THUYẾT MINH SẢN PHẨM SÁNG TẠO

ĐĂNG KÝ THÔNG TIN SẢN PHẨM DỰ THI

1	Đăng ký Bảng dự thi	Bảng D1	Bảng D2	Bảng D3
		☐	☐	☒
2	Sản phẩm dự thi	Phần mềm	SP tích hợp phần cứng	
		☒	☐	
3	Tên SPST dự thi	SignMeet – Hệ thống hỗ trợ liên lạc dành cho người khiếm thính		
4	Ngôn ngữ lập trình hoặc nền tảng	Python, JavaScript		
5	Cấu hình cài đặt	Các trình duyệt trên máy tính sử dụng nhân Chromium		

THÔNG TIN TÁC GIẢ (NHÓM TÁC GIẢ)

Số lượng thí sinh trong đội thi	1 người ☐	2 người ☒
Thí sinh thứ nhất (đội trưởng)		
Họ và tên:	Lê Quang Phúc	
Ngày/tháng/năm sinh:	08/05/2006	
Lớp, trường:	12A5 – Trường THPT chuyên Lê Quý Đôn	
Quận, huyện:	Sơn Trà, Đà Nẵng	
Điện thoại:	0935110820	
Email:	phuclequang2006@gmail.com	
Thí sinh thứ hai		
Họ và tên:	Huỳnh Tấn Phúc	
Ngày/tháng/năm sinh:	12/12/2006	
Lớp, trường:	12A5 – Trường THPT chuyên Lê Quý Đôn	
Quận, huyện:	Sơn Trà, Đà Nẵng	
Điện thoại:	0794395894	
Email:	htanphuc1212@gmail.com	
Giáo viên hoặc chuyên gia hướng dẫn		
Họ và tên:	Phạm Thị Thu Hà	
Đơn vị công tác:	Trường THPT chuyên Lê Quý Đôn, Sơn Trà, Đà Nẵng	
Chức vụ:	Giáo viên	
Điện thoại:	0765577094	
Email:	pttha195@gmail.com	

GIỚI THIỆU VỀ SẢN PHẨM

1. Ý tưởng của sản phẩm

Ngôn ngữ ký hiệu không chỉ là ngôn ngữ chính thức của cộng đồng người khiếm thính [1], mà còn là cầu nối trực tiếp giữa họ với xã hội nói chung. Việc sử dụng ngôn ngữ ký hiệu để truyền tải thông điệp giúp người khiếm thính hoà nhập với xã hội. Vì vậy, ngôn ngữ ký hiệu đóng vai trò quan trọng đối với cộng đồng người khiếm thính.

Tuy nhiên, số lượng người phiên dịch ngôn ngữ ký hiệu chuyên nghiệp hiện nay còn quá ít, chỉ chiếm khoảng 0,0004% so với số người sử dụng ngôn ngữ ký hiệu [2], khiến người khiếm thính gặp khó khăn khi thông hiểu các văn bản sử dụng văn phạm tiếng Việt. Hơn nữa, đặc trưng sử dụng ngôn ngữ hình thể khi giao tiếp đã gây không ít bất cập cho người khiếm thính khi truyền đạt thông tin với người có khả năng thính lực.

Từ ý nghĩa và sự cấp thiết của ngôn ngữ ký hiệu như đã đề cập, nhóm nghiên cứu quyết định tạo ra một hệ thống liên lạc có thể thay thế người phiên dịch truyền tải các cuộc hội thoại **từ tiếng Việt sang ngôn ngữ ký hiệu và ngược lại** với độ chính xác cao, giúp cho người khiếm thính dễ dàng tiếp cận thông tin và hoà nhập với xã hội.

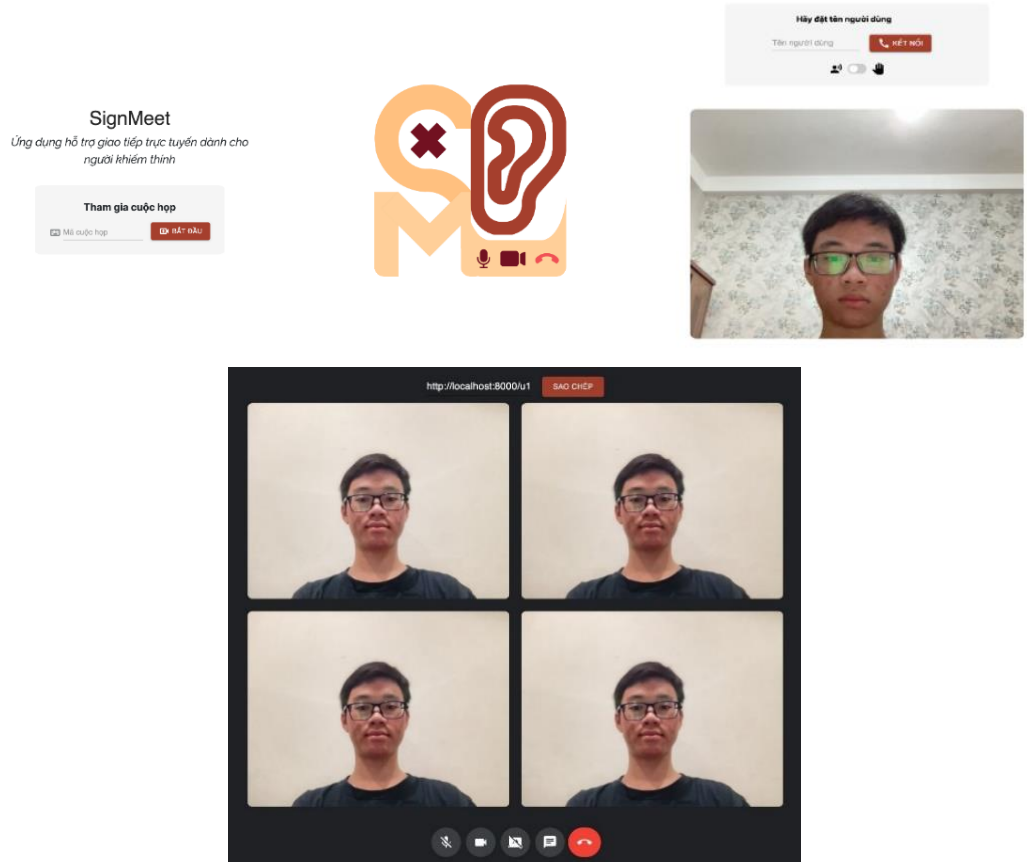
2. Giới thiệu tổng quan

Xây dựng **SignMeet** – một hệ thống hỗ trợ giao tiếp có thể **phiên dịch song ngữ tiếng Việt – ngôn ngữ ký hiệu** cho người khiếm thính.

MÔ TẢ SẢN PHẨM

1. Chức năng chính của sản phẩm

1.1. Giao diện chính của hệ thống



Hình 1. Tổng quan giao diện hệ thống

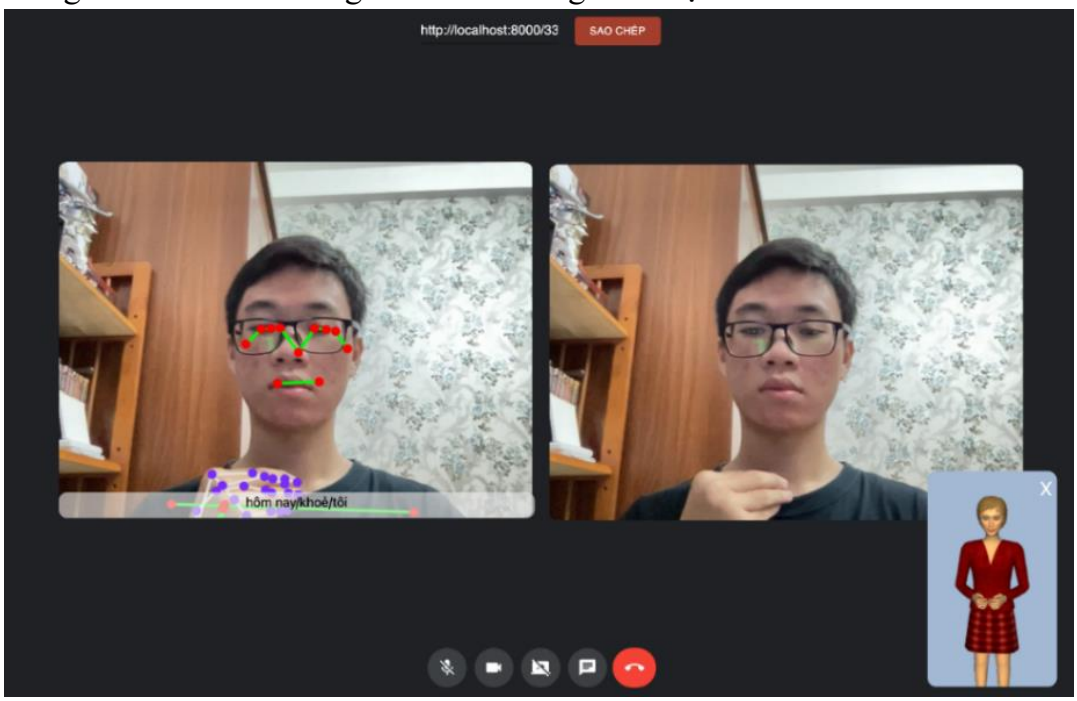
Khi truy cập trang web, người dùng có thể khởi tạo một cuộc họp mới, hoặc tham gia vào cuộc họp đã có sẵn thông qua ô nhập mã cuộc họp. Tiếp theo, người dùng phải đặt tên cá nhân và lựa chọn góc nhìn phù hợp để có thể tiếp tục tham gia vào cuộc họp.

