**ĐẠI HỌC BÁCH KHOA HÀ NỘI**



**BÁO CÁO**

**Automatic Speech Recognition**

**Nhóm 9:**

**Lê Đinh Thái Sơn - 20240831E**

**Lê Xuân Giao - 20210290**

**Hà Bùi Phúc - 20240834E**

**Trần Hồng Quân - 20240818E**

|  |  |  |
| --- | --- | --- |
| **Giảng viên hướng dẫn:** | Nguyễn Thị Thu Trang | |
| **Học phần:** | Tạo sinh âm thanh | |
|  | |
| **HÀ NỘI, 5/2025** | | |

**Mục lục**

[I. Giới thiệu 4](#_Toc198471278)

[1. Tổng quan về bài toán Nhận dạng tiếng nói tự động (ASR) cho tiếng Việt 4](#_Toc198471279)

[1.1. Khái niệm ASR 4](#_Toc198471280)

[1.2. Đặc thù ngôn ngữ tiếng Việt trong ASR 4](#_Toc198471281)

[2. Ứng dụng và tầm quan trọng của ASR đối với tiếng Việt 4](#_Toc198471282)

[2.1. Ứng dụng trong thực tế 4](#_Toc198471283)

[2.2. Tầm quan trọng với tiếng Việt 5](#_Toc198471284)

[3. Mục tiêu của báo cáo 5](#_Toc198471285)

[3.1. Đánh giá hiệu quả mô hình ASR hiện đại cho tiếng Việt 5](#_Toc198471286)

[3.2. Thực nghiệm fine-tuning mô hình Whisper-small với kỹ thuật LoRA 5](#_Toc198471287)

[3.3. So sánh hiệu suất với mô hình PhoWhisper 5](#_Toc198471288)

[3.4. Phân tích các yếu tố ảnh hưởng đến hiệu quả nhận dạng 5](#_Toc198471289)

[II. Kiến thức nền tảng (Background) 6](#_Toc198471290)

[1. Tổng quan pipeline của hệ thống ASR 6](#_Toc198471291)

[2. Mô hình sử dụng trong hệ thống 6](#_Toc198471292)

[3. Các yếu tố ảnh hưởng đến kết quả nhận dạng 7](#_Toc198471293)

[III. Tập dữ liệu sử dụng (Dataset) 7](#_Toc198471294)

[1. Tập huấn luyện VLSP 2020 7](#_Toc198471295)

[2. Tập kiểm thử VIVOS 8](#_Toc198471296)

[3. Tập kiểm thử Public và Private (đến từ cuộc thi đánh giá mô hình ASR) 8](#_Toc198471297)

[3.1. Biểu đồ phân bố thời lượng (duration\_sec) 9](#_Toc198471298)

[3.2. Biểu đồ phân bố năng lượng âm (rms\_energy) 10](#_Toc198471299)

[3.3. Bảng thống kê đặc trưng âm thanh 11](#_Toc198471300)

[IV. Tiền xử lý và cài đặt mô hình (Preprocessing & Model Setup) 11](#_Toc198471301)

[1. Tiền xử lý dữ liệu 11](#_Toc198471302)

[2. Cài đặt mô hình 12](#_Toc198471303)

[V. Thực nghiệm và Đánh giá (Experiment & Evaluation) 12](#_Toc198471304)

[1. Môi trường thực nghiệm 12](#_Toc198471305)

[2. Đánh giá hiệu suất 13](#_Toc198471306)

[3. Kết quả thực nghiệm 13](#_Toc198471307)

[VI. Thảo luận kết quả (Discussion) 14](#_Toc198471308)

[Kết quả thực nghiệm đưa ra nhiều điểm cần lưu ý: 14](#_Toc198471309)

[Phân tích lỗi 16](#_Toc198471310)

[VII. Kết luận và Hướng phát triển (Conclusion & Future Work) 16](#_Toc198471311)

[VIII. Tài liệu tham khảo (References) 17](#_Toc198471312)

