

**XÂY DỰNG CÔNG CỤ SPEECH TO TEXT**

**BẰNG WHISPER**

**MÔN HỌC: XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

**GIÁO VIÊN HƯỚNG DẪN:** PhD. Đinh Điền

**THÀNH VIÊN NHÓM:** 22C01016 - Tống Lương Hương Quỳnh

22C01026 - Nguyễn Ngọc Thảo Uyên



**MỤC LỤC**



[**1. Giới Thiệu 3**](#_9neim29297n1)

[**2. Tổng Quan về Whisper Audio To Text (WATT) 4**](#_2w0nrslev7ho)

[**3. Chi tiết về Whisper OpenAI 5**](#_y9kys125ruhc)

[3.1. Whisper OpenAI là gì 5](#_63d9i789u1s)

[3.2. Các tính năng chính của Whisper OpenAI 7](#_kej1zqdbbea5)

[3.3. Kiến trúc mô hình Whisper OpenAI 7](#_7brdcmd538j1)

[3.4. Ưu điểm và Nhược điểm của Whisper OpenAI 8](#_v2mkdf2lk63k)

[a. Ưu Điểm 8](#_vqj6mavr9ovz)

[b. Nhược Điểm 9](#_6cfew64ebpsy)

[3.5. Cách sử dụng Whisper OpenAI 10](#_pg390xwy8cxr)

[a. Sử dụng mã nguồn mở: 10](#_f7bug994jhtu)

[b. Sử dụng Whisper API: 10](#_ssqk5yzbgky9)

[**4. Ngữ liệu huấn luyện 10**](#_uh40uclmpig6)

[**5. Quá trình tinh chỉnh( fine tune) 12**](#_89e8n4nv9pxy)

[5.1. Các bước fine-tune 12](#_fsnmc1rqjzfw)

[5.2. Thông số 13](#_hhrvuofh6rpf)

[5.3. Tỉ lệ WER 13](#_no52124i4vb7)

[5.4. Kết quả fine-tune 14](#_qyj7yprovgw4)

[**6. Kiểm tra 15**](#_of0gl4g9egq9)

[6.1. Kết quả sử dụng Whisper API 15](#_ewen5iyqkgk4)

[a. Transcript audio Tiếng Anh 15](#_i8mt8hjm8gp)

[b. Transcribe audio Tiếng Việt 15](#_2cuxmktga7qv)

[c. Translate audio Tiếng Việt sang Tiếng Anh 15](#_gjxhd82fjhm5)

[6.2. Kết quả sử dụng mô hình fine-tune 15](#_245o3ncdgex6)

[**7. Những giới hạn của công cụ và hướng phát triển 16**](#_xfdpr0agfuy7)

[7.1. Những giới hạn của công cụ 16](#_evhsdhinpl1x)

[7.2. Hướng phát triển 16](#_ci9pr25k9ugf)

[**REFERENCES 17**](#_thulgcg40h53)

# 

# 

# **1. Giới Thiệu**

Trong thời đại số hóa hiện nay, sự tiến bộ trong lĩnh vực trí tuệ nhân tạo đã mở ra những cánh cửa mới cho nhiều công nghệ đột phá. Một trong những lĩnh vực nổi bật được đón nhận tích cực là chuyển đổi giọng nói thành văn bản - công nghệ Speech to Text. Speech to text có nhiều ứng dụng trong đời sống, hỗ trợ con người trong công việc ví dụ như chuyển đổi nội dung cuộc họp, tạo phụ đề cho video(youtube), chuyển nội dung podcast thành văn bản, phiên âm các cuộc gọi, hay tạo ra một vài ứng dụng hỗ trợ người kém hoặc không nghe được.

Để hỗ trợ các bài toán Speech to Text, hiện nay có nhiều công cụ đã được phát triển, có thể kể đến như Amazon Transcribe - một dịch vụ nhận dạng giọng nói tự động của Amazon Web Services (AWS), Microsoft Azure Speech to Text - một tính năng của dịch vụ Speech trong Azure Cognitive Services, Google Cloud Speech-to-Text - một dịch vụ nhận dạng giọng nói của Google Cloud, IBM Watson Speech to Text - một dịch vụ của IBM Watson AI Platform. Trong báo cáo này, nhóm xây dựng công cụ Speech to text dựa trên Whisper, một công cụ phổ biến và mạnh mẽ.

Whisper, một hệ thống phát triển bởi OpenAI, là một mô hình nhận dạng giọng nói tiên tiến dựa trên công nghệ học sâu (Deep Learning). Sự đột phá của Whisper nằm ở khả năng nắm bắt và hiểu rõ ngữ cảnh trong những đoạn hội thoại phức tạp. Được đào tạo trên một lượng dữ liệu âm thanh khổng lồ, Whisper đã tiến xa hơn các công cụ tương tự trước đây và mang lại chất lượng chuyển đổi từ giọng nói thành văn bản đáng kinh ngạc.

Với sự phát triển nhanh chóng của Whisper, việc xây dựng công cụ Speech to Text trở nên dễ dàng hơn bao giờ hết. Được hỗ trợ bởi Whisper, công cụ này có thể chuyển đổi âm thanh từ nguồn ghi âm thành văn bản một cách nhanh chóng và chính xác. Điều đáng chú ý là Whisper không chỉ cung cấp cho chúng ta một công cụ chuyển đổi đơn thuần. Mô hình này còn có khả năng tiếp tục học tập và cải thiện qua thời gian. Nhờ vào việc kết hợp các phương pháp học máy tiên tiến, Whisper có thể hiểu và thích nghi với các ngữ cảnh ngôn ngữ mới, từ đó cung cấp những kết quả tốt hơn theo thời gian. Điều này cho phép Whisper trở thành một công cụ ngôn ngữ linh hoạt và đáng tin cậy cho các ứng dụng Speech to Text. Sự phát triển tiếp tục của Whisper hứa hẹn mở ra nhiều cánh cửa mới cho ứng dụng của công nghệ này trong tương lai gần.