1.2. Góc nhìn từ người dùng

Vì hệ thống được thiết kế dưới dạng một cuộc họp nên nhóm nghiên cứu sẽ chia ra thành hai góc nhìn của hai đối tượng, tương ứng với hai tính năng chính của hệ thống.

1.2.1. Góc nhìn dành cho người khiếm thính (ngôn ngữ ký hiệu → tiếng Việt)

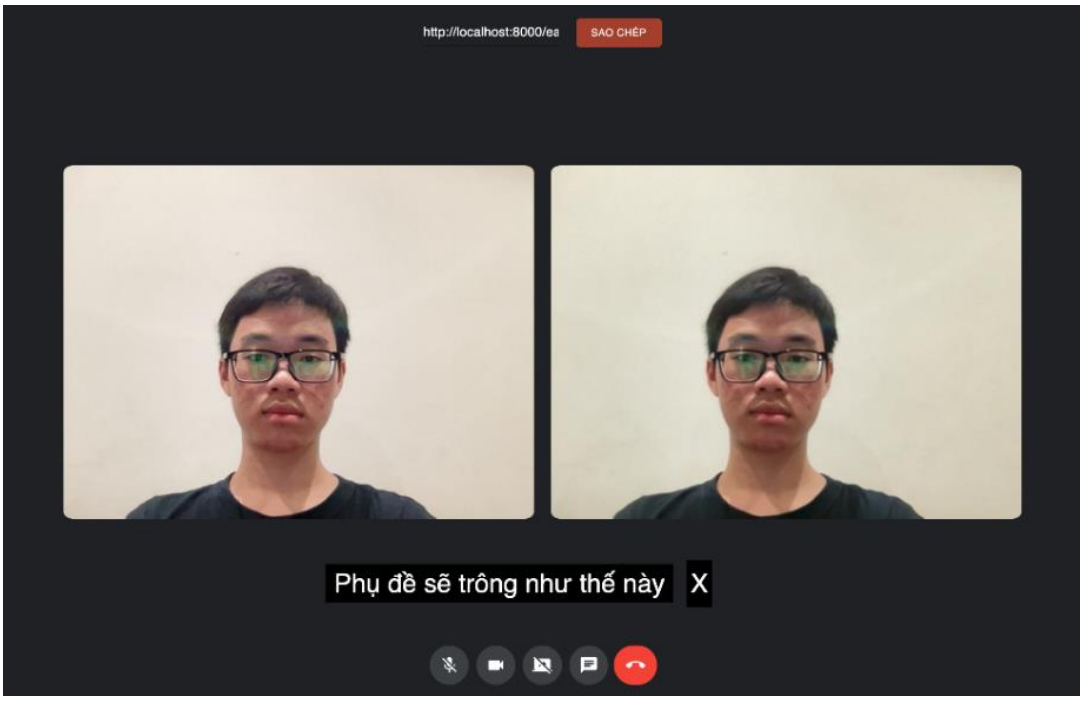
Dưới góc nhìn này, người dùng có thể biểu diễn các cử chỉ, động tác để hệ thống phiên dịch từ ngôn ngữ ký hiệu sang tiếng Việt. Dữ liệu sau khi phiên dịch sẽ được gửi sang các góc nhìn dành cho người có khả năng thính lực.



Hình 2. Giao diện từ góc nhìn dành cho người khiếm thính

1.2.2. Góc nhìn dành cho người có khả năng thính lực (tiếng Việt → ngôn ngữ ký hiệu)

Dưới góc nhìn này, một phụ đề sẽ được tích hợp trực tiếp vào hệ thống. Phụ đề này sẽ biểu diễn các câu, từ đã được phiên dịch từ ngôn ngữ ký hiệu sang tiếng Việt.



Hình 3. Giao diện từ góc nhìn dành cho người có khả năng thính lực

2. Mô tả chi tiết sản phẩm

- Ngôn ngữ lập trình: Python, JavaScript.
- Nền tảng để phát triển hệ thống: Website.

2.1. Hệ thống phiên dịch từ tiếng Việt sang ngôn ngữ ký hiệu

2.1.1. Mô hình phiên dịch tiếng Việt – ngôn ngữ ký hiệu

2.1.1.1. Thu thập dữ liệu huấn luyện

Dựa trên sự đa dạng các nguyên tắc cấu tạo câu trong tiếng Việt và ngôn ngữ ký hiệu, nhóm tác giả từ nghiên cứu [3] đã sinh ra bộ dữ liệu song ngữ gồm **10.000** cặp câu tiếng Việt – ngôn ngữ ký hiệu với sự thẩm định bán tự động bởi các chuyên gia ngôn ngữ. Nhóm nghiên cứu đã tận dụng bộ dữ liệu trên và chia thành tập huấn luyện và tập kiểm tra với tỉ lệ **80:20**. Các quy tắc chuẩn hoá khác được thực hiện theo quy trình tiền xử lý dữ liệu từ nghiên cứu [4].

2.1.1.2. Huấn luyện mô hình

Nhóm nghiên cứu quyết định sử dụng **các mô hình tiền huấn luyện**[1] **dựa trên kiến trúc Transformer** [5] – được huấn luyện trên một lượng dữ liệu khổng lồ, giúp phân cụm từ vựng và áp dụng các nguyên tắc chuyển đổi tốt hơn.

Ba mô hình tiền huấn luyện tiếng Việt bao gồm **BARTpho** [4]**, vELECTRA** [6] **và ViT5** [7] được lựa chọn để huấn luyện trong **15 epochs**[2]. Hiệu suất các mô hình được đánh giá thông qua thang đo **BLEU**[3] trên tập kiểm tra. Sau thời gian huấn luyện trung bình **2 giờ 30 phút**, kết quả cho ra từ cả ba mô hình tiền xử lý **đều tốt hơn** so với các mô hình ban đầu của nhóm tác giả [3].

STT	Mô hình dịch máy	Điểm BLEU
1	vELECTRA	0,90
2	ViT5	0,92
3	**BARTpho**	**0,94**

Bảng 1. So sánh điểm BLEU giữa các mô hình tiền huấn luyện trên tập kiểm tra

Mô hình **BARTpho** với số điểm cao nhất là **0,94** được nhóm lựa chọn sử dụng.

2.1.2. Mô hình khôi phục dấu câu

2.1.2.1. Xây dựng dữ liệu huấn luyện

Nhóm nghiên cứu kết hợp các tập dữ liệu từ nghiên cứu [9] và **ViCapPunc** [10] với độ uy tín cao, cho ra tổng **1.325.350 câu**. Với tính chất đơn giản và nhấn mạnh vào từ khoá đặc trưng của ngôn ngữ ký hiệu, nhóm quyết định chỉ khôi phục **dấu chấm (.) và dấu phẩy (,)**, các dấu câu khác được chuẩn hoá thành **dấu chấm (.)**.

Bộ dữ liệu được chia thành tập huấn luyện, đánh giá và kiểm tra với tỉ lệ **60:20:20**. Các quy tắc chuẩn hoá khác thực hiện theo quy trình tiền xử lý dữ liệu từ nghiên cứu [4].

Ngoài ra, các mô hình tiền huấn luyện thường sử dụng **các kỹ thuật tách từ**[4] để giảm kích thước mảng từ vựng quá lớn trong quá trình học. Hơn nữa, các mô hình tiền

---

[1] **Mô hình tiền huấn luyện (pre-trained model):** là mô hình đã được huấn luyện trước đó với một bộ dữ liệu lớn hoặc với các phương pháp tối tân giúp giảm công sức huấn luyện mô hình từ đầu.

[2] **Epoch:** là một thuật ngữ được sử dụng để mô tả một lần duyệt qua toàn bộ tập huấn luyện của mô hình.