# I. Giới thiệu

## 1. Tổng quan về bài toán Nhận dạng tiếng nói tự động (ASR) cho tiếng Việt

### 1.1. Khái niệm ASR

Nhận dạng tiếng nói tự động (Automatic Speech Recognition – ASR) là bài toán xử lý ngôn ngữ tự nhiên nhằm chuyển đổi tín hiệu âm thanh thành chuỗi văn bản tương ứng. Hệ thống ASR thường bao gồm các thành phần chính như: tiền xử lý tín hiệu, trích xuất đặc trưng, mô hình hóa ngôn ngữ, và giải mã đầu ra.

### 1.2. Đặc thù ngôn ngữ tiếng Việt trong ASR

Tiếng Việt là ngôn ngữ đơn âm, có hệ thống thanh điệu phong phú (6 thanh), và có độ đa dạng cao về vùng miền trong cách phát âm. Ngoài ra, mỗi âm tiết có thể mang nhiều nghĩa khác nhau tùy vào dấu thanh, do đó mô hình ASR cần có khả năng phân biệt chính xác từng thanh điệu, âm vị, và ngữ cảnh để đảm bảo độ chính xác nhận dạng.

## 2. Ứng dụng và tầm quan trọng của ASR đối với tiếng Việt

### 2.1. Ứng dụng trong thực tế

ASR đang được ứng dụng rộng rãi trong nhiều lĩnh vực, bao gồm:

* Giao tiếp với trợ lý ảo (như Google Assistant, Siri)
* Chuyển lời nói thành văn bản (speech-to-text) trong các hệ thống nhập liệu
* Tạo phụ đề tự động cho video/hội nghị
* Hỗ trợ người khuyết tật trong giao tiếp
* Hệ thống call center và chăm sóc khách hàng tự động

### 2.2. Tầm quan trọng với tiếng Việt

Tuy nhiên, phần lớn hệ thống ASR hiện tại vẫn còn thiếu tối ưu khi áp dụng cho tiếng Việt. Việc phát triển và tinh chỉnh các mô hình ASR hiệu quả dành riêng cho tiếng Việt sẽ giúp tăng độ bao phủ ngôn ngữ trong các ứng dụng AI, đảm bảo tính công bằng ngôn ngữ và nâng cao trải nghiệm người dùng.

## 3. Mục tiêu của báo cáo

### 3.1. Đánh giá hiệu quả mô hình ASR hiện đại cho tiếng Việt

Chúng em đánh giá khả năng nhận dạng tiếng Việt của các mô hình ASR hiện đại, với tiêu chí chính là độ chính xác (chỉ số WER – Word Error Rate) trên dữ liệu kiểm thử.

### 3.2. Thực nghiệm fine-tuning mô hình Whisper-small với kỹ thuật LoRA

Nhằm nâng cao hiệu suất mà vẫn tiết kiệm tài nguyên huấn luyện, chúng em thực hiện tinh chỉnh mô hình Whisper-small bằng kỹ thuật LoRA – một phương pháp tinh chỉnh hiệu quả tham số (Parameter-Efficient Fine-Tuning).

### 3.3. So sánh hiệu suất với mô hình PhoWhisper

Chúng em sử dụng mô hình PhoWhisper-small – một mô hình Whisper đã được tinh chỉnh sẵn cho tiếng Việt – để làm điểm so sánh với mô hình fine-tuned của nhóm.

### 3.4. Phân tích các yếu tố ảnh hưởng đến hiệu quả nhận dạng

Bên cạnh đánh giá định lượng, chúng em cũng phân tích các yếu tố có thể ảnh hưởng đến hiệu quả nhận dạng tiếng nói như: sự không đồng nhất về miền dữ liệu (domain mismatch), nhiễu âm, đặc trưng giọng nói tự nhiên, và độ nhạy của tiếng Việt đối với sai sót âm tiết hoặc thanh điệu.

