DESIGN AND BUILD A VIETNAMESE SIGN LANGUAGE TRANSLATION APPLICATION

Tran Vu Hoang[1*], Le Quoc Dat[1], Huynh Dinh Hiep[2], Doan Manh Cuong[3]
[1] *Ho Chi Minh City University of Technology and Education*
[2] *South Telecommunication & Software JSC*
[3] *TNU - University of Information and Communication Technology*

ARTICLE INFO

Received: 06/3/2025
Revised: 16/6/2025
Published: 27/6/2025

KEYWORDS

Vietnamese sign language translation
AlphaPose
SMPL
PhoWhisper
Blender Python API

ABSTRACT

In the rapidly developing technological era today, artificial intelligence applications worldwide are significantly contributing to economic and social development. Accompanying the swift advancement of society is the ever-changing influx of information, which poses a considerable challenge for those with limited access to information, language barriers, or disabilities in keeping up with new information. In this study, we propose a method to design and develop a translation software for the hearing-impaired, incorporating sign language based on natural language processing, deep learning models, and computer vision. The goal is to design a system that can convert information in the form of text or audio into short videos represented in sign language. After undergoing experimentation, the system has met all the specified requirements. The system can convert a text or audio file into a video that can be understood by the hearing-impaired, with a rendering time of approximately 20 seconds per word (phrase).

THIẾT KẾ XÂY DỰNG PHẦN MỀM PHIÊN DỊCH NGÔN NGỮ KÝ HIỆU TIẾNG VIỆT

Trần Vũ Hoàng[1*], Lê Quốc Đạt[1], Huỳnh Đình Hiệp[2], Đoàn Mạnh Cường[3]
[1]*Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh*
[2]*Công ty Cổ phần Phần mềm Viễn thông miền Nam*
[3]*Trường Đại học Công nghệ Thông tin và Truyền thông - ĐH Thái Nguyên*

THÔNG TIN BÀI BÁO

Ngày nhận bài: 06/3/2025
Ngày hoàn thiện: 16/6/2025
Ngày đăng: 27/6/2025

TỪ KHÓA

Phiên dịch ngôn ngữ ký hiệu tiếng Việt
AlphaPose
SMPL
PhoWhisper
Blender Python API

TÓM TẮT

Trong thời đại công nghệ phát triển nhanh chóng hiện nay, các ứng dụng sử dụng trí tuệ nhân tạo nói chung trên thế giới đang góp phần không nhỏ đến sự phát triển kinh tế - xã hội. Đi cùng với sự phát triển nhanh chóng của xã hội là lượng thông tin thay đổi hàng ngày, hàng giờ thế nên đối với người tiếp nhận thông tin bị hạn chế, gặp phải rào cản ngôn ngữ hay người khiếm khuyết thì việc cập nhật những thông tin mới là một điều tương đối khó khăn. Trong nghiên cứu này, chúng tôi đề xuất phương pháp thiết kế xây dựng phần mềm phiên dịch dành cho người khiếm thính, kết hợp ngôn ngữ ký hiệu dựa vào xử lý ngôn ngữ tự nhiên, mô hình học sâu và thị giác máy tính. Mục tiêu là thiết kế hệ thống có chức năng chuyển đổi được thông tin dưới dạng văn bản hoặc âm thanh thành các video ngắn biểu diễn bằng ngôn ngữ ký hiệu. Sau khi trải qua thực nghiệm, hệ thống đáp ứng tất cả các yêu cầu đã đề ra. Hệ thống có thể chuyển đổi một văn bản hoặc tệp âm thanh thành một video giúp người khiếm thính hiểu được và thời gian kết xuất video đạt tốc độ khoảng 20s/ từ (cụm từ).

DOI: https://doi.org/10.34238/tnu-jst.12232

[*] Corresponding author. *Email: hoangtv@hcmute.edu.vn*

1. Giới thiệu

Trong bối cảnh đời sống và xã hội phát triển nhanh chóng, thông tin thay đổi liên tục theo thời gian. Việc tiếp cận và cập nhật thông tin mới trở thành thách thức đối với những người hạn chế về ngôn ngữ hoặc khuyết tật. Theo Tổng cục Thống kê năm 2016 [1], có khoảng 0,24% tổng số người khuyết tật từ 18 tuổi trở lên bị khiếm thính theo bộ công cụ WG-SS (khoảng 14.000 người) và 1,37% theo bộ công cụ WG-ES (khoảng 880.721 người). Vì việc thiếu đi khả năng nghe nói ảnh hưởng lớn đến khả năng ghi nhớ và tiếp thu ngôn ngữ thông thường nên hầu hết cách để họ giao tiếp là ngôn ngữ ký hiệu. Tuy nhiên, ở Việt Nam, ngôn ngữ ký hiệu chưa có sự đồng nhất trong việc giảng dạy và sử dụng, bên cạnh đó sự thiếu hỗ trợ từ cộng đồng nghe-nói cũng là một vấn đề cần phải quan tâm. Do đó, một hệ thống đóng vai trò phiên dịch giúp hỗ trợ truyền tải thông tin đến nhóm người khiếm thính là cần thiết.