[3] **BLEU (Bilingual Evaluation Understudy):** là thang đo được sử dụng trong bài toán dịch máy khi so sánh một bản dịch với một hay nhiều bản dịch tham khảo. Điểm BLEU có giá trị 0 – 1, với số điểm càng lớn, mô hình càng cho ra bản dịch tốt. [8]

[4] **Kỹ thuật tách từ (tokenization):** là quá trình tách một cụm từ, câu, đoạn văn, một hoặc nhiều tài liệu văn bản thành các đơn vị nhỏ hơn. Mỗi đơn vị nhỏ hơn này được gọi là **token**. [11]

huấn luyện được áp dụng trong bài toán này hầu hết sử dụng **thuật toán mã hoá dựa trên từ phụ**[5] (**WordPiece** [12] và **Byte-Pair Encoding** [13]) để tách từ, dẫn đến việc cần thay đổi cách đặt các token đầu ra cho dữ liệu. Cụ thể:

Câu gốc	mọi giấy tờ**,** công việc được xử lý có hệ thống**.**
Tách từ	mọi giấy tờ công việc được xử lý có hệ th @@ống
Đầu ra	O O COMMA O O O O O O O PERIOD CUT

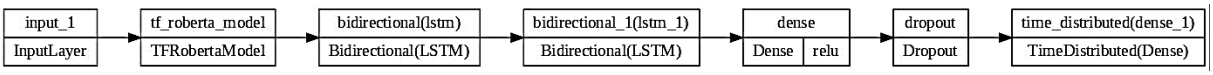
Bảng 2. Câu đầu vào mẫu và các bước tiền xử lý tách từ

Dấu chấm và dấu phẩy tương ứng với từ được đánh dấu token lần lượt là PERIOD và COMMA. Một từ không tồn tại dấu câu được đánh dấu bằng token O. Nếu có từ phụ bị tách ra khỏi một từ thì đánh dấu từ phụ khởi đầu của từ gốc bằng token PERIOD/COMMA/O tương ứng, còn các từ phụ còn lại đánh dấu bằng token CUT.

Điều này có ý nghĩa rất lớn khi có thể đảm bảo câu đầu ra sau xử lý sẽ giống hệt với câu gốc mà chỉ thay đổi về mặt dấu câu.

2.1.2.2. Xây dựng kiến trúc

Trong bài toán dự đoán dấu câu tiếng Việt, việc sử dụng các mô hình tiền huấn luyện vẫn rất cần thiết. Tuy nhiên, thay vì sử dụng các mô hình đầy đủ Encoder-Decoder/seq2seq Transformer như BARTpho hay ViT5, nhóm nghiên cứu chỉ sử dụng các mô hình tinh giản **Encoder-only Transformer**[6] có hỗ trợ tiếng Việt, cụ thể bao gồm **mBERT** [14], **viBERT** [6]**, mDeBERTa** [15] và **PhoBERT** [16]. Điều này giúp mô hình giảm độ phức tạp đi đáng kể mà vẫn đảm bảo tính chính xác.



Hình 4. Cấu trúc mô hình khôi phục dấu câu tiếng Việt

2.1.2.3. Huấn luyện mô hình

Các mô hình huấn luyện trong **5 epochs** với thời gian trung bình **7 giờ 40 phút**. Hiệu suất của các mô hình được đánh giá thông qua chỉ số **precision (P), recall (R) và macro F1-score (F1)**[7] trên tập kiểm tra.

Mô hình	**COMMA**			**PERIOD**			**Overall**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
mBERT	0,72	0,70	**0,71**	0,81	**0,86**	**0,84**	0,77	0,78	**0,78**
viBERT	0,69	0,69	0,69	0,81	0,84	0,82	0,75	0,77	0,76
mDeBERTa	0,58	**0,80**	0,67	**0,84**	0,83	**0,84**	0,71	**0,82**	0,76
PhoBERT	**0,73**	0,69	**0,71**	0,83	0,85	**0,84**	**0,78**	0,77	**0,78**

Bảng 3. Kết quả các mô hình trên tập kiểm tra[8]

Thông qua kết quả, nhóm nghiên cứu đề xuất sử dụng mô hình **PhoBERT** vì có độ chính xác cao nhất, đồng thời có độ phức tạp nhỏ, cho phép hệ thống hoạt động tốt trong nhiều điều kiện.

---

[5] **Thuật toán mã hoá dựa trên từ phụ (subword-based tokenization algorithm)** sẽ chia câu thành các từ khóa phụ: [“Let”, “us”, “learn”, “token”, **“@@ization.”**]. [11]

[6] **Encoder-only Transformer**: là biến thể từ kiến trúc Transformer [5] khi chỉ có lớp encoder. Lớp encoder có nhiệm vụ tạo ra embedding vector cho mỗi token trong câu đầu vào. Sau đó, các embedding vector sẽ được sử dụng để dự đoán kết quả đầu ra.

[7] **F1-score, precision và recall:** là các chỉ số đánh giá khả năng của mô hình trong việc dự đoán các lớp phân loại. Tương tự như thang đo BLEU, điểm F1-score, precision và recall có giá trị 0 – 1, với số điểm càng lớn, mô hình càng cho ra kết quả tốt. [17]

[8] Hai token O và CUT của các mô hình đều đạt điểm tuyệt đối nên không cần thiết đề cập trong bảng.

2.1.3. Mô hình tóm tắt câu

2.1.3.1. Xây dựng dữ liệu huấn luyện ViSenSum

	source_text	target_text
0	Thời điểm ấy , hiểu biết về bệnh ung thư của c...	Khi đó, kiến thức về ung thư còn ít, nhắc đến ...
1	Ru rú ở trong nhà Các nhà khoa học tại ĐH Toro...	Ru rú ở trong nhà khi mùa đông nhưng cảm thấy ...
2	" Nếu chuỗi cung ứng được tổ chức chặt chẽ tất...	Rau sạch đạt được nếu chuỗi cung ứng được tổ c...
3	" Stress kéo dài cộng với nhiều đêm mất ngủ sẽ...	Stress kéo dài và mất ngủ làm một người khoẻ m...
4	Phát hiện của nhóm nghiên cứu từ Đại học Penns...	Các bệnh liên quan đến stress như lo lắng, trầ...
...	...	...
9995	Theo số liệu của GLOBOCAN 2018 , có khoảng 500...	Có 500.000 ca mới và 450.000 ca tử vong do ung...
9996	Đắn đo có nên điều trị , Dung quyết định dành ...	Dung quyết định chăm sóc con trước khi phẫu th...
9997	Với trẻ lớn , biểu hiện có thể chỉ là cảm giác...	Biểu hiện ở trẻ lớn có thể là cảm giác khó thở...
9998	Nước này đã đóng cửa bệnh viện dã chiến cuối c...	Nước đã đóng cửa bệnh viện dã chiến ở Vũ Hán
9999	- Nuốt đau lan lên tai .	Nuốt đau lan lên tai.

10000 rows × 2 columns

*Hình 5. Tập dữ liệu tóm tắt câu **ViSenSum***

Do thiếu hụt các bộ dữ liệu tóm tắt câu, tập dữ liệu 5.000 bài báo sức khoẻ từ VnExpress [18] đã được tận dụng lại. Với mỗi bài viết, các câu sẽ được tách ra, tổng hợp và đưa qua mô hình ngôn ngữ lớn OpenAI GPT-4[9] để tạo tập dữ liệu huấn luyện. Cuối cùng, nhóm nghiên cứu đã tạo và kiểm duyệt được **10.000 câu**. Các quy tắc chuẩn hoá văn bản được thực hiện tương tự nghiên cứu [4]. Bộ dữ liệu được chia thành tập huấn luyện và tập kiểm tra với tỉ lệ **80:20**.

2.1.3.2. Huấn luyện mô hình

Dựa trên nghiên cứu [7], tác vụ tóm tắt văn bản được thực hiện tốt nhất bởi mô hình tiền huấn luyện **ViT5**, theo sau là **BARTpho**. Các mô hình được huấn luyện trong **15 epochs** với thời gian trung bình **3 giờ**. Dưới đây là kết quả khi huấn luyện hai mô hình trên với tập dữ liệu **ViSenSum** bằng **thang đo ROUGE[10]**, cụ thể là **ROUGE-L**:

STT	Mô hình	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
1	BARTpho	0,7301	**0,5691**	0,6536	0,6531
2	ViT5	**0,7532**	0,5583	**0,6814**	**0,6810**

*Bảng 4. Kết quả hai mô hình với tập kiểm tra khi được huấn luyện trên tập dữ liệu **ViSenSum***

ViT5 có hiệu suất tốt hơn BARTpho về điểm ROUGE-L, nhưng thời gian thực thi trung bình của ViT5 lại lâu hơn đáng kể. Sự khác biệt này có thể được giải thích bởi độ phức tạp của mô hình, khi $ViT5_{base}$ có đến **310 triệu tham số** [7], trong khi $BARTpho_{syllable}$ chỉ có **132 triệu tham số** [4]. Do đó, nhóm nghiên cứu ưu tiên sử dụng BARTpho, mặc dù điểm ROUGE-L của BARTpho thấp hơn ViT5 0,0278 điểm.

