ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN

BÁO CÁO ĐỒ ÁN
MÔN DS313.O21 - XỬ LÝ THÔNG TIN GIỌNG NÓI

Đề tài: Fine-Tuning ASR Models: Whisper and PhoWhisper on Vietnamese YouTube Dataset

GVHD: ThS. Nguyễn Thành Luân

Nhóm sinh viên thực hiện: 08

1. Trần Huỳnh Duy Đăng	MSSV: 21521922
2. Nông Đức Thắng	MSSV: 21522593
3. Nguyễn Đạt Tuấn	MSSV: 21522754
4. Lê Hữu Tài	MSSV: 21522562
5. Nguyễn Phú An	MSSV: 21521807

Tp. Hồ Chí Minh, 05/2024

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

……, ngày…….....tháng……năm 2024

Người nhận xét

(Ký tên và ghi rõ họ tên)

BẢNG PHÂN CÔNG, ĐÁNH GIÁ THÀNH VIÊN:

Bảng 1: Bảng phân công, đánh giá thành viên

Họ và tên	MSSV	Phân công	Đánh giá
Nông Đức Thắng	21522593	Tuần 1: Nghiên cứu chủ đề Tuần 2: Tìm kiếm thông tin Tuần 3: Thực nghiệm	Tuần 1: tốt Tuần 2: tốt Tuần 3: tốt
Nguyễn Phú An	21521807	Tuần 1: Nghiên cứu chủ đề Tuần 2: Tìm kiếm thông tin Tuần 3: Đánh giá, bổ sung	Tuần 1: tốt Tuần 2: tốt Tuần 3: tốt
Trần Huỳnh Duy Đăng	21521922	Tuần 1: Nghiên cứu chủ đề Tuần 2: Tìm kiếm thông tin Tuần 3: Đánh giá, bổ sung	Tuần 1: tốt Tuần 2: tốt Tuần 3: tốt
Lê Hữu Tài	21522562	Tuần 1: Nghiên cứu chủ đề Tuần 2: Tìm kiếm thông tin Tuần 3: Thực nghiệm	Tuần 1: tốt Tuần 2: tốt Tuần 3: tốt
Nguyễn Đạt Tuấn	21522754	Tuần 1: Nghiên cứu chủ đề Tuần 2: Tìm kiếm thông tin Tuần 3: Đánh giá, bổ sung	Tuần 1: tốt Tuần 2: tốt Tuần 3: tốt

LỜI MỞ ĐẦU

Giọng nói là một trong những phương tiện giao tiếp cơ bản và quan trọng nhất của con người. Khả năng nhận dạng và hiểu biết về thông tin được truyền đạt qua giọng nói không chỉ là yếu tố quan trọng trong giao tiếp hàng ngày mà còn có ứng dụng sâu rộng trong nhiều lĩnh vực, đặc biệt là trong ngành Khoa học Dữ liệu. Trong bối cảnh thế giới ngày càng chuyển đổi sang môi trường số hóa, nhu cầu sử dụng công nghệ xử lý thông tin giọng nói ngày càng tăng cao. Đặc biệt, trong môi trường đại học, việc áp dụng công nghệ này có thể đem lại nhiều lợi ích, từ việc tăng cường trải nghiệm học tập cho sinh viên đến việc tối ưu hóa quá trình giảng dạy và hỗ trợ hệ thống quản lý giáo dục.

Đồ án này tập trung vào việc nghiên cứu và phát triển các phương pháp, kỹ thuật xử lý thông tin giọng nói phục vụ cho môi trường Đại học Công nghệ Thông tin - ĐHQG.HCM. Chúng tôi hy vọng rằng những nỗ lực này sẽ góp phần vào việc nâng cao chất lượng giáo dục và tạo ra những giải pháp công nghệ tiên tiến trong cộng đồng học thuật và sinh viên của trường. Một trong những mục tiêu chính của đồ án là giới thiệu đề tài “Fine-Tuning ASR Models: Whisper and PhoWhisper on Vietnamese YouTube Dataset” và phương pháp Whisper với PhoWhisper được đề xuất trong bài này. Việc này nhằm mục đích mang lại cái nhìn toàn diện và cập nhật nhất về những tiến bộ mới nhất trong lĩnh vực xử lý thông tin giọng nói và áp dụng chúng vào môi trường đại học Công nghệ Thông tin - ĐHQG.HCM.

Đồ án cũng đặt ra mục tiêu tạo ra sự hiểu biết sâu sắc và ứng dụng thực tiễn của phương pháp Whisper và PhoWhisper trong việc cải thiện trải nghiệm học tập và giảng dạy tại trường Đại học Công nghệ Thông tin - ĐHQG.HCM. Chúng tôi hy vọng rằng đồ án này sẽ không chỉ là một đóng góp quan trọng vào lĩnh vực xử lý thông tin giọng nói mà còn là một công cụ hữu ích và cần thiết trong việc nâng cao chất lượng giáo dục tại trường.

Chúng tôi xin chân thành cảm ơn tới giảng viên Nguyễn Thành Luân đã tận tâm hướng dẫn chúng tôi từng bước nhỏ trong việc thực hiện đồ án. Chúc thầy thật nhiều sức khỏe để bước tiếp trong sự nghiệp của mình.

DANH MỤC CÁC BẢNG, HÌNH ẢNH

Danh mục các bảng:

Chương 3 - Mục 3

Dataset	Training size (hours)	Validation size (hours)	Test size (hours)	#syllables in training set (min – max | average)
CMV–Vi 14	3.04	0.41	1.35	1 – 14 | 7.55
VIVOS	13.94	0.98	0.75	2 – 30 | 13.25
viet_youtube_asr_corpus_v2	110	None	22	1 – 9 | 3

Bảng 1: So sánh dung lượng giữa các bộ dữ liệu tiếng Việt

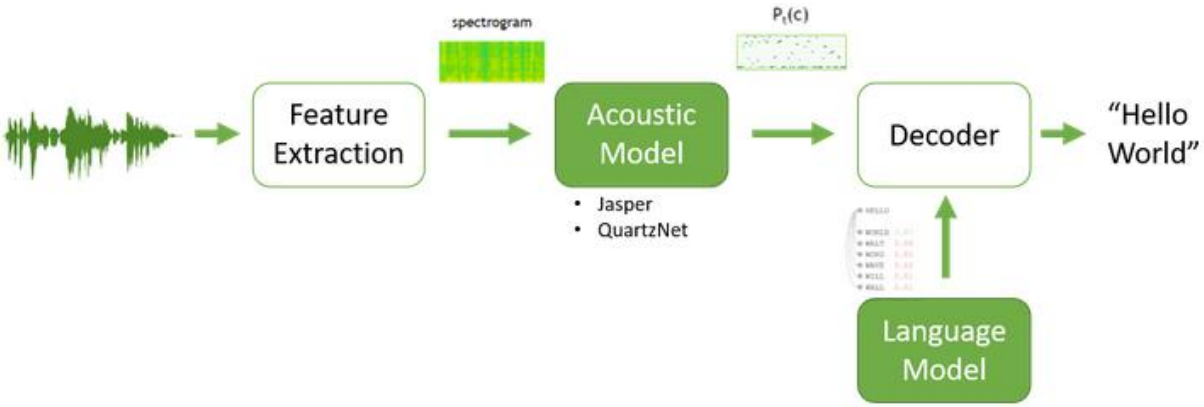
Chương 3 - Mục 5

Model	#paras	WER	CER	BLEU
Whisper(original)	74M	49.9	27.81	30.14
Whisper(fine-tuning)		6.88	3.07	86.14
PhoWhisper(original)		19.74	8.07	71.83
PhoWhisper(fine-tuning)		**4.76**	**2.09**	**90.38**

Bảng 2: Kết quả độ đo (%) của mô hình trên bộ dữ liệu "viet_youtube_asr_corpus_v2"

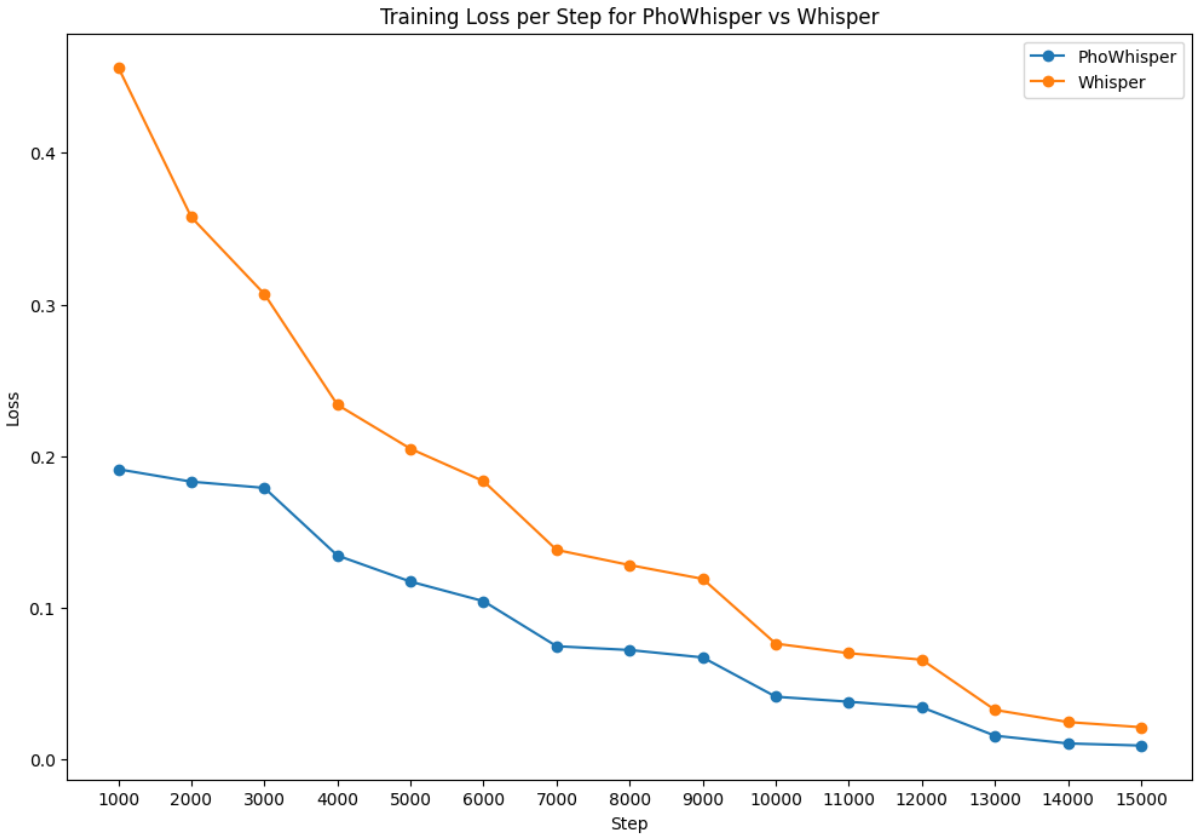
Danh mục hình ảnh:

Chương 3 - Mục 4.1



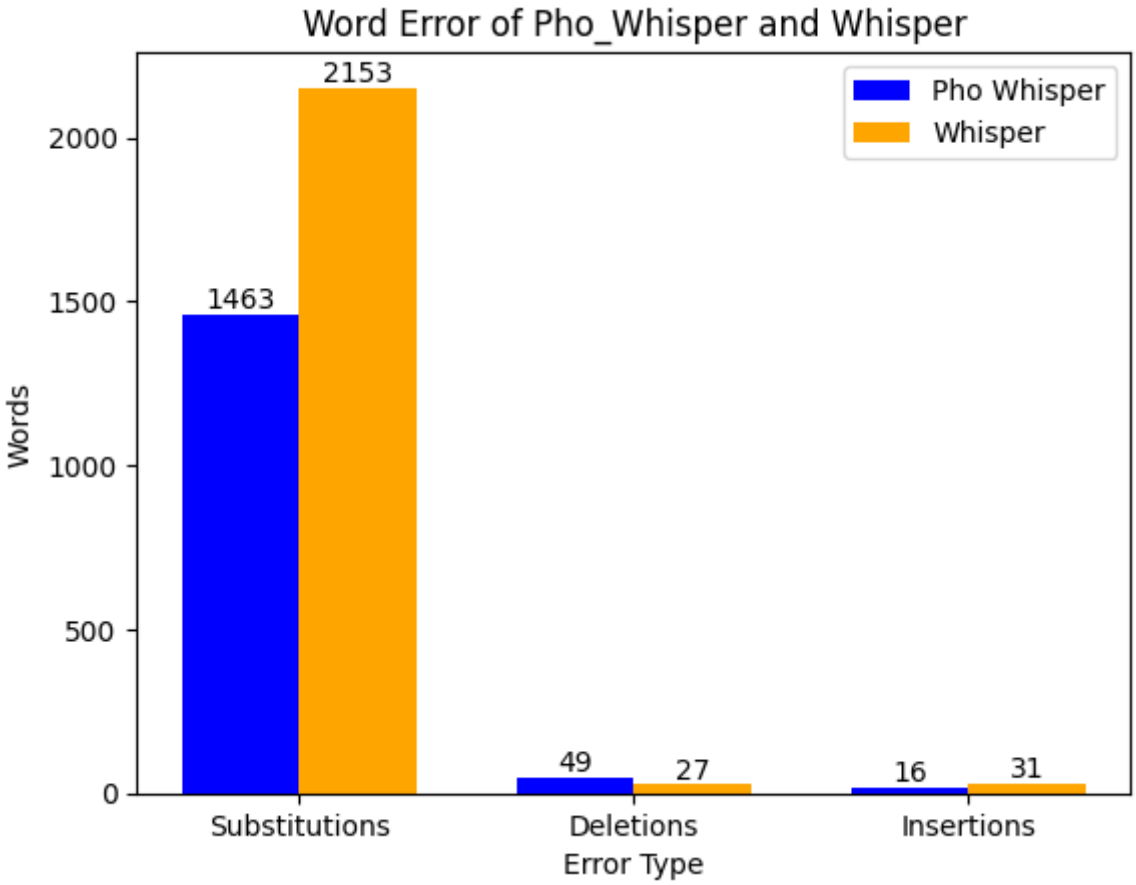
Hình 1: Tổng quan quy trình thực hiện

Chương 3 - Mục 6.1



Hình 2: Đường biểu diễn giá trị Loss của mô hình qua mỗi bước

Chương 3 - Mục 6.2



Hình 3: Biểu đồ thể hiện số từ mỗi loại lỗi trong mỗi mô hình

MỤC LỤC

Chương 1: GIỚI THIỆU CHUNG

Giọng nói là một phương tiện truyền đạt thông tin quan trọng và phổ biến nhất trong giao tiếp của con người. Không chỉ là cách chúng ta truyền đạt ý định và suy nghĩ, giọng nói còn phản ánh bản sắc cá nhân và văn hóa. Điều này tạo nên sự đa dạng và phong phú trong giọng nói của mỗi người. Con người tạo ra giọng nói và lời nói thông qua một quá trình phức tạp bắt đầu từ việc hít khí vào hệ thống hô hấp, tiếp tục qua hệ thống thanh nhạc trong đó các dây thanh quản rung lên để tạo ra âm thanh cơ bản. Các cơ quan phát âm như lưỡi, hàm dưới, hàm trên, môi và ống tiêu hóa được sử dụng để điều chỉnh và hình thành các âm tiếng khác nhau. Người nói cũng điều chỉnh ngữ điệu và cách lên âm để truyền đạt ý nghĩa và cảm xúc.