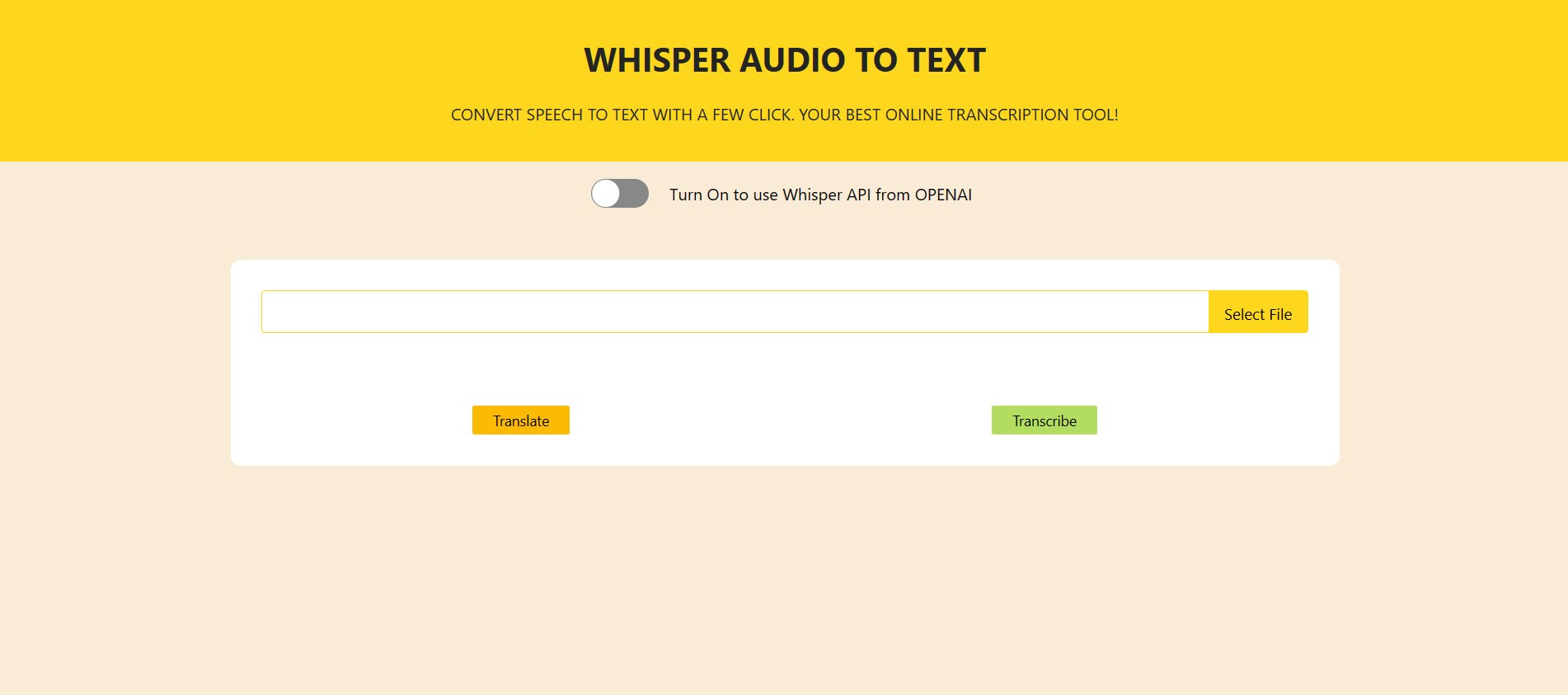
# **2. Tổng Quan về Whisper Audio To Text (WATT)**

Trong báo cáo này, nhóm xây dựng Công cụ Web - Whisper Audio To Text (WATT), công cụ được xây dựng để sử dụng mô hình Whisper của OpenAI nhằm chuyển đổi giọng nói thành văn bản. Người dùng có thể tải lên một file audio lên Công cụ Web. Sau đó, Whisper sẽ sử dụng mô hình đã huấn luyện để nhận dạng giọng nói trong file audio và chuyển đổi thành văn bản. Kết quả được hiển thị lên giao diện để người dùng xem.

Công cụ WATT cung cấp hai chức năng chính:

1. Chức năng Transcribe: Phiên âm đoạn audio
2. Chức năng Translate: Người dùng có thể chọn chức năng dịch văn bản. Kết quả được hiển thị lên giao diện là văn bản bằng Tiếng Anh của đoạn audio.

Giao diện của Web sẽ hiển thị như hình bên dưới:



Ở đây người dùng có thể lựa chọn sử dụng model của OpenAI hoặc sử dụng phiên bản Whisper API.

# **3. Chi tiết về Whisper OpenAI**

## **3.1. Whisper OpenAI là gì**

Whisper được phát triển bởi OpenAI, công ty của nhiều ứng dụng công nghệ nổi bật hiện nay như Chat GPT hay DALL-E . Nó là hệ thống nhận dạng giọng nói tự động (ASR) được đào tạo dựa trên 680.000 giờ dữ liệu được giám sát đa ngôn ngữ và đa tác vụ được thu thập từ web.

