TÓM TẮT

Đề tài "Hệ thống phiên dịch lời nói tiếng Việt thành ngôn ngữ kí hiệu cho người mất khả năng thính lực" được thực hiện tại "trung tâm nghiên cứu và giáo dục người khiếm thính (CED)", từ tháng 7/2020 đến nay

- Nghiên cứu về người Mất thính lực và cách giao tiếp với họ
- Nghiên cứu về ngôn ngữ kí hiệu
- Nghiên cứu công nghệ "Speech to text"
- Nghiên cứu công nghệ "Xử lí ngôn ngữ tự nhiên" trên nền tảng tiếng Việt
- Nghiên cứu phương pháp xây dựng đồ hoạ 3D bằng ngôn ngữ Python

Kết quả thu được:

- ✓ Đưa ra thuật toán giúp nhập văn bản bằng lời nói hoặc thủ công từ bàn phím
- ✓ Xây dựng dữ liệu tương đương giữa ngôn ngữ tiếng Việt và ngôn ngữ kí hiệu
- ✓ Xử lí được dữ liệu lời thoại đầu vào, từ đó đưa ra được các từ khoá cần sử dụng trong việc giao tiếp bằng ngôn ngữ kí hiệu
- ✓ Từ các từ khoá được tạo, tiến hành sử dụng đồ hoạ 3D để mô phỏng ngôn ngữ kí hiệu

MỤC LỤC

DANH SÁCH HÌNH VẼ VÀ ĐỒ THỊ

CHƯƠNG 1
ĐẶT VẤN ĐỀ

1.1 Tính cấp thiết của đề tài

Năm 2010, thế giới có khoảng 250 triệu người điếc. Con số này tăng lên khoảng 360 triệu vào năm 2015. Điều đó cho thấy số lượng người có vấn đề về thính giác ngày càng tăng (theo bà Suchira Prasansuk, chủ tích hội thính học thế giới) [1]. Ở Việt Nam, con số này là 7,3 triệu người vào năm 2017 [2].

Với đặc thù của người Mất thính lực là khả năng nghe hầu như không có, khả năng nói bị ảnh hưởng nặng nề nên hầu như người Mất thính lực không thể giao tiếp bằng lời nói với người bình thường. Từ đó, ngôn ngữ ký hiệu ra đời giúp người Mất thính lực có thể giao tiếp với người khác. Tuy nhiên, trở ngại lớn nhất của họ trong giao tiếp chính là người bình thường không thể hiểu ngôn ngữ ký hiệu này.

Mặc dù đã có một số nỗ lực ở Việt Nam để giúp người Mất thính lực có thể học tập và làm việc như người bình thường, thực tế họ vẫn gặp rất nhiều khó khặn. Khi đi vào các cơ quan công cộng, người Mất thính lực thường gặp trở ngại trong giao tiếp, đặc biệt với những người Mất thính lực không biết chữ. Các dịch vụ thuê người thông dịch cho người Mất thính lực có chi phí quá cao, không phù hợp với điều kiện tài chính của đại đa số người Mất thính lực.

Do số lượng người Mất thính lực ngày càng tăng, việc đáp ứng nhu cầu giao tiếp của họ với cộng đồng ngày càng được quan tâm. Cụ thể, Đài truyền hình Việt Nam (VTV) có một chương trình riêng vào mỗi buổi sáng dành cho người Mất thính lực. Gần đây nhất, đài đã bổ sung một phiên dịch ở khung trái màn hình tivi để giúp người Mất thính lực có thể tiếp thu thông tin hàng ngày. Tuy nhiên điều này khá tốn kinh phí khi nên VTV chỉ có thể hỗ trợ vào khung giờ thời sự. Các đài truyền hình khác vẫn không thể làm điều tương tự vì chi phí quá cao.

