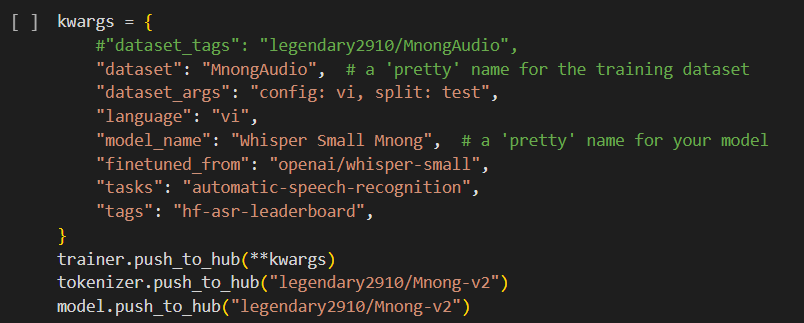


Hình 23. Huấn luyện mô hình

Output:



Hình 24. Kết quả huấn luyện





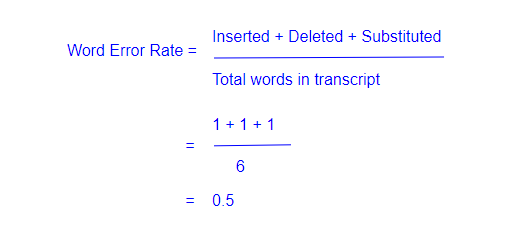
Hình 25. Đưa kết quả lên Hugging Face Hub

## Đánh giá

Trong hệ thống ASR thường sử dụng metric là Word Error Rate (WER) cho bài toán này. WER so sánh đầu ra dự đoán với transcript mục tiêu, từng từ hoặc từng ký tự với ký tự để phát hiện sự khác biệt giữa hai câu.

Sự khác biệt này có thể bao gồm các từ có trong transcript nhưng không được dự đoán (Deletion), một từ không có trong transcript nhưng lại được thêm vào dự đoán (Insertion), hoặc một từ được thay thế giữa dự đoán và transcript (Substitution).

Công thức tính metric đo tỉ lệ của số từ khác biệt giữa transcript và dự đoán.

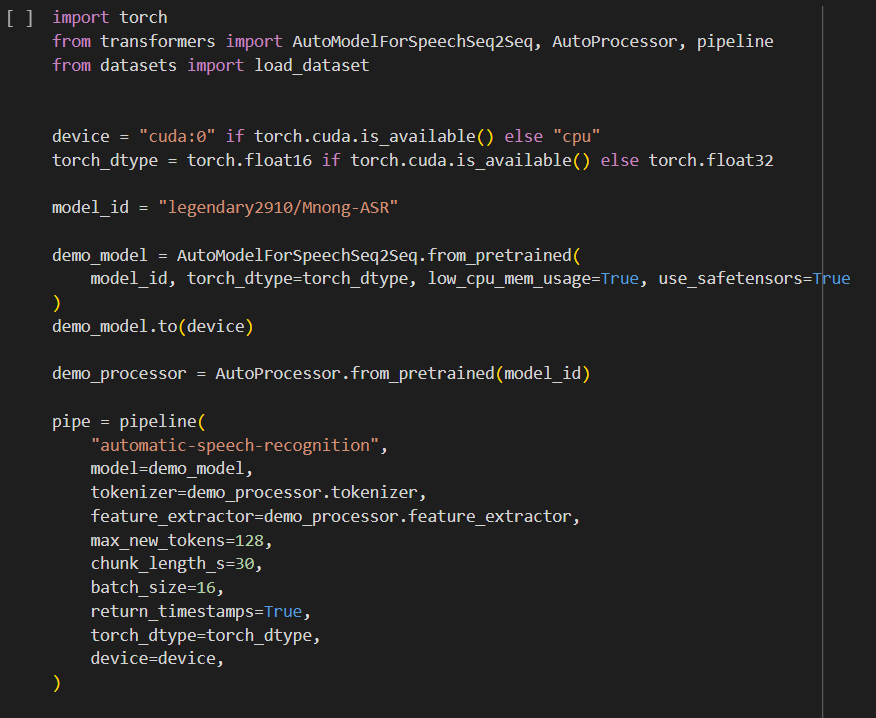


Hình 26. Công thức tính WER

Hiện tại có thể thấy ở hình 24, kết quả huấn luyện mô hình còn chưa tốt với WER khá cao. Nguyên nhân có thể đến từ việc dữ liệu dù đa dạng, nhưng không có quá nhiều giọng đọc cho cùng một câu, và tất cả các dữ liệu đều khá khác nhau.