2.1.4. Mô hình nhận dạng giọng nói

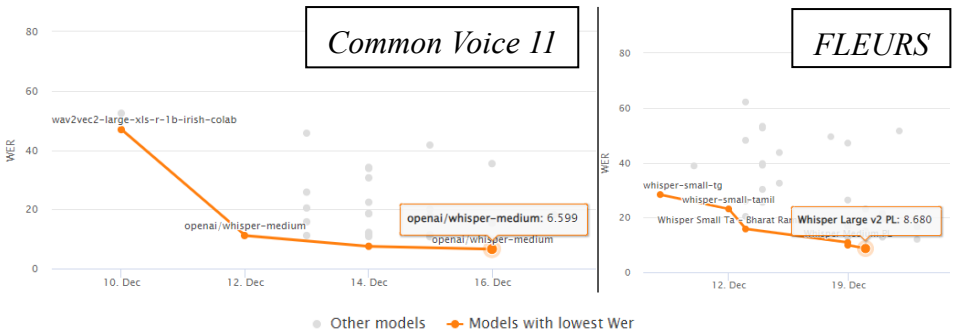
Nhóm nghiên cứu thực hiện khảo sát sự hiệu quả giữa các mô hình nhận dạng giọng nói đã tồn tại trên thị trường. Độ chính xác của mô hình sẽ được đánh giá thông qua thang đo WER[11] trên hai bộ dữ liệu âm thanh đa ngôn ngữ cho quá trình so sánh, bao gồm **Common Voice 11** [22] và **FLEURS** [23].

---

[9] **OpenAI GPT-4:** là mô hình ngôn ngữ lớn đa phương thức thứ tư trong loạt mô hình nền tảng GPT được tạo bởi OpenAI. [19]

[10] **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** là tập hợp các chỉ số đánh giá hiệu suất của các mô hình xử lý ngôn ngữ tự nhiên, thường được sử dụng trong tác vụ tóm tắt văn bản. Tương tự như thang đo BLEU, điểm ROUGE có giá trị 0 – 1, với số điểm càng lớn, mô hình càng cho ra bản tóm tắt tốt. [20]

[11] **WER (Word Error Rate):** là một chỉ số được sử dụng để đo lường độ chính xác của các mô hình nhận dạng giọng nói. Trái ngược với BLEU, thang đo WER có số điểm từ 0-100, với số điểm càng thấp thì hệ thống nhận dạng càng chính xác. [21]



*Hình 6. So sánh các mô hình nhận dạng giọng nói trên bộ dữ liệu **Common Voice 11** và **FLEURS** [24] [25]*

Có thể thấy rằng, trên cả hai tập dữ liệu kiểm định, **mô hình Whisper** của OpenAI cho ra **số điểm tốt nhất** với lần lượt là **6,599** và **8,680**. Chính vì vậy, nhóm nghiên cứu quyết định lựa chọn mô hình Whisper cho dự án lần này.

2.1.5. Kỹ thuật lượng tử hoá (quantization) mô hình

Nhóm nghiên cứu nhận thấy rằng tập dữ liệu huấn luyện cho Whisper có sự mất cân bằng về dữ liệu ngôn ngữ. Cụ thể từ nghiên cứu [26], trong tổng số 680.000 giờ âm thanh, chỉ có 117.000 giờ thuộc tập đa ngôn ngữ, bao gồm tiếng Việt. Sự mất cân bằng này có thể dẫn đến việc mô hình nhận dạng kém khi xử lý các ngôn ngữ ít dữ liệu hơn.

Giải pháp phù hợp nhất là tăng kích thước mô hình. Tuy nhiên, điều này sẽ dẫn đến việc tăng yêu cầu về phần cứng và độ trễ khi truy xuất. Vì vậy, **kỹ thuật lượng tử hoá (quantization)** đã được áp dụng, giúp giảm kích thước các mô hình học sâu, đồng thời giảm độ trễ và tăng tốc độ truy xuất, lên **mô hình kích thước lớn nhất large-v2**.

Cụ thể, nhóm sử dụng công cụ faster-whisper [27] và CTranslate2 [28] để chuyển các tham số từ kiểu dữ liệu số thực 32-bit về kiểu số nguyên 8-bit. Phương pháp thật sự đem lại hiệu quả khi so sánh với mô hình gốc trên đoạn video thời lượng 16 phút:

Mô hình	Thời gian thực thi	VRAM tối đa	RAM tối đa
Whisper large-v2	5 phút 36 giây	12.047 MB	10.020 MB
faster-whisper (large-v2)	57 giây	3.082 MB	3.561 MB

Bảng 5. So sánh trước và sau khi lượng tử hoá mô hình (thực nghiệm trên NVIDIA Tesla T4)

2.1.6. Thuật toán VAD

Âm thanh đầu vào của hệ thống có thể từ đa dạng các video, và có thể lẫn nhiều tạp âm khác nhau. Chính vì vậy, nhóm nghiên cứu quyết định ứng dụng **thuật toán VAD (Voice Activity Detection)** để phát hiện sự hiện diện của giọng nói trong một đoạn âm thanh. Thuật toán giúp xác định và loại bỏ tiếng ồn xung quanh, chẳng hạn như tiếng ồn giao thông hoặc tiếng gõ phím,... và loại bỏ các đoạn âm thanh không có tiếng để cải thiện chất lượng âm thanh của giọng nói và giảm thiểu tiêu tốn tài nguyên xử lý.

Cụ thể, nhóm nghiên cứu sử dụng **Silero VAD** [29], một mô hình mạng nơ-ron học sâu được đào tạo trên một tập dữ liệu lớn đa ngôn ngữ gồm âm thanh giọng nói và tiếng ồn. Với kích thước mô hình rất nhỏ, tốc độ truy xuất nhanh và độ chính xác cao, mô hình này được nhóm lựa chọn để tích hợp vào mô hình nhận dạng giọng nói nhằm tăng độ chính xác của hệ thống.

2.1.7. Kỹ thuật trượt cửa sổ (window sliding) và thuật toán LocalAgreement

Nhóm nghiên cứu nhận thấy rằng, trong các bài toán thời gian thực, âm thanh được coi là một mảng dữ liệu có kích thước thay đổi theo thời gian. Trong khi đó, các mô hình học máy yêu cầu dữ liệu đầu vào có kích thước cố định để xử lý. Chính vì vậy,

kỹ thuật trượt cửa sổ được sử dụng nhằm chia âm thanh thành các đoạn có cùng kích thước **S**.

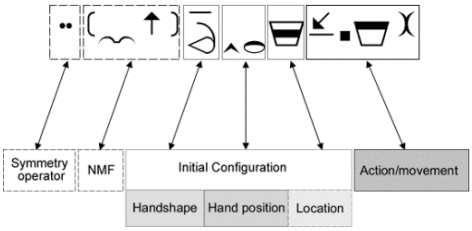
Tuy nhiên khi sử dụng kỹ thuật trượt cửa sổ, với một đoạn hội thoại, một từ khi nói có thể bị "chia đôi" khi trùng lặp vào đúng khung thời gian cắt đoạn, gây mất mát sự liền mạch của âm thanh. Hơn nữa, các đoạn âm thanh thường có sự liên kết nội dung chặt chẽ, nhưng thuật toán lại hoạt động đơn lẻ trên từng đoạn âm thanh được yêu cầu mà không ảnh hưởng bởi các câu sinh ra trước đó, gây ra hiệu tượng ảo giác (hallucination)[12] khá thường xuyên. Tất cả những nhược điểm trên khiến độ chính xác của mô hình nhận dạng âm thanh kém đi rất nhiều.

Chính vì vậy, nhóm quyết định cải tiến sang **thuật toán LocalAgreement** [30], một biến thể của thuật toán trượt cửa sổ nhưng "thông minh" hơn. Quy tắc của thuật toán như sau: Nếu bản âm thanh thứ N được gửi qua, bản âm thanh thứ $N-1$ được ghép chung lại và thông dịch sang văn bản. Sau đó, thuật toán sẽ tìm xâu con tiền tố chung dài nhất giữa xâu thứ N mới được sinh ra và xâu thứ $N-1$. Đoạn xâu con trùng khớp sẽ được đánh dấu là đúng và gửi kết quả cho các mô hình khác xử lý. Cùng lúc đó, các bản dịch đúng sẽ được coi là ngữ cảnh chính của đoạn âm thanh và đưa vào **"init prompt"**[13] của Whisper để tăng tỉ lệ chính xác của các văn bản dự đoán tiếp theo.

2.1.8. Biểu diễn ngôn ngữ ký hiệu thông qua nhân vật hoạt hình 3D

2.1.8.1. Xây dựng bộ từ điển phiên dịch tiếng Việt sang ngôn ngữ ký hiệu

Với các yêu cầu về thành tố của ngôn ngữ ký hiệu, nhóm quyết định sử dụng **Hệ thống ngôn ngữ ký hiệu Hamburg (HamNoSys).** Đây là một hệ thống phiên âm cho hầu hết các ngôn ngữ ký hiệu trên thế giới, với sự tương ứng trực tiếp giữa các biểu tượng và các thành tố cần thiết của một ký hiệu [32].



Hình 7. Cấu trúc hệ thống

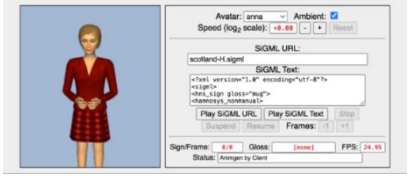
2.558 bản dịch các từ tiếng Việt sang HamNoSys được xây dựng từ nghiên cứu [33] đã được thu thập và sử dụng để chuyển đổi văn bản sang HamNoSys nhằm truyền tải nội dung cho người khiếm thính.