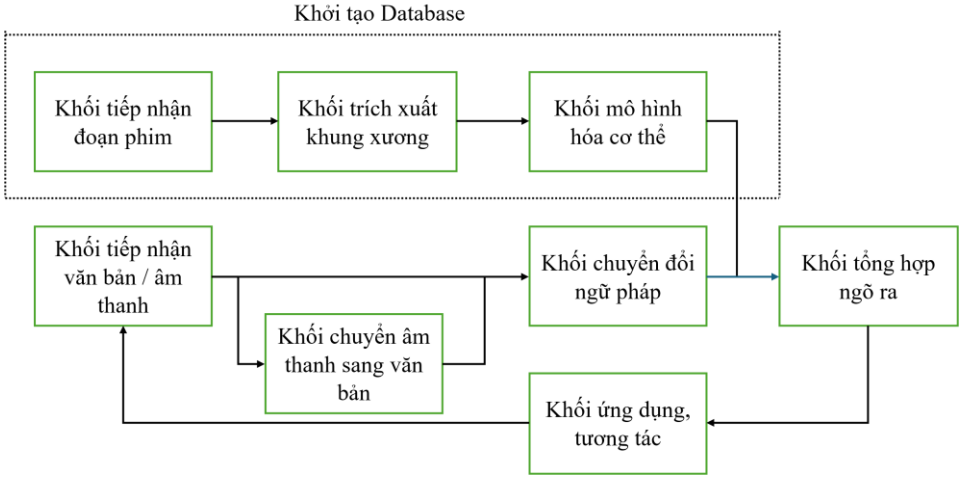
Với mục tiêu đó, đã và đang có rất nhiều sản phẩm ra đời nhằm giúp cho việc giao tiếp bằng ngôn ngữ ký hiệu trở nên đơn giản và thuận tiện hơn như Google Live Transcribe [2] - Phiên dịch thân thiện của người khiếm thính. Đây là một ứng dụng di động miễn phí được các kỹ sư Google tạo ra nhằm hỗ trợ người khiếm thính có thể giao tiếp tốt hơn. Tuy nhiên, ứng dụng này chỉ đơn thuần là chuyển giọng nói thành văn bản theo thời gian thực, từ đó người dùng có thể xem và phản hồi. Trong khi đó, ngôn ngữ ký hiệu có ngữ pháp và cú pháp riêng, khác với ngôn ngữ viết. Với người khiếm thính bẩm sinh, ngôn ngữ ký hiệu là ngôn ngữ mẹ đẻ, trong khi ngôn ngữ viết có thể được học sau này và có thể không dễ tiếp thu như đối với người nghe. Do đó, gần đây, Hand Talk [3] do công ty Acesso para Todos phát hành đã có thể biểu diễn được văn bản thành hoạt ảnh 3D. Ứng dụng này sẽ tự động dịch văn bản và âm thanh sang Ngôn ngữ ký hiệu Brazil (Libras) và Ngôn ngữ ký hiệu của Mỹ (ASL) thông qua trí thông minh nhân tạo. Tuy nhiên, ứng dụng này được phát triển ở nước ngoài nên có sự khác biệt về ngữ pháp so với ở Việt Nam, mặc dù một số mô hình AI có thể học cách chuyển đổi giữa các hệ thống ngôn ngữ ký hiệu khác nhau, nhưng điều này cần thu thập một lượng dữ liệu lớn để huấn luyện. Trong nước gần đây cũng có các nghiên cứu liên quan như: găng tay chuyển ngữ giành giải nhất cuộc thi Khoa học Kỹ thuật cấp quốc gia (ViSEF) cho học sinh trung học [4]. Găng tay hỗ trợ người dùng trong việc dịch ngôn ngữ ký hiệu sang lời nói bằng tiếng Việt nhưng không thể chuyển theo hướng ngược lại từ ngôn ngữ tiếng Việt sang thủ ngữ. Do đó, nhóm nghiên cứu của Đại học Cần Thơ [5] đã thiết kế phần mềm chuyển bản tin thành ngôn ngữ ký hiệu biểu diễn dưới dạng video 2D. Tuy nhiên, các ứng dụng này vẫn cần sự hỗ trợ của phần cứng và vẫn còn phụ thuộc một số yếu tố con người trong lúc vận hành. Trong khi đó, các nghiên cứu gần nhất về ngôn ngữ ký hiệu [6] - [8] thì lại tập trung vào việc nhận diện các ký tự đơn, điều này không thực tế vì ngôn ngữ ký hiệu được thể hiện theo ý nghĩa của từng từ hoặc cụm từ chứ không phải là ghép từng ký tự đơn lại với nhau. Chính vì lẽ đó, trong nghiên cứu [9], Yu Liu và cộng sự đã đề xuất sử dụng kỹ thuật DETR [10] mới nhất để nhận diện ngôn ngữ cử chỉ theo từng từ dựa vào video, tuy nhiên đề xuất này cần một lượng lớn dữ liệu để huấn luyện, và nhóm tác giả chỉ có thể thử nghiệm được trên chín từ thông dụng. Những khảo sát trên cho thấy có rất ít các nghiên cứu đi theo hướng ngược lại là sinh ra video thủ ngữ từ văn bản hoặc âm thanh, nếu có cũng đa phần là ở các ngôn ngữ thông dụng như tiếng Anh, tiếng Đức... Bên cạnh đó, việc thiếu thốn về dữ liệu huấn luyện cũng là một thử thách của hướng nghiên cứu này.