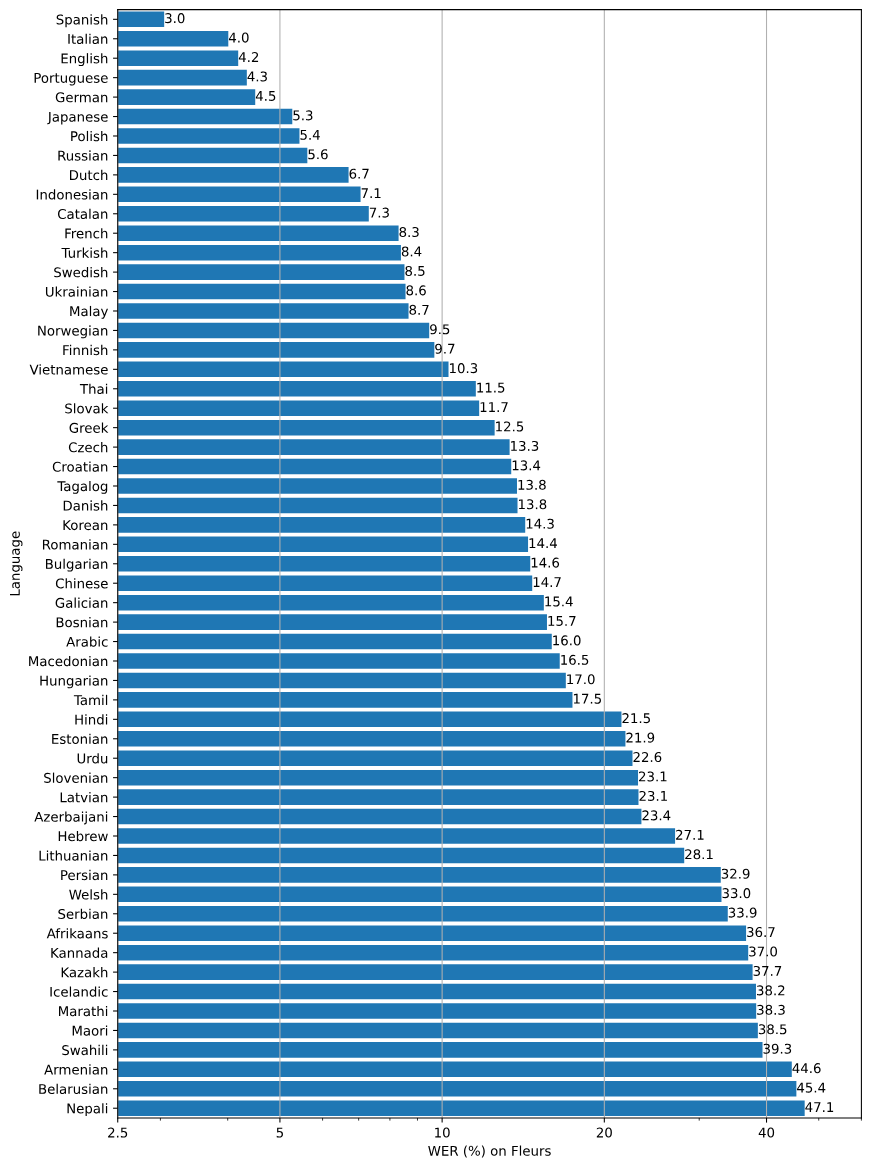
Whisper sử dụng một tập dữ liệu lớn và đa dạng giúp cải thiện độ chắc chắn cho các dấu trọng âm, tiếng ồn xung quanh và ngôn ngữ kỹ thuật. Hơn nữa, Whisper cho phép phiên âm bằng nhiều ngôn ngữ, cũng như dịch từ các ngôn ngữ đó sang tiếng Anh. Whisper là mô hình nguồn mở và mã logic làm nền tảng cho việc xây dựng các ứng dụng hữu ích và cho nghiên cứu sâu hơn về xử lý giọng nói mạnh mẽ.

Whisper đi kèm với nhiều mô hình cơ sở dữ liệu khác nhau. Có năm kích thước mô hình khác nhau, mang lại sự cân đối giữa tốc độ và độ chính xác. Các mô hình Whisper thể hiện trong bảng dưới đây:

| **Size** | **Parameters** | **English-only model** | **Multilingual model** | **Required VRAM** | **Relative speed** |
| --- | --- | --- | --- | --- | --- |
| tiny | 39 M | tiny.en | tiny | ~1 GB | ~12x |
| base | 74 M | base.en | base | ~1 GB | ~16x |
| small | 244 M | small.en | small | ~2 GB | ~6x |
| medium | 769 M | medium.en | medium | ~5 GB | ~2x |
| large | 1550 M | N/A | large | ~10 GB | 1x |

5 kích thước mô hình của Whisper từ nhỏ nhất đến lớn nhất tương ứng là tiny, base, small, medium và large. ‘Tiny’ là mô hình có kích thước nhỏ nhất với yêu cầu bộ nhớ ít nhất nên có tốc độ xử lý nhanh nhất, tuy nhiên kết quả khi dùng mô hình tiny lại kém nhất trong số các mô hình. Ngược lại ‘large’ là mô hình phức tạp nhất, đồng thời cũng có tốc độ xử lý chậm nhất, nhưng lại cho ra kết quả tốt nhất. Khi người dùng sử dụng Whisper, cần cân nhắc giữa hai yếu tố tốc độ xử lý và độ chính xác của mô hình.