# II. Kiến thức nền tảng (Background)

## 1. Tổng quan pipeline của hệ thống ASR

Hệ thống nhận dạng tiếng nói tự động (ASR) trong bài được xây dựng theo pipeline gồm bốn bước chính:

* **Tiền xử lý**: Dữ liệu âm thanh được chuẩn hóa về định dạng .wav, mono channel, tần số lấy mẫu 16kHz. Các tệp bị lỗi, thiếu nhãn hoặc không đạt yêu cầu được loại bỏ trước khi đưa vào huấn luyện hoặc suy luận. Các file wav được tiền xử lý trim silence, normalize volume, chuẩn hóa độ dài thành nhiều file nhỏ độ dài 5s.
* **Trích xuất đặc trưng**: Đặc trưng âm thanh được trích xuất dưới dạng **log-Mel spectrogram**. Quá trình này được thực hiện tự động bởi kiến trúc mô hình Whisper trong thư viện Hugging Face Transformers.
* **Mô hình hóa**: Hai mô hình được sử dụng:
  + Whisper-small: Mô hình gốc do OpenAI huấn luyện trên dữ liệu đa ngôn ngữ. Trong báo cáo này, mô hình được tiếp tục **fine-tuned thêm trên tập dữ liệu VLSP 2020** với 2 và 4 epoch để đánh giá hiệu quả huấn luyện bổ sung.
  + vinai/PhoWhisper-small: Là mô hình Whisper được tinh chỉnh cho tiếng Việt bởi nhóm VINAI. Được sử dụng để suy luận trực tiếp trên tập kiểm thử VIVOS.
* **Giải mã**: Văn bản được tạo ra từ đầu ra của mô hình bằng **kỹ thuật giải mã tích hợp sẵn trong pipeline của thư viện Hugging Face Transformers**, bao gồm greedy decoding hoặc beam search.

## 2. Mô hình sử dụng trong hệ thống

* **Whisper-small**: Là mô hình encoder-decoder dựa trên kiến trúc Transformer, do OpenAI huấn luyện trên tập dữ liệu lớn, đa ngôn ngữ và đa nhiệm. Trong hệ thống này, Whisper-small được sử dụng để thực hiện suy luận trực tiếp trên tập dữ liệu kiểm thử VIVOS mà không tinh chỉnh thêm. Mô hình này đóng vai trò tham chiếu để so sánh hiệu quả với mô hình đã được fine-tune.
* **PhoWhisper-small**: Là phiên bản mô hình Whisper được tinh chỉnh sẵn (pretrained fine-tuned) cho tiếng Việt do nhóm VINAI phát hành. Trong bài này, mô hình này tiếp tục được **fine-tune thêm trên tập dữ liệu VLSP 2020** (không phải VIVOS) trong 2 và 4 epoch để đánh giá khả năng cải thiện hiệu suất nhận dạng.
* **Đánh giá mô hình**: Cả hai mô hình đều được tích hợp và đánh giá bằng cách sử dụng pipeline từ thư viện Hugging Face Transformers. Quá trình đánh giá được thực hiện bằng cách cho pipeline nhận đầu vào là các file âm thanh .wav và đầu ra là văn bản đã được mô hình giải mã.

## 3. Các yếu tố ảnh hưởng đến kết quả nhận dạng

Hiệu quả nhận dạng tiếng nói tiếng Việt trong mô hình chịu ảnh hưởng bởi một số yếu tố đặc trưng ngôn ngữ và dữ liệu:

* **Giọng vùng miền:** Các biến thể về phát âm giữa các vùng miền gây khó khăn cho mô hình nếu dữ liệu huấn luyện chưa đa dạng.
* **Thanh điệu và âm tiết đơn lập:** Tiếng Việt có hệ thống thanh điệu phong phú, dễ gây nhầm lẫn giữa các từ chỉ khác nhau ở dấu.
* **Tiếng ồn và hội thoại:** Những đoạn ghi âm có nhiều tạp âm hoặc hội thoại giữa nhiều người làm giảm độ chính xác nhận dạng.