Hiện nay có rất nhiều mô hình được huấn luyện sẵn cho phép chuyển đổi từ giọng nói sang văn bản được thiết kế đặc biệt cho tiếng Việt ra đời như: wav2vec2-base-vietnamese-250h [11] và PhoWhisper [12]. Những mô hình này ngày càng có độ chính xác cao và cung cấp nhiều phiên bản đáp ứng được với những phần cứng khác nhau, điều này giúp đơn giản hóa bài toán sinh ra video thủ ngữ từ âm thanh thành bài toán sinh ra thủ ngữ từ văn bản. Bên cạnh đó, việc hiểu thủ ngữ có thể thực hiện một cách đơn giản hơn dựa vào các mô hình nhận diện khung xương người được đề xuất gần đây như AlphaPose [13] mà không cần phải huấn luyện lại. Do đó, trong nghiên cứu này, chúng tôi đề xuất xây dựng phần mềm phiên dịch dành cho người khiếm thính ở Việt

Nam, kết hợp xử lý ngôn ngữ tự nhiên và các mô hình học sâu đã được huấn luyện sẵn này nhằm mục đích giải quyết vấn đề hạn chế về dữ liệu huấn luyện. Từ đó, hệ thống sẽ giúp người khiếm thính thuận tiện hơn khi tiếp nhận sự hỗ trợ ở các cơ sở công cộng như bệnh viện, siêu thị, nhà hàng, khách sạn,.... Đóng góp chính của bài báo bao gồm các nội dung như bên dưới:

- Xây dựng phần mềm chuyển đổi văn bản hoặc âm thanh thành video ngôn ngữ ký hiệu cho người Việt.
- Hệ thống phiên dịch theo ý nghĩa của từ và cụm từ thay vì từng ký tự riêng lẻ trong những nghiên cứu gần nhất.
- Hệ thống có thể tận dụng các mô hình đã được huấn luyện sẵn với các mục đích khác nhau mà không cần phải mất thời gian huấn luyện lại trên tập dữ liệu thu được.

2. Hệ thống đề xuất



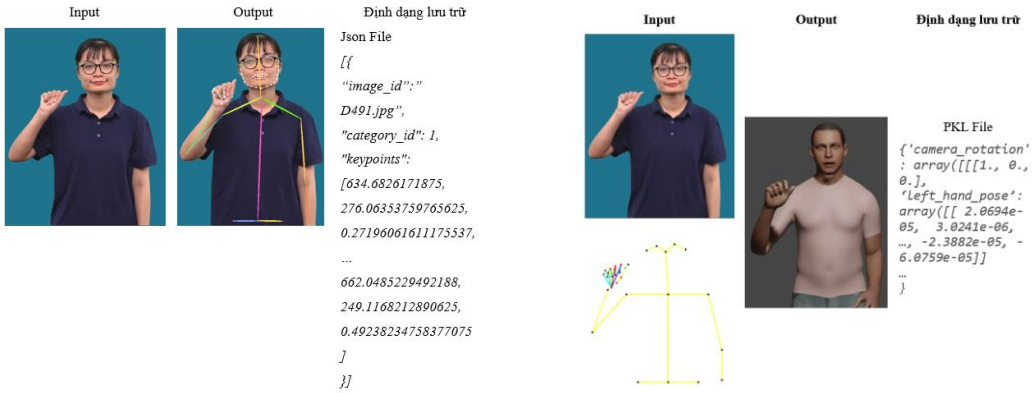
Hình 1. *Sơ đồ khối hệ thống*

Sơ đồ khối hệ thống được biểu diễn như Hình 1, thông tin về từng khối như sau:

- Khối tiếp nhận đoạn phim: chia các đoạn phim, từ dự án "Nâng cao chất lượng giáo dục học sinh khiếm thính cấp tiểu học thông qua ngôn ngữ ký hiệu (QIPEDC)" [14], thành các nhóm frame làm đầu vào cho khối trích xuất khung xương.
- Khối trích xuất khung xương: sử dụng mô hình phát hiện các bộ phận trên cơ thể làm đầu vào cho khối mô hình hóa cơ thể. Dựa theo thông tin từ Bảng 1, chúng tôi chọn AlphaPose [13] là mô hình chính để trích xuất khung xương cơ thể người. Quá trình này được mô tả như Hình 2a.
- Khối mô hình hóa cơ thể: tạo những ảnh mô phỏng tư thế con người dựa trên tọa độ khung xương trích xuất được và lưu trữ lại. Với đầu vào là khung xương và ảnh đối tượng, khối này sẽ tạo ra bản thể mô phỏng dưới dạng lưới và lưu trữ lại trong database như Hình 2b. Với yêu cầu về một hệ thống có thể mô hình hóa chi tiết cơ thể đặc biệt là vùng khớp tay và ngón tay, chúng tôi đã khoanh vùng được các phiên bản của SMPL [15] từ đó đưa ra kết quả so sánh như Bảng 2. Để mô hình hóa được toàn bộ cơ thể bao gồm phần thân, tay, mặt với độ lỗi giữa các khớp xương thấp, chúng tôi đã quyết định chọn model SMPL-X cho khối mô hình hóa cơ thể.
- Khối tiếp nhận văn bản/ âm thanh: đưa văn bản hoặc âm thanh từ người dùng vào hệ thống.