Tỉ lệ nhận diện sai của whisper phụ thuộc vào loại ngôn ngữ, bảng dưới đây thể hiện thống kê tỷ lệ nhận diện sai của từng loại ngôn ngữ khi sử dụng mô hình lớn nhất (large) của Whisper. Trong đó, Tiếng Việt có tỉ lệ nhận diện sai là 10.3, Tiếng Anh là 4.2. Tiếng Tây Ban Nha có tỉ lệ nhận diện sai thấp nhất là 3.0.



Hiện nay Whisper đã cho ra mắt phiên bản 2 của mô hình ‘large’ là ‘large-v2’, được mô tả có hiệu suất và tốc độ xử lý tốt hơn nhiều so với mô hình ‘large’.

## **3.2. Các tính năng chính của Whisper OpenAI**

Whisper là một công cụ chuyển đổi giọng nói thành văn bản do OpenAI phát triển. Dưới đây là một số tính năng chính của công cụ Whisper:

* **Chuyển đổi giọng nói thành văn bản:** Whisper có khả năng chuyển đổi âm thanh từ các tệp âm thanh hoặc luồng giọng nói thành văn bản tương ứng. Điều này rất hữu ích khi bạn muốn chuyển đổi nhanh chóng các tệp âm thanh hoặc khi bạn muốn sử dụng giọng nói để nhập liệu vào các ứng dụng hoặc hệ thống.
* **Hỗ trợ nhiều ngôn ngữ:** Whisper hỗ trợ nhiều ngôn ngữ khác nhau, bao gồm tiếng Anh, tiếng Tây Ban Nha, tiếng Pháp, tiếng Đức, tiếng Ý, tiếng Bồ Đào Nha và nhiều ngôn ngữ khác. Điều này cho phép bạn chuyển đổi giọng nói từ các ngôn ngữ khác nhau thành văn bản.
* **Hỗ trợ ngữ cảnh:** Whisper có khả năng hiểu và xử lý ngữ cảnh của giọng nói. Điều này có nghĩa là nó có thể giải quyết các câu truy vấn phức tạp, sử dụng thông tin từ các câu trước đó để hiểu ý nghĩa chung và chuyển đổi thành văn bản phù hợp.
* **Hỗ trợ ứng dụng thời gian thực:** Whisper có thể được tích hợp vào các ứng dụng và hệ thống thời gian thực để cung cấp chuyển đổi giọng nói thành văn bản ngay lập tức. Điều này hữu ích trong các trường hợp như chatbot trò chuyện trực tiếp, ghi chú giọng nói và nhiều ứng dụng khác đòi hỏi sự tương tác ngay lập tức.
* **Điều chỉnh đầu vào và đầu ra:** Whisper cho phép bạn tùy chỉnh các thông số đầu vào và đầu ra để đáp ứng các yêu cầu cụ thể. Bạn có thể thiết lập các thông số như định dạng âm thanh đầu vào (wav, mp3, ogg), tốc độ mẫu âm thanh, độ nhạy micro, và các thông số khác để tối ưu hóa quá trình chuyển đổi giọng nói thành văn bản.
* **Hỗ trợ API:** Whisper cung cấp API để tích hợp vào các ứng dụng và dịch vụ của bạn. Điều này cho phép bạn tận dụng khả năng chuyển đổi giọng nói thành văn bản của Whisper trong các ứng dụng của riêng bạn.

## **3.3. Kiến trúc mô hình Whisper OpenAI**

Whisper hỗ trợ nhiều ngôn ngữ khác nhau, tuy nhiên độ chính xác cao nhất ở ngôn ngữ tiếng Anh. Mô hình hoạt động dựa theo trình tự của *Transformer*. Kiến trúc của Whisper *bao gồm nhiều lớp encoder và decoder*, cho phép mô hình biểu diễn thông tin từ âm thanh đầu vào và suy luận ra văn bản tương ứng.



* Đầu tiên, âm thanh đầu vào được chia thành các đoạn 30 giây, được chuyển đổi thành phổ log-Mel. *(Biểu đồ phổ log-Mel (log-Mel spectrogram) là một biểu đồ được sử dụng để mô tả phổ âm thanh của một tín hiệu âm thanh. Nó được sử dụng phổ biến trong xử lý âm thanh và học máy âm thanh để biểu diễn thông tin tần số và thời gian của âm thanh)*
* Tiếp theo dữ liệu âm thanh sẽ đi qua 2 lớp Conv1D với hàm kích hoạt GELU (*Gaussian Error Linear Units*) để chuẩn hóa và trích xuất đặc trưng.
* Để bổ sung thông tin vị trí cho dữ liệu, mô hình sử dụng mã hóa vị trí hình sin (*Sinusoidal Positional Encoding* ). Điều này giúp mô hình nhận biết vị trí tương đối của các phần tử trong câu hoặc đoạn văn bản.
* Tiếp theo dữ liệu sẽ đưa vào khối mã hóa transformer (chính là các *Encoder blocks*). Ở đây mô hình sẽ học và trích xuất thông tin ngữ nghĩa từ của dữ liệu đầu vào.
* Sau khi đi qua các encoder blocks, đầu ra của encoder block trên cùng được đưa qua 1 lớp cross attention để mô hình tập trung vào các phần quan trọng trong thông tin đầu vào.
* Đầu ra của cross attention được đưa vào các *decoder blocks* (bộ giải mã Transformer) , tại đây mô hình sẽ tạo ra thông tin đầu ra. Đặc biệt các decoder blocks có sử dụng 1 *learned positional encoding* để mã hóa thông tin vị trí cho các phần tử trong câu.