Trong xã hội hiện đại, giọng nói không chỉ được sử dụng trong giao tiếp trực tiếp mà còn trong các công nghệ trợ giúp như trợ lý ảo, hệ thống nhận dạng giọng nói, và tổng hợp giọng nói. Do đó, học xử lý giọng nói là quan trọng vì nó cho phép chúng ta hiểu và tương tác với cách con người giao tiếp tự nhiên nhất. Công nghệ xử lý giọng nói đang phát triển mạnh mẽ, mang lại nhiều tiện ích và ứng dụng hứa hẹn trong nhiều lĩnh vực từ y tế đến giáo dục và giao thông.

Môn học “Xử lý thông tin giọng nói” là một lĩnh vực tập trung vào việc phân tích, hiểu và xử lý thông tin từ giọng nói. Chúng ta sẽ được học về các khía cạnh như nhận dạng người nói, chuyển đổi giọng nói thành văn bản, phân tích ngữ cảnh và cảm xúc từ giọng nói, tổng hợp giọng nói, và các ứng dụng thực tiễn của xử lý thông tin giọng nói. Quá trình xử lý giọng nói yêu cầu người nghe nghe một luồng lời nói liên tục, chia luồng thành các đơn vị lời nói khác nhau, giải mã tất cả âm thanh có ý nghĩa trong đó và cuối cùng hiểu được thông điệp dự định của nó. Trong lĩnh vực máy học, các bài toán liên quan đến xử lý giọng nói như Automatic Speech Recognition (ASR), Text To Speech (TTS), Speech To Text (STT), và Speech Translate (ST),.. đang trở nên ngày càng quan trọng. ASR tập trung vào việc nhận dạng và chuyển đổi giọng nói thành văn bản, đây là cơ sở cho nhiều ứng dụng như trợ lý ảo và hệ thống điều khiển bằng giọng nói. TTS, ngược lại, tạo ra giọng nói tự nhiên từ văn bản và được sử dụng trong sách nói và hệ thống đọc tin tức tự động. STT chuyển đổi giọng nói thành văn bản, hữu ích cho việc

ghi chú và tạo văn bản từ hội thoại. Cuối cùng, ST dịch giọng nói từ một ngôn ngữ sang một ngôn ngữ khác, mở ra nhiều cơ hội trong việc giao tiếp đa ngôn ngữ.

Ở môn học này, chúng tôi hướng đến chủ đề Automatic Speech Recognition. Automatic Speech Recognition (ASR), hay nhận dạng giọng nói tự động, là công nghệ chuyển đổi giọng nói của con người thành văn bản bằng cách sử dụng các kỹ thuật học máy và xử lý tín hiệu. Hệ thống ASR hoạt động thông qua các bước: thu nhận âm thanh, xử lý tín hiệu để loại bỏ nhiễu, trích xuất đặc trưng âm thanh (như MFCCs), và sử dụng các mô hình học máy như HMMs, RNNs hoặc DNNs để nhận dạng và giải mã tín hiệu thành văn bản. Ứng dụng của ASR rất đa dạng, bao gồm trợ lý ảo (Siri, Google Assistant), hệ thống điều khiển bằng giọng nói, chuyển đổi giọng nói thành văn bản, dịch tự động, và hỗ trợ giáo dục. Tuy nhiên, ASR đối mặt với nhiều thách thức như môi trường ồn ào, đa dạng giọng nói, tốc độ và ngữ cảnh nói nhanh, cũng như cập nhật từ vựng và ngôn ngữ mới. Mặc dù còn nhiều thách thức, ASR hứa hẹn mang lại nhiều lợi ích to lớn trong nhiều lĩnh vực của đời sống và công nghệ.

Automatic Speech Recognition for Vietnamese là một trong những bài toán cơ bản và quan trọng của lĩnh vực Automatic Speech Recognition. Bài toán ASR cho tiếng Việt đối diện với nhiều thách thức đặc biệt. Sự đa dạng về ngữ điệu, âm sắc và cách phát âm tùy thuộc vào vùng miền tạo ra những biến thể ngôn ngữ phong phú, làm tăng độ phức tạp của việc nhận diện và hiểu biết giọng nói. Ngoài ra, môi trường ồn ào, sự thay đổi trong ngữ cảnh và ngôn ngữ cũng đều ảnh hưởng đến hiệu suất của hệ thống ASR. Tuy nhiên, qua sự phát triển của công nghệ và nghiên cứu, ASR cho tiếng Việt đã và đang ngày càng được cải thiện về độ chính xác và hiệu suất. Các kỹ thuật tiên tiến trong lĩnh vực học máy và xử lý tín hiệu đã được áp dụng để tạo ra các hệ thống ASR mạnh mẽ hơn, đáp ứng được nhu cầu ngày càng cao của thị trường và người dùng. ASR cho tiếng Việt không chỉ là một công cụ hữu ích trong việc tạo ra các ứng dụng trợ lý ảo, hệ thống điều khiển bằng giọng nói mà còn là một lĩnh vực nghiên cứu quan trọng trong việc phát triển ngôn ngữ và giao tiếp tự nhiên. Việc giới thiệu và ứng dụng các giải pháp ASR cho tiếng Việt không chỉ mang lại lợi ích công nghệ mà còn đóng góp vào sự phát triển và bảo tồn của văn hóa và ngôn ngữ Việt Nam.

Chương 2: AUTOMATIC SPEECH RECOGNITION FOR VIETNAMESE

1. Giới thiệu bài toán "Automatic Speech Recognition for Vietnamese"

Bài toán "Automatic Speech Recognition for Vietnamese" tập trung vào việc phát triển các hệ thống nhận dạng giọng nói tự động và chính xác cho tiếng Việt. Trong bối cảnh ngày nay, với sự gia tăng của ứng dụng trí tuệ nhân tạo và giao tiếp tự nhiên, ASR cho tiếng Việt trở thành một lĩnh vực nghiên cứu quan trọng. Bài toán này đặt ra nhiều thách thức đối với cộng đồng nghiên cứu, bao gồm sự đa dạng về ngữ điệu, âm sắc, và biến thể ngôn ngữ tùy thuộc vào vùng miền. Điều này yêu cầu sự phát triển và ứng dụng các thuật toán và mô hình học máy tiên tiến để nhận diện và hiểu biết giọng nói một cách chính xác. Mục tiêu của bài toán là tạo ra các hệ thống ASR có khả năng hoạt động tốt trong nhiều điều kiện môi trường và đa dạng ngôn ngữ, từ giọng nói hàng ngày đến các ứng dụng chuyên biệt như y tế hoặc tài chính. ASR cho tiếng Việt không chỉ giúp cải thiện trải nghiệm người dùng trong việc tương tác với máy tính và các thiết bị thông minh, mà còn đóng vai trò quan trọng trong việc phát triển các ứng dụng giao tiếp tự nhiên và hỗ trợ ngôn ngữ đa dạng.

2. Giới thiệu bài báo

Để đáp ứng bài toán "Automatic Speech Recognition for Vietnamese", chúng tôi đã tiến hành tham khảo từ nhiều nguồn tài liệu khác nhau, bao gồm các bài báo khoa học, tài liệu hướng dẫn và công trình nghiên cứu trước đó trong lĩnh vực xử lý ngôn ngữ tự nhiên và nhận dạng giọng nói. Trong quá trình nghiên cứu, chúng tôi đã tập trung vào việc phát triển và sử dụng hai mô hình chính là Whisper và PhoWhisper đến từ 2 bài báo dưới đây.

2.1. Bài báo "Robust Speech Recognition via Large-Scale Weak Supervision"

2.1.1. Giới thiệu chung

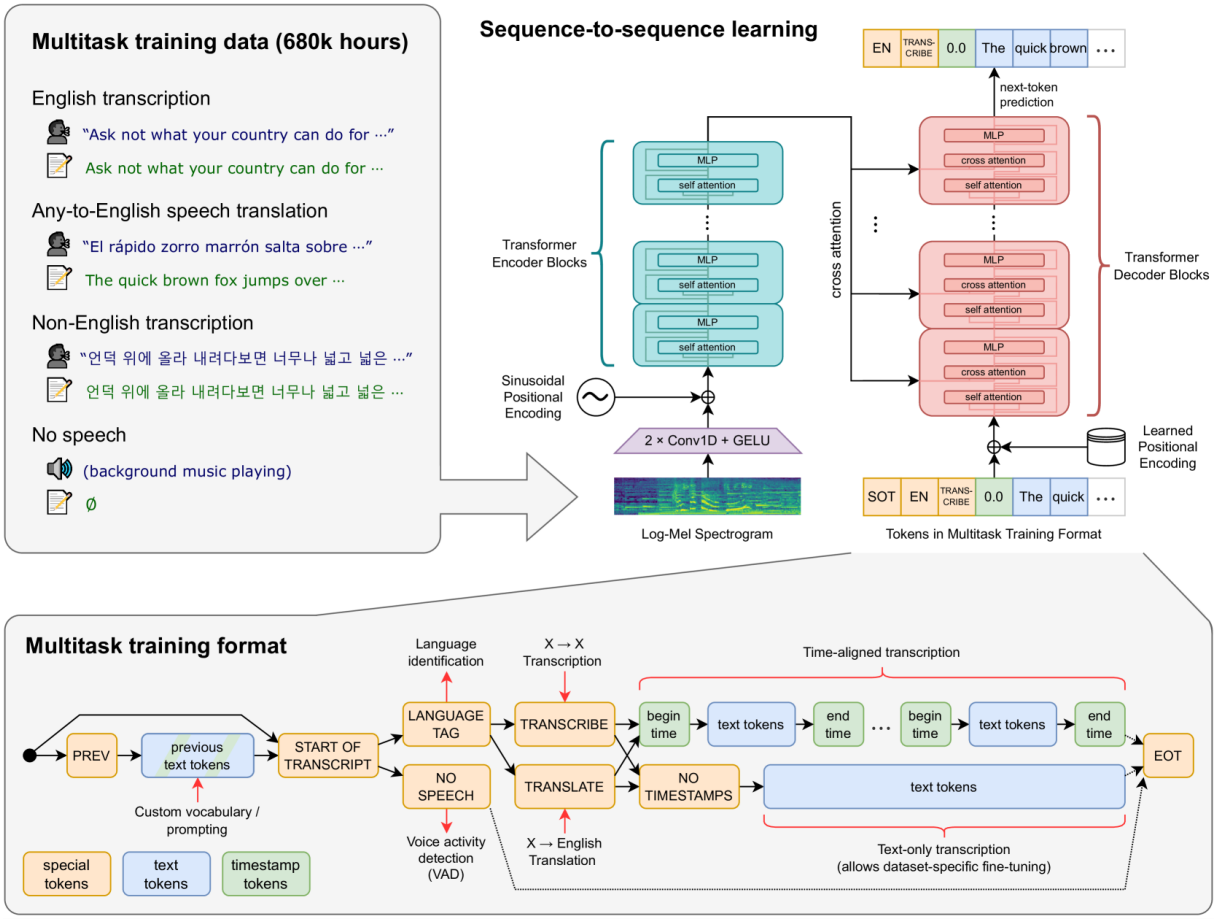
Các tiến bộ gần đây trong nhận dạng giọng nói đều nhờ vào các kỹ thuật tiền huấn luyện không giám sát như Wav2Vec 2.0, có thể học từ lượng lớn dữ liệu âm thanh không được gắn nhãn. Tuy nhiên, trong khi các phương pháp này có triển vọng, chúng thiếu một bộ

giải mã mạnh mẽ cho các nhiệm vụ thực tế như nhận dạng giọng nói, yêu cầu việc điều chỉnh tinh chỉnh có thể phức tạp và cụ thể cho tập dữ liệu. Để vượt qua điều này, một số nhà nghiên cứu đã tập trung vào việc tiền huấn luyện có giám sát trên các tập dữ liệu đa dạng, trong khi các nhà nghiên cứu khác đã khám phá các phương pháp giám sát yếu để tạo ra các tập dữ liệu lớn hơn. Bài báo giới thiệu Whisper, một phương pháp mở rộng nhận dạng giọng nói giám sát yếu đến 680.000 giờ âm thanh được gắn nhãn, đạt được kết quả chất lượng cao mà không cần phải điều chỉnh tinh chỉnh. Bài báo cũng mở rộng phương pháp này cho cài đặt đa ngôn ngữ và đa nhiệm, đưa mã và mô hình của Bài báo vào sử dụng cho nghiên cứu tiếp theo.