Bảng 1. *So sánh AP AlphaPose và mô hình khác trên tập COCO test-dev 2015 [13]*

Method	AP @0,5:0,95	AP @0,5	AP @0,75	AP medium	AP large
OpenPose (CMU-Pose)	61,8	84,9	67,5	57,1	68,2
Detectron (Mask R-CNN)	67,0	88,0	73,1	62,2	75,6
AlphaPose	**73,3**	**89,2**	**79,1**	**69,0**	**78,6**



(a) Khối trích xuất khung xương *(b) Khối mô hình hóa cơ thể*

Hình 2. *Đầu vào, đầu ra và định dạng lưu trữ*

- Khối chuyển âm thanh sang văn bản: nếu đầu vào là âm thanh sẽ được chuyển đổi thành văn bản trước. Ở đây chúng tôi chọn mô hình PhoWhisper [12] dựa vào khảo sát tại Bảng 3.

- Khối chuyển đổi ngữ pháp: chuyển đổi ngữ pháp ngôn ngữ nói viết sang ngôn ngữ ký hiệu làm đầu vào cho khối dựng tổng hợp ngõ ra. Phần mã của khối chuyển đổi ngữ pháp được xây dựng dựa trên đặc điểm nghiên cứu về ngữ pháp của ngôn ngữ ký hiệu [16]. Với ví dụ "Tôi rất yêu động vật", đại từ "Tôi" sẽ được ưu tiên ở vị trí đầu tiên, trạng từ "rất" sẽ được loại bỏ, danh từ "động vật" sẽ được đưa lên trước động từ "yêu" nhằm nhấn mạnh câu như được thể hiện trong Hình 3. Để thực hiện được tác vụ này, các phương pháp truyền thống thường tách thành nhiều bước như: tách từ (tokenize), phân đoạn từ, gắn thẻ (part-of-speech), chuyển đổi ngữ pháp. Hiệu suất của từng bước sẽ ảnh hưởng đến hiệu suất tổng thể của hệ thống. Do đó, chúng tôi đề xuất huấn luyện mô hình chuyên dụng cho việc xử lý ngôn ngữ tự nhiên là Transformer để thực hiện nhiệm vụ này một cách trực tiếp. Ở đây, chúng tôi lựa chọn sử dụng mô hình ViT5 [17] dựa vào khảo sát tại Bảng 4. Vì mục tiêu chỉ là thay đổi cấu trúc câu mà không làm thay đổi ý nghĩa và mô hình không tốn quá nhiều thời gian để sinh ra câu mới, nên một mô hình đơn giản như ViT5 là lựa chọn phù hợp.

Bảng 2. *So sánh sai số khớp của SMPL-X và mô hình khác [15]*

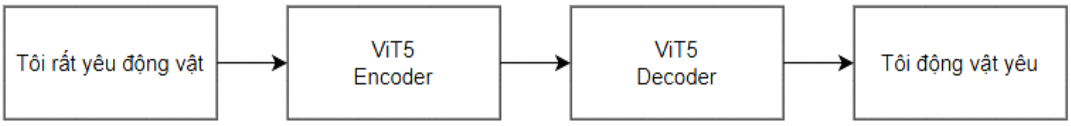
Mô hình	Các điểm khớp	Sai số khớp
SMPL	Cơ thể	63,5
SMPL-H	Cơ thể + Bàn tay + Khuôn mặt	71,7
SMPL-X	Cơ thể + Bàn tay + Khuôn mặt	**62,6**

Bảng 3. *So sánh khả năng nhận diện giọng nói tiếng Việt của các mô hình [12]*

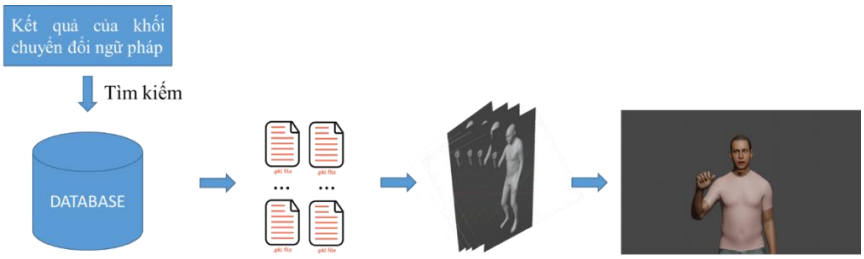
Mô hình	Tỷ lệ lỗi từ			
	CMV–Vi	**VIVOS**	**VLSP Task-1**	**VLSP Task-2**
wav2vec2-base-vietnamese-250h	102,04	10,83	21,02	50,35
wav2vec2-base-vi-vlsp2020	103,71	9,09	16,82	44,91
$\text{PhoWhisper}_{\text{tiny}}$	19,05	10,41	20,74	49,85
$\text{PhoWhisper}_{\text{small}}$	**11,08**	**6,33**	**15,93**	**32,96**