Whisper được đào tạo về các tác 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 tác vụ này được biểu diễn chung dưới dạng một chuỗi mã thông báo được bộ giải mã dự đoán, cho phép một mô hình duy nhất thay thế nhiều giai đoạn khác nhau của quy trình xử lý giọng nói truyền thống. Định dạng đào tạo đa nhiệm sử dụng một tập hợp các mã thông báo đặc biệt đóng vai trò là công cụ xác định nhiệm vụ hoặc mục tiêu phân loại.

Để mô hình có thể thực hiện nhiều nhiệm vụ từ 1 đầu vào âm thanh duy nhất, cần có 1 định dạng đơn giản để xác định nhiệm vụ và thông tin điều kiện dưới dạng *một chuỗi các mã thông báo đầu vào* cho bộ giải mã. Quy trình là:

* Đầu tiên bộ giải mã nhận 1 đoạn âm thanh làm đầu vào để xử lý.
* Sau khi có dữ liệu âm thanh, bộ giải mã dự đoán ngôn ngữ được nói trong đoạn âm thanh bằng cách sử dụng 1 mã thông báo duy nhất. Trong trường hợp không có âm thanh thì mô hình đưa ra mã thông báo là ‘SILENCE’.
* Tiếp theo bộ giải mã dự đoán nhiệm vụ cụ thể được thực hiện trên đoạn âm thanh qua mã thông báo là ‘TRANSCRIPTION’ hay ‘TRANSLATION’.
* Tiếp theo bộ giải mã sẽ quyết định có dự đoán thông tin thời gian( timestamps) không bằng cách sử dụng mã ‘TIMESTAMP’. TIMESTAMP được sử dụng để để đánh dấu thời gian bắt đầu và kết thúc của các phần trong đoạn âm thanh, giúp tạo ra một ghi chú hoặc trình bày văn bản 1 cách đồng bộ với âm thanh.
* Cuối cùng, bộ giải mã sẽ bắt đầu dự đoán nội dung từng phần của đoạn âm thanh, bao gồm các đoạn văn bản, và nếu được yêu cầu cũng dự đoán thông tin thời gian (timestamps) tương ứng với phần đó. Khi đoạn âm thanh kết thúc hoặc các điều kiện kết thúc khác được đáp ứng, bộ giải mã đưa ra mã thông báo ‘END’ để kết thúc dự đoán.

## **3.4. Ưu điểm và Nhược điểm của Whisper OpenAI**

### **Ưu Điểm**

* **Độ chính xác cao:** Whisper được đào tạo trên một lượng lớn dữ liệu giọng nói, cho phép nó chuyển đổi giọng nói thành văn bản với độ chính xác cao. Điều này giúp cải thiện trải nghiệm người dùng và tăng tính khả dụng của công cụ.
* **Hỗ trợ nhiều ngôn ngữ:** Whisper hỗ trợ nhiều ngôn ngữ khác nhau, cho phép người dùng chuyển đổi giọng nói từ các ngôn ngữ khác nhau thành văn bản. Điều này làm cho công cụ phù hợp cho một đối tượng người dùng toàn cầu.
* **Hỗ trợ ngữ cảnh:** Whisper có khả năng hiểu và xử lý ngữ cảnh của giọng nói, giúp nâng cao độ chính xác và sự hiểu biết của kết quả chuyển đổi. Nó có thể hiểu ý nghĩa chung và sử dụng thông tin từ các câu trước đó để đưa ra văn bản phù hợp.
* **Hỗ trợ API:** Whisper cung cấp API cho phép nhà phát triển tích hợp công cụ vào các ứng dụng và dịch vụ của mình. Điều này mở ra nhiều khả năng ứng dụng và tùy chỉnh cho người dùng.

### **Nhược Điểm**

* **Hạn chế về độ chính xác:** Mặc dù Whisper có độ chính xác cao, nhưng vẫn có thể gặp phải sai sót trong quá trình chuyển đổi giọng nói thành văn bản. Độ chính xác có thể bị ảnh hưởng bởi chất lượng âm thanh đầu vào và ngữ cảnh của ngôn ngữ.
* **Hạn chế về ngôn ngữ và lĩnh vực:** Dù đã hỗ trợ nhiều ngôn ngữ, Whisper vẫn có thể gặp khó khăn khi xử lý các ngôn ngữ đặc thù hoặc chuyên ngành. Nếu ngôn ngữ hoặc lĩnh vực của bạn không được hỗ trợ tốt, độ chính xác có thể giảm đi.
* **Yêu cầu dữ liệu huấn luyện lớn:** Whisper đòi hỏi một lượng lớn dữ liệu giọng nói để huấn luyện và đạt được độ chính xác cao. Việc thu thập và xử lý dữ liệu này có thể tốn thời gian và tài nguyên.
* **Giới hạn về khả năng tương tác:** Whisper tập trung vào một nhiệm vụ chuyển đổi giọng nói thành văn bản, nghĩa là nó không cung cấp khả năng tương tác phức tạp như nhận diện ngôn ngữ tự nhiên hoặc xử lý yêu cầu phức tạp hoặc yêu cầu tương tác nâng cao.
* **Phụ thuộc vào chất lượng âm thanh đầu vào:** Whisper đòi hỏi chất lượng âm thanh đầu vào tốt để đạt được độ chính xác tốt nhất. Nếu âm thanh đầu vào không rõ ràng, bị nhiễu hay chất lượng thấp, độ chính xác của kết quả chuyển đổi có thể bị ảnh hưởng.
* **Yêu cầu kiến ​​thức kỹ thuật để vận hành và chi phí cao nếu sử dụng nhiều.**

***Lưu ý rằng*** các ưu điểm và nhược điểm này có thể thay đổi theo thời gian và với sự phát triển của công nghệ. Điều quan trọng là cân nhắc những yêu cầu và mục tiêu cụ thể của bạn khi sử dụng công cụ Whisper để xác định xem nó phù hợp cho nhu cầu của bạn hay không.