2.1.2. Phương pháp chính

Trọng tâm nghiên cứu của bài báo là khám phá khả năng của tiền huấn luyện có giám sát quy mô lớn trong nhận dạng giọng nói. Do đó, bài báo sử dụng kiến trúc có sẵn để tránh gây nhiễu kết quả với các cải tiến mô hình. Bài báo chọn Transformer encoder-decoder vì tính ổn định và khả năng mở rộng của nó. Âm thanh được tái mẫu thành 16,000 Hz và biểu diễn dưới dạng Mel spectrogram log-magnitude với 80 kênh trên các cửa sổ 25 mili giây và khoảng cách 10 mili giây. Để chuẩn hóa, đầu vào được điều chỉnh trong khoảng từ -1 đến 1 với trung bình xấp xỉ bằng 0 trên tập dữ liệu tiền huấn luyện. Bộ mã hóa xử lý biểu diễn đầu vào này với hai lớp tích chập và hàm kích hoạt GELU. Nhúng vị trí được thêm vào trước khi áp dụng các khối mã hóa. Transformer sử dụng khối dư tiền kích hoạt, và một lớp chuẩn hóa cuối cùng cho đầu ra của bộ mã hóa. Bộ giải mã sử dụng nhúng vị trí học được và các biểu diễn token đầu vào-đầu ra liên kết. Bộ mã hóa và giải mã có cùng chiều rộng và số lượng khối transformer.

Bài báo sử dụng bộ token hóa văn bản BPE cấp byte của GPT-2 cho các mô hình tiếng Anh và điều chỉnh từ vựng cho các mô hình đa ngôn ngữ để tránh phân mảnh quá mức ở các ngôn ngữ khác.



Một mô hình Transformer sequence-to-sequence được huấn luyện trên nhiều nhiệm vụ xử lý giọng nói khác nhau, bao gồm nhận dạng giọng nói đa ngôn ngữ, dịch giọng nói, nhận dạng ngôn ngữ nói, và phát hiện hoạt động giọng nói. Tất cả các nhiệm vụ này đều được biểu diễn chung dưới dạng một chuỗi các token mà bộ giải mã cần dự đoán, cho phép một mô hình duy nhất thay thế nhiều giai đoạn khác nhau của một quy trình xử lý giọng nói truyền thống. Định dạng huấn luyện đa nhiệm sử dụng một tập hợp các token đặc biệt đóng vai trò làm chỉ định nhiệm vụ hoặc mục tiêu phân loại.

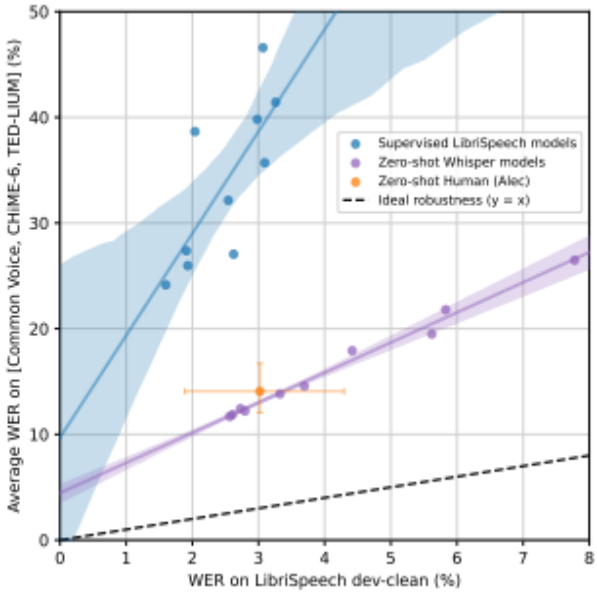
2.1.3. Kết quả thực nghiệm

2.1.3.1. English Speech Recognition

Năm 2015, Deep Speech 2 báo cáo hệ thống nhận dạng giọng nói đạt hiệu suất tương đương con người khi phiên âm LibriSpeech test-clean, nhưng bảy năm sau, WER trên LibriSpeech test-clean đã giảm từ 5.3% xuống còn 1.4% (Zhang et al., 2021), thấp hơn nhiều so với tỷ lệ lỗi của con người là 5.8%. Tuy nhiên, các mô hình nhận dạng giọng nói huấn luyện trên LibriSpeech vẫn còn lỗi cao hơn nhiều so với con người khi sử dụng trong các bối cảnh khác.

Sự khác biệt này có thể do sự khác nhau trong khả năng mà con người và máy móc được đo lường. Con người thường thực hiện nhiệm vụ với ít hoặc không có giám sát, đo lường khả năng tổng quát hóa ngoài phân phối, trong khi các mô hình máy học được đánh giá sau khi huấn luyện với lượng lớn dữ liệu giám sát từ phân phối đánh giá, đo lường khả năng tổng quát hóa trong phân phối.

Các mô hình Whisper, được huấn luyện trên một phân phối âm thanh đa dạng và đánh giá trong bối cảnh zero-shot, có thể phù hợp với hành vi của con người hơn. Bài báo đo lường tính ổn định tổng thể và tính ổn định hiệu quả, tức là sự khác biệt trong hiệu suất giữa tập dữ liệu tham chiếu và các tập dữ liệu ngoài phân phối.



Dataset	wav2vec 2.0 Large (no LM)	Whisper Large V2	RER (%)
LibriSpeech Clean	**2.7**	**2.7**	0.0
Artie	24.5	**6.2**	74.7
Common Voice	29.9	**9.0**	69.9
Fleurs En	14.6	**4.4**	69.9
Tedlium	10.5	**4.0**	61.9
CHiME6	65.8	**25.5**	61.2
VoxPopuli En	17.9	**7.3**	59.2
CORAAL	35.6	**16.2**	54.5
AMI IHM	37.0	**16.9**	54.3
Switchboard	28.3	**13.8**	51.2
CallHome	34.8	**17.6**	49.4
WSJ	7.7	**3.9**	49.4
AMI SDM1	67.6	**36.4**	46.2
LibriSpeech Other	6.2	**5.2**	16.1
Average	29.3	**12.8**	55.2

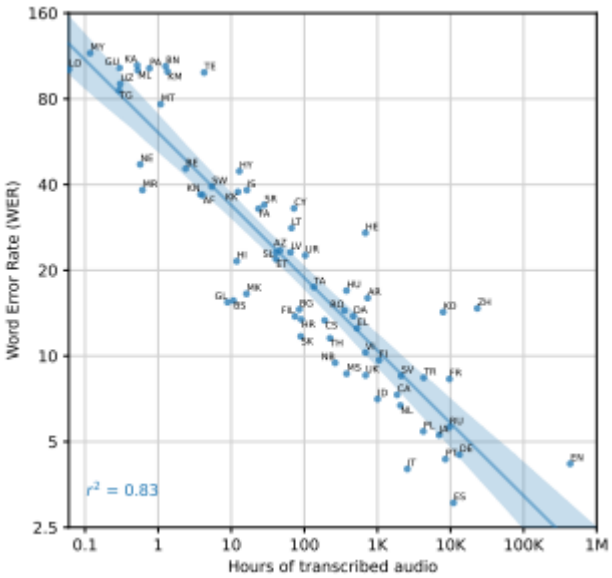
Kết quả chính cho thấy mô hình Whisper zero-shot tốt nhất có WER 2.5 trên LibriSpeech test-clean, tương đương với các mô hình huấn luyện có giám sát hiện đại, nhưng vượt trội hơn nhiều trên các bộ dữ liệu khác. Ngay cả mô hình Whisper zero-shot nhỏ nhất cũng cạnh tranh được với các mô hình có giám sát tốt nhất khi đánh giá trên các bộ dữ liệu khác, và gần như tương đương với con người về độ chính xác và tính ổn định. Phát hiện này gợi ý cần nhấn mạnh các đánh giá zero-shot và ngoài phân phối của các mô hình, đặc biệt khi so sánh với con người, để tránh đánh giá quá cao khả năng của các hệ thống máy học.

2.1.3.2. Multi-lingual Speech Recognition

Model	MLS	VoxPopuli
VP-10K + FT	-	15.3
XLS-R (1B)	10.9	10.6
mSLAM-CTC (2B)	9.7	9.1
Maestro	-	**8.1**
Zero-Shot Whisper	**7.3**	13.6

Để so sánh với các nghiên cứu trước về nhận dạng giọng nói đa ngôn ngữ, bài báo đưa ra kết quả trên hai bộ đánh giá ít dữ liệu: Multilingual LibriSpeech (MLS) và VoxPopuli.

Whisper hoạt động tốt trên Multilingual LibriSpeech, vượt trội hơn XLS-R, mSLAM và Maestro trong bối cảnh zero-shot. Tuy nhiên, bài báo sử dụng một chuẩn hóa văn bản đơn giản, do đó không thể so sánh trực tiếp hoặc khẳng định hiệu suất SOTA. Trên VoxPopuli, Whisper lại kém hiệu quả hơn so với các nghiên cứu trước và chỉ vượt qua baseline VP-10K+FT. Sự kém hiệu quả này có thể do các mô hình khác sử dụng VoxPopuli làm nguồn chính cho dữ liệu huấn luyện không giám sát và VoxPopuli có nhiều dữ liệu giám sát hơn.



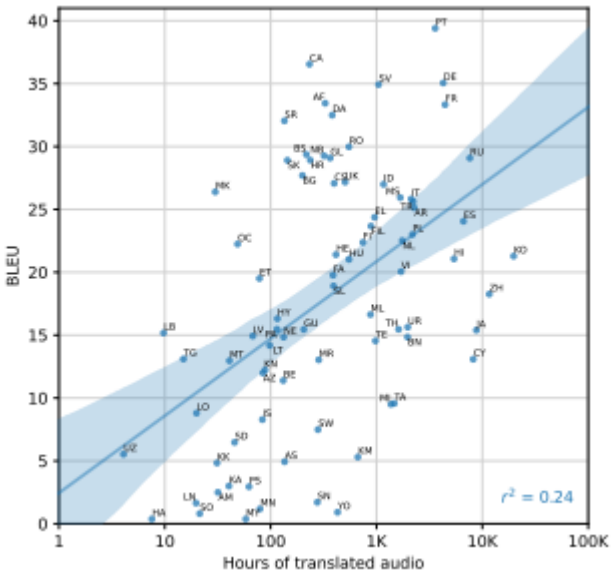
Hai bộ đánh giá này chỉ bao gồm 15 ngôn ngữ, chủ yếu thuộc ngữ hệ Ấn-Âu và nhiều ngôn ngữ có nguồn tài nguyên cao, nên không đủ để nghiên cứu khả năng đa ngôn ngữ của Whisper, vốn được huấn luyện trên 75 ngôn ngữ. Do đó, bài báo cũng báo cáo hiệu suất trên bộ dữ liệu Fleurs để nghiên cứu quan hệ giữa lượng dữ liệu huấn luyện và hiệu suất zero-shot. Kết quả cho thấy mối tương quan mạnh mẽ giữa tỷ lệ lỗi từ và lượng dữ

liệu huấn luyện, với tỷ lệ lỗi giảm một nửa khi lượng dữ liệu tăng gấp 16 lần. Những ngôn ngữ có hiệu suất thấp hơn mong đợi thường có chữ viết độc đáo và khác biệt về ngữ hệ như Hebrew, Telugu, Chinese và Korean. Điều này có thể do khoảng cách ngôn ngữ, tokenizer byte-level BPE không phù hợp hoặc chất lượng dữ liệu khác nhau.

2.1.3.3. Translation

X → English	High	Mid	Low	All
XMEF-X	34.2	20.2	5.9	14.7
XLS-R (2B)	36.1	27.7	15.1	22.1
mSLAM-CTC (2B)	37.8	29.6	18.5	24.8
Maestro	**38.2**	31.3	18.4	25.2
Zero-Shot Whisper	36.2	**32.6**	**25.2**	**29.1**

Bài báo nghiên cứu khả năng dịch của mô hình Whisper bằng cách đo hiệu suất trên tập X→en của CoVoST2. Kết quả cho thấy Whisper đạt BLEU 29.1 trong bối cảnh zero-shot, lập kỷ lục mới mà không sử dụng dữ liệu huấn luyện của CoVoST2. Thành công này có được nhờ 68,000 giờ dữ liệu dịch X→en trong tập huấn luyện, lớn hơn nhiều so với 861 giờ của CoVoST2. Whisper đặc biệt hiệu quả với chúng tôi ngôn ngữ ít tài nguyên, vượt mSLAM 6.7 BLEU, nhưng không vượt trội Maestro và mSLAM ở chúng tôi ngôn ngữ nhiều tài nguyên.



Để phân tích thêm, bài báo sử dụng bộ dữ liệu Fleurs như một tập dịch thuật. Kết quả cho thấy có mối tương quan giữa lượng dữ liệu huấn luyện dịch thuật và điểm BLEU zero-shot, nhưng hệ số tương quan thấp hơn so với nhận dạng giọng nói, chỉ đạt 0.24.

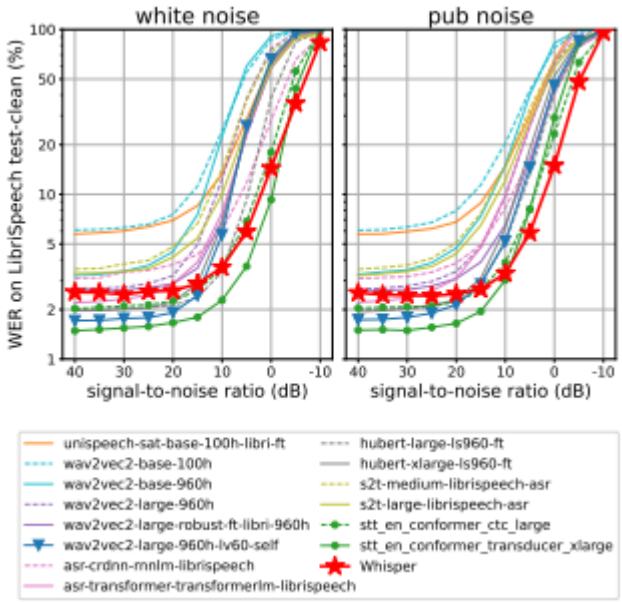
Điều này có thể do dữ liệu huấn luyện nhiều lỗi. Ví dụ, tiếng Welsh chỉ đạt 13 BLEU dù có 9,000 giờ dữ liệu dịch, nhưng phần lớn là âm thanh tiếng Anh bị phân loại nhầm.