# III. Tập dữ liệu sử dụng (Dataset)

## 1. Tập huấn luyện VLSP 2020

* Tập dữ liệu VLSP 2020 bao gồm các đoạn âm thanh tiếng Việt được ghi âm chuẩn, có kèm bản ghi chú (transcripts) chính xác.
* Dữ liệu được sử dụng để huấn luyện mô hình Whisper-small với mục tiêu tăng khả năng nhận dạng tiếng Việt chuẩn.
* Dữ liệu huấn luyện được lựa chọn từ các đoạn audio và bản text tương ứng trong file /kaggle/input/prompt-for-vlsp2020/output.txt, cắt từ chỉ mục 30001 đến 45000.

## 2. Tập kiểm thử VIVOS

* Tập kiểm thử VIVOS bao gồm các đoạn hội thoại tiếng Việt, có giọng nói vùng miền đa dạng và môi trường thực tế hơn, tập trung vào khả năng đánh giá hiệu quả mô hình trên dữ liệu ngoài huấn luyện.
* File /kaggle/input/vivos-vietnamese/vivos/test/prompts.txt được sử dụng để lấy thông tin đường dẫn âm thanh và transcript cho việc đánh giá.
* Tập kiểm thử có đặc điểm phức tạp hơn, có tiếng ồn và các yếu tố khó trong tiếng nói tự nhiên.

## 3. Tập kiểm thử Public và Private (đến từ cuộc thi đánh giá mô hình ASR)

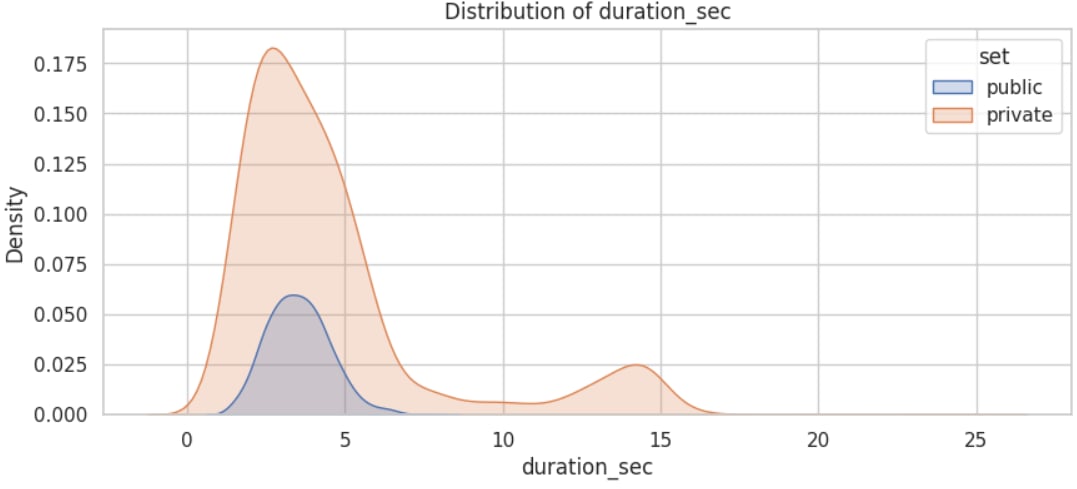
Ngoài các tập dữ liệu huấn luyện và kiểm thử truyền thống như **VLSP 2020** và **VIVOS**, mô hình cũng được đánh giá trên hai tập kiểm thử **Public** và **Private**, đến từ một **cuộc thi đánh giá mô hình nhận dạng tiếng nói tự động (ASR - Automatic Speech Recognition)**.

Hai tập này được sử dụng để đo lường khả năng tổng quát hóa và độ chính xác của mô hình trong môi trường thực tế, bao gồm các đoạn ghi âm mang tính thử thách cao hơn, được thu từ các nguồn không có trong tập huấn luyện.

