# ĐẠI HỌC QUỐC GIA TP HCM

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

🙡★🙣



# Đề cương nghiên cứu

Ứng dụng mô hình Whisper Large v3 trong nhận dạng giọng nói tiếng Việt

IT2021.CH190: Xử lý tín hiệu số nâng cao

Giảng viên: TS. Hà Minh Tân

Học viên thực hiện:

Văn Đức Ngọ 240101020

Lê Thanh Dũng 240101074

Nguyễn Hoàng Hải 240101008

Trần Quốc Huy 230101048

# 

# **Abstract**

Whisper Large v3, do OpenAI phát triển, là một trong những mô hình nhận dạng giọng nói tiên tiến nhất hiện nay, hỗ trợ đa ngôn ngữ với độ chính xác cao. Tuy nhiên, mô hình vẫn bị ảnh hưởng bởi nhiễu và khoảng lặng trong dữ liệu đầu vào, làm giảm hiệu suất nhận dạng, đặc biệt khi áp dụng vào tiếng Việt trong môi trường thực tế.

Nghiên cứu này triển khai và đánh giá Whisper Large v3 trên tiếng Việt, thử nghiệm trên hai loại dữ liệu: dữ liệu gốc có nhiễu và dữ liệu đã qua xử lý (lọc nhiễu và loại bỏ khoảng lặng). Việc so sánh kết quả giúp xác định mức độ ảnh hưởng của các phương pháp tiền xử lý đến độ chính xác nhận dạng giọng nói.

Ngoài ra, nghiên cứu còn thực hiện fine-tune mô hình bằng cách tiền huấn luyện (pre-train) trên tập dữ liệu tiếng Việt có chứa nhiễu và đã được xử lý, giúp mô hình thích nghi tốt hơn với đặc điểm âm thanh của tiếng Việt trong các điều kiện khác nhau. Quá trình fine-tune dự kiến sẽ cải thiện khả năng xử lý nhiễu, nâng cao độ chính xác nhận dạng và giảm chỉ số Word Error Rate (WER).

Kết quả mong đợi là giảm đáng kể WER sau khi lọc nhiễu và loại bỏ khoảng lặng, giúp cải thiện độ chính xác nhận dạng giọng nói tiếng Việt. Điều này góp phần nâng cao tính ứng dụng thực tế của hệ thống trong các lĩnh vực như chuyển đổi giọng nói thành văn bản, điều khiển bằng giọng nói và các công nghệ hỗ trợ khác. Nghiên cứu này sẽ đóng góp vào việc tối ưu hóa Whisper Large v3 cho tiếng Việt, tạo nền tảng cho các nghiên cứu tiếp theo về nhận dạng giọng nói trong môi trường có nhiễu.

# **1. Giới Thiệu**

Nhận dạng giọng nói (Automatic Speech Recognition - ASR) là một lĩnh vực quan trọng trong trí tuệ nhân tạo, ứng dụng rộng rãi trong trợ lý ảo, dịch thuật tự động và chuyển đổi giọng nói thành văn bản. Sự phát triển của deep learning đã nâng cao độ chính xác của ASR, trong đó Whisper Large v3 của OpenAI là một trong những mô hình tiên tiến nhất hỗ trợ đa ngôn ngữ, bao gồm tiếng Việt. Tuy nhiên, mô hình vẫn gặp khó khăn khi xử lý dữ liệu có nhiễu và khoảng lặng, làm giảm độ chính xác trong môi trường thực tế.

Chất lượng dữ liệu đầu vào ảnh hưởng lớn đến hiệu suất hệ thống nhận dạng giọng nói (ASR). Trong thực tế, tiếng ồn môi trường như giao thông, hội thoại nền hoặc máy móc có thể làm tăng tỷ lệ lỗi từ (WER). Ngoài ra, các khoảng lặng dư thừa cũng gây ảnh hưởng tiêu cực. Vì vậy, tiền xử lý như lọc nhiễu và loại bỏ khoảng lặng là cần thiết để nâng cao độ chính xác của ASR.

Nghiên cứu này tập trung vào hai hướng chính nhằm tối ưu hóa hiệu suất của Whisper Large v3 trong nhận dạng giọng nói tiếng Việt.