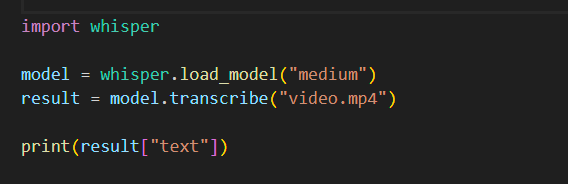
## **3.5. Cách sử dụng Whisper OpenAI**

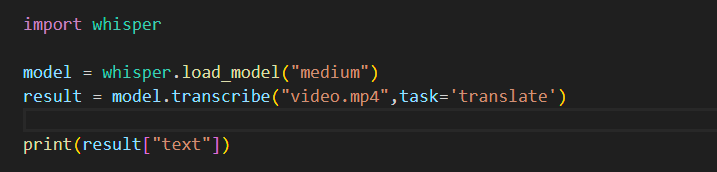
Hiện nay, người dùng có thể sử dụng Whisper theo 2 cách: sử dụng như một mã nguồn mở, và sử dụng phiên bản Whisper API.

### **Sử dụng mã nguồn mở**

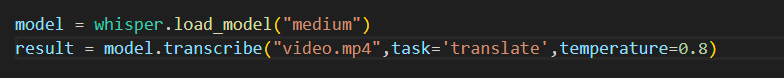
Người dùng có thể tải mã nguồn mở của Whisper trên trang GITHUB của OpenAI. Khi sử dụng Whisper, người dùng sẽ lựa chọn các yếu tố sau:

* model size: có tổng cộng 9 mô hình Whisper, tương đương 5 size là tiny, base, small, medium và large
* task: có 2 task là ‘transcript’ và ‘translate’





Ngoài ra có thể kể đến parameter ‘temperature’. Temperature sẽ quyết định độ ngẫu nhiên trong kết quả trả ra.

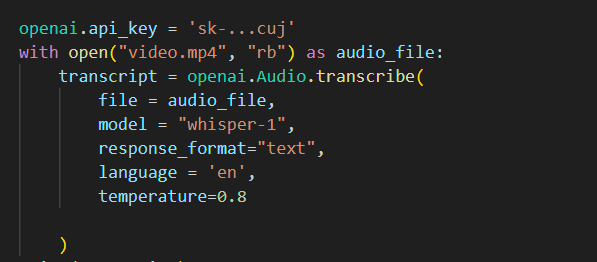


Cùng một audio, khi thay đổi temperature thì kết quả sẽ cho ra khác nhau. Thực hiện translate một audio Tiếng Việt sang Tiếng Anh với temperature là 0 và 0.8 thì kết quả như sau:

* Với temperature = 0: The attack was carried out by the Sia Tiu and Ia Kato police in Dak Lak. The plane crash was reported to have killed two people. The Vietnamese owner was shot to death in Taiwan. All information will be updated in the hot news article on November 11.
* Với temperature = 0.8: The attack was carried out on the safety of Ia Tiu, Taekwondo and Ia Kato in Daklak province, when the vehicle on board crashed into the truck, causing the number of casualties, including two deaths. The owner and Vietnamese were shot dead in Taiwan. More information will be updated in the 30 second hot news article on the 11th.

### **Sử dụng Whisper API**

Whisper API sử dụng mô hình ‘large-v2’, phiên bản 2 của mô hình ‘large’ của Whisper. Whisper API thích hợp để phát triển các ứng dụng với độ chính xác cao hơn và thời gian tính toán nhanh hơn nhiều so với các mô hình khác. Theo mặc định, API Whisper chỉ hỗ trợ các tệp nhỏ hơn 25 MB, vì vậy đối với các tệp audio có kích thước lớn hơn 25 MB, người dùng có thể sử dụng gói Python mã nguồn mở PyDub để tách âm thanh ra từng đoạn có kích thước nhỏ hơn 25 MB, rồi sau đó thực hiện việc phiên âm hay phiên dịch theo yêu cầu.



# **4. Ngữ liệu huấn luyện**

Trong báo cáo này, nhóm sử dụng bộ dữ liệu VIVOS để tiến hành fine-tune mô hình trên tiếng Việt. VIVOS là một tập dữ liệu tiếng Việt miễn phí, bao gồm 15 giờ ghi âm tiếng nói được chuẩn bị cho nhiệm vụ Nhận dạng Tiếng Nói Tự Động tiếng Việt. Tập dữ liệu này được chuẩn bị bởi AILAB, một phòng thí nghiệm khoa học máy tính thuộc Đại học Khoa học Tự Nhiên - Đại học Quốc gia TP.HCM.



VIVOS bao gồm 12420 dòng dữ liệu, bao gồm các trường dữ liệu:

* **speaker\_id**: id cho người đã ghi âm
* **path**: đường dẫn của file audio
* **audio**: có dạng dictionary chứa đường dẫn đến tệp âm thanh đã tải xuống, mảng âm thanh đã giải mã và tốc độ lấy mẫu. Lưu ý rằng khi truy cập cột âm thanh: dataset[0]["audio"], tệp âm thanh sẽ tự động được giải mã và lấy mẫu lại thành dataset.features["audio"].sampling\_rate. Việc giải mã và lấy mẫu lại một số lượng lớn tệp âm thanh có thể mất một khoảng thời gian đáng kể. Do đó, điều quan trọng trước tiên là truy vấn chỉ mục mẫu trước cột "audio", tức là tập dữ liệu[0]["audio"] phải luôn được ưu tiên hơn tập dữ liệu["audio"][0].
* **sentence:** câu người dùng nói