* **Tập kiểm thử Public**: Là phần được công bố công khai trong giai đoạn đầu cuộc thi, cho phép các nhóm phát triển mô hình có thể thử nghiệm và tinh chỉnh hệ thống nhận dạng của mình.
* **Tập kiểm thử Private**: Được giữ kín đến giai đoạn cuối cuộc thi để đánh giá công bằng giữa các mô hình. Tập này chứa dữ liệu đa dạng và khó hơn, bao gồm tiếng ồn, giọng nói khác biệt và cách phát âm tự nhiên hơn.

Cả hai tập đều đã được xử lý và trích xuất thông tin đặc trưng như: độ dài đoạn âm thanh (duration), năng lượng RMS, số kênh và tần số mẫu. Những đặc trưng này được phân tích chi tiết thông qua biểu đồ và bảng thống kê ở các hình bên dưới.

### 3.1. Biểu đồ phân bố thời lượng (duration\_sec)



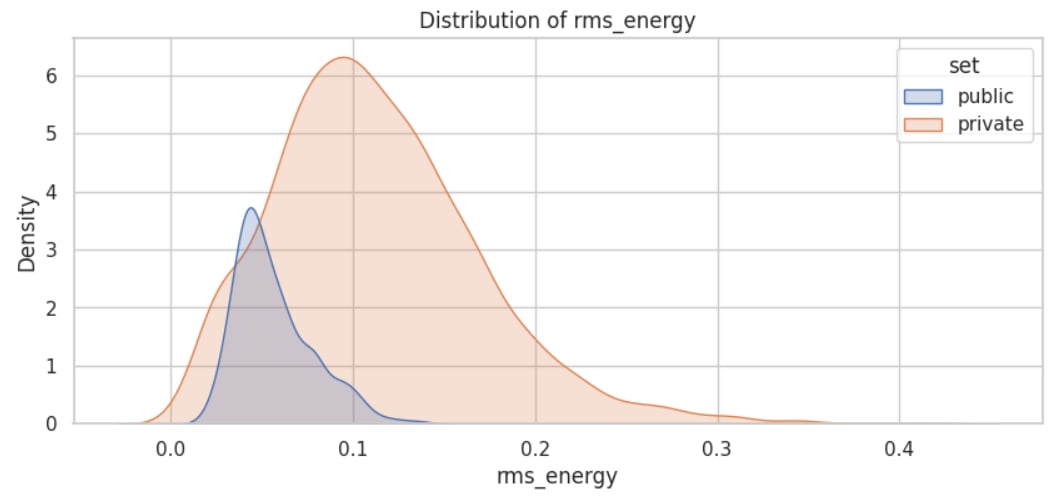
**Quan sát:**

* Tập public (màu xanh) có thời lượng chủ yếu nằm trong khoảng **2–5 giây**, với mật độ cao nhất khoảng 3–4 giây.
* Tập private (màu cam) có phân bố rộng hơn và nghiêng về phía thời lượng dài hơn, với **đỉnh phân bố lớn nhất gần 3 giây**, và có nhiều mẫu kéo dài đến hơn 15 giây.

**Ý nghĩa:**

* Tập private chứa nhiều đoạn âm thanh dài hơn, điều này có thể ảnh hưởng đến độ chính xác của mô hình nếu mô hình chưa được huấn luyện tốt với đoạn dài.
* Tính đa dạng về thời lượng trong tập private cũng có thể là nguyên nhân gây khó khăn cho các mô hình.

### ****3.2. Biểu đồ phân bố năng lượng âm (****rms\_energy****)****



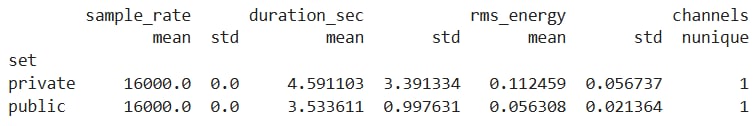
**Quan sát:**

* Tập public có năng lượng âm RMS tập trung chặt trong khoảng 0.03–0.06, với giá trị đỉnh quanh **0.04**.
* Tập private có mức năng lượng RMS **cao hơn đáng kể**, với đỉnh khoảng **0.09**, và phân bố trải rộng hơn đến 0.3–0.4.