* Lọc nhiễu và khoảng lặng. Nhóm sẽ chạy mô hình trên dữ liệu gốc, dữ liệu có nhiễu, sau đó áp dụng các kỹ thuật lọc nhiễu và loại bỏ khoảng lặng để cải thiện chất lượng đầu vào, sau đó so sánh kết quả trước và sau khi lọc nhiễu để đánh giá hiệu suất cải thiện.
* Fine-Turn mô hình với dữ liệu tiếng Việt có nhiễu và dữ liệu đã được lọc nhiễu, sau đó chạy thử nghiệm với mô hình đã pre-train để đánh giá sự cải thiện về độ chính xác.

Để đánh giá hiệu quả của các phương pháp trên, nghiên cứu sử dụng Word Error Rate (WER) làm thước đo chính. Chỉ số WER phản ánh mức độ sai sót trong kết quả nhận dạng, do đó, mục tiêu của nghiên cứu là giảm WER bằng cách tối ưu hóa chất lượng dữ liệu đầu vào và cải thiện khả năng học của mô hình thông qua pre-training.

Kết quả mong đợi từ nghiên cứu này là xây dựng một mô hình Whisper Large v3 được tối ưu hóa tốt hơn cho tiếng Việt, giúp nâng cao độ chính xác nhận dạng giọng nói trong các điều kiện môi trường thực tế. Những đóng góp của nghiên cứu không chỉ giúp cải thiện hiệu suất nhận dạng giọng nói mà còn mở rộng khả năng ứng dụng của mô hình trong các hệ thống thực tế như trợ lý ảo, công nghệ chuyển đổi giọng nói thành văn bản tự động, và các ứng dụng giao tiếp bằng giọng nói trong môi trường có nhiễu.

# **2. Mục tiêu của dự án**

Nghiên cứu này nhằm triển khai, thử nghiệm và tối ưu mô hình Whisper Large v3 cho nhận dạng giọng nói tiếng Việt, với ba mục tiêu chính:

* Triển khai và đánh giá Whisper Large v3 trên dữ liệu tiếng Việt, đồng thời phân tích tác động của nhiễu đến độ chính xác nhận dạng.
* Cải thiện chất lượng đầu vào bằng cách áp dụng các phương pháp lọc nhiễu và loại bỏ khoảng lặng.
* Pre-train mô hình Whisper trên dữ liệu tiếng Việt có nhiễu và đã được xử lý, sau đó đánh giá hiệu suất bằng Word Error Rate (WER) để đo lường tác động của việc lọc nhiễu và pre-training.

# **3. Methodology**

## **3.1 Chuẩn bị dữ liệu**

1. **Thu thập dữ liệu**

Dữ liệu sẽ được lấy từ bộ dữ liệu tiếng Việt từ FPT Open Speech Dataset. Bộ dữ liệu giọng nói từ FPT AI. Chia dữ liệu thành 2 bộ dữ liệu là Train và Test. Sau đó tăng cường dữ liệu bằng cách biến đổi speed của dữ liệu.

Định dạng file: .wav (chuẩn 16kHz hoặc 44.1kHz, 16-bit PCM).

### **Xử lý Dữ liệu**

Sau khi thu thập, dữ liệu cần được xử lý để đảm bảo chất lượng đầu vào tốt nhất cho mô hình.

* Chuyển đổi định dạng: chuyển đổi về dạng WAV để tốt hơn cho xử lý âm thanh
* Chuẩn hóa mức âm lượng: Sử dụng sox để chuẩn hóa âm lượng
* Cắt đoạn giọng nói và loại bỏ im lặng: Có thể dùng công cụ VAD (Voice Activity Detection) từ Webrtc hoặc Silero VAD để phát hiện và loại bỏ khoảng lặng.

1. **Triển khai Phương pháp Lọc Nhiễu và Khoảng Lặng**

* Phương pháp Lọc Nhiễu
  + Fast Fourier Transform (FFT) - Biến đổi Fourier nhanh: Dùng để phân tích phổ tần số và loại bỏ các thành phần nhiễu không mong muốn.
  + Spectral Subtraction: Được sử dụng phổ biến trong xử lý âm thanh để ước lượng và loại bỏ nhiễu nền.
* Xử lý Khoảng Lặng
  + Voice Activity Detection (VAD): Dùng để phát hiện phần nào trong âm thanh có giọng nói và loại bỏ phần không có giọng.
  + WebRTC VAD thư viện phổ biến cho xử lý khoảng lặng
  + Sox - Auto Trim Silence: nhanh chóng loại bỏ khoảng lặng
  + Energy-based Silence Removal: Dựa trên mức năng lượng tín hiệu để quyết định phần nào là khoảng lặng