2.1.3.4. Language Identification

Language ID	Fleurs
w2v-bert-51 (0.6B)	71.4
mSLAM-CTC (2B)	**77.7**
Zero-shot Whisper	64.5

Để đánh giá khả năng nhận dạng ngôn ngữ, bài báo sử dụng bộ dữ liệu Fleurs. Hiệu suất zero-shot của Whisper kém hơn 13.6% so với các mô hình giám sát trước đây. Tuy nhiên, Whisper gặp bất lợi vì không có dữ liệu huấn luyện cho 20 trong số 102 ngôn ngữ của Fleurs, giới hạn độ chính xác tối đa ở mức 80.4%. Trên 82 ngôn ngữ trùng lặp, mô hình Whisper tốt nhất đạt độ chính xác 80.3%.

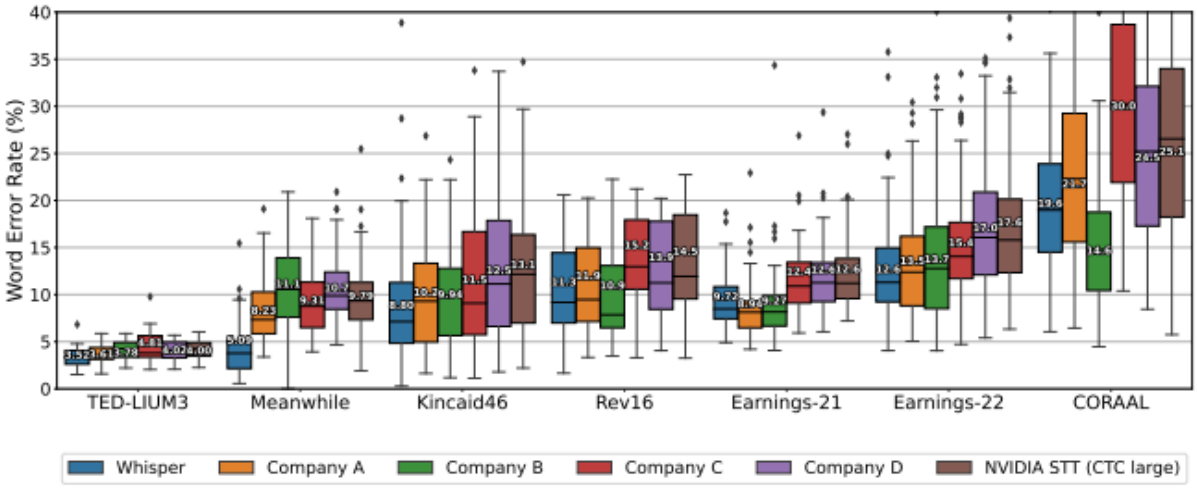
2.1.3.5. Robustness to Additive Noise



Bài báo kiểm tra độ bền của mô hình Whisper và 14 mô hình huấn luyện trên LibriSpeech bằng cách đo WER khi thêm white noise hoặc pub noise từ Audio Degradation Toolbox. Trong số 14 mô hình, 12 mô hình được tiền huấn luyện và/hoặc tinh chỉnh trên LibriSpeech, và 2 mô hình NVIDIA STT huấn luyện trên tập dữ liệu hỗn hợp bao gồm LibriSpeech. Kết quả cho thấy nhiều mô hình vượt trội hơn Whisper trong điều kiện tiếng ồn thấp (40 dB SNR), nhưng tất cả các mô hình nhanh chóng giảm hiệu

suất khi tiếng ồn tăng, và kém hơn Whisper dưới pub noise với SNR dưới 10 dB. Điều này chứng tỏ khả năng chống ồn của Whisper, đặc biệt trong môi trường tự nhiên như pub noise.

2.1.3.6. **Long-form Transcription**



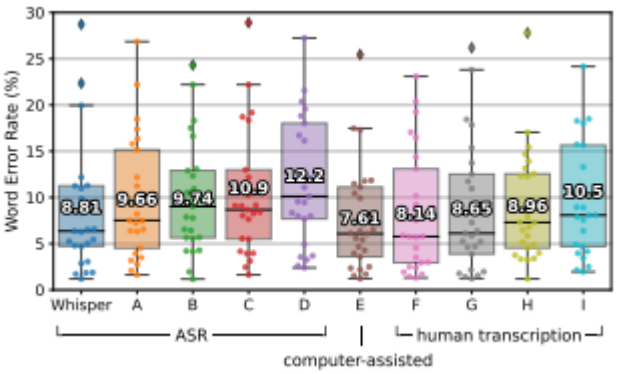
Mô hình Whisper được huấn luyện trên các đoạn âm thanh 30 giây và không thể xử lý các đầu vào dài hơn. Điều này không gây vấn đề với các bộ dữ liệu học thuật chứa các câu ngắn nhưng lại thách thức khi áp dụng thực tế yêu cầu phiên âm âm thanh kéo dài. Bài báo phát triển chiến lược phiên âm âm thanh dài bằng cách liên tiếp phiên âm các đoạn âm thanh 30 giây và dịch chuyển cửa sổ dựa trên dấu thời gian dự đoán của mô hình. Để phiên âm âm thanh dài một cách đáng tin cậy, bài báo sử dụng tìm kiếm chùm tia và lập lịch nhiệt độ dựa trên độ lặp lại và xác suất dự đoán của mô hình.

Bài báo đánh giá hiệu suất phiên âm dài hạn trên bảy bộ dữ liệu bao gồm các đoạn ghi âm có độ dài và điều kiện ghi âm khác nhau: phiên bản dài của TED-LIUM3, các đoạn khó từ The Late Show with Stephen Colbert, các video/podcast được sử dụng làm tiêu chuẩn ASR trên các blog trực tuyến (Rev16 và Kincaid46), ghi âm cuộc gọi báo cáo thu nhập và các cuộc phỏng vấn đầy đủ từ Corpus of Regional African American Language (CORAAL).

Kết quả so sánh với các mô hình mã nguồn mở và 4 dịch vụ ASR thương mại, cho thấy Whisper vượt trội hơn trên hầu hết các bộ dữ liệu, đặc biệt là trên tập Meanwhile với

nhiều từ ít phổ biến. Bài báo cũng lưu ý rằng một số hệ thống ASR thương mại có thể đã được huấn luyện trên một số bộ dữ liệu công khai này, do đó kết quả có thể không phản ánh chính xác độ bền tương đối của các hệ thống.

2.1.3.7. Comparison with Human Performance



Do giọng nói mơ hồ hoặc không rõ ràng cũng như lỗi gán nhãn, mỗi bộ dữ liệu có mức độ sai sót không thể giảm thiểu khác nhau, khiến việc đánh giá tiềm năng cải thiện của mỗi bộ dữ liệu chỉ dựa trên chỉ số WER từ các hệ thống ASR trở nên khó khăn. Để đo lường độ gần gũi của hiệu suất Whisper với hiệu suất của con người, bài báo chọn 25 bản ghi âm từ bộ dữ liệu Kincaid46 và sử dụng 5 dịch vụ để thu thập bản chép từ các chuyên gia, trong đó một dịch vụ cung cấp chép văn bản hỗ trợ máy tính và bốn dịch vụ còn lại hoàn toàn do con người thực hiện. Các bản ghi âm bao gồm các điều kiện ghi âm khác nhau như phát sóng có kịch bản và không kịch bản, cuộc gọi điện thoại và VoIP, và các cuộc họp. Kết quả cho thấy phân bố WER cho từng ví dụ và WER tổng hợp trên 25 bản ghi, trong đó dịch vụ hỗ trợ máy tính có WER tổng hợp thấp nhất, thấp hơn Whisper 1,15%, và hiệu suất hoàn toàn do con người thực hiện chỉ nhỉnh hơn Whisper một chút. Kết quả này cho thấy hiệu suất ASR tiếng Anh của Whisper không hoàn hảo nhưng rất gần với độ chính xác của con người.

2.1.4. Ưu điểm

Whisper có nhiều ưu điểm nổi bật. Khi mở rộng mô hình Whisper, các phiên bản lớn hơn đã tiến bộ đều đặn và đáng tin cậy trong việc giảm các lỗi liên quan đến nhận thức, như nhầm lẫn giữa các từ có âm thanh giống nhau. Độ bền vững của Whisper một phần nhờ vào bộ giải mã mạnh mẽ, hoạt động như một mô hình ngôn ngữ điều kiện âm thanh.

2.1.5. Hạn chế

Whisper vẫn tồn tại một số hạn chế, đặc biệt trong việc chép văn bản dài. Nhiều lỗi khó khắc phục và không giống lỗi của con người, như lặp lại liên tục, không chép được các từ đầu hoặc cuối của đoạn âm thanh, hoặc tạo ra bản chép hoàn toàn không liên quan đến âm thanh thực tế. Hiệu suất nhận dạng giọng nói của Whisper trên nhiều ngôn ngữ vẫn còn kém, và phân tích cho thấy hiệu suất này phụ thuộc nhiều vào lượng dữ liệu huấn luyện cho mỗi ngôn ngữ. Vì bộ dữ liệu huấn luyện hiện tại chủ yếu là tiếng Anh do thiên lệch trong quá trình thu thập dữ liệu từ các phần tiếng Anh trên internet, hầu hết các ngôn ngữ khác có ít hơn 1000 giờ dữ liệu huấn luyện. Hiện vẫn chưa rõ lợi ích của Whisper đến từ việc huấn luyện bộ mã hóa, bộ giải mã hay cả hai. Whisper khác biệt rõ rệt so với các hệ thống nhận dạng giọng nói hiện đại do thiếu các phương pháp huấn luyện không giám sát hoặc tự học. Mặc dù bài báo không thấy cần thiết phải sử dụng các phương pháp này để đạt được hiệu suất tốt, có thể kết quả sẽ được cải thiện hơn nếu kết hợp chúng.

2.1.6. Kết luận

Whisper cho thấy việc mở rộng tiền huấn luyện yếu có giám sát đã bị đánh giá thấp trong nghiên cứu nhận diện giọng nói. Kết quả của bài báo, đạt được mà không cần các kỹ thuật tự giám sát hay tự huấn luyện, cho thấy rằng việc huấn luyện trên một tập dữ liệu giám sát lớn và đa dạng, cùng với việc tập trung vào khả năng chuyển đổi không huấn luyện trước, có thể cải thiện đáng kể độ bền vững của hệ thống nhận diện giọng nói.

2.2. Bài báo "PhoWhisper: Automatic Speech Recognition for Vietnamese"

2.2.1. Giới thiệu chung

Trong nghiên cứu này, bài báo trình bày một phân tích về việc áp dụng Whisper cho tiếng Việt. Bài báo tiếp tục cải thiện mô hình Whisper đa ngôn ngữ trên một tập dữ liệu lớn của tiếng Việt, tạo ra một mô hình mới gọi là PhoWhisper. Kết quả thử nghiệm của bài báo cho thấy PhoWhisper vượt trội so với các mô hình cơ sở trước đó trên nhiều bộ dữ liệu kiểm tra tiếng Việt khác nhau. Bài báo công khai PhoWhisper và hy vọng rằng

nó sẽ là một nền tảng mạnh mẽ cho nghiên cứu và ứng dụng ASR tiếng Việt trong tương lai.

2.2.2. Phương pháp chính

Bài báo đã phát triển năm phiên bản của PhoWhisper: PhoWhisper-tiny, PhoWhisper-base, PhoWhisper-small, PhoWhisper-medium và PhoWhisper-large, dựa trên các mô hình đa ngôn ngữ Whisper-tiny, Whisper-base, Whisper-small, Whisper-medium và Whisper-large-v2. Bài báo tinh chỉnh các mô hình trên một tập dữ liệu luyện tập ASR lớn gồm 844 giờ âm thanh, bao gồm dữ liệu từ Common Voice, VIVOS, VLSP 2020, và dữ liệu riêng của bài báo, với sự đa dạng về giọng địa phương từ 63 tỉnh thành ở Việt Nam. Bài báo cũng tăng cường tính đồng nhất của mô hình với tiếng ồn tự nhiên từ Piczak và thêm tiếng ồn vào một nửa tập dữ liệu huấn luyện. Sử dụng transformers, bài báo huấn luyện các mô hình với tốc độ học phù hợp và thực hiện tổng cộng 48,000 bước cập nhật.

2.2.3. Kết quả thực nghiệm

Table 2: Results on Vietnamese ASR benchmarks. "#paras" denotes the number of parameters.

Model	#paras	Word Error Rate			
		CMV–Vi	VIVOS	VLSP Task-1	VLSP Task-2
wav2vec2-base-vietnamese-250h	95M	102.04	10.83	21.02	50.35
wav2vec2-base-vi-vlsp2020	95M	103.71	9.90	16.82	44.91
wav2vec2-large-vi-vlsp2020	317M	101.41	8.61	15.18	36.75
$\text{PhoWhisper}_{\text{tiny}}$	39M	19.05	10.41	20.74	49.85
$\text{PhoWhisper}_{\text{base}}$	74M	16.19	8.46	19.70	43.01
$\text{PhoWhisper}_{\text{small}}$	244M	11.08	6.33	15.93	32.96
$\text{PhoWhisper}_{\text{medium}}$	769M	8.27	4.97	14.12	26.85
$\text{PhoWhisper}_{\text{large}}$	1.55B	**8.14**	**4.67**	**13.75**	**26.68**

Bài báo so sánh các mô hình của bài báo với các mô hình cơ sở dựa trên "wav2vec2" từ nghiên cứu trước của Nguyen (2021). Các mô hình cơ sở này được đạt được bằng cách đầu tiên tiền huấn luyện mô hình Wav2Vec2.0 "base" và "large" trên 13K giờ âm thanh không gắn nhãn từ YouTube tiếng Việt, sau đó tinh chỉnh chúng bằng cách sử dụng hơn 240 giờ dữ liệu luyện tập có nhãn từ thách thức VLSP 2020 ASR.