2.1.8.2. Thuật toán chuyển đổi dạng so khớp

Sử dụng bộ từ điển đã đề cập bên trên, nhóm nghiên cứu đã phát triển một thuật toán so khớp nhằm chuyển đổi các từ trong văn bản đầu ra sang HamNoSys. Nếu một từ không tồn tại trong từ điển, nó sẽ được tách ra thành các kí tự đơn lẻ. Cuối cùng, văn bản sau khi chuyển về HamNoSys sẽ được lưu dưới dạng **tập tin SiGML** [34] vào cơ sở dữ liệu, nhằm phục vụ cho các mục đích sử dụng sau này.

2.1.8.3. Mô hình JASigning

JASigning là hệ thống biểu diễn ngôn ngữ ký hiệu thông qua mô hình nhân vật ảo [35]. Hệ thống nhận vào dữ liệu dưới dạng tập tin **SiGML** nhằm biểu diễn ngôn ngữ ký hiệu theo hệ thống **HamNoSys** đã đề cập ở trên. Nhóm nghiên cứu tích hợp trực tiếp hệ thống này vào ứng dụng như một công cụ biểu diễn ngôn ngữ ký hiệu.

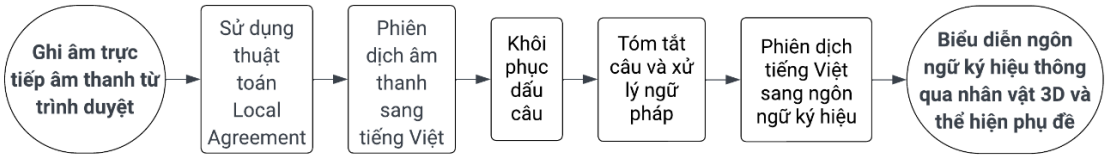


Hình 8. Giao diện JASigning

[12] **Hiện tượng ảo giác (hallucination):** xảy ra khi một mô hình AI sinh ra kết quả sai lệch hoặc gây hiểu nhầm khi không hay biết về sự vật, sự việc được nhắc đến.

[13] **init prompt (lệnh khởi tạo):** là một chuỗi các mã thông tin được cung cấp như dữ liệu trước đó, thường là một chuỗi các mã thông tin được giải mã trong phân đoạn âm thanh trước phân đoạn đang được xử lý, với mục tiêu là duy trì đầu ra thống nhất giữa các phân đoạn bằng cách theo dõi thông tin đã được phiên âm trước đó. [31]

2.1.9. Sơ đồ phương thức hoạt động

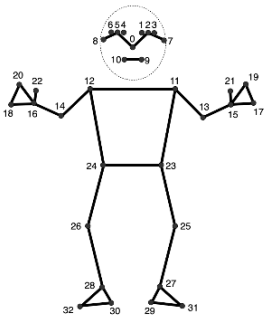


Hình 9. Sơ đồ phương thức hoạt động Hệ thống phiên dịch tiếng Việt sang ngôn ngữ ký hiệu

2.2. Hệ thống phiên dịch từ ngôn ngữ ký hiệu sang tiếng Việt

2.2.1. Toạ độ hoá các bộ phận cơ thể con người

Ngôn ngữ hình thể trong ngôn ngữ ký hiệu là một tập hợp các cử chỉ biểu diễn cho một từ hoặc một câu. Để cụ thể hoá các hành động đó, nhóm nghiên cứu sử dụng **MediaPipe**[14] để thiết lập các toạ độ điểm lên các phần quan trọng của cơ thể con người. Điều này sẽ giúp hệ thống trích xuất các lịch sử hành động dễ dàng hơn.



Hình 10. Các toạ độ điểm trên cơ thể

2.2.2. Lệnh ghi hình

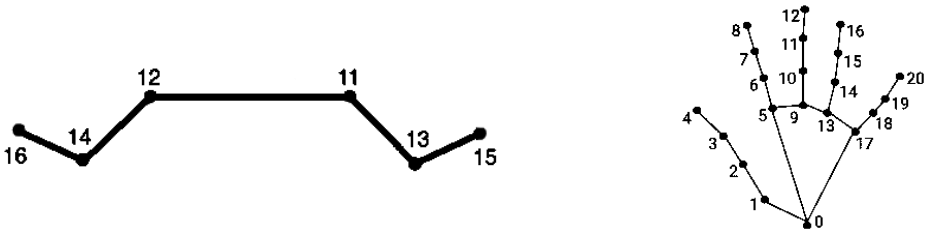
Ta nhận thấy rằng mỗi cử chỉ biểu diễn đều có thời gian bắt đầu và kết thúc. Để hệ thống biết được thời điểm bắt đầu hoặc kết thúc ghi hình, nhóm đã thiết lập một cử chỉ cố định để ra lệnh bằng cách **đưa đầu ngón giữa tay phải lên môi trong nửa giây**.

2.2.3. Mô hình nhận diện ngôn ngữ ký hiệu thông qua cử chỉ cơ thể

2.2.3.1. Xây dựng bộ dữ liệu

Theo tìm hiểu của nhóm, các cử chỉ trong ngôn ngữ ký hiệu hầu hết đến từ vùng nửa thân trên của cơ thể. Hơn nữa, một số người khiếm thính còn sử dụng các điệu bộ trên khuôn mặt hoặc dùng khẩu hình miệng để phỏng theo cách phát âm của từ cần biểu đạt [37]. Nhưng trong bài toán này, những sự thay đổi trên khuôn mặt là rất nhỏ và dễ trùng lặp với nhiều từ khác nhau, gây mất tính đặc trưng, làm nhiễu bộ dữ liệu và khó khăn trong việc dự đoán.

Với các kết luận trên, nhóm nghiên cứu quyết định chỉ trích xuất **48 toạ độ điểm** quan trọng, bao gồm 6 toạ độ phần thân trên của cơ thể và 21 toạ độ trên mỗi bàn tay.



Hình 11. Các toạ độ được trích xuất

Nhóm đã chọn ra **20 cụm từ/câu** đặc trưng gồm: tôi, bạn, xin chào, khoẻ, và, cảm ơn, tên, hôm qua, hôm nay, ngày mai, là gì, khi nào, ở đâu, tại sao, là ai, như thế nào, có, không, tốt, xấu.

Với mỗi cụm từ/câu, nhóm nghiên cứu tự thu thập **100 mẫu** là mảng của 48 toạ độ điểm trích xuất từ MediaPipe trong mỗi khung hình, với thời lượng dao động từ **5 ~ 30 khung hình**. Các mẫu động tác được tham khảo từ trang tudienngonngukyhieu.com.

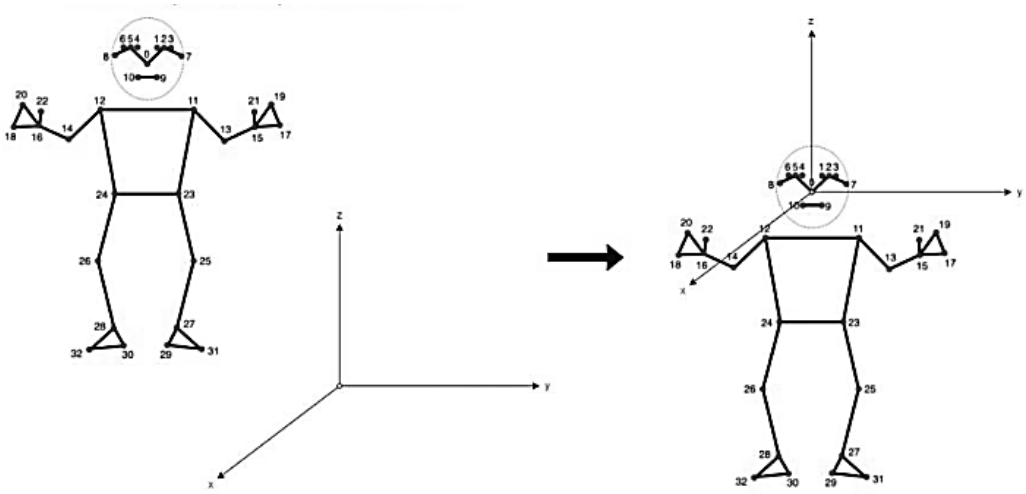
[14] **MediaPipe:** là một thư viện mã nguồn mở do Google phát triển, cung cấp nhiều giải pháp học máy đa nền tảng, có thể tùy chỉnh và hiệu quả cao. Nó được thiết kế để xử lý các tác vụ xử lý dữ liệu thời gian thực, đặc biệt là liên quan đến việc phân tích hình ảnh và video. [36]

Bộ dữ liệu được chia thành tập huấn luyện, đánh giá và kiểm tra với tỉ lệ **60:20:20**.

2.2.3.2. Chuẩn hoá các điểm toạ độ về hệ trục toạ độ Oxyz

Dễ nhận thấy rằng với tính năng vẽ khung xương, toạ độ của các điểm được cung cấp bởi MediaPipe là các toạ độ đã được **chuẩn hoá theo tỉ lệ khung hình** của thiết bị. Điều này vô hình trung đã dẫn đến tình trạng nhiễu dữ liệu đầu vào, khiến mô hình khó dự đoán nếu người dùng sử dụng các loại thiết bị khác nhau.