Từ thực tế nêu trên, em nhận thấy rằng việc đưa ra một sản phẩm giúp người Mất thính lực dễ dàng hơn trong giao tiếp với chi phí thấp là điều hết sức cần thiết. Ứng dụng công nghệ "Xử lí ngôn ngữ tự nhiên" và các công cụ trong lĩnh vực trí tuệ nhân tạo khác, em đã nghiên cứu thành công dự án "Hệ thống phiên dịch lời nói tiếng Việt thành ngôn ngữ kí hiệu cho người mất khả năng thính lực", với mong muốn rút ngắn khoảng cách với người điếc, khiếm thính. Nhóm người điếc, khiếm thính là nhóm thiểu số đã chịu nhiều thiệt thòi trong xã hội. Em hy vọng rằng đề tài sẽ mang đến một giải pháp khả thi giúp nâng cao chất lượng cuộc sống đáng kể cho người Mất thính lực.

1.2 Ý nghĩa khoa học và thực tiễn của đề tài

Về khoa học, dự án tạo ra công cụ để từ tiếng Việt có thể chuyển sang ngôn ngữ kí hiệu, giúp phát triển các dự án khác cho người điếc, khiếm thính.

Về thực tiễn, công cụ này có thể ứng dụng trên các kênh truyền hình, các khu vực công cộng, giúp người điếc, khiếm thính có thể tiếp thu các nội dung bên ngoài và giảm bớt thiệt thòi cho họ.

1.3 Mục tiêu nghiên cứu của đề tài

- Xây dựng thuật toán chuyển tiếng Việt thành văn bản
- Rút gọn văn bản trên
- Chuyển văn bản rút gọn thành ngôn ngữ kí hiệu
- Mô phỏng ngôn ngữ kí hiệu trên công nghệ 3D

1.4 Đối tượng và phạm vi nghiên cứu

1.4.1 Đối tượng nghiên cứu

Người khiếm thính, người điếc

Ngôn ngữ lập trình python, công nghệ xử lí ngôn ngữ tự nhiên, chuyển giọng nói thành văn bản, công nghệ 3D

1.4.2 Phạm vi nghiên cứu

Nghiên cứu các đối tượng trên phạm vi địa bàn thành phố Hồ Chí Minh

Nghiên cứu thư viện speech_recognition, underthesea, MMD

1.5 Phương pháp nghiên cứu

Nghiên cứu lý thuyết:

- Phương pháp phân tích và tổng hợp lý thuyết
- Phương pháp phân loại và hệ thống hoá lý thuyết
- Phương pháp mô hình hóa
- Phương pháp giả thuyết

Nghiên cứu thực nghiệm:

- Phương pháp quan sát
- Phương pháp chuyên gia
- Phương pháp thực nghiệm khoa học
- Phương pháp phân tích và tổng kết kinh nghiệm

CHƯƠNG 2
TỔNG QUAN ĐỀ TÀI

2.1 Tổng quan về người Mất thính lực

2.1.1 Khả năng của người Mất thính lực

Khiếm thính là tình trạng một người hoặc một động vật có thính giác kém trong khi cá thể khác cùng một loài có thể nghe thấy âm thanh đó dễ dàng [3] [4]. Bệnh do nhiều yếu tố khác nhau, bao gồm tuổi tác, tiếng ồn, bệnh tật, hóa chất và các chấn thương vật lý.

Người Điếc đó là những người không nghe được và không thể nói chuyện được. Thuật ngữ tiếng Anh thì phân biệt rõ từ Deaf (danh từ chung) - viết hoa - dùng chỉ người Điếc. Ngược lại, từ deaf (tính từ) - viết thường – dùng để nói về việc mất thính lực. [3]

Người nghe kém (Hard of Hearing – HoH) được phân biệt như sau: đó là những người bị suy giảm thính lực, nghe khó khăn nhưng vẫn có thể nói chuyện được. Đa số người nghe kém phát hiện bệnh sau một thời gian nghe nói được bình thường.

Cũng có người điếc, do được can thiệp sớm, nên có thể nghe được, dù ít, và đặc biệt là nói chuyện được. Nếu một người nghe kém có thể đọc được tín hiệu môi/ khẩu hình miệng (lip reading) tốt thì khó có thể phân biệt được đó là người nghe kém. Nhưng không phải người nghe kém nào cũng có thể đọc được tín hiệu môi trong tất cả mọi trường hợp, mọi tình huống, để nắm bắt thông điệp từ người khác, và vì họ cũng nói chuyện được bình thường, nên khó ai đoán được khó khăn trong giao tiếp của họ để mà giúp đỡ.