Bảng 2 trình bày các kết quả Word Error Rate (WER) thu được cho PhoWhisper và các mô hình cơ sở. PhoWhispersmall, PhoWhispermedium và PhoWhisperlarge của bài báo vượt trội hơn so với tất cả các mô hình cơ sở dựa trên "wav2vec2". Trong khi đó, PhoWhispertiny và PhoWhisperbase còn lại cạnh tranh với "wav2vec2-base-vi-

vlsp2020" và hiệu suất tốt hơn so với "wav2vec2 base-vietnamese-250h". Ở đây, PhoWhisperlarge thiết lập một điểm mới WER tốt nhất trên mỗi bộ dữ liệu thử nghiệm.

2.2.4. Ưu điểm

PhoWhisper được đánh giá cao về tính ổn định, đạt được thông qua việc điều chỉnh mô hình Whisper trên tập dữ liệu lên tới 844 giờ bao gồm nhiều giọng địa phương của tiếng Việt. Nghiên cứu thực nghiệm của bài báo chứng minh hiệu suất tiên tiến của PhoWhisper trên các tập dữ liệu thử nghiệm tiêu biểu về ASR tiếng Việt. Đặc biệt, PhoWhisper cũng cho thấy khả năng xử lý hiệu quả ngôn ngữ và giọng địa phương đa dạng trong tiếng Việt, từ miền Bắc đến miền Nam, từ các giọng địa phương đến giọng nói chính thống, tạo ra một sự linh hoạt đáng kể trong ứng dụng thực tế.

2.2.5. Hạn chế

PhoWhisper chỉ có thể thực hiện trên các bộ dữ liệu tiếng Việt, điều này là một hạn chế đáng lưu ý của mô hình. Sự hạn chế này có nghĩa là PhoWhisper không thể tự tin trong việc xử lý các nhiệm vụ hoặc dữ liệu ở các ngôn ngữ khác, giới hạn khả năng ứng dụng của nó trong các ngữ cảnh đa ngôn ngữ hoặc đa văn hóa. Điều này đặt ra thách thức trong việc mở rộng tính ứng dụng của PhoWhisper ra ngoài phạm vi tiếng Việt.

2.2.6. Kết luận

Trong bài báo này, bài báo đã trình bày một nghiên cứu kinh nghiệm về các mô hình dựa trên Whisper, đặc biệt là PhoWhisper, cho ASR tiếng Việt. Các kết quả thực nghiệm của bài báo cho thấy hiệu suất tiên tiến của PhoWhisper. Bài báo hy vọng rằng nghiên cứu của bài báo và việc công bố PhoWhisper sẽ mở ra cơ hội cho những tiến bộ và sự hợp tác tiếp theo trong lĩnh vực đang phát triển này.

Chương 3: FINE-TUNING ASR MODELS: WHISPER AND PHOWHISPER ON VIETNAMESE YOUTUBE DATASET

Các hệ thống Nhận diện Giọng nói Tự động (ASR) đã có những bước tiến vượt bậc trong những năm gần đây. Tuy nhiên, để xây dựng một hệ thống ASR chính xác, cần phải có quá trình huấn luyện rộng rãi trên các bộ dữ liệu lớn, đặc biệt đối với các ngôn

ngữ có nguồn tài nguyên hạn chế như tiếng Việt. Trong nghiên cứu này, chúng tôi giới thiệu hai mô hình tinh chỉnh là Whisper[1] và PhoWhisper[2], áp dụng trên bộ dữ liệu tiếng Việt. Cả hai mô hình đều đạt được kết quả tốt, đánh dấu một bước tiến quan trọng trong việc đạt được khả năng chuyển đổi ngôn ngữ chính xác và chất lượng cao.

1. Giới thiệu chung

Các hệ thống Nhận diện Giọng nói Tự động (Automatic Speech Recognition - ASR) đã đạt được những tiến bộ vượt bậc trong những năm gần đây. Sự phát triển của các mô hình học máy và xử lý ngôn ngữ tự nhiên đã biến ASR thành một công nghệ không thể thiếu trong nhiều ứng dụng đời sống hàng ngày và các lĩnh vực chuyên môn. Tuy nhiên, để xây dựng một hệ thống ASR chính xác và hiệu quả, cần có quá trình huấn luyện mở rộng trên các bộ dữ liệu lớn, đặc biệt đối với các ngôn ngữ có nguồn tài nguyên hạn chế như tiếng Việt.

Tiếng Việt là một ngôn ngữ có nhiều đặc thù riêng, bao gồm hệ thống âm vị phong phú và sự phức tạp của thanh điệu, đặt ra thách thức lớn cho việc xây dựng các hệ thống ASR chính xác. Ngoài ra, nguồn tài nguyên ngôn ngữ cho tiếng Việt thường hạn chế hơn so với các ngôn ngữ phổ biến khác như tiếng Anh hay tiếng Trung, khiến việc thu thập và xử lý dữ liệu trở nên khó khăn hơn.

Trong nghiên cứu này, chúng tôi tiến hành tinh chỉnh hai mô hình là Whisper[1] và PhoWhisper[2] nhằm cải thiện khả năng nhận diện giọng nói tiếng Việt. Whisper, được phát triển bởi OpenAI, là một mô hình ASR đa ngôn ngữ, được huấn luyện trên một bộ dữ liệu lớn và đa dạng, bao gồm nhiều ngôn ngữ và giọng nói khác nhau. Whisper đã chứng minh được khả năng nhận diện giọng nói ấn tượng trên nhiều ngôn ngữ, bao gồm cả tiếng Việt. PhoWhisper là phiên bản tinh chỉnh của Whisper, được tối ưu hóa đặc biệt cho ngôn ngữ tiếng Việt. Bằng cách sử dụng bộ dữ liệu tiếng Việt phong phú, PhoWhisper có thể nhận diện và chuyển đổi giọng nói tiếng Việt một cách chính xác và hiệu quả hơn.

Kết quả của quá trình tinh chỉnh cho thấy cả hai mô hình, Whisper[1] và PhoWhisper[2], đều đạt được kết quả xuất sắc trong việc nhận diện giọng nói tiếng Việt. PhoWhisper đặc biệt nổi bật với hiệu suất cao hơn nhờ vào sự tối ưu hóa đặc biệt cho ngôn ngữ và giọng nói tiếng Việt.

Việc phát triển và tinh chỉnh các mô hình ASR như Whisper[1] và PhoWhisper[2] trên bộ dữ liệu tiếng Việt không chỉ giúp cải thiện khả năng nhận diện giọng nói tiếng Việt mà còn đóng góp vào sự phát triển chung của công nghệ ASR. Với những bước tiến này, chúng tôi hy vọng sẽ tiếp tục nâng cao chất lượng và độ chính xác của các hệ thống ASR, hỗ trợ nhiều hơn nữa cho các ứng dụng thực tế và người dùng.

2. Các công trình liên quan

Automatic Speech Recognition bắt đầu với các hệ thống đơn giản đáp ứng một số âm thanh hạn chế và đã phát triển thành các hệ thống phức tạp có thể đáp ứng linh hoạt ngôn ngữ tự nhiên. Automatic Speech Recognition (ASR) đang ngày càng phát triển và trở thành một phần không thể thiếu trong nhiều lĩnh vực khác nhau, chẳng hạn như phương pháp dựa trên nhận dạng giọng nói tự động (ASR) tiên tiến cho liệu pháp nói của bệnh nhân mắc chứng mất ngôn ngữ *Norezmi Jamal, Shahnoor Shanta, Farhanahani Mahmud, MNAH Sha'abani SEPTEMBER 14 2017*[3] hay như trong hệ thống đánh giá đọc tự động mô phỏng một chuyên gia ngôn ngữ nói trong tình huống học tiếng Anh Preeti RAO*, Prakhar SWARUP tại Proceedings of the 24th International Conference on Computers in Education. India: Asia-Pacific Society for Computers in Education, để đạt được dự đoán tự động mạnh mẽ về sự lưu loát trong việc đọc và độ chính xác trong giải mã từ, mà còn được sử dụng nhận dạng giọng nói tự động trong bệnh thoái hóa thần kinh *Benjamin G. Schultz Published: 04 May 2021*[4]

Roddy Cowie, Available online 17 December 2002.[5] đã sử dụng nhiều phương pháp trong nghiên cứu của mình để khám phá mối quan hệ giữa tiếng nói và cảm xúc. Ông cũng đã nghiên cứu về cách mà cảm xúc được thể hiện thông qua âm điệu, giọng điệu, và các thuật ngữ ngôn ngữ trong tiếng nói. Ông đã đóng góp đáng kể vào việc hiểu và phát triển các phương pháp mô tả và phân tích cảm xúc trong tiếng nói.

Emmanuel Vincent, Shinji Watanabe 25 April 2016[6] sử dụng tập dữ liệu CHIME-3 với các môi trường ồn ào và microphone, bài viết phân tích và tiến hành thí nghiệm mới để đánh giá tác động của các yếu tố khác nhau như môi trường tiếng ồn, số lượng và vị trí của microphones, hoặc dữ liệu mô phỏng so với dữ liệu. Kết quả cho thấy, trừ việc sử dụng MVDR beamforming, hầu hết các thuật toán hoạt động ổn định trên cả dữ liệu thực và mô phỏng, và có thể hưởng lợi từ việc huấn luyện trên dữ liệu mô phỏng. Thêm vào đó, việc huấn luyện trên các môi trường và microphones khác nhau ít ảnh hưởng

đến hiệu suất của ASR, và chỉ có số lượng microphones có tác động đáng kể. Dựa trên kết quả này, giới thiệu Thách thức Tách biệt và Nhận dạng Tiếng nói CHIME-4, tái khám phá tập dữ liệu CHIME-3.

Shipra J. Arora, December 2012 đã giới thiệu bài báo nhằm mục đích thực hiện một cuộc đánh giá tài liệu về Nhận Dạng Giọng Nói Tự Động (ASR), khám phá những tiến bộ được đạt được trong những năm gần đây để làm sáng tỏ sự tiến bộ đã đạt được trong lĩnh vực nghiên cứu này, mục tiêu chính của bài báo đánh giá này là để làm sáng tỏ những tiến bộ đã đạt được trong ASR qua các ngôn ngữ khác nhau và khám phá các quan điểm công nghệ về ASR ở các quốc gia khác nhau. Hơn nữa, nó cũng nhằm so sánh và phân biệt các kỹ thuật được sử dụng ở các giai đoạn khác nhau của việc nhận dạng giọng nói và xác định các chủ đề nghiên cứu trong lĩnh vực phức tạp này.

Tian Kexin, Huang Yongming, 2019 Chinese Automation Congress (CAC)[7] đã đề xuất một hệ thống nhận diện chỉ dẫn đỗ xe khẩn cấp dựa trên nhận diện giọng nói và nhận diện cảm xúc giọng nói. Hệ thống đầu tiên trích xuất vector đặc trưng của tín hiệu giọng nói, sau đó sử dụng máy vector hỗ trợ (SVM) để nhận diện cảm xúc của giọng nói. Khi cảm xúc bất thường, thuật toán so khớp thời gian động (DWT) được sử dụng để khớp mẫu chỉ dẫn đỗ xe. Kết quả kiểm tra cho thấy hệ thống có thể thực hiện việc nhận diện giọng nói của các chỉ dẫn đỗ xe rất tốt.

Đorđe T. Grozdić; Slobodan T. Jovičić, December 2017[8] phát triển một phương pháp hiệu quả cho nhận diện giọng nói thì thầm, nghiên cứu này đầu tiên phân tích các đặc tính âm học của giọng nói thì thầm, giải quyết các vấn đề của nhận diện giọng nói thì thầm trong các điều kiện không phù hợp, và sau đó đề xuất các đặc trưng cepstral mới và phương pháp tiền xử lý dựa trên bộ mã hóa tự động giảm nhiễu sâu (DDAE) nhằm cải thiện nhận diện giọng nói thì thầm. Kết quả thực nghiệm xác nhận rằng các đặc trưng cepstral dựa trên năng lượng Teager, đặc biệt là TECCs, là những mô tả giọng nói thì thầm tốt hơn và mạnh mẽ hơn so với các hệ số cepstral tần số Mel truyền thống (MFCC). Khung mới dựa trên DDAE và đặc trưng TECC, cải thiện đáng kể độ chính xác nhận diện giọng nói thì thầm và vượt trội hơn so với MFCC và GMM-HMM (mô hình Markov ẩn mật độ hỗn hợp Gaussian) truyền thống, dẫn đến cải thiện độ chính xác nhận diện giọng nói thì thầm tuyệt đối 31%. Tỷ lệ nhận diện từ đạt được trong kịch bản bình thường/thì thầm là 92.81%.

Reza Lotfidereshgi, Philippe Gournay, 2017 IEEE International Conference on Acoustics, Speech and Signal Processing[9] bài báo này trình bày một phương pháp hoạt động trực tiếp trên tín hiệu giọng nói, do đó tránh được bước trích xuất đặc trưng đầy vấn đề. Hơn nữa, phương pháp này kết hợp những ưu điểm của source-filter model cổ điển trong sản xuất giọng nói của con người với những ưu điểm của liquid state machine (LSM), mới được giới thiệu, một mạng nơ-ron lấy cảm hứng sinh học (SNN). Phương pháp này được chứng minh là cung cấp hiệu suất phân loại rất tốt trên Cơ sở dữ liệu Giọng nói Cảm xúc Berlin (Emo-DB). Đây dường như là một khung rất hứa hẹn để giải quyết hiệu quả nhiều vấn đề khác trong xử lý giọng nói.

