THÔNG TIN CHUNG

- Link YouTube video của báo cáo (tối đa 5 phút):
  https://www.youtube.com/watch?v=CkoxAwQSZdE

- Link slides (dạng .pdf đặt trên Github của nhóm):
  https://github.com/chihqqc20-subrosa/Research-Methodology/blob/main/Chi%20Ho%C3%A0ng%20Quy%20Qu%E1%BB%B3nh%20-%20CS2205.SEP2025.DeCuong.FinalReport.Template.Slide.pdf

- Link Github:
  https://github.com/chihqqc20-subrosa/Research-Methodology

- Họ và Tên:
  Hoàng Quy Quỳnh Chi
- MSHV: 250201003

- Lớp: CS2205.RM
- Tự đánh giá (điểm tổng kết môn): 9.5/10
- Số buổi vắng: 0

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

VI-SIGN: HỆ THỐNG CHUYỂN ĐỔI TIẾNG NÓI TIẾNG VIỆT SANG NGÔN NGỮ KÝ HIỆU DỰA TRÊN MÔ HÌNH NGÔN NGỮ LỚN CHO NGƯỜI KHIẾM THÍNH

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

VI-SIGN: A LARGE LANGUAGE MODEL–ASSISTED VIETNAMESE SPEECH-TO-SIGN LANGUAGE TRANSLATION SYSTEM FOR DEAF COMMUNICATION

TÓM TẮT *(Tối đa 400 từ)*

Trong lĩnh vực chuyển đổi tiếng nói sang ngôn ngữ ký hiệu, Ngôn ngữ ký hiệu Việt Nam (VSL) vẫn chưa được nghiên cứu và khai thác đầy đủ. Nguyên nhân chính là thiếu các tập dữ liệu đa phương thức được căn chỉnh chặt chẽ, đồng thời nhiều hướng tiếp cận trước đây phụ thuộc vào giám sát gloss. Bên cạnh đó, đặc thù tiếng Việt như tính thanh điệu và cú pháp linh hoạt làm tăng độ phức tạp khi ánh xạ từ lời nói sang chuyển động ký hiệu liên tục.

Các phương pháp hiện có tại Việt Nam chủ yếu dựa trên tra cứu từ điển hoặc hoạt hình ký hiệu thủ công. Cách tiếp cận này thường đảm bảo nội dung "dịch gì" nhưng chưa mô hình hóa tốt "dịch như thế nào" để tự nhiên trong giao tiếp. Do đó, chuyển động sinh ra thường rời rạc, thiếu liên tục và chưa bám sát nhịp điệu lời nói. Ngoài ra, các yếu tố phi thủ công như nét mặt, hướng nhìn và cử động đầu - những thành phần quan trọng để truyền tải ngữ nghĩa và cảm xúc - chưa được thể hiện đầy đủ.

Để khắc phục các hạn chế trên, luận văn đề xuất Vi-Sign, hệ thống chuyển đổi tiếng nói/văn bản tiếng Việt sang ngôn ngữ ký hiệu theo hướng không cần giám sát gloss và có khả năng nhận biết ngữ điệu. Vi-Sign kết hợp căn chỉnh văn bản–video ở mức đơn vị từ vựng, trích xuất đặc trưng chuyển động và mô-đun chuyển văn bản thành tín hiệu điều khiển với sự hỗ trợ của mô hình ngôn ngữ lớn nhằm dự đoán nhịp điệu và mức biểu cảm. Các tín hiệu này được tinh chỉnh theo thời gian để tạo chuyển động ký hiệu mượt và đồng bộ với lời nói.