Bộ dữ liệu chia làm 2 tập train và test, được ghi lại thông tin dưới bảng sau:

|  | Train | Test |
| --- | --- | --- |
| Speakers | 46 | 19 |
| Utterances | 11660 | 760 |
| Duration | 14:55 | 00:45 |
| Unique Syllables | 4617 | 1692 |

# **5. Quá trình tinh chỉnh( fine tune)**

Mặc dù Whisper có hỗ trợ dịch giọng nói cho Tiếng Việt, tuy nhiên tỷ lệ nhận diện sai vẫn còn khá cao, đặc biệt đối với tiếng nói mang âm sắc tiếng địa phương. Vì vậy để mô hình đạt độ chính xác cao hơn đối với âm thanh là ngôn ngữ Tiếng Việt, nhóm thực hiện fine tune trên bộ dữ liệu VIVOS. Ở đây nhóm sẽ thực hiện fine tune Whisper trên mô hình ‘base’, chạy trên GPU của Google Collab.

## **5.1. Các bước fine-tune**

Nhóm thực hiện fine-tune Whisper cho bộ dữ liệu ASR, cụ thể là VIVOS, bằng Hugging Face Transformers. Để thực hiện fine-tune, nhóm đã thực hiện các bước sau đây:

1. Tải dữ liệu từ Hugging Face.
2. Thực hiện Feature Extraction, Tokenizer dữ liệu..
3. Huấn luyện:

* Xác định bộ đối chiếu dữ liệu( data collator): bộ đối chiếu dữ liệu lấy dữ liệu đã được xử lý trước của chúng tôi và chuẩn bị sẵn các tensor PyTorch cho mô hình.
* Số liệu đánh giá: trong quá trình đánh giá, chúng tôi muốn đánh giá mô hình bằng cách sử dụng số liệu tỷ lệ lỗi từ (WER). Chúng ta cần định nghĩa một hàm compute\_metrics để xử lý việc tính toán này.
* Tải pre-trained checkpoint : chúng tôi cần tải pre-trained checkpoint và định cấu hình chính xác cho việc đào tạo.
* Xác định cấu hình đào tạo( training configuration): điều này sẽ được Hugging Face huấn luyện và sử dụng để xác định lịch trình đào tạo.
* Thực hiện quá trình huấn luyện.

## **5.2. Thông số**

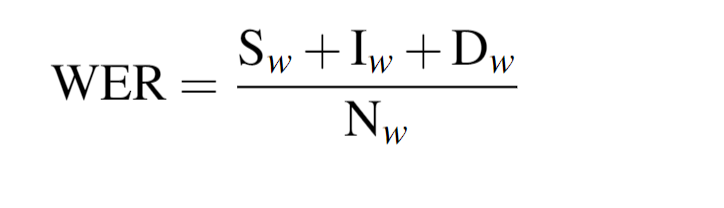
Tại bước định nghĩa Training Configuration, nhóm đã định nghĩa các thông số sau:

* **learning rate:** 1e-5
* **max\_steps**( số step huấn luyện dữ liệu): 2000
* **warmup\_steps**(số lượng bước huấn luyện ban đầu mà learning rate sẽ tăng dần từ 0 lên giá trị tối đa): 200
* **per\_device\_eval\_batch\_size**( batch size của một lần đánh giá): 8
* **per\_device\_train\_batch\_size**( batch size của mỗi bước huấn luyện): 8
* **save\_steps**( số bước huấn luyện cần thực hiện trước khi lưu lại trạng thái): 1000,
* **eval\_steps**( số lượng bước thực hiện trước khi thực hiện đánh giá): 200

Tất cả các thông số này đóng vai trò quan trọng trong việc định hình quá trình huấn luyện của mô hình và ảnh hưởng đến hiệu suất và tốc độ hội tụ của mô hình.

## **5.3. Tỉ lệ WER**

WER( Word Error Rate) là tỉ lệ lỗi từ, thường được áp dụng cho việc phiên âm các đoạn văn và câu chứa các từ có nghĩa. WER hoạt động ở cấp độ từ. WER tính tỉ lệ số từ sự khác biệt giữa transcript và prediction. Công thức tính WER như sau:



Trong đó:

* Sw: số từ bị thay thế
* Iw: số từ được cho thêm vào
* Dw: số từ bị xóa
* Nw: Tổng từ trong transcript

## **5.4. Kết quả fine-tune**

Dưới đây là kết quả quá trình fine-tune Whisper trên bộ dữ liệu VIVOS khi sử dụng mô hình ‘base’ của Whisper với thông số được đề cập phía trên:



Có thể thấy WER của mô hình đã được cải thiện đáng kể sau quá trình fine-tune, từ khoảng 72 tại bước huấn luyện số 200 xuống còn khoảng 29 sau quá trình huấn luyện. Quá trình và kết quả fine-tune sẽ được lưu dưới dạng file .ipynb đính kèm.

# **6. Kiểm tra**

Để hiển thị kết quả, nhóm sẽ thực hiện bôi đỏ tại những đoạn text sai so với audio. Những kết quả dưới đây dựa vào quá trình thực nghiệm, kiểm tra của nhóm.

## **6.1. Kết quả sử dụng Whisper API**

### **Transcript audio Tiếng Anh**

Hello. This is 6 Minute English from BBC Learning English. I'm Neil. And I'm Rob. Now, I'm sure most of us have interacted with a chatbot. These are bits of computer technology that respond to text with text, or respond to your voice. You ask it a question and usually it comes up right.