3. Bộ dữ liệu

Bộ dữ liệu mà chúng tôi sử dụng mang tên "viet_youtube_asr_corpus_v2" được thu thập từ nền tảng mạng xã hội YouTube. Bộ dữ liệu này bao gồm hơn 100 giờ nội dung âm thanh, phản ánh sự đa dạng và phong phú của ngôn ngữ tiếng Việt qua nhiều thể loại khác nhau như giải trí, giáo dục, tin tức và nhiều lĩnh vực khác. Bộ dữ liệu đã được công khai trên HuggingFace, một nền tảng uy tín dành cho các nhà nghiên cứu và phát triển trí tuệ nhân tạo, giúp dễ dàng tiếp cận và sử dụng cho mục đích nghiên cứu và học tập.

Với tổng dung lượng lên tới 17,6GB, bộ dữ liệu này không chỉ cung cấp một lượng lớn thông tin âm thanh mà còn bao gồm các đoạn hội thoại được chuyển đổi thành văn bản một cách chính xác. Điều này tạo điều kiện thuận lợi cho việc nghiên cứu và phát triển các ứng dụng nhận diện giọng nói, dịch máy, và nhiều ứng dụng khác trong lĩnh vực xử lý ngôn ngữ tự nhiên.

Dataset	**Training size (hours)**	**Validation size (hours)**	**Test size (hours)**	**#syllables in training set (min – max | average)**
CMV–Vi 14	3.04	0.41	1.35	1 – 14 | 7.55
VIVOS	13.94	0.98	0.75	2 – 30 | 13.25
viet_youtube_asr_corpus_v2	110	None	22	1 – 9 | 3

Bảng 1: So sánh dung lượng giữa các bộ dữ liệu tiếng Việt

Bảng 1 cho thấy rõ ràng rằng dung lượng của tập huấn luyện và tập kiểm tra trong bộ dữ liệu mà chúng tôi sử dụng lớn hơn nhiều so với các bộ dữ liệu tiếng Việt khác, chẳng hạn như Vietnamese Common Voice và VIVOS. Sự khác biệt này rất quan trọng, vì nó cung cấp cho chúng tôi một lượng lớn dữ liệu âm thanh đa dạng để huấn luyện mô hình, giúp cải thiện độ chính xác và khả năng tổng quát hóa của các mô hình nhận diện giọng nói và xử lý ngôn ngữ tự nhiên. Thông tin dữ liệu chi tiết ở phụ lục **A**

Mặc dù dung lượng tổng thể của bộ dữ liệu lớn, mỗi mẫu dữ liệu trung bình chỉ kéo dài khoảng 3 giây. Điều này có nghĩa là các đoạn âm thanh ngắn, giúp cho quá trình huấn luyện mô hình trở nên dễ dàng hơn. Thời gian ngắn của mỗi mẫu dữ liệu giúp giảm tải tính toán, tối ưu hóa hiệu suất huấn luyện và làm cho việc xử lý và phân tích dữ liệu nhanh chóng hơn. Đồng thời, nó cũng giảm thiểu nguy cơ quá tải bộ nhớ và tăng cường khả năng song song hóa trong quá trình huấn luyện. Điều này đảm bảo rằng chúng tôi có thể tận dụng tối đa lợi thế của bộ dữ liệu lớn mà không gặp phải các vấn đề về hiệu suất hoặc tài nguyên, đồng thời vẫn đảm bảo chất lượng cao cho các mô hình phát triển.

4. Phương pháp chính

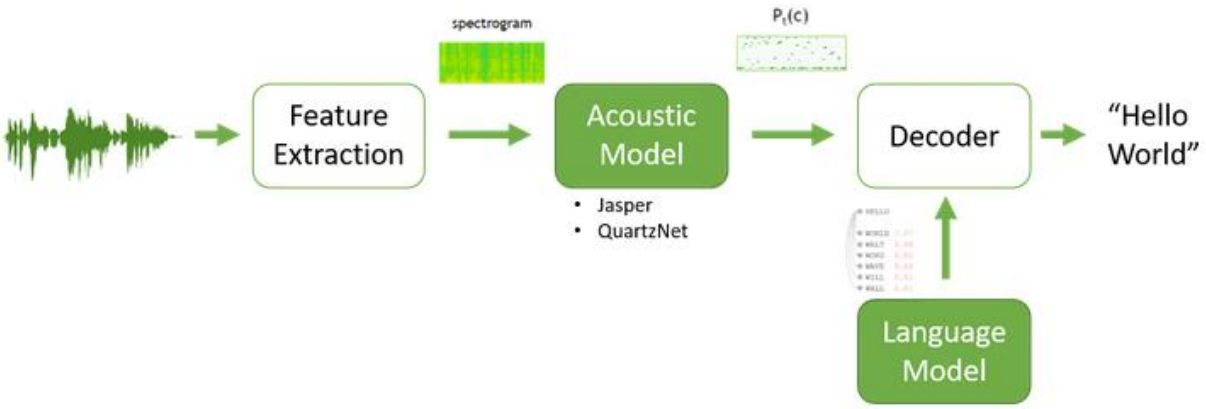
4.1. Tổng quan quy trình

Quá trình chúng tôi thực hiện bao gồm nhiều bước chi tiết và phức tạp để chuyển đổi các tệp âm thanh thành văn bản một cách chính xác. Đầu vào của hệ thống là các tệp âm thanh từ bộ dữ liệu đã được chuẩn bị kỹ lưỡng. Các tệp âm thanh này được đưa vào các mô hình ASR và trải qua nhiều giai đoạn xử lý.

Trước tiên, các tệp âm thanh được tiền xử lý để loại bỏ nhiễu và chuẩn hóa mức âm thanh, đảm bảo chất lượng tín hiệu tốt nhất cho quá trình nhận diện. Sau đó, các mô hình ASR, như Whisper[1] và PhoWhisper[2], tiếp nhận tín hiệu âm thanh đã được tiền xử lý và bắt đầu quá trình phân tích. Các mô hình này sử dụng các kỹ thuật học sâu tiên tiến để phân đoạn tín hiệu âm thanh và nhận diện các đặc trưng âm vị của ngôn ngữ.

Tiếp theo, các mô hình chuyển đổi các đặc trưng âm vị này thành các từ và cụm từ bằng cách áp dụng các thuật toán nhận dạng mẫu và xử lý ngôn ngữ tự nhiên. Quá trình này bao gồm việc giải mã các chuỗi tín hiệu âm thanh và so khớp chúng với các từ trong từ điển ngôn ngữ đã được mô hình học trước đó.

Kết quả của quá trình này là các dòng văn bản được tạo ra từ các tệp âm thanh ban đầu. Văn bản đầu ra này sau đó được kiểm tra và hiệu chỉnh để đảm bảo tính chính xác và độ tin cậy. Quá trình này không chỉ đòi hỏi sự chính xác và hiệu quả của các mô hình ASR mà còn yêu cầu sự phối hợp chặt chẽ giữa các giai đoạn tiền xử lý và hậu xử lý để đảm bảo chất lượng đầu ra tốt nhất.



Hình 1: Tổng quan quy trình thực hiện

4.2. Model

Như đã đề cập ở mục 4.1, chúng tôi sử dụng hai mô hình ASR tiên tiến: Whisper[1] và PhoWhisper[2].

Whisper[1], được phát triển bởi OpenAI, là một mô hình nhận diện giọng nói tự động tiên tiến, có khả năng chuyển đổi giọng nói thành văn bản với độ chính xác cao và hiệu quả vượt trội. Mô hình này đã được huấn luyện trên một bộ dữ liệu đa ngôn ngữ khổng lồ, cho phép nó hoạt động tốt trên nhiều ngôn ngữ khác nhau, bao gồm cả tiếng Việt. Bộ dữ liệu phong phú và đa dạng giúp Whisper nắm bắt được các đặc trưng ngôn ngữ và giọng nói khác nhau, tăng cường khả năng nhận diện và chuyển đổi chính xác.

PhoWhisper[2] là phiên bản tinh chỉnh của Whisper, được tối ưu hóa đặc biệt cho tiếng Việt. Mặc dù Whisper đã có hiệu suất tốt đối với nhiều ngôn ngữ, PhoWhisper tập trung vào việc khai thác các sắc thái và đặc thù của tiếng Việt để cải thiện hiệu suất. PhoWhisper sử dụng dữ liệu tiếng Việt phong phú và các phương pháp huấn luyện chuyên biệt để đảm bảo mô hình có thể nhận diện và chuyển đổi giọng nói tiếng Việt một cách chính xác nhất.

Cả Whisper[1] và PhoWhisper[2] đều có năm phiên bản: Tiny, Base, Small, Medium, và Large. Tiny: phiên bản nhỏ nhất, yêu cầu ít tài nguyên tính toán nhất nhưng hiệu suất

không cao. Base: phiên bản cơ bản, cân bằng giữa hiệu suất và yêu cầu tài nguyên. Small: phiên bản nhỏ, cải thiện hiệu suất so với Tiny và Base nhưng vẫn cần ít tài nguyên hơn Medium và Large. Medium: phiên bản trung bình, cung cấp hiệu suất tốt với yêu cầu tài nguyên trung bình. Large: phiên bản lớn nhất, mang lại hiệu suất cao nhất nhưng đòi hỏi nhiều tài nguyên tính toán nhất.

Trong nghiên cứu này, chúng tôi tập trung huấn luyện trên phiên bản Base của cả Whisper[1] và PhoWhisper[2]. Phiên bản Tiny không mang lại hiệu suất tốt do hạn chế về kích thước mô hình và khả năng học hỏi. Ngược lại, các phiên bản Medium và Large mặc dù có hiệu suất cao nhưng đòi hỏi nguồn tài nguyên tính toán rất lớn, vượt quá khả năng môi trường hiện tại của chúng tôi. Thông tin chi tiết ở phụ lục **B**

Phiên bản Base cung cấp một sự cân bằng tốt giữa chất lượng nhận diện giọng nói và chi phí tài nguyên. Nó đủ mạnh để đạt được kết quả chính xác mà không đòi hỏi quá nhiều tài nguyên tính toán. Bằng cách lựa chọn phiên bản Base, chúng tôi đảm bảo rằng mô hình có thể hoạt động hiệu quả trong môi trường tính toán hiện tại, đồng thời cung cấp khả năng nhận diện giọng nói tiếng Việt với độ chính xác cao.

4.3. Độ đo đánh giá

Trong nghiên cứu của chúng tôi, chúng tôi áp dụng ba chỉ số quan trọng để đánh giá hiệu suất của các mô hình nhận diện giọng nói tự động (ASR): Tỷ lệ Lỗi từ (WER), Tỷ lệ Lỗi ký tự (CER), và Điểm BLEU. Mỗi chỉ số này cung cấp cái nhìn sâu sắc từ các góc độ khác nhau về hiệu suất của mô hình, giúp chúng tôi đánh giá một cách toàn diện và chính xác.

WER[10], viết tắt của "Word Error Rate", là một chỉ số chuẩn trong lĩnh vực nhận diện giọng nói tự động. Chỉ số này đo lường tỷ lệ phần trăm các từ bị nhận diện sai so với tổng số từ trong bản phiên âm gốc. WER cung cấp một cái nhìn tổng quan về độ chính xác của mô hình ở mức từ vựng, thể hiện khả năng của hệ thống trong duy trì ngữ nghĩa chính xác của toàn bộ đoạn văn bản.

$$WER = \frac{Substitutions + Deletions + Insertions}{Number\ of\ words\ in\ the\ reference}$$

CER, viết tắt của "Character Error Rate", là một chỉ số tương tự WER nhưng được áp dụng ở mức độ ký tự. Chỉ số này đo lường tỷ lệ phần trăm các ký tự bị nhận dạng sai so

với tổng số ký tự trong bản phiên âm gốc. CER đặc biệt hữu ích khi đánh giá các ngôn ngữ có hệ thống ký tự phức tạp hoặc nhiều biến thể, như tiếng Việt. Nó cung cấp cái nhìn chi tiết hơn về độ chính xác của mô hình, giúp chúng tôi cải thiện các yếu tố nhỏ nhưng quan trọng trong quá trình nhận diện.

$$CER = \frac{Substitutions + Deletions + Insertions}{Number\ of\ characters\ in\ the\ reference}$$

BLEU[11], viết tắt của "Bilingual Evaluation Understudy", là một chỉ số đánh giá phổ biến trong lĩnh vực dịch máy và hiện cũng được áp dụng cho ASR. Điểm BLEU đo lường chất lượng của văn bản được tạo ra bằng cách so sánh nó với một hoặc nhiều văn bản tham chiếu. Chỉ số này tính toán độ tương đồng giữa các n-gram của văn bản dự đoán và văn bản tham chiếu, với các điều chỉnh để phạt việc lặp lại n-gram không tự nhiên. BLEU giúp đánh giá mức độ tương đồng về ngữ nghĩa giữa văn bản nhận diện và văn bản gốc, cung cấp cái nhìn toàn diện về chất lượng ngôn ngữ của đầu ra.

4.4. Chi tiết huấn luyện

Trong nghiên cứu này, chúng tôi đã thực hiện huấn luyện hai mô hình, Whisper (base) và PhoWhisper (base), bằng cách sử dụng GPU P100 để đảm bảo hiệu suất và tốc độ huấn luyện tối ưu. Chúng tôi tiến hành đóng băng một số lớp của mô hình và sử dụng kích thước batch là 64. Để giảm tải bộ nhớ trên mỗi GPU và xử lý các batch lớn hơn hiệu quả hơn, chúng tôi áp dụng kỹ thuật tích lũy gradient với số bước tích lũy là 2. Tốc độ học (learning rate) là một tham số quan trọng trong quá trình tối ưu hóa của mô hình, và chúng tôi thiết lập nó ở mức 5e-4 cho cả hai mô hình để đảm bảo sự cân bằng giữa tốc độ hội tụ và tính ổn định của mô hình. Tổng cộng, chúng tôi đã thực hiện khoảng 15,000 bước cập nhật, tương đương với 5 epochs, để tối ưu hóa việc sử dụng tài nguyên trong quá trình huấn luyện.