Luận văn đồng thời xây dựng tập dữ liệu khoảng 6.000 câu dạng tiếng nói–văn bản–ngôn ngữ ký hiệu để phục vụ huấn luyện và đánh giá. Kết quả dự kiến cho thấy cải thiện về độ mượt, độ ổn định và mức đồng bộ ngữ điệu–cử chỉ. Qua đó, Vi-Sign hướng tới việc nâng cao khả năng tiếp cận giao tiếp cho cộng đồng người khiếm thính trong giáo dục và các bối cảnh thông tin công cộng.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Trong các bối cảnh giáo dục, y tế, môi trường làm việc và truyền thông công cộng, người khiếm thính tại Việt Nam vẫn gặp nhiều khó khăn trong việc tiếp cận thông tin bằng lời nói. Nguyên nhân chủ yếu là sự hạn chế trong việc cung cấp phiên dịch ngôn ngữ ký hiệu cho các hoạt động giảng dạy, dịch vụ và thông tin công cộng. VSL hiện hiếm khi được sử dụng một cách chính thức và rộng rãi trong các bối cảnh nói trên. Thực trạng đó làm gia tăng khoảng cách tiếp cận thông tin và hạn chế cơ hội hòa nhập xã hội của người khiếm thính. Vì vậy, việc nghiên cứu các giải pháp công nghệ hỗ trợ chuyển đổi tiếng nói tiếng Việt sang ngôn ngữ ký hiệu có ý nghĩa xã hội và thực tiễn rõ rệt.

Những năm gần đây, các nghiên cứu về chuyển đổi tiếng nói và văn bản sang ngôn ngữ ký hiệu đã đạt được nhiều tiến bộ đối với các ngôn ngữ ký hiệu có nguồn lực lớn như ASL, BSL và DGS. Tuy nhiên, các hướng tiếp cận này chưa được áp dụng hiệu quả cho VSL do sự khác biệt về ngôn ngữ, tính thanh điệu của tiếng Việt và hạn chế về dữ liệu. Các nghiên cứu hiện có tại Việt Nam chủ yếu dựa trên quy tắc gloss đơn giản, video ký hiệu rời rạc hoặc hoạt hình ký hiệu thủ công, vốn khó tái hiện chuyển động liên tục, nhịp điệu và yếu tố biểu cảm. Đồng thời, phần lớn các phương pháp này phụ thuộc vào gloss supervision hoặc dữ liệu ghi chuyển động chuyên dụng, vốn khan hiếm đối với VSL. Do đó, vẫn còn thiếu một hệ thống chuyển đổi tiếng nói và văn bản sang VSL có khả năng sinh chuyển động ký hiệu tự nhiên, đồng bộ ngữ điệu và không cần giám sát gloss.

Một thách thức cốt lõi trong bài toán này là chất lượng và mức độ căn chỉnh của các tài nguyên hiện có cho VSL chưa đáp ứng yêu cầu mô hình hóa theo câu. Các nguồn dữ liệu phổ biến chủ yếu gồm video ký hiệu đơn lẻ, chưa có liên kết chặt với câu tiếng Việt, thông tin ngữ điệu và quỹ đạo chuyển động chi tiết. Hạn chế này làm giảm khả năng huấn luyện các mô hình sinh chuyển động ký hiệu liên tục theo hướng đầu-cuối và khó đánh giá một cách nhất quán. Do đó, luận văn xây dựng một tập dữ liệu tiếng Việt dạng tiếng nói–văn bản–ngôn ngữ ký hiệu với âm thanh chất lượng cao và căn chỉnh ở mức đơn vị từ vựng. Tập dữ liệu này đóng vai trò nền tảng cho việc học mối liên hệ giữa nội dung ngôn ngữ và chuyển động ký hiệu.

Trên cơ sở đó, luận văn đề xuất Vi-Sign, một hệ thống chuyển đổi tiếng nói tiếng Việt sang ngôn ngữ ký hiệu theo hướng không cần giám sát gloss và có khả năng nhận biết ngữ điệu. Đầu vào của hệ thống là tiếng nói hoặc văn bản tiếng Việt, trong khi đầu ra là chuỗi chuyển động ngôn ngữ ký hiệu Việt Nam biểu diễn dưới dạng quỹ đạo tư thế liên tục. Hệ thống tích hợp xử lý tiếng nói và mô hình ngôn ngữ lớn để dự đoán các tín hiệu điều khiển phục vụ sinh chuyển động. Các tín hiệu này được kết hợp với mạng tinh chỉnh chuyển động theo thời gian nhằm đảm bảo tính liên tục và ổn định của ký hiệu. Cách tiếp cận này góp phần cải thiện khả năng tiếp cận thông tin bằng lời nói cho người khiếm thính trong giáo dục và truyền thông công cộng.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