### **Transcribe audio Tiếng Việt**

Diễn biến vụ tấn công trụ sở công an xã Ia Tiêu và Ia Kato tại Đắk Lắk. Rít súng camera, vụ xe máy phóng như tên bay gây tai nạn 2 người chết. Chủ quán chóng người Việt ở Đài Loan bị bắn chết. Tất cả thông tin sẽ được cập nhật trong bản tin 30Giây Nóng tối 11-6.

### **Translate audio Tiếng Việt sang Tiếng Anh**

The incident took place at the Ia Tieu and Ia Kato police station in Daklak. The camera was shot down, the motorbike was shot like a flying arrow, causing the death of two people. The owner of the Vietnamese restaurant in Taiwan was shot to death. All information will be updated in the 30s news at 11pm June.

## **6.2. Kết quả sử dụng mô hình fine-tune**

Đối với mô hình fine-tune sẽ chỉ kiểm tra chức năng phiên âm cho audio Tiếng Việt. Dưới đây là kết quả

DÌN BÍNH PHỤ THỨNG CÔNG TRỘ SỰ CÔNG AN SẠI YÊN TÊN PHẢI YÊN CÁC TUÔN TỀ ĐẠT LẮT TRÍCH XUỐNG CAMAI RA VỤ XE MẤY PHÓNG DƯƠI TÊN BAY GÌ TÀI ĐẠN HÀNG GÙI CHẾT CHỐ QUAN CHÓNG NGƯỜI VIỆC Ở ĐẠI LOAN BỊ BẮNG CHẾT TỚI CA THÔNG TÊN SỰ CẬM NHỮC CHO BẠN TÊN BAN MƯỜI GÌ ĐỘNG TỐI MỪA MỘT THA SỐNG

***Nhận xét:*** *Mặc dù fine-tune trên bộ dữ liệu VIVOS thì tỉ lệ nhận diện sai WER có giảm, tuy nhiên khi tiến hành kiểm tra trên một đoạn audio thì có thể thấy kết quả trả về không tốt. Có thể do bộ dữ liệu chưa đủ, cũng như thời gian huấn luyện hay lựa chọn mô hình huấn luyện nên dẫn đến kết quả không tốt*.

**6.3. So sánh tỉ lệ WER giữa các model size của Whisper**

Ở đây nhóm sẽ thực hiện chức năng transcribe, so sánh 5 size model của Whisper trên 3 audio khác nhau: audio tiếng Anh, audio tiếng Việt và audio tiếng Việt có nhiễu( tại 1 cuộc họp). Kết quả so sánh được ghi lại như sau:

|  | tiếng Việt | tiếng Anh | tiếng Việt  có nhiễu |
| --- | --- | --- | --- |
| tiny | 0,8676 | 0,0185 | 1 |
| base | 0,6296 | 0 | 0,9414 |
| small | 0,2647 | 0,0370 | 0,8031 |
| medium | 0,1296 | 0,0294 | 0,4095 |
| large | 0,0555 | 0,0294 | 0,1329 |

Từ kết quả trên có thể nhận xét:

* Đối với những audio có nhiễu thì tỉ lệ nhận diện sai WER cao hơn những audio không có nhiễu.
* Đối với audio tiếng Anh, không có sự chênh lệch nhiều về kết quả, chọn model size phù hợp, không nên chọn model có kích thước quá lớn.
* Tiếng Việt tốt hơn hẳn khi sử dụng model size là medium và large.

# **7. Những giới hạn của công cụ và hướng phát triển**

Công cụ web chuyển giọng nói thành văn bản đang ngày càng phát triển và đóng vai trò quan trọng trong việc chuyển đổi giọng nói thành văn bản một cách hiệu quả. Tuy nhiên, vẫn còn một số hạn chế cần được cải thiện, và những hướng phát triển trong tương lai sẽ giúp nâng cao hiệu suất và khả năng sử dụng của công cụ này.

## **7.1. Những giới hạn của công cụ**

Công cụ vẫn chưa đạt được độ chính xác tuyệt đối, đặc biệt đối với các giọng nói phức tạp hoặc phát âm không rõ ràng. Điều này có thể gây ra những sai sót nhỏ trong quá trình chuyển đổi và dẫn đến hiểu lầm thông tin. Công cụ chỉ hỗ trợ một số ngôn ngữ chính, điều này giới hạn khả năng sử dụng của người dùng có nhu cầu với ngôn ngữ khác. Ngoài ra, khi phải xử lý các đoạn giọng nói dài và phức tạp có thể làm giảm hiệu suất và độ trễ trong việc hiển thị kết quả, gây phiền toái cho người dùng.

## **7.2. Hướng phát triển**

Trong tương lai, nhóm muốn thực hiện huấn luyện mô hình phiên âm Tiếng Việt tốt hơn. Để đạt được mong muốn này, cần nghiên cứu thêm các bộ dữ liệu Tiếng Việt khác, đồng thời nghiên cứu việc kết hợp Whisper với việc sử dụng các kiến trúc Deep Learning để đạt được hiệu suất cao cũng như tối ưu chi phí về mặt thời gian hay tính toán.

# **REFERENCES**

[1] “Hướng dẫn sử dụng công cụ speech-to-text Whisper OpenAI - AI ChatGPT.” *AI ChatGPT vn - AIChatGPT.vn*, 26 May 2023, https://aichatgpt.vn/huong-dan-su-dung-cong-cu-speech-to-text-whisper-openai/.

[2] OpenAI. “OpenAI Official.” *Introducing Whisper*, 22 June 2023, https://openai.com/research/whisper.