Bảng 4. *So sánh khả năng tạo sinh ngôn ngữ tiếng Việt của các mô hình [17]*

Mô hình	ROUGE-1	ROUGE-2	ROUGE-L
PhoBERT2PhoBERT	60,37	29,12	39,44
mBERT2mBERT	59,67	27,36	36,73
mBART	59,81	28,28	38,71
BARTpho	61,14	30,31	40,15
ViT5	**63,37**	**34,24**	**43,55**



Hình 3. *Khối chuyển đổi ngữ pháp*



Hình 4. *Quá trình tạo ra video của khối tổng hợp ngõ ra*

- Khối dựng tổng hợp ngõ ra: kết hợp văn bản đã được chuyển đổi ngữ pháp và kho dữ liệu mô hình của cơ thể để dựng đoạn phim biểu diễn cho văn bản đầu vào. Sau khi nhận được văn bản đã được chuyển đổi theo ngữ pháp của ngôn ngữ ký hiệu, khối sẽ truy xuất vào từ điển để lấy các tệp đã mô hình hóa trong database, cuối cùng sử dụng phần mềm đồ họa Blender để kết hợp các file lại và tạo ra video. Khối tổng hợp đầu ra được tự động hóa bằng Blender Python API, do đó chúng ta có thể kết xuất được video ngõ ra như mô tả ở Hình 4.

3. Kết quả triển khai thực nghiệm

3.1. Tiêu chí đánh giá

Để đánh giá độ chính xác của khối chuyển đổi ngữ pháp chúng tôi sử dụng tỷ lệ lỗi từ (WER) được tính theo phương trình (1).

$$WER = \frac{S+D+I}{N} \quad (1)$$

trong đó, S là số lần thay thế, D là số lần xóa, I là số lần chèn cần thiết để chuyển một câu thành câu khác và N là tổng số ký tự trong câu.

Ngoài ra, để đánh giá độ chính xác của mô hình 3D được tạo ra, chúng tôi sử dụng các mô hình trích xuất khung xương để xác định độ tương đồng giữa các điểm đặc trưng trên cơ thể trích xuất từ hình ảnh gốc và hình ảnh đầu ra. Chúng tôi đề xuất hai phương pháp đánh giá chính:

Khoảng cách Euclidean (D_E): với hai điểm đặc trưng tương đồng A và B có tọa độ (X_A,Y_A) và (X_B,Y_B) tương ứng trên hình gốc và hình ảnh tái tạo, D_E được tính theo công thức (2).

$$D_E(A,B) = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2} \quad (2)$$

Khoảng cách tương đối theo điểm tham chiếu (D_R): đánh giá độ chính xác dựa trên vị trí tương đối của các điểm khung xương, giúp giảm thiểu ảnh hưởng của tỷ lệ và góc nhìn. Chúng tôi sẽ đặt lại gốc tọa độ tại điểm ngực $C(X_{Centre},Y_{Centre})$ trên cơ thể rồi tính lại tọa độ các điểm từ tọa độ mới, sau đó so sánh sự khác biệt giữa hình đầu vào với hình tái tạo trên tọa độ mới này như được thể hiện trong công thức (3).

$$D_R = |D_E(A,C_A) - D_E(B,C_B)| \quad (3)$$

3.2. Dữ liệu thử nghiệm

Để đánh giá khối chuyển đổi ngữ pháp giữa tiếng Việt và ngôn ngữ ký hiệu, nhóm sử dụng tập dữ liệu Corpus-Vie-VSL-10k [18] bao gồm 10.000 câu tiếng Việt. Mỗi câu trong tập dữ liệu này được chú thích với các nhãn tương ứng như được thể hiện trong Bảng 5, giúp cải thiện độ chính xác của các mô hình dịch và hỗ trợ người khiếm thính trong giao tiếp hàng ngày.

Bảng 5. *Mô tả tập dữ liệu Corpus-Vie-VSL-10k*

Câu đầu vào	Nhãn
Tôi 19 tuổi.	Tôi tuổi 19.
Bạn tên gì?	Bạn tên gì?
Ai biết bơi?	Biết bơi ai?
Con gà ăn gì?	Con gà ăn gì?
Mít thì ngọt.	Mít ngọt.

Để đánh giá hiệu suất chuyển đổi từ âm thanh sang giọng nói với ngôn ngữ là tiếng Việt, nhóm sử dụng tập dữ liệu LSVSC [19]. Đây là một bộ ngữ liệu tiếng Việt quy mô lớn, bao gồm 100,5 giờ ghi âm với đa dạng về chủ đề như tin tức, đọc sách, sách nói, phim ảnh, thể thao, chăm sóc sức khỏe, giao thông và du lịch.