**Ý nghĩa:**

* Năng lượng RMS cao hơn cho thấy tập private có **âm lượng lớn hơn hoặc nhiễu nhiều hơn**, ảnh hưởng đến khả năng nhận dạng nếu mô hình không được huấn luyện với các điều kiện âm thanh tương tự.
* Đây là một nguyên nhân tiềm năng gây ra lỗi trong quá trình nhận dạng của PhoWhisper-small, vốn đã fine-tuned với dữ liệu sạch hơn (VLSP 2020).

### ****3.3. Bảng thống kê đặc trưng âm thanh****



**So sánh chi tiết:**

* Private có thời lượng trung bình cao hơn (4.59s so với 3.53s) và độ lệch chuẩn lớn hơn ⇒ nhiều đoạn dài và không đồng đều.
* Năng lượng trung bình private gấp **2 lần** so với public, cho thấy **khả năng chứa nhiều nhiễu nền** hoặc **giọng nói lớn hơn**, gây khó cho mô hình.

# IV. Tiền xử lý và cài đặt mô hình (Preprocessing & Model Setup)

## 1. Tiền xử lý dữ liệu

* Với mô hình Whisper-small:
  + Dữ liệu âm thanh được đọc từ các file WAV và chuyển về dạng mảng sóng âm thanh.
  + Sử dụng WhisperFeatureExtractor để trích xuất đặc trưng log-Mel spectrogram với tần số mẫu 16 kHz.
  + Văn bản transcript được chuyển về chữ thường và mã hóa token thông qua WhisperTokenizer.
  + Dữ liệu được định dạng lại dưới dạng Dataset của Hugging Face để phục vụ cho huấn luyện.
* Với mô hình PhoWhisper-small:
  + Không thực hiện huấn luyện mà sử dụng pipeline transformers.pipeline để nhận dạng trực tiếp từ file âm thanh.
  + Văn bản đầu ra được chuẩn hóa về chữ thường và loại bỏ dấu chấm câu cuối câu để chuẩn hóa kết quả.

## 2. Cài đặt mô hình

* **Whisper-small**:
  + Tải mô hình và các bộ xử lý liên quan (feature extractor, tokenizer, processor) từ Hugging Face hub: openai/whisper-small.
  + Áp dụng kỹ thuật PEFT với LoRA (sử dụng thư viện peft) để fine-tuning mô hình trên dữ liệu tiếng Việt.
  + Cấu hình huấn luyện với Seq2SeqTrainer, thiết lập các tham số như batch size, learning rate, số epoch, fp16,…
  + Định nghĩa data collator tùy chỉnh để chuẩn bị batch dữ liệu với padding và mã hóa phù hợp cho mô hình.
* **PhoWhisper-small**:
  + Sử dụng mô hình tiền huấn luyện vinai/PhoWhisper-small dưới dạng pipeline ASR của thư viện transformers, phục vụ inference trực tiếp.
  + Triển khai mô hình trên GPU nếu có (device=0), hoặc CPU (device=-1) tùy môi trường.

# V. Thực nghiệm và Đánh giá (Experiment & Evaluation)

## 1. Môi trường thực nghiệm

* Mô hình được huấn luyện và đánh giá trên môi trường GPU với thư viện PyTorch và Transformers.
* Các tham số huấn luyện của Whisper-small: batch size 16, learning rate 5e-5, fine-tuning 3 epochs.
* PhoWhisper-small chỉ thực hiện inference trên tập dữ liệu test mà không huấn luyện lại.

## 2. Đánh giá hiệu suất

* Sử dụng chỉ số WER (Word Error Rate) làm thước đo chính để đánh giá chất lượng nhận dạng của mô hình trên tập kiểm thử VIVOS.
* Công thức tính WER là tỉ lệ sai lệch về từ giữa bản dịch tự động và bản transcript gốc.