## Thực nghiệm và kết quả

Vì mô hình đã được lưu lại trên Huggingface Hub, ta có thể tạo ra một ví dụ đơn giản cho việc sử dụng mô hình trong tương lai. Bằng cách này, ta có thể sử dụng mô hình để thực hiện phiên dịch những file âm thanh trong bộ dữ liệu sử dụng để đánh giá. Qua đó có thể quan sát được kết quả ra sao trong quá trình huấn luyện.



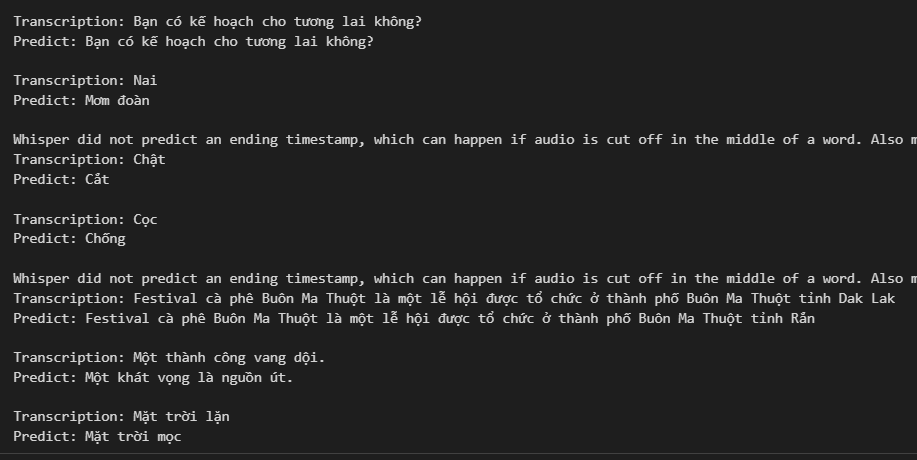
Hình 27. Ví dụ đơn giản sử dụng mô hình đã qua huấn luyện



Hình 28. Kết quả khi thực hiện kiểm tra trên bộ dữ liệu sử dụng để kiểm định

Ở phần chụp lại kết quả trên, phần “Transcription” hiển thị cho phần dịch đúng tương ứng với file âm thanh, phần “Predict” hiển thị kết quả mà mô hình đã dự đoán ý nghĩa của câu nói có trong file âm thanh. Ta có thể thấy mô hình dự đoán những từ ta đã sử dụng trong quá trình huấn luyện với độ chính xác cao.

Vì sự hạn chế của bộ dữ liệu, vẫn có những từ ngữ mà mô hình chưa được học xuất hiện trong bộ dữ liệu đánh giá, khiến cho kết quả dự đoán bị sai đi hoặc sai hoàn toàn.



Hình 29. Kết quả khi thực hiện kiểm tra trên bộ dữ liệu sử dụng để kiểm định

# KẾT LUẬN

Đây là một đề tài mới lạ đối với nhóm, yêu cầu nhiều sự tâm huyết cũng như thời gian dành cho đề tài. Việc phiên dịch tiếng nói của những dân tộc ít người có thể giúp cho công tác giáo dục tiếng nói và trong sinh hoạt hàng ngày của người dân tộc M’Nông nói riêng và người dân tộc thiểu số nói chung trở nên thuận tiện hơn. Việc chưa hề có bất kì dự án nào tương tự cũng gây nên không ít khó khăn cho nhóm khi thực hiện đề tài.

Qua việc huấn luyện trên, ta có thể thấy được bất cập của bộ dữ liệu ảnh hưởng đến chất lượng của kết quả ra sao. Bộ dữ liệu chỉ có sự tham gia của 3 giọng đọc, số lượng dữ liệu dạng từ vựng lại chiếm quá nhiều trong bộ dữ liệu; những câu dài được sử dụng để phiên dịch sang tiếng M’Nông cho quá trình thu âm lại quá phức tạp cho một ngôn ngữ thô sơ như tiếng M’Nông, làm cho việc phiên dịch trở nên khó khăn. Những lý do trên đã khiến cho kết quả của mô hình không được như mong đợi.

Ngoài ra, về mặt kỹ thuật cũng xảy ra rất nhiều vấn đề trong quá trình thực hiện. Từ xử lý việc mô hình tự động phát hiện đa ngôn ngữ, cho đến cấu hình lại mô hình, lưu trữ mô hình sau khi huấn luyện,…

Tuy có nhiều khó khăn như thế, xong kết quả có được cũng có một vài dấu hiệu khả quan, như việc mô hình đã có thể dịch được khá sát những từ vựng đã được sử dụng trong quá trình huấn luyện. Đây sẽ là nền tảng để nhóm khắc phục những nhược điểm trên nhằm cải thiện kết quả, cũng như hy vọng vào việc đề tài này có thể đem lại nhiều những giá trị thực tiễn hơn để phục vụ cộng đồng, phục vụ đất nước.

# 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.

[4] Towards Datascience. Foundations of NLP Explained BLEU Score and WER Metrics