Bên cạnh đó, để đánh giá độ chính xác của những hình ảnh được sinh ra, chúng tôi tự tạo một tập dữ liệu khoảng 3.000 hình ảnh cho mỗi nhãn [20]. Như được thể hiện trong Hình 5, hình bên trái là ảnh gốc, còn hình bên phải là ảnh được tạo ra từ mô hình. Không phải lúc nào quá trình tạo ảnh cũng hoàn hảo, có những khung hình bị sai lệch so với bản gốc. Vì vậy, chúng tôi tính toán độ tương đồng giữa hai bức ảnh, xác định những trường hợp không đạt yêu cầu và xử lý chúng một cách hiệu quả. Bằng cách này, chúng tôi không chỉ đảm bảo chất lượng dữ liệu mà còn nâng cao độ chính xác của mô hình, giúp hệ thống hoạt động ổn định hơn.

Hình 5. *Mô tả tập dữ liệu [20]*

3.3. Kết quả thực nghiệm của hệ thống

3.3.1. Đánh giá khối chuyển đổi ngữ pháp

Chúng tôi chia dữ liệu huấn luyện và đánh giá theo tỷ lệ 8:2 trên tập Corpus-Vie-VSL-10k. Thông qua kết quả trong Bảng 6, tỷ lệ lỗi từ (WER) hiện tại đã cải thiện được đáng kể, sai số chỉ còn 2% so với phương pháp truyền thống. Chúng tôi đã tiến hành phân tích và tìm thấy những nguyên nhân chính gây lỗi cao trong mô hình "Xử lý theo phương pháp truyền thống" là do chứa quá nhiều bước xử lý trung gian. Điều này sẽ gây ra việc tích lũy và tăng dần sai số qua từng bước. Chính vì lẽ đó, việc thiết kế một hệ thống toàn diện (end-to-end) không thông qua các bước trung gian sẽ mang lại hiệu suất tốt hơn.

Bảng 6. *WER của mô hình chuyển đổi ngữ pháp trên tập dữ liệu Corpus-Vie-VSL-10k*

Mô hình	WER (%)
Phương pháp đề xuất (ViT5)	**2,0**
Xử lí theo phương pháp truyền thống [16]	60,68

3.3.2. Đánh giá mô hình chuyển đổi giọng nói sang văn bản

Từ kết quả so sánh trong Bảng 7, chúng ta có thể nhận ra rằng $\text{Whisper}_{\text{small}}$ là mô hình tổng quát, được thiết kế cho nhiều ngôn ngữ, bao gồm tiếng Việt. Tuy nhiên, do chưa được tinh chỉnh chuyên sâu cho tiếng Việt, hiệu suất của mô hình này chưa đạt mức tối ưu trên bộ dữ liệu đánh giá. wav2vec2-base-vietnamese-250h có hiệu suất tốt hơn các mô hình Whisper gốc, nhưng vẫn

chưa vượt qua PhoWhisper$_{small}$ về độ chính xác. PhoWhisper$_{small}$ đạt WER thấp nhất (23,61%), khẳng định rằng sự lựa chọn mô hình PhoWhisper làm mô hình chuyển đổi giọng nói chính để xử lý âm thanh tiếng Việt đã mang lại hiệu suất tốt.

Bảng 7. *Bảng so sánh mô hình chuyển đổi giọng nói sang văn bản trên bộ dữ liệu LSVSC*

Mô hình	WER (%)
wav2vec2-base-vietnamese-250h [11]	24,13
Whisper$_{small}$ [21]	31,89
PhoWhisper$_{small}$ (**phương pháp đề xuất**)	**23,61**

3.3.3. Đánh giá mô hình dựa trên so sánh khung xương giữa ảnh đầu ra và ảnh thực tế

Trong Bảng 8, chúng tôi đánh giá trên toàn bộ 133 điểm khớp của cơ thể và nhận thấy rằng một số điểm có thể nằm ngoài khung hình. Kết quả cho thấy AlphaPose vẫn xử lý tốt hơn so với MMPose và DWPose, với sai số trên cả hai phương pháp tính khoảng cách đều thấp hơn. Điều này chứng tỏ AlphaPose có độ chính xác cao và khả năng định vị các điểm khớp ổn định, ngay cả trong những tình huống khó. AlphaPose đặc biệt hiệu quả trong các trường hợp cơ thể không xuất hiện đầy đủ trong khung hình, giúp giảm đáng kể sai số tổng thể. Ngược lại, MMPose và DWPose mặc dù mới ra đời nhưng tỏ ra kém ổn định với sai số dao động lớn. Bên cạnh đó, độ lệch khung xương giữa ảnh đầu ra và ảnh thực tế rất nhỏ trên cả hai phương pháp đánh giá cũng chứng minh sự hiệu quả của phương pháp mô hình hóa cơ thể SMPL-X.