1. Xây dựng bộ dữ liệu tiếng Việt dạng Tiếng nói–Văn bản–Ngôn ngữ ký hiệu (Speech–Text–Sign) gồm khoảng 6.000 câu, có căn chỉnh theo đơn vị từ vựng, âm thanh chất lượng cao và quỹ đạo tư thế đa phương thức, phục vụ huấn luyện và đánh giá mô hình sinh chuyển động VSL liên tục.
2. Thiết kế và phát triển hệ thống Vi-Sign chuyển đổi tiếng nói/văn bản tiếng Việt sang chuyển động ngôn ngữ ký hiệu liên tục, theo hướng không cần giám sát gloss và có xét ngữ điệu, kết hợp mô-đun dự đoán tín hiệu điều khiển dựa trên mô hình ngôn ngữ lớn và mạng tinh chỉnh chuyển động theo thời gian.
3. Đánh giá hiệu quả hệ thống thông qua các tiêu chí về dự đoán nhịp/ngữ điệu, độ mượt, tính ổn định và độ đồng bộ giữa ngữ điệu và cử chỉ, đồng thời kiểm tra tính nhất quán hình học và cấu trúc của chuyển động sinh ra.

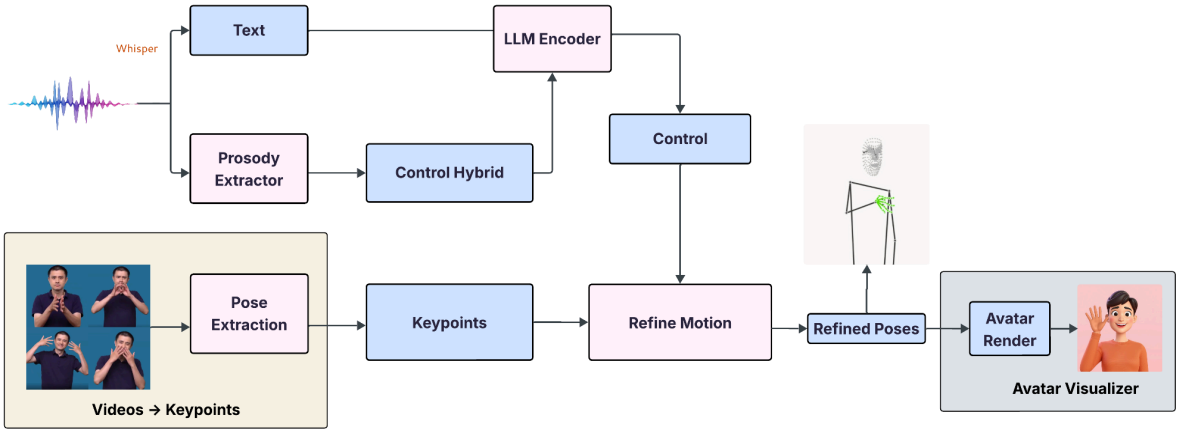
NỘI DUNG VÀ PHƯƠNG PHÁP

1. Nội dung

Luận văn nghiên cứu bài toán chuyển đổi tiếng nói tiếng Việt sang ngôn ngữ ký hiệu Việt Nam theo hướng sinh chuyển động ký hiệu liên tục. Nhu cầu nghiên cứu xuất phát từ việc người khiếm thính khó tiếp cận thông tin bằng lời nói do thiếu phiên dịch trong giáo dục và truyền thông công cộng. Khoảng trống nghiên cứu hình thành vì VSL thiếu dữ liệu đa phương thức được căn chỉnh, gây khó khăn cho việc huấn luyện mô hình sinh chuyển động tự nhiên. Do đó, luận văn tập trung đồng thời vào xây dựng dữ liệu và phát triển hệ thống để thu hẹp khoảng trống này. Kết quả hướng tới là chuyển động ký hiệu đúng nội dung và tự nhiên về nhịp điệu, biểu cảm mà không cần giám sát gloss.

Luận văn xây dựng tập dữ liệu Speech–Text–Sign khoảng 6.000 câu vì dữ liệu căn chỉnh là điều kiện tiên quyết cho huấn luyện và đánh giá. Phần video ký hiệu được thu thập từ các nguồn công khai và được thực hiện bởi một số người thực hiện khác nhau, nhằm phản ánh khác biệt phong cách ký hiệu và tăng khả năng tổng quát hóa của mô hình. Âm thanh được thu ở chất lượng cao nhằm phản ánh chính xác ngữ điệu của tiếng nói. Việc căn chỉnh theo đơn vị từ vựng giúp liên kết nội dung với chuyển động ở mức chi tiết phù hợp cho mô hình học. Quỹ đạo tư thế đa phương thức được lưu trữ vì chuyển động ký hiệu phụ thuộc đồng thời vào tay, cơ thể và khuôn mặt, làm nền tảng để hệ thống Vi-Sign sinh chuyển động ký hiệu liên tục và đồng bộ.