Một người nghe kém nếu được trang bị máy trợ thính và các dụng cụ hỗ trợ (Technical devices) tốt, họ sẽ là người không khuyết tật. Còn một người Điếc, nếu được can thiệp sớm với sự hỗ trợ của máy trợ thính có thể nghe và nói chuyện được, họ sẽ là người nghe kém. Cho nên, thuật ngữ Điếc hay nghe kém chỉ là sự định nghĩa chung. [4]

Theo Tiến sĩ Akio Suemori thuộc Liên Đoàn Người Điếc Nhật Bản, chuyên viên của Liên Đoàn Người Điếc Thế Giới (World Federation of the Deaf-WFD) thì người nghe kém với người điếc được phân biệt qua việc giáo dục. Nếu với người điếc,

ngôn ngữ ký hiệu được dùng để giáo dục, thì với người nghe kém giáo viên có thể dùng ngôn ngữ nói.

Hình 2.1 Trẻ em có vấn đề về thính lực được giáo dục sớm

(Nguồn: Trợ thính Cát Tường)

Theo Hiệp hội Điếc Quốc Gia Hoa Kỳ: "Cộng đồng người Điếc và Nghe kém rất đa dạng, có sự khác nhau rất lớn về nguyên nhân và mức độ mất thính lực, độ tuổi phát bệnh, nền tảng giáo dục, phương pháp giao tiếp, và sự cảm nhận về việc mất thích lực như thế nào? Một người tự gắn cho mình thuật ngữ về sự mất thính lực như thế nào là chuyện cá nhân và có thể phản ánh một sự xác nhận với cộng đồng hay chỉ đơn thuần là việc phản ánh sự mất thính lực ảnh hưởng đến khả năng giao tiếp của họ như thế nào.

Trên thế giới, nhất là ở các nước phát triển, hai thuật ngữ trên được phân biệt rất rõ ràng qua các tên gọi như World Federation of the Deaf (Liên Đoàn Người Điếc Thế Giới), … Liên Đoàn Khiếm thính Quốc tế (International Federation of Hard of Hearing People) hay Liên Đoàn Khiếm thính Trẻ Quốc tế (IHOHYP) … Trong khi tại Việt Nam, và cũng như ở hầu hết các nước Châu Á khác, chỉ có các hội, chi hội hoặc câu lạc bộ của người Điếc. Người nghe kém không lập thành nhóm riêng mà tham gia sinh hoạt chung với người Điếc hoặc sống hòa nhập. [3]

Vậy có thể thấy, việc giao tiếp sẽ giúp người mất khả năng khiếm thính phát triển tư duy, hoà nhập được với cuộc sống.

2.1.2 Phương pháp giao tiếp của người Mất thính lực

Ở người Điếc, thị giác và xúc giác là phương tiện chính để cảm nhận và định hướng không gian. Hơn nữa, việc sử dụng ngôn ngữ kí hiệu từ lâu đã góp phần tạo nên văn hóa giao tiếp của họ. Đây được coi là cơ sở để bố trí không gian sử dụng dành cho đối tượng này.

Khi giao tiếp, người Điếc thường phải sắp xếp không gian thành một vòng tròn để tất cả mọi người có thể có tầm nhìn đủ tốt để trò chuyện với nhau. Trong cuộc sống hàng ngày, họ cũng cần tối ưu hóa những khoảng trống giữa các phòng, đặt gương và đèn ở những vị trí phù hợp nhất định để tăng khả năng nhận thức về hình ảnh với con người và sự vật xung quanh. Do đó để người Điếc sử dụng không gian một cách tiện nghi, cần có giải pháp về mặt kiến trúc nhằm thỏa mãn nhu cầu đặc thù của họ.