Bảng 8. *Tính trên tất cả các điểm trên cơ thể của tập dữ liệu chúng tôi tự thu thập*

Phương pháp	AlphaPose (phương pháp đề xuất)	MMPose [22]	DWPose [23]
Khoảng cách Euclidean	**0,1299**	0,4099	0,1768
Khoảng cách tương đối theo điểm tham chiếu	**0,0713**	0,1691	0,3227

3.3.4. Thời gian xử lí

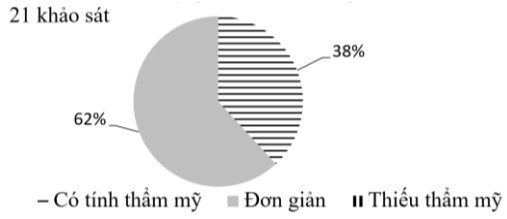
Bảng 9 thể hiện tốc độ xử lý được đo khi chạy trên phần cứng GeForce RTX3090 và Intel core i9-10900K CPU 3.7GHz x20 với cài đặt ngõ ra: độ phân giải 1920x1080 và tỉ lệ khung hình 60 fps. Kết quả cho thấy thời gian xử lý cao chủ yếu là do khối tổng hợp ngõ ra. Để giảm thời gian này, chúng ta có thể thay đổi các yếu tố sau: giảm số khung hình kết xuất, giảm độ phân giải của video ngõ ra, thay đổi phần cứng sử dụng cho quá trình kết xuất hình ảnh.

Bảng 9. *Thời gian xử lý của từng khối*

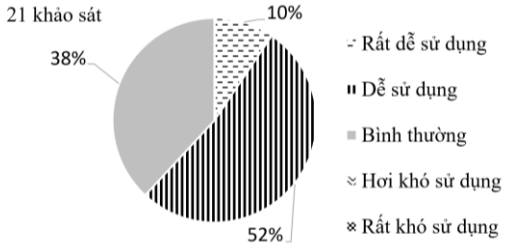
Tên khối	Khối tiếp nhận văn bản/ âm thanh	Khối chuyển âm thanh thành văn bản	Khối chuyển đổi ngữ pháp	Khối tổng hợp ngõ ra
Thời gian (s)	0,588	1,0522	0,1255	19,368
Tổng (s)		21,1337		

3.3.5. Khảo sát trải nghiệm người dùng

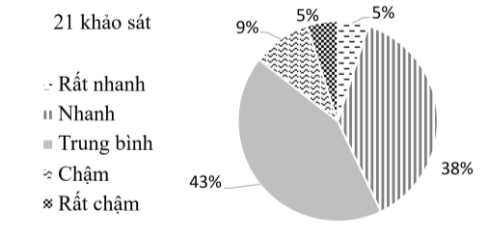
Chúng tôi đã thực hiện khảo sát trên 21 người đã từng trải nghiệm hệ thống với độ tuổi từ 20 đến 32, mục đích là lấy ý kiến đóng góp để hoàn thiện hệ thống được tốt hơn. Nội dung khảo sát xoay quanh các vấn đề về tính thẩm mỹ, tính tiện dụng giao diện của hệ thống, hoạt động của hệ thống và cảm nhận của người dùng về hệ thống. Thông qua các kết quả thể hiện trên Hình 6, nhóm tổng hợp lại ý kiến chung như sau: hầu hết người dùng cảm thấy hệ thống có giao diện còn đơn giản tuy nhiên dễ sử dụng, tốc độ khi hoạt động là trung bình và tổng quan thì hệ thống được đánh giá là hay và hữu ích.



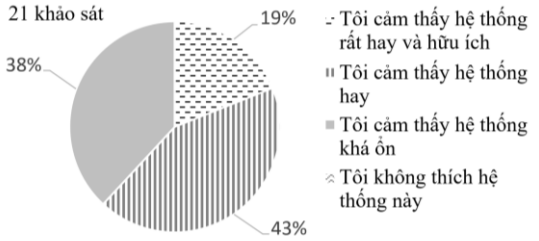
(a) Anh. Chị cảm thấy giao diện hệ thống như thế nào?



(b) Sau khi trải nghiệm hệ thống anh/chị cảm thấy hệ thống có dễ sử dụng không?



(c) Sau khi trải nghiệm hệ thống anh/chị cảm thấy tốc độ hoạt động hệ thống như thế nào?



(d) Cảm nhận chung về hệ thống của anh/chị?

Hình 6. *Kết quả khảo sát*

4. Kết luận

Trong bài báo này, chúng tôi đã đề xuất xây dựng và phát triển một phần mềm chuyển đổi văn bản hoặc giọng nói thành các video giúp người sử dụng ngôn ngữ ký hiệu có thể hiểu một cách dễ dàng. Phần mềm đã giải quyết được yêu cầu đặt ra về việc phiên dịch theo từ, cụm từ thay vì từng kí tự đơn lẻ. Bên cạnh đó, hệ thống có thể tận dụng các mô hình đã được huấn luyện sẵn mà không cần phải huấn luyện lại quá nhiều trên tập dữ liệu mới. Tuy nhiên, hệ thống còn có tốc độ xử lý chưa cao vì việc tổng hợp ngõ ra tốn quá nhiều thời gian. Trong tương lai, chúng tôi sẽ tiếp tục cải thiện về vấn đề này bằng cách áp dụng các kỹ thuật kết xuất hình ảnh gọn nhẹ hơn.