2. Phương pháp



Hình 1. Sơ đồ pipeline tổng quát của hệ thống Vi-Sign cho bài toán chuyển đổi tiếng nói tiếng Việt sang ngôn ngữ ký hiệu liên tục.

Luận văn áp dụng phương pháp nghiên cứu thực nghiệm kết hợp xử lý dữ liệu đa phương thức và học sâu để giải quyết bài toán chuyển đổi tiếng nói tiếng Việt sang ngôn ngữ ký hiệu liên tục. Cách tiếp cận được xây dựng dựa trên pipeline tổng thể của hệ thống Vi-Sign (Hình 1), trong đó mỗi bước xử lý đảm nhiệm một vai trò chức năng cụ thể. Việc mô tả phương pháp theo pipeline giúp làm rõ luồng xử lý và mối quan hệ nhân quả giữa các thành phần. Phương pháp được thiết kế ở mức trừu tượng nhằm đảm bảo tính khả thi trong giai đoạn đề cương. Các chi tiết cài đặt cụ thể sẽ được trình bày trong luận văn hoàn chỉnh.

Dữ liệu đầu vào của hệ thống bao gồm tiếng nói hoặc văn bản tiếng Việt và video ngôn ngữ ký hiệu. Các nguồn dữ liệu này được chuẩn hóa theo ba kênh tiếng nói–văn bản–video nhằm học mối liên hệ giữa nội dung ngôn ngữ và biểu đạt ký hiệu. Video ký hiệu được chuyển đổi sang dạng biểu diễn keypoints để giảm độ phức tạp và thuận lợi cho mô hình hóa chuỗi chuyển động. Tiếng nói được xử lý để trích xuất thông tin ngữ điệu vì nhịp điệu và sự nhấn mạnh ảnh hưởng trực tiếp đến cách ký hiệu. Dữ liệu sau đó được căn chỉnh theo đơn vị từ vựng nhằm đảm bảo tính nhất quán giữa nội dung và chuyển động.

Trên cơ sở dữ liệu đã căn chỉnh, hệ thống tiến hành dự đoán các tín hiệu điều khiển ở mức trừu tượng từ thông tin văn bản và ngữ điệu. Các tín hiệu này đại diện cho nhịp, tốc độ và mức độ biểu cảm của chuyển động ký hiệu. Về mặt lý luận, mô hình LLM được sử dụng như một cơ chế suy luận ngữ cảnh ở cấp câu, giúp nắm bắt trọng tâm thông tin và xu hướng nhấn mạnh của tiếng Việt, đồng thời hỗ trợ chuyển đổi cấu trúc câu sang dạng phù hợp hơn với đặc thù ngữ pháp của VSL. Nhờ đó, hệ thống hình thành ánh xạ từ nội dung ngôn ngữ sang tín hiệu điều khiển mà không cần giám sát gloss và tăng khả năng tổng quát hóa trong điều kiện dữ liệu hạn chế.

Từ các tín hiệu điều khiển và biểu diễn keypoints, hệ thống sinh chuỗi chuyển động ký hiệu thông qua quá trình tinh chỉnh theo thời gian. Bước tinh chỉnh nhằm đảm bảo chuyển động liên tục, mượt mà và ổn định theo nhịp lời nói. Các ràng buộc về tính liên tục theo thời gian và tính nhất quán hình học được áp dụng để hạn chế hiện tượng giật cục hoặc biến dạng chuyển động. Kết quả của bước này là chuỗi tư thế ký hiệu đã được tinh chỉnh và được xem là đầu ra chính của hệ thống.