## 3. Kết quả thực nghiệm

Kết quả đo lường bằng chỉ số WER (Word Error Rate) được tổng hợp như sau:

**Với mô hình Whisper-small:**

Ảnh có chứa hàng, Sơ đồ, văn bản, biểu đồ

Nội dung do AI tạo ra có thể không chính xác.

* Mô hình gốc Whisper-small chưa được fine-tune đạt WER 30.38% trên tập huấn luyện và 25.47% trên tập kiểm thử, cho thấy mô hình đã có khả năng tổng quát tốt dù chưa tinh chỉnh với tiếng Việt.
* Sau khi fine-tune mô hình PhoWhisper-small trong 2 epoch trên tập VLSP 2020, kết quả WER giảm rõ rệt còn 10.33% (train) và 17.06% (test). Điều này chứng minh rằng việc tinh chỉnh mô hình trên dữ liệu tiếng Việt giúp cải thiện đáng kể khả năng nhận dạng.
* Tuy nhiên, khi tiếp tục huấn luyện đến 4 epoch, mô hình có hiện tượng **overfitting** nghiêm trọng: chỉ số WER tăng rất cao, mô hình hoàn toàn dự đoán sai đầu ra, cho thấy hiệu suất suy giảm mạnh. Điều này cho thấy việc fine-tuning cần được dừng đúng lúc để tránh mô hình học quá mức đặc điểm nhiễu của dữ liệu huấn luyện.

**Kết quả:**

|  |  |  |
| --- | --- | --- |
| Mô hình | WER trên public test | WER trên private test |
| |  | | --- | | Whisper-small |  |  | | --- | |  | |  | 73.34% |
| |  | | --- | | PhoWhisper-small (không làm sạch) |  |  | | --- | |  | | 6.78% | 40.45% |
| PhoWhisper-small (có làm sạch) |  | 38.8% |

=> Kết quả cho thấy PhoWhisper-small đạt hiệu suất nhận dạng tốt hơn nhiều so với Whisper-small trên bộ dữ liệu kiểm thử public và test, với WER trên private test thấp hơn gần 1.8 lần. Quá trình làm sạch cơ bản giúp nâng cao chất lượng

# VI. Thảo luận kết quả (Discussion)

## Kết quả thực nghiệm đưa ra nhiều điểm cần lưu ý:

* Whisper-small cho kết quả kém hơn PhoWhisper-small trên bộ dữ liệu kiểm thử VIVOS:

Mô hình Whisper-small chưa được fine-tune chưa có kết quả trên tập public test, nhưng trên tập private test đạt WER rất cao là 73.34%. Trong khi đó, PhoWhisper-small sau fine-tune đạt WER rất tốt trên tập public test (6.78%) và thấp hơn đáng kể trên tập private test (40.45%). Điều này cho thấy PhoWhisper-small có khả năng nhận dạng tiếng Việt được cải thiện rõ rệt so với mô hình gốc Whisper-small.

**Nguyên nhân kết quả thấp có thể bao gồm:**

* *Overfitting:* PhoWhisper-small có thể bị overfitting trên tập huấn luyện VLSP 2020, vốn có điều kiện thu âm chuẩn hóa và giọng đọc rõ ràng, khác biệt nhiều so với dữ liệu hội thoại và tiếng ồn trong tập kiểm thử private test (VIVOS).
* *Thiếu đa dạng dữ liệu huấn luyện:* Bộ dữ liệu VLSP 2020 có thể chưa đủ phong phú về giọng vùng miền, tốc độ nói và mức độ nhiễu, khiến mô hình khó khái quát hóa khi áp dụng trên dữ liệu thực tế đa dạng hơn.
* *Hiệu ứng fine-tune quá mức:* Mặc dù PhoWhisper-small hoạt động rất tốt trên tập public test, chỉ số WER trên private test vẫn còn khá cao, thể hiện sự khác biệt lớn giữa các tập dữ liệu và có thể do fine-tuning chưa tối ưu hoặc dữ liệu test private khó hơn nhiều.