Lời cám ơn

Nghiên cứu này được tài trợ bởi Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh với mã số đề tài T2024-147.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] General Statistics Office of Viet Nam, "Vietnam National survey on people with disabilities 2016," 2019. [Online]. Available: https://www.gso.gov.vn/en/data-and-statistics/2019/03/vietnam-national-survey-on-people-with-disabilities-2016. [Accessed Aug. 16, 2024].

[2] S. Savla, "Real-time Continuous Transcription with Live Transcribe," *Google Research*, 2019. [Online]. Available: https://research.google/blog/real-time-continuous-transcription-with-live-transcribe/. [Accessed Aug. 16, 2024]

[3] HandTalk, "Discover the largest Sign Language translation platform in the world," *HandTalk*, 2024. [Online]. Available: https://www.handtalk.me/en. [Accessed Aug. 16, 2024].

[4] A. Le, "Going hand in glove with sign language," *Viet Nam News*, April 09, 2017. [Online]. Available: https://vietnamnews.vn/sunday/features/374025/going-hand-in-glove-with-sign-language.html. [Accessed Aug. 16, 2024].

[5] L. D. Quach, H. D. K. Nguyen, and C. N. Nguyen, "Converting the Vietnamese Television News into 3D Sign Language Animations for the Deaf," *4th EAI International Conference on Industrial Networks and Intelligent Systems*, Da Nang, Vietnam, 2018, pp. 155-163.

[6] R. K. Pathan, M. Biswas, S. Yasmin, *et al.*, "Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network," *Sci. Rep.*, vol.13, 2023, Art. no. 16975.

[7] J. Zhang, X. Bu, Y. Wang, H. Dong, Y. Zhang, and H. Wu, "Sign language recognition based on dual-path background erasure convolutional neural network," *Sci. Rep.*, vol. 14, 2024, Art. no. 11360.

[8] N. F. Attia, M. T. F. S. Ahmed, and M. A. M. Alshewimy, "Efficient deep learning models based on tension techniques for sign language recognition," *Intelligent Systems with Applications*, vol. 20, 2023, Art. no. 200284.

[9] Y. Liu, P. Nand, M. A. Hossain, M. Nguyen, and W. Q. Yan, "Sign language recognition from digital videos using feature pyramid network with detection transformer," *Multimedia Tools and Applications*, vol. 82, pp. 21673–21685, 2023.

[10] Y. Li, N. Miao, L. Ma, F. Shuang, and X. Huang, " Transformer for object detection: Review and benchmark," *Engineering Applications of Artificial Intelligence*, vol. 126, 2023, Art. no. 107021.

[11] T. B. Nguyen, "Vietnamese end-to-end speech recognition using wav2vec 2.0," 2021. [Online]. Available: https://github.com/vietai/ASR. [Accessed Mar. 3, 2025].

[12] T.-T. Le, T. L. Nguyen, and Q. D. Nguyen, "PhoWhisper: Automatic Speech Recognition for Vietnamese," *Proceedings of the ICLR 2024 Tiny Papers track*, 2024, pp. 1-3.

[13] H. Fang *et al.*, "AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, pp. 7157-7173.

[14] QIPEDC Project, "Vietnam Quality Improvement of Primary Education for Deaf Children Project," 2022. [Online]. Available: https://qipedc.moet.gov.vn/. [Accessed Aug. 16, 2024].

[15] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A Skinned Multi-Person Linear Model," *ACM Transactions on Graphics*, vol. 34, pp. 1-16, 2015.

[16] T. M. H. Nguyen, T. T. L. Hoang, and X. L. Vu, "Guidelines for Word Unit Recognition in Vietnamese Text," (in Vietnamese), 2009. [Online]. Available: https://www.jaist.ac.jp/~bao/VLSP-text/Mar2009/SP82_%20Baocaokythuat_2009thang3.pdf. [Accessed Aug. 16, 2024].

[17] L. Phan, H. Tran, H. Nguyen, and T. H. Trinh, "ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation," *in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 2022, pp. 136–142.

[18] T. B. D. Nguyen, "Parallel-Corpus-Vie-VSL," 2024. [Online]. Available: https://github.com/BichDiep/Parallel-Corpus-Vie-VSL. [Accessed Mar. 24, 2025]

[19] T. T. L. Tran, H.-G. Kim, M. H. La, and V. S. Pham, "Automatic Speech Recognition of Vietnamese for a New Large-Scale Corpus," *Electronics*, vol. 13, no. 5, Art. no.977, 2024.

[20] Q. D. Le, "Datasets," 2025. [Online]. Available: https://drive.google.com/drive/folders/1XO69X3Kjlr6m_Y37EmIJEGUz_GQ_Lybu?usp=drive_link. [Accessed Mar. 3, 2025].

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, pp. 28492-28518, 2023.

[22] MMPose Contributors, "MMPOSE: OpenMMLab Pose Estimation Toolbox and Benchmark," 2024. [Online]. Available: https://github.com/open-mmlab/mmpose. [Accessed Mar. 3, 2025].

[23] Z. Yang, A. Zeng, C. Yuan, and Y. Li, "Effective Whole-body Pose Estimation with Two-stages Distillation," *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4210-4220.