5. Kết quả thực nghiệm

Model	**#paras**	**WER**	**CER**	**BLEU**
Whisper(original)	74M	49.9	27.81	30.14
Whisper(fine-tuning)		6.88	3.07	86.14

PhoWhisper(original)		19.74	8.07	71.83
PhoWhisper(fine-tuning)		**4.76**	**2.09**	**90.38**

Bảng 2: Kết quả độ đo (%) của mô hình trên bộ dữ liệu "viet_youtube_asr_corpus_v2"

Khi quan sát bảng 2, chúng ta thấy rằng cả bốn mô hình đều thể hiện khả năng chuyển đổi âm thanh thành văn bản với hiệu suất đáng kể. Các chỉ số đánh giá bao gồm WER (Word Error Rate), CER (Character Error Rate), và BLEU (Bilingual Evaluation Understudy) cung cấp cái nhìn toàn diện về độ chính xác và chất lượng của văn bản được tạo ra từ âm thanh.

Mô hình PhoWhisper, sau khi được fine-tuning, đã nổi bật với kết quả xuất sắc trên tất cả các chỉ số đánh giá. Cụ thể, mô hình này đạt điểm WER (Word Error Rate) và CER (Character Error Rate) thấp nhất, cùng với điểm BLEU (Bilingual Evaluation Understudy) cao nhất trong số bốn mô hình, chứng tỏ đây là lựa chọn tốt nhất cho bộ dữ liệu này. Điểm WER thấp cho thấy mô hình này ít mắc lỗi trong quá trình nhận diện từ ngữ từ âm thanh, và điểm CER thấp chỉ ra rằng mô hình ít mắc lỗi khi nhận diện các ký tự riêng lẻ. Điểm BLEU cao cho thấy văn bản được tạo ra có chất lượng tốt, phù hợp với văn bản gốc về ngữ cảnh và cú pháp. Ngược lại, mô hình Whisper ban đầu cho kết quả chưa thực sự tốt với điểm WER gần 50% (49.9), điểm CER cao và điểm BLEU chỉ trên 30% (30.14). Điều này cho thấy rằng mô hình Whisper ban đầu, do chưa được tiếp xúc nhiều với dữ liệu tiếng Việt, đã mắc nhiều lỗi trong quá trình nhận diện và tạo văn bản, dẫn đến chất lượng văn bản được tạo ra không cao.

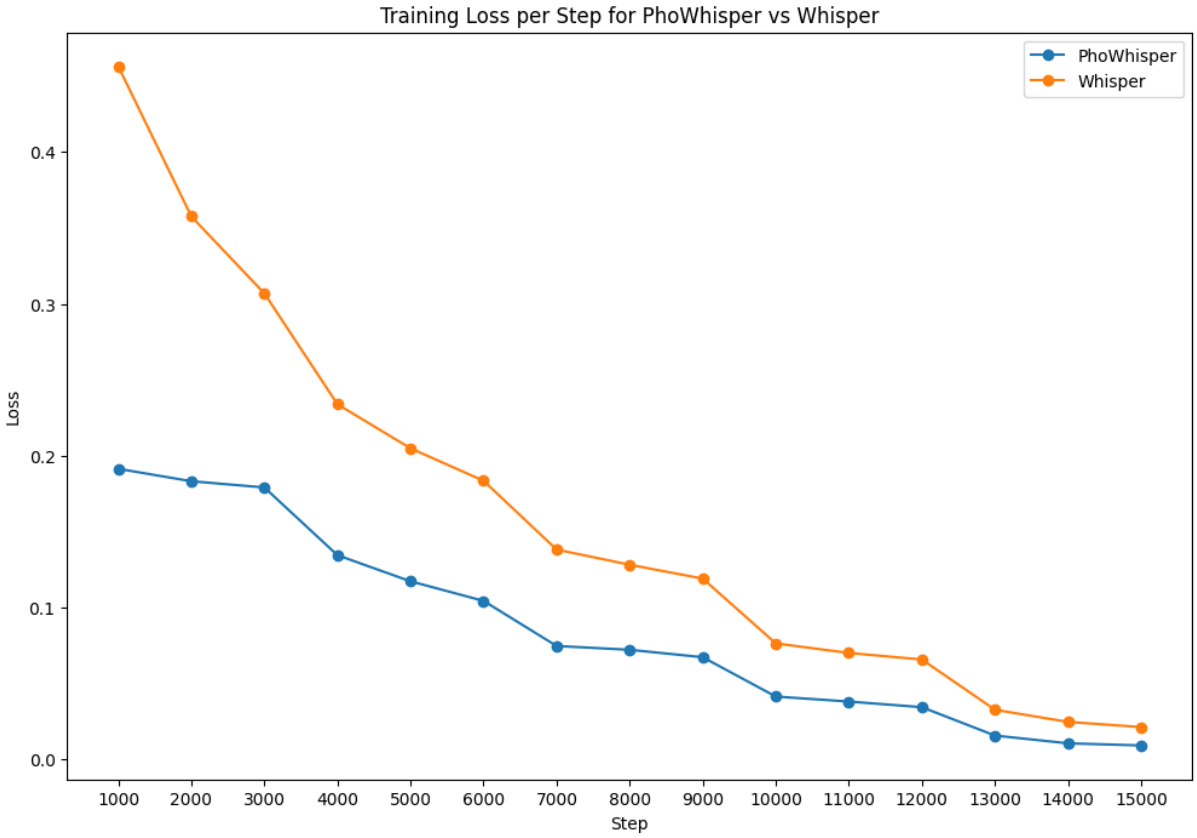
Sau quá trình fine-tuning, mô hình Whisper đã cải thiện hiệu suất đáng kể. Điểm WER giảm mạnh từ gần 50% (49.9) xuống còn gần 7% (6.88), và điểm BLEU tăng từ 30% (30.14) lên 86% (86.14). Đây là một sự tiến bộ ấn tượng, cho thấy việc fine-tuning đã giúp mô hình cải thiện đáng kể khả năng nhận diện và tạo văn bản. Điểm WER giảm mạnh nghĩa là mô hình đã trở nên chính xác hơn nhiều trong việc nhận diện từ ngữ từ âm thanh, và điểm BLEU tăng cao cho thấy chất lượng văn bản được tạo ra đã được nâng cao rõ rệt, phù hợp hơn với ngữ cảnh và cú pháp của văn bản gốc.

Tương tự, mô hình PhoWhisper cũng cho thấy sự cải thiện vượt bậc khi được huấn luyện tiếp trên lượng lớn dữ liệu mới. Ban đầu, phiên bản gốc của mô hình đã có hiệu suất khá tốt với điểm WER gần 20% (19.74) và điểm BLEU trên 70% (71.83), những kết quả này khá ổn định và chấp nhận được cho bộ dữ liệu. Tuy nhiên, sau quá trình fine-tuning, mô hình PhoWhisper đạt hiệu suất ấn tượng với điểm WER chỉ còn gần 5% (4.76) và điểm BLEU trên 90% (90.38). Điều này cho thấy việc tinh chỉnh đã giúp mô hình trở nên cực kỳ chính xác và hiệu quả trong việc nhận diện và tạo văn bản từ âm thanh. Điểm WER rất thấp và điểm BLEU rất cao của mô hình PhoWhisper sau fine-tuning chứng minh rằng đây là mô hình tốt nhất cho bộ dữ liệu này, vượt trội hơn cả phiên bản gốc và các mô hình khác.

Từ kết quả thu được, ta thấy rằng mô hình PhoWhisper cho thấy hiệu suất vượt trội hơn nhiều so với mô hình Whisper. Điều này có thể là do PhoWhisper được tối ưu hóa đặc biệt cho việc nhận diện giọng nói tiếng Việt, làm cho nó trở thành một công cụ mạnh mẽ và hiệu quả hơn trong bối cảnh này. Khả năng của PhoWhisper trong việc nhận diện chính xác và tạo ra văn bản chất lượng cao từ âm thanh tiếng Việt cho thấy sự tiềm năng to lớn của mô hình này trong các ứng dụng thực tế, nơi mà sự chính xác và chất lượng của văn bản là cực kỳ quan trọng. Điều này cũng mở ra những cơ hội mới cho việc phát triển các công nghệ nhận diện giọng nói và xử lý ngôn ngữ tự nhiên trong tiếng Việt, góp phần nâng cao trải nghiệm người dùng và hiệu quả của các ứng dụng trong lĩnh vực này.

6. Phân tích và thảo luận

6.1. Hiệu suất huấn luyện



Hình 2: Đường biểu diễn giá trị Loss của mô hình qua mỗi bước

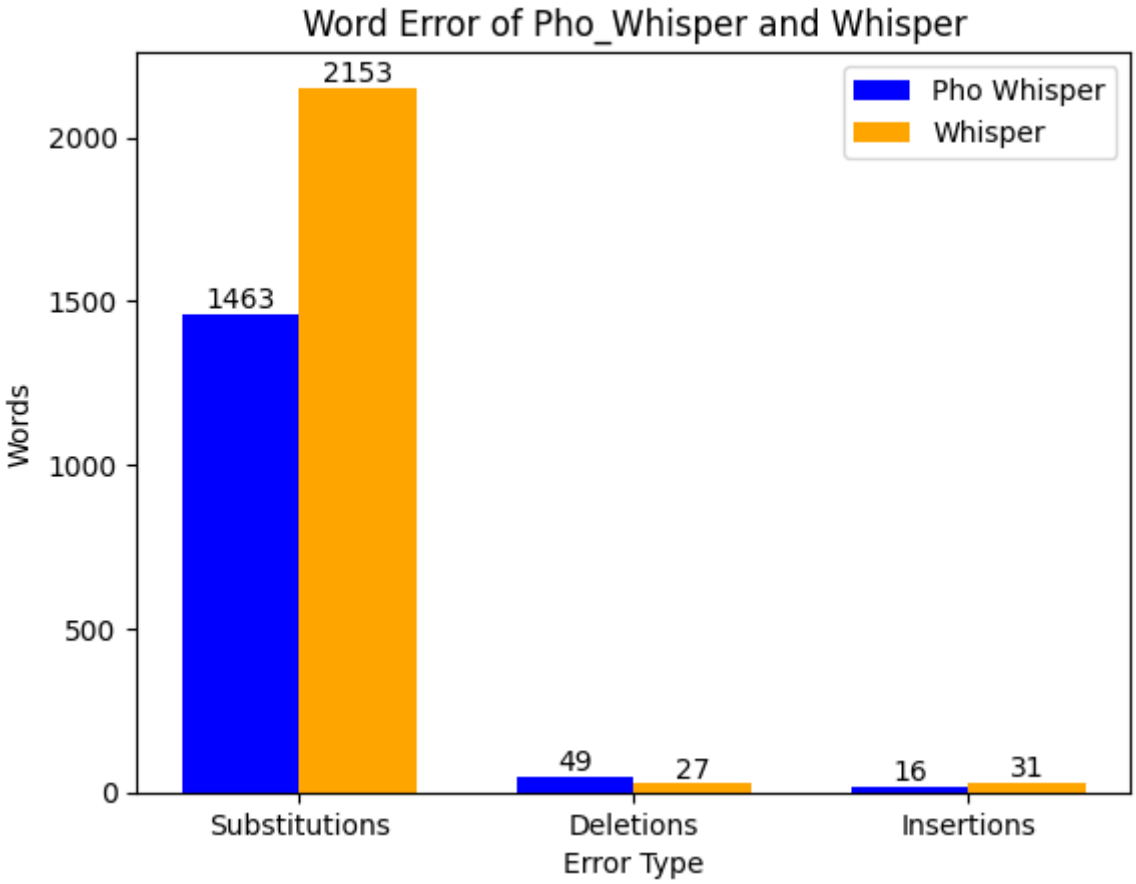
Hình 2 minh họa đường biểu diễn giá trị loss qua từng bước, cho thấy sự giảm đều đặn qua các epoch. Điều này chứng tỏ rằng cả hai mô hình đang học một cách hiệu quả từ dữ liệu, liên tục cải thiện và dần dần tiến đến điểm tối ưu. Đường biểu diễn loss giảm dần là dấu hiệu tích cực, phản ánh khả năng của các mô hình trong việc tối ưu hóa các tham số và giảm thiểu sai số.

Trong giai đoạn đầu của quá trình huấn luyện, độ dốc của đường loss của mô hình Whisper khá mạnh. Điều này là do mô hình đang bắt đầu học và thực hiện các điều chỉnh lớn đối với các tham số để tối ưu hóa hiệu suất. Những thay đổi lớn trong các epoch đầu tiên là bình thường và mong đợi khi mô hình cố gắng nắm bắt các đặc trưng cơ bản của dữ liệu.

Hiệu suất của hai mô hình sau khi fine-tune trên bộ dữ liệu mà chúng tôi đề cập có sự cải thiện rõ rệt so với khi fine-tune trên những bộ dữ liệu ASR tiếng việt khác, điều này

liên quan nhiều đến cấu trúc của bộ dữ liệu mà chúng tôi sử dụng. Như đã đề cập tại mục 3, đầu tiên ảnh hưởng lớn nhất đến hiệu suất chính là độ dài của mỗi audio của bộ dữ liệu chỉ trung bình khoảng 3 giây cộng thêm lượng dữ liệu phục vụ việc training khá lớn lên tới 110h nên dẫn đến mô hình nhận diện hiệu quả trên những audio ngắn. Ngược lại, theo như thử nghiệm của chúng tôi, mô hình nhận diện rất tệ trên những đoạn audio dài hơn 7 giây, đây cũng chính là hạn chế của mô hình khi huấn luyện trên bộ dữ liệu có thời lượng mỗi audio ngắn như thế này. Thứ hai, theo như chúng tôi khảo sát, bộ dữ liệu này chỉ chứa những từ ngữ chính quy, thông dụng, nên việc đánh giá trên tập test này cũng trở nên chính xác hơn.

6.2. Phân tích lỗi từ



Hình 3: Biểu đồ thể hiện số từ mỗi loại lỗi trong mỗi mô hình

Trong tổng số 32,113 từ trong tập kiểm tra, hệ thống nhận dạng giọng nói (ASR) của mô hình Whisper có số lượng từ bị phiên âm sai thành một từ khác cao hơn đáng kể so với PhoWhisper. Điều này cho thấy Whisper có xu hướng mắc lỗi khi nhận diện và phiên âm từ, dẫn đến việc tạo ra các từ không chính xác. Mặc dù mô hình nhận dạng các

từ này, chúng lại bị hiểu sai và chuyển đổi thành từ khác, làm giảm độ chính xác và chất lượng của văn bản.