Cuối cùng, chuyển động ký hiệu được trực quan hóa thông qua mô hình nhân vật ảo để phục vụ đánh giá và minh họa kết quả. Việc đánh giá được thực hiện bằng cách kết hợp các thước đo định lượng và quan sát định tính. Các tiêu chí đánh giá tập trung vào độ mượt, độ ổn định và mức đồng bộ giữa ngữ điệu và cử chỉ. Kết quả đánh giá cho phép phân tích đóng góp của từng thành phần trong pipeline và thảo luận tính khả thi cũng như tiềm năng mở rộng của hệ thống Vi-Sign.

KẾT QUẢ MONG ĐỢI

1. Luận văn dự kiến xây dựng một tập dữ liệu tiếng Việt dạng Speech–Text–Sign có cấu trúc rõ ràng, âm thanh chất lượng cao và căn chỉnh theo đơn vị từ vựng. Tập dữ liệu này nhằm bổ sung nguồn tài nguyên đa phương thức còn thiếu cho nghiên cứu sinh chuyển động ngôn ngữ ký hiệu Việt Nam theo hướng liên tục. Dữ liệu được chuẩn hóa để hỗ trợ huấn luyện, đánh giá và so sánh mô hình một cách nhất quán. Bộ dữ liệu cũng tạo điều kiện cho việc tái lập thí nghiệm và mở rộng nghiên cứu trong tương lai. Đây là kết quả nền tảng phục vụ toàn bộ các mục tiêu còn lại của luận văn.

2. Hệ thống Vi-Sign được kỳ vọng tạo ra chuyển động ký hiệu liên tục từ tiếng nói hoặc văn bản tiếng Việt theo hướng không cần giám sát gloss. Chuyển động sinh ra dự kiến mượt hơn và bám sát nhịp lời nói nhờ mô hình hóa ngữ điệu và sử dụng tín hiệu điều khiển. Hệ thống đồng thời được kỳ vọng phản ánh tốt hơn các yếu tố phi thủ công như nét mặt, hướng nhìn và cử động đầu nhằm tăng khả năng truyền đạt ngữ nghĩa và cảm xúc. Tính khả thi của hướng tiếp cận được thể hiện qua khả năng vận hành theo pipeline hoàn chỉnh từ đầu vào đến đầu ra. Kết quả này là cơ sở để xem xét tiềm năng ứng dụng trong các bối cảnh giáo dục và thông tin công cộng.

3. Về đánh giá, luận văn dự kiến kết hợp thước đo định lượng và quan sát định tính để đo chất lượng chuyển động ký hiệu sinh ra. Các tiêu chí trọng tâm gồm độ mượt, độ ổn định theo thời gian và mức đồng bộ giữa ngữ điệu và cử chỉ, đồng thời kiểm tra tính nhất quán hình học và cấu trúc chuyển động. Kết quả đánh giá sẽ làm rõ đóng góp của từng thành phần trong hệ thống thông qua so sánh với các baseline ablation. Cụ thể, hệ thống sẽ được đối chiếu với (i) phiên bản không dùng tín hiệu điều khiển ngữ điệu/biểu cảm (text-only control) và (ii) phiên bản không có mô-đun tinh chỉnh chuyển động theo thời gian (không refinement). Trên cơ sở đó, luận văn cung cấp bằng chứng thực nghiệm cho hiệu quả thiết kế và định hướng các cải tiến tiếp theo.

TÀI LIỆU THAM KHẢO

[1]. Danielle Bragg, Oscar Koller, Maryam Afsar, et al.:
Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. CHI 2019.

[2]. Dat Quoc Nguyen, Anh Tuan Nguyen:
PhoBERT: Pre-trained Language Models for Vietnamese. Findings of EMNLP 2020: 1037–1046.

[3]. Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, Richard Bowden:
Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. CVPR 2020: 10023–10033.

[4]. Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, Richard Bowden:
Text2Sign: Towards Sign Language Production Using Neural Machine Translation and GANs. IJCV 129 (2021): 2757–2778.

[5]. MediaPipe Team:
MediaPipe Holistic. URL: https://developers.google.com/mediapipe
(accessed 2025-11-13).

[6]. Shaojie Bai, J. Zico Kolter, Vladlen Koltun:
An Empirical Evaluation of Convolutional and Recurrent Networks for Sequence Modeling. arXiv:1803.01271 (2018).

[7]. QIPEDC:
Vietnamese Sign Language Dictionary. URL: https://qipedc.moet.gov.vn/dictionary
(accessed 2025-11-13).