Chính vì lẽ đó, nhóm nghiên cứu đề xuất giải pháp bằng cách **chuẩn hoá các toạ độ điểm về hệ trục toạ độ Oxyz**, trong đó toạ độ của điểm đầu tiên sẽ được chuẩn hoá về gốc toạ độ. Sau đó, ta trừ tất cả các điểm toạ độ cho toạ độ của điểm đầu tiên [38]. Việc chuẩn hoá được thể hiện như hình sau:



Hình 12. Chuẩn hoá điểm toạ độ về gốc toạ độ

Với 6 điểm toạ độ **phần thân trên**, nhóm lấy điểm gốc là **điểm số 16**; còn với 21 điểm trên **mỗi bàn tay**, nhóm lấy điểm gốc là **điểm số 0**.

2.2.3.3. Chuẩn hoá khoảng cách giữa các điểm về tỉ lệ cố định

Ta cũng biết rằng, với mỗi người khác nhau ta lại có **tỉ lệ cơ thể khác nhau**. Điều này cũng sẽ dẫn tới một số khác biệt trong việc nhận dạng từ mô hình đã huấn luyện.

Chính vì vậy, nhóm nghiên cứu đề xuất giải pháp bằng cách **chuẩn hoá về một tỉ lệ cố định** [38] theo các bước sau:

a. Tính toạ độ điểm trung tâm giữa **N** điểm mốc.

$$center = \left(\frac{1}{N}\sum_{i=0}^{N} x^i ; \frac{1}{N}\sum_{i=0}^{N} y^i ; \frac{1}{N}\sum_{i=0}^{N} z^i\right)$$

b. Đối với mỗi điểm trong **N** điểm mốc, tiến hành tính khoảng cách so với điểm trung tâm sử dụng **khoảng cách Euclid trong không gian nhiều chiều**.

$$distance^i = \sqrt{(x^i - center_x)^2 + \left(y^i - center_y\right)^2 + (z^i - center_z)^2}$$

c. Sau khi có khoảng cách, tính **hệ số tỉ lệ chuẩn hoá**.

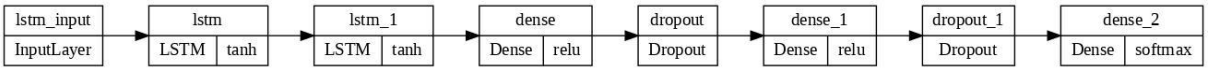
$$scale\_factor^i = \frac{1.0}{distance^i}$$

d. Cuối cùng, chuẩn hoá tỉ lệ các điểm trong **N** điểm mốc bằng cách nhân toạ độ điểm với hệ số tỉ lệ chuẩn hoá.

$$point^i = (x^i * scale\_factor^i;\ y^i * scale\_factor^i;\ z^i * scale\_factor^i)$$

2.2.3.4. Xây dựng kiến trúc

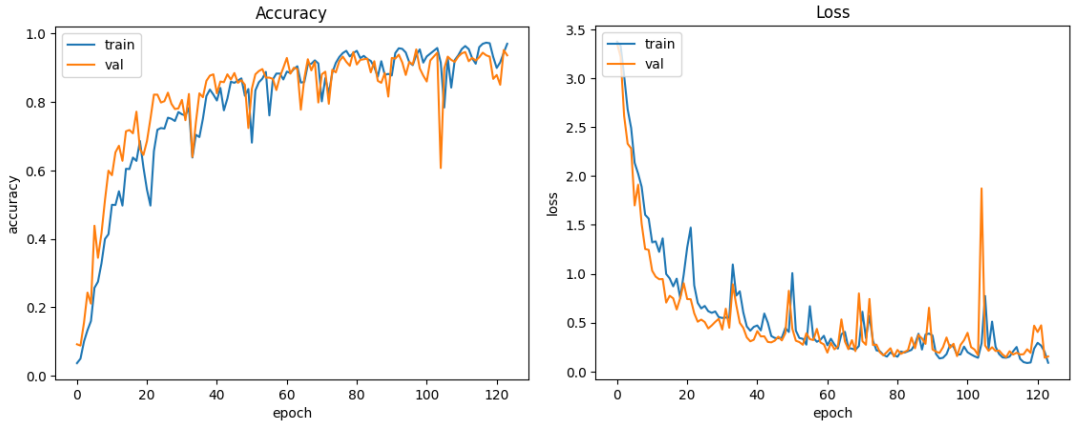
Đầu vào của hệ thống là một chuỗi mảng thông tin từ các khung hình liên tiếp nhau, tức có nghĩa đây là bài toán có tính phụ thuộc thời gian. Như vậy, nhóm đã đề xuất một kiến trúc đơn giản sử dụng **mạng LSTM**[15] để xử lý chuỗi dữ liệu này.



Hình 13. Cấu trúc mô hình nhận diện ngôn ngữ ký hiệu thông qua cử chỉ cơ thể

2.2.3.5. Huấn luyện mô hình

Mô hình huấn luyện trong **125 epochs** với thời gian trung bình **43 phút**. Hiệu suất của mô hình được đánh giá thông qua **precision (P), recall (R) và F1-score (F1)** trên tập kiểm tra.



Bảng 6. Tính chính xác (Accuracy) và độ mất mát (Loss) của mô hình trên tập huấn luyện và đánh giá

	precision	recall	f1-score	support
và	0.85	0.97	0.91	30
xấu	1.00	0.85	0.92	33
khoẻ	1.00	1.00	1.00	30
tốt	1.00	1.00	1.00	33
xin chào	0.93	0.43	0.59	30
như thế nào	0.84	0.90	0.87	30
tôi	0.76	0.67	0.71	33
tên	0.91	0.91	0.91	33
không	0.86	0.94	0.90	33
cảm ơn	0.77	0.82	0.79	33
hôm nay	1.00	1.00	1.00	30
ngày mai	1.00	1.00	1.00	33
cái gì	0.88	1.00	0.94	30
khi nào	1.00	0.94	0.97	33
ở đâu	1.00	1.00	1.00	33
là ai	0.94	0.94	0.94	33
tại sao	0.72	1.00	0.84	33
có	0.86	0.97	0.91	33
hôm qua	1.00	1.00	1.00	39
bạn	0.92	0.77	0.84	30
accuracy			0.91	645
macro avg	0.91	0.90	0.90	645
weighted avg	0.91	0.91	0.90	645

Hình 14. Độ chính xác của mô hình trên tập kiểm tra

[15] **Mạng LSTM (Long Short-Term Memory):** là một loại mạng nơ-ron nhân tạo được thiết kế để xử lý các dữ liệu có tính phụ thuộc thời gian (ví dụ như văn bản, âm thanh, video). LSTM có thể lưu trữ thông tin trong thời gian dài hơn so với các loại mạng RNN (Recurrent Neural Network) truyền thống, giúp nó trở nên hiệu quả hơn trong việc xử lý các chuỗi dữ liệu dài và phức tạp.

2.2.3.6. Tối ưu hoá mô hình

Để hoạt động trên máy khách có các cấu hình khác nhau, mô hình cần được thu gọn nhất có thể mà vẫn đảm bảo được tính chính xác. Chính vì vậy, nhóm nghiên cứu quyết định chuyển đổi mô hình sang định dạng **TensorFlow Lite**, một phiên bản nhỏ gọn của TensorFlow cho các thiết bị di động. Kết quả trung bình cho ra **độ trễ ~16,72ms** khi khởi chạy thực tế trên mẫu máy HP Pavilion 15 với bộ vi xử lý Intel Core i7-7500U.

Ngoài ra để triển khai lên trang web, nhóm còn sử dụng thư viện **TensorFlow.js**, một thư viện giúp thực hiện các tác vụ học máy và trí tuệ nhân tạo trong JavaScript.

2.2.4. Dự đoán câu hoàn chỉnh

Hai đặc điểm quan trọng nhất của ngôn ngữ ký hiệu là tính giản lược và có điểm nhấn [37]. Chính vì nguyên nhân đó, các cụm từ/câu được mô hình dự đoán và ghép lại thành một đoạn văn bản hoàn chỉnh, vẫn khác mơ hồ về mặt ngữ nghĩa cho người có khả năng thính lực hiểu được.

Nhận thấy vấn đề trên, nhóm nghiên cứu tận dụng API từ mô hình ngôn ngữ lớn **OpenAI GPT-4**, sử dụng **kỹ thuật prompt engineering**[16] để giúp hoàn thiện về nghĩa cho văn bản đầu ra. Cụ thể, nhóm sử dụng **kỹ thuật few-shot**[17] với mẫu lệnh như sau:

"Bạn hãy đóng vai một phiên dịch viên từ ngôn ngữ ký hiệu sang tiếng Việt. Bạn sẽ được cho một danh sách các cụm từ/câu theo thứ tự mà người khiếm thính đã biểu diễn. Hãy gợi ý một cách sắp xếp chúng thành một câu có nghĩa và nếu có thể, hãy thêm những từ hoặc dấu câu mà bạn cho là cần thiết để câu đầu ra trở nên dễ hiểu hơn.