## **3.2 Kiểm thử trên Whisper Large v3**

* Chạy mô hình trên tập dữ liệu gốc để lấy kết quả nhận dạng cơ bản.
* Áp dụng kỹ thuật lọc nhiễu, loại bỏ khoảng lặng và thử nghiệm lại mô hình.
* Chạy mô hình trên dữ liệu đã lọc nhiễu.
* Dùng chỉ số WER để đánh giá hiệu suất của mô hình Whisper Large v3 trên hai bộ dữ liệu: Dữ liệu gốc và dữ liệu đã được lọc nhiễu để kiểm tra độ hiệu quả của bộ lọc.

## **3.3 Fine-Tune mô hình**

Pre-trained model sẽ được fine tuning trên bộ dữ liệu train của bộ dữ liệu Tiếng Việt đã được lọc nhiễu. Mục đích là tăng cường hiệu suất của mô hình trên dataset tiếng Việt.

* Mô hình Whisper large V3 - 1.54B
* Dataset: <https://www.kaggle.com/datasets/thinh127/fpt-open-speech-dataset-fosd-vietnamese/data>

## **3.4 Đánh giá hiệu suất của Fine-Tune model**

Chạy mô hình trên bộ dữ liệu Tiếng Việt đã được lọc nhiễu, đánh giá hiệu suất của mô hình bằng chỉ số WER trước và sau khi fine-tuning model.

## **3.5 Viết báo cáo và tài liệu hướng dẫn**

Tổng hợp kết quả nghiên cứu.

Viết hướng dẫn chi tiết để triển khai và sử dụng mô hình

# **4. Tiến độ thực hiện**

| Giai đoạn | Mô tả công việc |
| --- | --- |
| Tuần 1 - 2 | Tìm hiểu mô hình Whisper và phương pháp lọc nhiễu.  Chuẩn bị và xử lý dữ liệu.  Kiểm thử mô hình với dữ liệu gốc và nhiễu. |
| Tuần 3 - 4 | Lọc nhiễu, chạy lại mô hình và đánh giá kết quả.  Pre-train mô hình trên tập dữ liệu tiếng Việt  Đánh giá hiệu suất của mô hình đã pre-train. |
| Tuần 5 - 6 | Hoàn thiện báo cáo và trình bày kết quả. |

# **5. Phân chia công việc của các thành viên**

| Thành viên | Công việc phụ trách |
| --- | --- |
| Văn Đức Ngọ | Tìm hiểu mô hình Whisper và triển khai code ứng dụng Whisper  Tìm hiểu các phương pháp lọc nhiễu.  Viết báo cáo, tài liệu hướng dẫn sử dụng |
| Lê Thanh Dũng | Chuẩn bị và xử lý dữ liệu.  Triển khai phương pháp lọc nhiễu và khoảng lặng.  Viết báo cáo, tài liệu hướng dẫn sử dụng |
| Nguyễn Hoàng Hải | Thực hiện thử nghiệm và đánh giá mô hình với dữ liệu gốc và dữ liệu đã được lọc nhiễu  Viết báo cáo, tài liệu hướng dẫn sử dụng |
| Trần Quốc Huy | Pre-train mô hình trên tập dữ liệu tiếng Việt.  Đánh giá hiệu suất của mô hình đã pre-train.  Viết báo cáo, tài liệu hướng dẫn sử dụng |