Ngoài ra, Whisper cũng có số lượng từ không được nói ra nhưng lại được phiên âm cao hơn đáng kể so với PhoWhisper. Đây là lỗi nghiêm trọng khi mô hình thêm từ không tồn tại trong đoạn âm thanh gốc, gây ra nhiễu và sai lệch trong văn bản. Sự hiện diện của các từ thừa này không chỉ làm giảm độ tin cậy của mô hình mà còn ảnh hưởng tiêu cực đến các ứng dụng thực tế, nơi độ chính xác và sự tin cậy là vô cùng quan trọng.

Tuy nhiên, PhoWhisper lại có tỷ lệ lỗi cao hơn trong việc không phiên âm các từ có trong đoạn âm thanh gốc. Đây là trường hợp khi mô hình bỏ sót từ, không phiên âm chúng dù chúng có trong đoạn âm thanh. Mặc dù PhoWhisper vượt trội trong việc giảm thiểu lỗi phiên âm sai và thêm từ không tồn tại, nhưng vẫn gặp khó khăn trong việc đảm bảo tất cả các từ trong đoạn âm thanh được nhận diện và phiên âm đầy đủ.

Để cải thiện cả hai mô hình, cần có các biện pháp điều chỉnh và tối ưu hóa cụ thể. Đối với Whisper, cần tập trung vào việc giảm thiểu lỗi phiên âm sai và thêm từ không tồn tại. Điều này có thể đạt được bằng cách cải thiện thuật toán nhận dạng và tăng cường dữ liệu huấn luyện với các mẫu âm thanh đa dạng và phức tạp hơn. Các kỹ thuật như học sâu (deep learning) và học chuyển giao (transfer learning) có thể được áp dụng để nâng cao khả năng nhận diện của mô hình.

Đối với PhoWhisper, cần giảm thiểu lỗi bỏ sót từ bằng cách cải thiện độ nhạy và khả năng nhận diện của mô hình đối với các từ trong đoạn âm thanh. Sử dụng mô hình ngôn ngữ mạnh mẽ hơn hoặc tích hợp thêm các lớp mạng nơ-ron để xử lý các đặc trưng phức tạp của âm thanh có thể giúp giải quyết vấn đề này.

6.3. Thách thức

Trong quá trình huấn luyện mô hình, chúng tôi đã phải đối mặt với một số thách thức đáng kể. Một trong những hạn chế chính là tài nguyên tính toán. Do giới hạn về tài nguyên, chúng tôi chỉ có thể huấn luyện mỗi mô hình trong 5 epoch. Điều này đặt ra một giới hạn lớn cho khả năng tối ưu hóa và tinh chỉnh các tham số của mô hình, làm giảm cơ hội để mô hình học sâu và cải thiện hiệu suất qua nhiều vòng huấn luyện.

Ngoài ra, dữ liệu huấn luyện chủ yếu nằm trong khoảng thời gian từ 2 đến 3 giây. Điều này dẫn đến một thách thức khác: khả năng nhận diện các đoạn audio dài hơn của mô hình chưa thực sự tốt. Khi mô hình được huấn luyện chủ yếu trên các đoạn âm thanh ngắn, nó có thể không đủ khả năng nắm bắt và phân tích các đặc trưng của những đoạn audio dài hơn, dẫn đến việc giảm hiệu quả trong việc nhận diện và xử lý các âm thanh phức tạp và kéo dài.

Mặc dù mô hình Whisper đã đạt được những kết quả tích cực trên bộ dữ liệu hiện tại, vẫn còn tồn tại một số vấn đề cần được giải quyết. Tỉ lệ lỗi từ của mô hình, tức là số lượng từ bị nhận diện sai hoặc không chính xác, vẫn còn khá cao. Điều này chỉ ra rằng mặc dù mô hình hoạt động tốt trong nhiều trường hợp, nhưng vẫn cần có sự cải thiện để đạt được độ chính xác cao hơn và giảm thiểu các lỗi nhận diện. Việc cải thiện mô hình Whisper trong tương lai là rất cần thiết.

7. Kết luận

Trong báo cáo này, chúng tôi đã trình bày các nghiên cứu thực nghiệm về hai mô hình nhận dạng giọng nói tự động (ASR) là Whisper và PhoWhisper dành cho tiếng Việt. Nghiên cứu được thực hiện trên bộ dữ liệu từ YouTube bằng tiếng Việt, với mục tiêu đánh giá và so sánh hiệu suất của hai mô hình này. Kết quả thu được cho thấy mô hình PhoWhisper đạt hiệu suất vượt trội, với các điểm đánh giá chất lượng cao hơn so với kỳ vọng ban đầu. Mô hình PhoWhisper đã chứng minh khả năng nhận diện giọng nói tiếng Việt một cách hiệu quả và chính xác, vượt qua những thách thức mà mô hình Whisper còn gặp phải. Điều này không chỉ khẳng định tiềm năng của PhoWhisper trong việc ứng dụng vào thực tế mà còn mở ra nhiều hướng đi mới cho nghiên cứu và phát triển các công nghệ nhận dạng giọng nói tiếng Việt.

Chúng tôi tin rằng nghiên cứu được trình bày trong báo cáo này sẽ đóng góp quan trọng vào sự phát triển và ứng dụng của các mô hình Whisper và PhoWhisper trong tương lai. Những kết quả tích cực này hy vọng sẽ thúc đẩy việc sử dụng rộng rãi các mô hình này trong các ứng dụng thực tế, từ các hệ thống trợ lý ảo đến các công cụ hỗ trợ học tập và làm việc. Hơn nữa, nghiên cứu này cũng mở đường cho những tiến bộ và hợp tác sâu

rộng hơn trong lĩnh vực nhận dạng giọng nói. Với sự phát triển không ngừng của công nghệ và sự quan tâm ngày càng lớn từ cộng đồng nghiên cứu, chúng tôi kỳ vọng rằng sẽ có nhiều công trình nghiên cứu và cải tiến mới dựa trên nền tảng của Whisper và PhoWhisper, góp phần nâng cao chất lượng và hiệu quả của các hệ thống ASR tiếng Việt.

Chương 4: KẾT LUẬN

Trong dự án này, chúng tôi đã tập trung vào một vấn đề đầy tiềm năng và thu hút sự quan tâm rộng rãi trong cộng đồng nghiên cứu: "Nhận diện Giọng nói Tự động cho tiếng Việt" (Automatic Speech Recognition for Vietnamese). Đây là một lĩnh vực nghiên cứu quan trọng, với tiềm năng ứng dụng rộng lớn trong nhiều lĩnh vực như giáo dục, y tế, dịch vụ khách hàng, và truyền thông. Đầu tiên, chúng tôi trình bày tổng quan về bài toán này, cùng với việc giới thiệu các mô hình từ những bài báo uy tín để cung cấp cơ sở lý thuyết vững chắc cho nghiên cứu của mình.

Chúng tôi đã sử dụng bộ dữ liệu “viet_youtube_asr_corpus_v2”, một tập dữ liệu phong phú được thu thập từ nền tảng mạng xã hội YouTube, chứa các đoạn âm thanh tiếng Việt. Bộ dữ liệu này không chỉ đa dạng về ngữ cảnh và nội dung mà còn phong phú về giọng điệu và phát âm, tạo nên một thách thức đáng kể cho các mô hình nhận diện giọng nói. Bộ dữ liệu này bao gồm hàng trăm giờ âm thanh từ nhiều nguồn khác nhau, giúp đảm bảo rằng các mô hình được huấn luyện trên nó có thể xử lý tốt các biến thể ngôn ngữ và âm thanh thực tế.

Tiếp theo, chúng tôi tiến hành thực nghiệm trên hai mô hình Whisper và PhoWhisper. Kết quả cho thấy PhoWhisper, một mô hình được tinh chỉnh từ Whisper, đạt được hiệu suất tốt nhất trong các bài kiểm tra của chúng tôi. Cụ thể, PhoWhisper đã thể hiện khả năng vượt trội trong việc chuyển đổi âm thanh thành văn bản tiếng Việt, chứng minh sự ưu việt của mô hình này khi áp dụng cho ngôn ngữ Việt Nam. Mô hình PhoWhisper không chỉ vượt qua Whisper mà còn vượt trội so với các mô hình nhận diện giọng nói khác từng được thử nghiệm trên cùng bộ dữ liệu.

Chúng tôi cũng đưa ra các phân tích và đánh giá chi tiết về ưu điểm và thách thức của từng mô hình. Whisper, mặc dù có tiềm năng lớn, vẫn gặp phải những hạn chế trong việc nhận diện giọng nói tiếng Việt một cách chính xác. Các vấn đề như sự khác biệt về giọng điệu, tốc độ nói, và tiếng ồn môi trường đều ảnh hưởng đáng kể đến hiệu suất của mô hình này. Trong khi đó, PhoWhisper đã khắc phục được nhiều vấn đề này nhờ vào việc tinh chỉnh và tối ưu hóa dựa trên đặc thù của ngôn ngữ. Việc sử dụng kỹ thuật fine-

tuning trên bộ dữ liệu tiếng Việt đã giúp PhoWhisper cải thiện độ chính xác và tính ổn định trong việc nhận diện giọng nói.

Để tiếp tục cải thiện hiệu suất của các mô hình ASR cho tiếng Việt, chúng tôi đề xuất một số hướng đi tương lai như tiếp tục tinh chỉnh và phát triển các mô hình hiện tại để đạt độ chính xác cao hơn, đặc biệt là trong các điều kiện âm thanh phức tạp. Điều này có thể bao gồm việc áp dụng các kỹ thuật học sâu tiên tiến hơn, như mô hình transformer, và việc sử dụng các kỹ thuật học không giám sát để khai thác tốt hơn dữ liệu âm thanh không gán nhãn. Bên cạnh đó, thu thập thêm dữ liệu từ nhiều nguồn khác nhau để làm giàu bộ dữ liệu huấn luyện, giúp mô hình học được nhiều đặc trưng âm thanh đa dạng hơn. Điều này đặc biệt quan trọng trong việc xử lý các biến thể ngôn ngữ và ngữ cảnh khác nhau, giúp mô hình có thể ứng dụng rộng rãi hơn trong thực tế. Ngoài ra, khuyến khích sự hợp tác giữa các nhà nghiên cứu, các tổ chức và doanh nghiệp để chia sẻ kiến thức, dữ liệu và công nghệ, nhằm thúc đẩy nhanh quá trình nghiên cứu và ứng dụng ASR trong thực tế. Các dự án hợp tác quốc tế và liên ngành có thể mang lại nhiều lợi ích, từ việc chia sẻ tài nguyên đến việc phát triển các ứng dụng tiên tiến.

Chúng tôi hy vọng rằng nghiên cứu này sẽ mở đường cho những tiến bộ và hợp tác sâu rộng hơn trong lĩnh vực nhận diện giọng nói, đặc biệt là đối với tiếng Việt. Việc phát triển các hệ thống ASR hiệu quả không chỉ góp phần nâng cao chất lượng cuộc sống mà còn mở ra nhiều cơ hội mới trong việc ứng dụng công nghệ vào các lĩnh vực như giáo dục, chăm sóc sức khỏe, và dịch vụ công. Nghiên cứu của chúng tôi mong muốn là nền tảng để các nhà nghiên cứu và phát triển tiếp tục xây dựng và hoàn thiện các giải pháp nhận diện giọng nói cho tương lai. Chúng tôi kỳ vọng rằng các kết quả đạt được sẽ thúc đẩy sự phát triển của công nghệ ASR, không chỉ giới hạn ở tiếng Việt mà còn mở rộng ra các ngôn ngữ khác, góp phần xây dựng một thế giới kỹ thuật số toàn diện và đa dạng hơn.

TÀI LIỆU THAM KHẢO

[1] Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." *International Conference on Machine Learning*. PMLR, 2023.

[2] Le, Thanh-Thien, and Dat Quoc Nguyen. "PhoWhisper: Automatic Speech Recognition for Vietnamese." *The Second Tiny Papers Track at ICLR 2024*.

[3] Jamal, Norezmi, et al. "Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: A review." *AIP Conference Proceedings*. Vol. 1883. No. 1. AIP Publishing, 2017.

[4] Schultz, Benjamin G., et al. "Automatic speech recognition in neurodegenerative disease." *International Journal of Speech Technology* 24.3 (2021): 771-779.

[5] Cowie, Roddy, and Randolph R. Cornelius. "Describing the emotional states that are expressed in speech." *Speech communication* 40.1-2 (2003): 5-32.

[6] Vincent, Emmanuel, et al. "An analysis of environment, microphone and data simulation mismatches in robust speech recognition." *Computer Speech & Language* 46 (2017): 535-557.

[7] Kexin, Tian, et al. "Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition." *2019 Chinese Automation Congress (CAC)*. IEEE, 2019.

[8] Grozdić, Đorđe T., and Slobodan T. Jovičić. "Whispered speech recognition using deep denoising autoencoder and inverse filtering." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12 (2017): 2313-2322.

[9] Lotfidereshgi, Reza, and Philippe Gournay. "Biologically inspired speech emotion recognition." *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017.

[10] Ali, Ahmed, and Steve Renals. "Word error rate estimation for speech recognition: e-WER." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2018.

[11] Chuang, Shun-Po, et al. "Worse wer, but better bleu? leveraging word embedding as intermediate in multitask end-to-end speech translation." arXiv preprint arXiv:2005.10678 (2020).

PHỤ LỤC

A. Bộ dữ liệu

- **CMV–Vi 14**: https://commonvoice.mozilla.org/en/datasets
- **VIVOS**: https://huggingface.co/datasets/vivos
- **viet_youtube_asr_corpus_v2**:https://huggingface.co/datasets/linhtran92/viet_youtube_asr_corpus_v2/viewer/default/train

B. Mô hình Whisper

Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

Bảng 3: Chi tiết kiến trúc của dòng mô hình Whisper