Ví dụ: "tôi/khoẻ/bạn/như thế nào" → "tôi rất khoẻ, còn bạn thì sao?"

"xin chào/hôm nay/như thế nào/bạn" → "xin chào, ngày hôm nay của bạn như thế nào?"

"hôm qua/ở đâu/bạn" → "ngày hôm qua bạn đã ở đâu?"

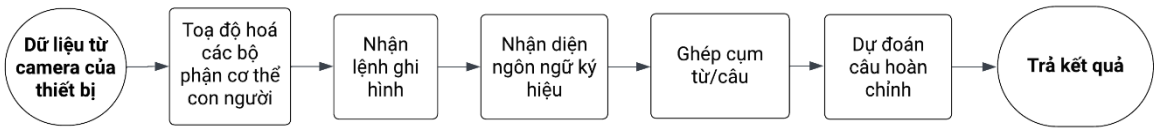
"là gì/tên/bạn" → "tên của bạn là gì?"

"cảm ơn/bạn" → "tôi cảm ơn bạn."

Hãy giúp tôi thực hiện với các cụm từ/câu sau: "**< các cụm từ/câu từ người dùng đưa vào khi biểu diễn xong, được chia cắt bằng dấu "/" >**"

Chỉ in ra một đáp án duy nhất. Bắt buộc sử dụng tiếng Việt."

2.2.5. Sơ đồ phương thức hoạt động



Hình 15. Sơ đồ phương thức hoạt động của Hệ thống phiên dịch từ ngôn ngữ ký hiệu sang tiếng Việt

2.3. Triển khai máy chủ

Sau khi đã hoàn thiện sáp nhập các mô hình và thuật toán liên quan, nhóm cần thiết lập máy chủ để các máy khách có thể giao tiếp với hệ thống.

Nhóm nghiên cứu sử dụng **giao thức WebSocket** – một giao thức truyền thông cho phép giao tiếp hai chiều thời gian thực giữa máy khách và máy chủ qua một kết nối TCP duy nhất. Nhờ đó, nhóm có thể gửi nhận thông tin trực tiếp từ các máy khách đến

---

[16] **Kỹ thuật prompt engineering:** là một kỹ thuật giúp điều chỉnh các yếu tố trong câu lệnh để tối ưu hóa đầu vào và khai thác tối đa hiệu quả của các mô hình ngôn ngữ lớn ta nhận được qua câu trả lời đầu ra.

[17] **Kỹ thuật few-shot:** đúng như tên gọi, là một kỹ thuật cung cấp cho mô hình một vài ví dụ ta mong muốn để mô hình có thể bắt chước quá trình xử lý trong câu trả lời đầu ra.

máy chủ để thực hiện xử lý liên tục với độ trễ vô cùng thấp, giúp dữ liệu đầu ra được ổn định và chính xác hơn.

Tuy nhiên, hệ thống chỉ mới thao tác được trên máy chủ cục bộ, hạn chế khả năng chia sẻ dữ liệu giữa máy chủ và các thiết bị khách với nhau, gọi là mô hình khách – chủ. Vì vậy, nhóm kết hợp **ngrok** và **Packetriot** giúp kết nối máy chủ cục bộ với internet thông qua một kết nối đường hầm (tunnel) an toàn.

3. Thực nghiệm và đánh giá sản phẩm

3.1. Thực nghiệm

Hệ thống được thẩm định bởi **32 người**, bao gồm giáo viên và học sinh tại Trung tâm GD&ĐT Ngôn ngữ ký hiệu và Hỗ trợ người Điếc Miền Trung – CDS Đà Nẵng.

3.1.1. Hệ thống phiên dịch từ tiếng Việt sang ngôn ngữ ký hiệu

Nhóm nghiên cứu cần đánh giá độ chính xác của các bản dịch **từ âm thanh sang tiếng Việt** và của các bản dịch **từ tiếng Việt sang ngôn ngữ ký hiệu**. Vì hệ thống chia đoạn âm thanh làm nhiều cửa sổ có cùng kích thước nên nhóm nghiên cứu đề xuất một phương pháp tính toán mới theo công thức sau đây:

$$Acc_1 = (1 - \frac{số\ lượng\ cửa\ sổ\ có\ bản\ dịch\ lỗi/không\ rõ\ nghĩa\ trung\ bình}{tổng\ số\ lượng\ cửa\ sổ\ trong\ video}) * 100\%$$

Khảo sát thực hiện với hình thức đánh giá trực tiếp độ chính xác thông qua **10 video bản tin** trên **kênh VTV24** thuộc nền tảng YouTube.

Video tham chiếu	**Độ chính xác bản dịch âm thanh (WER)**	**Số lượng cửa sổ**	**Số lượng cửa sổ có bản dịch lỗi/không rõ nghĩa trung bình**	**Độ chính xác Acc_1 (%)**
ĐBQH đề xuất lương giáo viên	9,6	61	13	78,69
Góc khuất về thu nhập khủng từ livestream bán hàng	9,4	105	18	82,86
Dệt may trên hành trình chuyển đổi xanh	9	87	12	86,21
Tình hình nhân đạo thảm khốc tại Gaza	9,8	93	17	81,72
Tuyển 1.000 sinh viên ngành bán dẫn năm 2024	7,6	92	21	77,17
Mỹ liên minh 40 quốc gia chăn đứng tài trợ cho tin tặc	7,1	17	2	88,24
Phức tạp tội phạm thanh thiếu niên trên đường phố	9	32	2	93,75
Đi tìm phương thuốc trường sinh	8,5	135	32	76,30
Điểm tuần: Hư cấu và hiện thực	7,3	151	31	79,47
Mỹ ban hành quy định về trí tuệ nhân tạo	8,3	12	0	100
Độ chính xác trung bình	**8,6**			**84,44**

Bảng 7. Thống kê tính chính xác các bản dịch theo thời gian thực từ âm thanh sang tiếng Việt (WER) và từ tiếng Việt sang ngôn ngữ ký hiệu (Acc_1) cho từng video tham chiếu

Như vậy, hệ thống có độ chính xác của các bản dịch từ âm thanh sang tiếng Việt khá cao với **8,6 điểm**, còn độ chính xác của các bản dịch từ tiếng Việt sang ngôn ngữ ký hiệu hoạt động với độ chính xác xuất sắc, lên đến **84,44%**.

3.1.2. Hệ thống phiên dịch từ ngôn ngữ ký hiệu sang tiếng Việt

Về phương pháp đánh giá, nhóm sẽ đưa ra **năm câu tham chiếu** để mỗi người tham gia khảo sát mô phỏng theo bằng cách ghép các cụm từ/câu bên trên bằng ngôn ngữ ký hiệu. Từ đó, số lần mô hình trả kết quả sai/không rõ nghĩa sẽ được thống kê lại và tính toán độ chính xác thông qua công thức sau đây:

$$Acc_2 = (1 - \frac{số\ lần\ mô\ hình\ trả\ kết\ quả\ sai/không\ rõ\ nghĩa}{số\ người\ tham\ gia\ khảo\ sát\ (32\ người)}) * 100\%$$

STT	Câu tham chiếu	Số lần trả kết quả sai/không rõ nghĩa	Độ chính xác Acc_2 (%)
1	hôm nay/như thế nào/bạn	3	90,625
2	xin chào/tôi/khoẻ/cảm ơn/bạn/như thế nào	7	78,125
3	hôm qua/ở đâu/bạn	5	84,375
4	bạn/là ai/tên/là gì/bạn	4	87,5
5	ngày mai/bạn/và/tôi/ở đâu	8	75,0
	Độ chính xác trung bình		**83,125**

Bảng 8. Thống kê tính chính xác các bản phiên dịch từ ngôn ngữ ký hiệu sang tiếng Việt (Acc_2) cho từng câu tham chiếu

Như vậy, trên thực tế, hệ thống có độ chính xác trung bình vào khoảng **83,125%**, thấp hơn khoảng **7,825%** so với khi thử nghiệm trên tập kiểm tra của mô hình.

3.2. Hạn chế

3.2.1. Hệ thống phiên dịch từ tiếng Việt sang ngôn ngữ ký hiệu

- Việc sử dụng phụ đề có thể không đảm bảo tốc độ và độ chính xác cao như ngôn ngữ ký hiệu, có thể bị trễ hoặc không tương thích hoàn toàn với nội dung đã truyền đạt.
- Nhân vật biểu diễn còn thiếu các khẩu hình miệng và biểu cảm của nhân vật.
- Vẫn còn một vài từ diễn tả sai nghĩa bởi sự thiếu hoàn chỉnh của bộ dữ liệu từ điển.

3.2.2. Hệ thống phiên dịch từ ngôn ngữ ký hiệu sang tiếng Việt

- Hệ thống đôi lúc nhận diện sai, nhất là khi thay đổi các điều kiện môi trường.
- Hiện tượng ảo giác của mô hình ngôn ngữ lớn vẫn hay xảy ra, khiến câu đầu ra bị lệch nghĩa so với nghĩa truyền đạt mong muốn.