**Whisper-small thể hiện hạn chế về khả năng tổng quát:**

Do không được fine-tune riêng cho tiếng Việt, Whisper-small đạt WER rất cao trên bộ dữ liệu private test (73.34%), cho thấy hiệu suất nhận dạng chưa đáp ứng được yêu cầu thực tế trong môi trường tiếng Việt đa dạng và có nhiều nhiễu.

## Phân tích lỗi

Phân tích lỗi đầu ra chỉ ra các dạng lỗi phổ biến như sau:

* **Tần suất nhầm âm cao**, đặc biệt ở những âm dễ gây nhầm do vùng miền: “b” ↔ “v”, “n” ↔ “l”, “t” ↔ “th”.



* **Lỗi bỏ từ và nối từ không chính xác**: Khi gặp câu dài hoặc nói nhanh, mô hình dễ bỏ cụm từ hoặc ghép từ sai lệch hoàn toàn.
* **Không xử lý tốt tiếng ồn**: Gặp nhiều lỗi khi có tiếng ồn nền hoặc khi người nói ngắt nghỉ không đều.
* **Không tinh chỉnh kết quả**: Nhiều kết quả mô hình trả ra sai chính tả



Tổng thể, Whisper-small cho thấy hiệu năng tốt hơn không chỉ ở WER mà còn ở chất lượng đầu ra – ổn định hơn, ít lỗi nghiêm trọng, và có khả năng khái quát hóa cao hơn với tiếng nói thực tế.

# VII. Kết luận và Hướng phát triển (Conclusion & Future Work)

Trong báo cáo này, chúng em đã tiến hành huấn luyện và đánh giá hai mô hình Whisper-small và PhoWhisper-small trên dữ liệu tiếng Việt VLSP 2020, đồng thời đánh giá hiệu suất trên bộ kiểm thử VIVOS. Kết quả cho thấy mô hình PhoWhisper-small đạt WER thấp hơn nhiều trên tập public test (6.78%) so với Whisper-small chưa được fine-tune (chưa có số liệu trên public test), nhưng trên tập private test, PhoWhisper-small vẫn cho WER thấp hơn (40.45%) so với Whisper-small (73.34%).

Từ kết quả thực nghiệm, có thể rút ra một số kết luận chính:

* Mô hình PhoWhisper-small cho thấy khả năng nhận dạng tốt trên tập public test, tuy nhiên hiệu suất trên tập private test vẫn còn hạn chế do dữ liệu thực tế phức tạp, đa dạng hơn.
* Mô hình Whisper-small được fine-tune riêng cho tiếng Việt nhưng vẫn chưa thể hiện khả năng tổng quát nhất định trên tập private test, thông qua WER vẫn còn khá cao.
* Hiện tượng overfitting và sự khác biệt về đặc tính dữ liệu giữa tập huấn luyện và kiểm thử là những thách thức chính cần được giải quyết để nâng cao hiệu quả của các mô hình nhận dạng tiếng Việt.

**Hướng phát triển trong tương lai**:

* Kết hợp thêm nhiều nguồn dữ liệu tiếng Việt đa dạng, bao gồm giọng nói vùng miền và dữ liệu hội thoại thực tế.
* Tinh chỉnh lại PhoWhisper-small với tập dữ liệu phù hợp hơn và áp dụng các kỹ thuật chống overfitting như augmentation, egularization.
* Thử nghiệm thêm các mô hình lớn hơn như Whisper-medium hoặc Whisper-large để đánh giá khả năng cải thiện WER.

# VIII. Tài liệu tham khảo (References)

1. Hugging Face. **vinai/PhoWhisper-small**. https://huggingface.co/vinai/PhoWhisper-small
2. VLSP 2020. Vietnamese Large Speech Corpus. <https://vlsp.org.vn>
3. VIVOS Corpus. A Free Vietnamese Speech Dataset. https://ailab.hcmus.edu.vn/vivos
4. OpenAI Whisper Model. <https://github.com/openai/whisper>
5. Transformers Library. https://huggingface.co/docs/transformers