# **6. Kết Quả Dự Kiến**

* Giảm đáng kể Word Error Rate (WER) sau khi lọc nhiễu và loại bỏ khoảng lặng
  + Việc áp dụng các phương pháp tiền xử lý như lọc nhiễu và loại bỏ khoảng lặng sẽ giúp cải thiện đáng kể chất lượng dữ liệu đầu vào.
  + WER được kỳ vọng giảm mạnh do mô hình nhận dạng có thể phân tích tín hiệu giọng nói rõ ràng hơn, hạn chế lỗi do nhiễu và khoảng lặng không cần thiết.
  + Kết quả sẽ được so sánh giữa các tập dữ liệu gốc, có nhiễu và đã qua xử lý để đánh giá mức độ hiệu quả của phương pháp cải tiến.
* Fine-Tune giúp mô hình cải thiện độ chính xác nhận dạng tiếng Việt
  + Mô hình Whisper Large v3 sẽ được fine-tune trên tập dữ liệu tiếng Việt chứa nhiễu và tập dữ liệu đã qua lọc nhiễu để tăng khả năng thích nghi với điều kiện thực tế.
  + Kết quả thử nghiệm với mô hình fine-tuned dự kiến sẽ có độ chính xác cao hơn so với mô hình gốc, đặc biệt là trong môi trường có nhiễu.
  + WER sẽ được đo lường trước và sau khi pre-train để đánh giá mức độ cải thiện của mô hình đối với nhận dạng giọng nói tiếng Việt.
* Hệ thống có thể triển khai thực tế trong môi trường có nhiễu.
  + Sau khi tối ưu hóa, hệ thống nhận dạng giọng nói có thể được áp dụng trong các môi trường thực tế như trung tâm chăm sóc khách hàng, điều khiển bằng giọng nói, hoặc trợ lý ảo.
  + Mô hình đã được huấn luyện với dữ liệu có nhiễu sẽ giúp tăng độ ổn định và khả năng hoạt động trong điều kiện thực tế, nơi tiếng ồn luôn tồn tại.
  + Các kết quả thu được từ nghiên cứu có thể đóng góp vào việc cải tiến các hệ thống nhận dạng giọng nói tiếng Việt trong nhiều ứng dụng khác nhau.

# **7. Tài Liệu Tham Khảo**

[1] OpenAI. (n.d.). Whisper Large v3. Hugging Face. Retrieved from <https://huggingface.co/openai/whisper-large-v3>

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762. Retrieved from <https://arxiv.org/abs/1706.03762>

[3] Elsevier. (n.d.). Speech processing for ASR. ScienceDirect. Retrieved from <https://www.sciencedirect.com/topics/computer-science/speech-recognition>

[4] Author Unknown. (n.d.). A study on the fast Fourier transform applications. American Journal of Biomedical Science & Research. Retrieved from <https://ajbsr.net/data/uploads/57363.pdf>

[5] Kondoz, A. M. (2013). Noise removal in speech processing using spectral subtraction. Scientific Research Publishing. Retrieved from <https://www.scirp.org/html/2-3400314_45989.htm>

[6] Hugging Face. (2023). Fine-tune Whisper for multilingual ASR with transformers. Retrieved from https://huggingface.co/blog/fine-tune-whisper

[7] Kynthesis. (2022). VIVOS: Vietnamese speech corpus for ASR (VNU-HCMUS). Kaggle. Retrieved from <https://www.kaggle.com/datasets/kynthesis/vivos-vietnamese-speech-corpus-for-asr>

[8] Thinh127. (2023). FPT Open Speech Dataset: Vietnamese speech dataset from FPT AI. Kaggle. Retrieved from <https://www.kaggle.com/datasets/thinh127/fpt-open-speech-dataset-fosd-vietnamese/data>

[9] Vincenzo Timmel, Claudio Paonessa, Reza Kakooee, Manfred Vogel, Daniel Perruchoud (2024). Fine-tuning Whisper on Low-Resource Languages for Real-World Applications. arXiv preprint arXiv:2412.15726. Retrieved from <https://arxiv.org/abs/2412.15726>

[9] Yunpeng Liu; Dan Qu (2024). Parameter-efficient fine-tuning of Whisper for low-resource speech recognition.

# **Kết luận**

Nghiên cứu này đã đề xuất và thực hiện kế hoạch tối ưu hóa mô hình Whisper Large v3 để nâng cao độ chính xác nhận dạng giọng nói tiếng Việt. Bằng cách áp dụng phương pháp lọc nhiễu và loại bỏ khoảng lặng, chúng tôi hướng đến việc cải thiện chất lượng đầu vào giúp mô hình hoạt động hiệu quả hơn trong môi trường thực tế. Ngoài ra, quá trình pre-training trên dữ liệu tiếng Việt giúp mô hình thích nghi tốt hơn với đặc thù của ngôn ngữ này.

Kết quả mong đợi là một hệ thống nhận dạng giọng nói có độ chính xác cao, giảm sai số WER, tối ưu thời gian xử lý và có khả năng triển khai thực tế. Hệ thống này có thể ứng dụng trong các lĩnh vực như trợ lý ảo, tổng đài tự động, nhận dạng giọng nói trong môi trường nhiễu và các ứng dụng khác yêu cầu nhận dạng tiếng Việt chính xác.