KẾT LUẬN

1. Kết luận chung

Nhóm nghiên cứu tin rằng **SignMeet** có thể hỗ trợ con người trong việc phiên dịch song ngữ tiếng Việt – ngôn ngữ ký hiệu cho người khiếm thính. Tuy nhiên, ứng dụng vẫn tồn tại một số hạn chế. Trong tương lai gần, nhóm sẽ cố gắng khắc phục các nhược điểm nêu trên và hoàn thiện ứng dụng, bổ sung và cải tiến nhiều chức năng để đem đến sự tiện nghi tốt nhất cho người khiếm thính.

2. Hướng phát triển của sản phẩm trong tương lai

- Phát triển và mở rộng ứng dụng sang nền tảng di động.
- Tập trung xây dựng, củng cố các bộ dữ liệu để cải thiện tính chính xác của mô hình.
- Tối ưu hoá ứng dụng, tăng tốc độ xử lý.

TÀI LIỆU THAM KHẢO

[1] Đ. T. Hiên, "Cơ sở của việc dạy học cho người khiếm thính bằng ngôn ngữ kí hiệu," *VNU Journal of Science: Education Research*, vol. 29, no. 2, Art. no. 2, Jun. 2013, Accessed: Jan. 11, 2024. [Online]. Available: https://js.vnu.edu.vn/ER/article/view/498

[2] "Phiên dịch viên Ngôn ngữ ký hiệu, 'nghề' kén người học bậc nhất," giaoduc.net.vn. Accessed: Oct. 14, 2023. [Online]. Available: https://giaoduc.net.vn/post-217305.gd

[3] T. B. D. Nguyen, T. N. Phung, and T. T. Vu, "A STUDY OF DATA AUGMENTATION AND ACCURACY IMPROVEMENT IN MACHINE TRANSLATION FOR VIETNAMESE SIGN LANGUAGE," *Journal of Computer Science and Cybernetics*, vol. 39, no. 2, Art. no. 2, Jun. 2023, doi: 10.15625/1813-9663/18025.

[4] N. L. Tran, D. M. Le, and D. Q. Nguyen, "BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese." arXiv, Jun. 27, 2022. Accessed: Aug. 06, 2023. [Online]. Available: http://arxiv.org/abs/2109.09701

[5] A. Vaswani *et al.*, "Attention Is All You Need." arXiv, Aug. 01, 2023. Accessed: Oct. 15, 2023. [Online]. Available: http://arxiv.org/abs/1706.03762

[6] V. B. The, O. T. Thi, and P. Le-Hong, "Improving Sequence Tagging for Vietnamese Text Using Transformer-based Neural Models." arXiv, Sep. 25, 2020. Accessed: Oct. 15, 2023. [Online]. Available: http://arxiv.org/abs/2006.15994

[7] L. Phan, H. Tran, H. Nguyen, and T. H. Trinh, "ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation." arXiv, May 26, 2022. Accessed: Oct. 01, 2023. [Online]. Available: http://arxiv.org/abs/2205.06457

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. doi: 10.3115/1073083.1073135.

[9] T. Pham, N. Nguyen, Q. Pham, H. Cao, and B. Nguyen, "Vietnamese Punctuation Prediction Using Deep Neural Networks," in *SOFSEM 2020: Theory and Practice of Computer Science*, A. Chatzigeorgiou, R. Dondi, H. Herodotou, C. Kapoutsis, Y. Manolopoulos, G. A. Papadopoulos, and F. Sikora, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 388–400. doi: 10.1007/978-3-030-38919-2_32.

[10] H. T. T. Uyen, N. A. Tu, and T. D. Huy, "Vietnamese Capitalization and Punctuation Recovery Models." arXiv, Jul. 04, 2022. Accessed: Aug. 06, 2023. [Online]. Available: http://arxiv.org/abs/2207.01312

[11] xdevlabs, "Tokenization là gì? Các kỹ thuật tách từ trong Xử lý ngôn ngữ tự nhiên," VinBigData. Accessed: Jan. 27, 2024. [Online]. Available: https://vinbigdata.com/chatbot/tokenization-la-gi-cac-ky-thuat-tach-tu-trong-xu-ly-ngon-ngu-tu-nhien.html

[12] "WordPiece tokenization - Hugging Face NLP Course." Accessed: Nov. 01, 2023. [Online]. Available: https://huggingface.co/learn/nlp-course/chapter6/6

[13] "Byte-Pair Encoding tokenization - Hugging Face NLP Course." Accessed: Nov. 01, 2023. [Online]. Available: https://huggingface.co/learn/nlp-course/chapter6/5

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019. Accessed: Nov. 01, 2023. [Online]. Available: http://arxiv.org/abs/1810.04805

[15] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing."

arXiv, Mar. 24, 2023. Accessed: Nov. 01, 2023. [Online]. Available: http://arxiv.org/abs/2111.09543

[16] D. Q. Nguyen and A. Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 1037–1042. doi: 10.18653/v1/2020.findings-emnlp.92.

[17] "F-score," *Wikipedia*. Sep. 09, 2023. Accessed: Nov. 01, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=F-score&oldid=1174585267

[18] "Vietnamese health news." Accessed: Oct. 15, 2023. [Online]. Available: https://www.kaggle.com/datasets/yuiikin/vietnamese-vnexpress-news

[19] "GPT-4," *Wikipedia*. Nov. 01, 2023. Accessed: Nov. 05, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=GPT-4&oldid=1182990565

[20] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. Accessed: Oct. 15, 2023. [Online]. Available: https://aclanthology.org/W04-1013

[21] "Word error rate," *Wikipedia*. Dec. 15, 2023. Accessed: Jan. 12, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Word_error_rate&oldid=1190029194

[22] "mozilla-foundation/common_voice_11_0 · Datasets at Hugging Face." Accessed: Jan. 12, 2024. [Online]. Available: https://huggingface.co/datasets/mozilla-foundation/common_voice_11_0

[23] "google/fleurs · Datasets at Hugging Face." Accessed: Jan. 12, 2024. [Online]. Available: https://huggingface.co/datasets/google/fleurs

[24] "Papers with Code - mozilla-foundation/common_voice_11_0 Benchmark (Automatic Speech Recognition)." Accessed: Jan. 12, 2024. [Online]. Available: https://paperswithcode.com/sota/automatic-speech-recognition-on-mozilla-66

[25] "Papers with Code - google/fleurs Benchmark (Automatic Speech Recognition)." Accessed: Jan. 12, 2024. [Online]. Available: https://paperswithcode.com/sota/automatic-speech-recognition-on-google-fleurs-16

[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv.org. Accessed: Jan. 12, 2024. [Online]. Available: https://arxiv.org/abs/2212.04356v1

[27] "SYSTRAN/faster-whisper." SYSTRAN, Jan. 12, 2024. Accessed: Jan. 12, 2024. [Online]. Available: https://github.com/SYSTRAN/faster-whisper

[28] "OpenNMT/CTranslate2." OpenNMT, Jan. 11, 2024. Accessed: Jan. 12, 2024. [Online]. Available: https://github.com/OpenNMT/CTranslate2

[29] A. Veysov, "snakers4/silero-vad." Mar. 18, 2024. Accessed: Mar. 18, 2024. [Online]. Available: https://github.com/snakers4/silero-vad

[30] D. Macháček, R. Dabre, and O. Bojar, "Turning Whisper into Real-Time Transcription System," in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, Bali, Indonesia: Association for Computational Linguistics, 2023, pp. 17–24. doi: 10.18653/v1/2023.ijcnlp-demo.3.

[31] D. Cochard, "Prompt Engineering in Whisper," axinc-ai. Accessed: May 03, 2024. [Online]. Available: https://medium.com/axinc-ai/prompt-engineering-in-whisper-6bb18003562d

[32] "Hamburg Notation System," *Wikipedia*. Oct. 27, 2023. Accessed: Nov. 01, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Hamburg_Notation_System&oldid=1182196933

[33] L. D. Quach and C.-N. Nguyen, "Converting the Vietnamese Television News into 3D Sign Language Animations for the Deaf," *Lecture Notes of the Institute for Computer Sciences*, vol. 257, Jan. 2019, doi: 10.1007/978-3-030-05873-9.

[34] R. Elliott, J. Glauert, V. Jennings, and R. Kennaway, "An Overview of the SiGML Notation and SiGMLSigning Software System".

[35] "JASigning - Virtual Humans." Accessed: Oct. 17, 2023. [Online]. Available: https://vh.cmp.uea.ac.uk/index.php/JASigning

[36] "MediaPipe," Google for Developers. Accessed: May 02, 2024. [Online]. Available: https://developers.google.com/mediapipe

[37] "Giáo trình ngôn ngữ kí hiệu thực hành.pdf." Accessed: Jan. 11, 2024. [Online]. Available: https://www.slideshare.net/man2017/gio-trnh-ngn-ng-k-hiu-thc-hnhpdf

[38] "Hội những anh em thích ăn Mì AI | Hi xin chào mọi người, | Facebook." Accessed: Jun. 09, 2024. [Online]. Available: https://www.facebook.com/groups/miaigroup/permalink/1648051782632754

Đà Nẵng, ngày 04 tháng 8 năm 2024

Chữ ký của tác giả/nhóm tác giả

Lê Quang Phúc