Ngoài ra, giải pháp liên quan đến thiết bị và ứng dụng (Hỗ trợ hàng ngày) phục vụ cho nhu cầu sinh hoạt hàng ngày góp phần đảm bảo cuộc sống độc lập của người Điếc trở nên dễ dàng hơn.

Tuy nhiên, không phải người mất khả năng thính lực nào cũng có điều kiện để sở hữu một máy trợ thính. Hơn nữa máy trợ thính chỉ có khả năng hỗ trợ một phần nhỏ cho người điếc. Vậy ngoài máy trợ thính, dự án đưa ra một thiết bị để giúp người mất khả năng thính lực có thể hiểu được mọi người nói.

2.1.3 Ngôn ngữ kí hiệu chuẩn Ngôn ngữ ký hiệu Việt Nam

Ngôn ngữ ký hiệu Việt Nam là tên gọi ba ngôn ngữ ký hiệu được phát triển bởi các cộng đồng khiếm thính tại Thành phố Hồ Chí Minh, Hà Nội, và Hải Phòng ở Việt Nam. Các ngôn ngữ này trực thuộc một khu vực cũng bao gồm các ngôn ngữ ký hiệu của Lào và Thái Lan, nhưng người ta chưa biết các ngôn ngữ này có liên quan với nhau. Các ngôn ngữ ký hiệu Việt Nam đã chịu ảnh hưởng từ ngôn ngữ ký hiệu Pháp. Các ngôn ngữ ký hiệu Thành phố Hồ Chí Minh và Hà Nội dùng chung vào khoảng 58% từ vựng cơ bản, trong khi các ngôn ngữ TPHCM và Hải Phòng dùng chung vào khoảng 54% từ vựng cơ bản. [5]

Từ những năm 2000, Việt Nam bắt đầu triển khai những nỗ lực của mình nhằm hoàn thiện và hệ thống hóa ngôn ngữ ký hiệu Việt Nam. Các câu lạc bộ, nhóm dạy, và sinh hoạt ngôn ngữ ký hiệu bắt đầu hình thành và nở rộ. Một số tài liệu khá công

phu xuất hiện như: bộ 3 tập Ký hiệu cho người điếc Việt Nam, từ điển ngôn ngữ ký hiệu Việt Nam, v.v. [6]

Cũng như ngôn ngữ nói, ngôn ngữ ký hiệu của từng quốc gia, thậm chí là từng khu vực trong một quốc gia rất khác nhau. Điều đó là do mỗi quốc gia, khu vực có lịch sử, văn hóa, tập quán khác nhau nên ký hiệu để biểu thị sự vật hiện tượng cũng khác nhau. Chẳng hạn, cùng chỉ tính từ màu hồng thì ở Hà Nội người ta xoa vào má (má hồng), còn tại Thành phố Hồ Chí Minh lại chỉ vào môi (môi hồng). Điều tương tự cũng diễn ra khi có sự khác biệt lớn hơn trên tầm quốc gia, dẫn tới sự khác biệt của hệ thống từ vựng và ngữ pháp ngôn ngữ ký hiệu giữa các nước.

Tuy nhiên, ký hiệu tất cả mọi nơi trên thế giới đều có những điểm tương đồng nhất định. Ví dụ: ký hiệu ‘uống nước’ thì nước nào cũng làm như nhau là giả bộ cầm cốc uống nước, ký hiệu ‘lái ô tô’ thì giả bộ cầm vô lăng ô tô quay quay, v.v. Mỗi người (dù bình thường hay câm điếc) đều có sẵn 30% kiến thức ngôn ngữ ký hiệu. Do ngôn ngữ ký hiệu phát triển hơn trong cộng đồng người khiếm thính, nên những người thuộc cộng đồng này của hai nước khác nhau có thể giao tiếp với nhau tốt hơn hai người bình thường nhưng mà không biết ngoại ngữ. [7]

Hai đặc điểm quan trọng nhất của ngôn ngữ kí hiệu là tính giản lược và có điểm nhấn:

- Ví dụ:

Bình thường: Anh có khỏe không ạ?

Ngôn ngữ kí hiệu: “KHỎE không”?

Do tính giản lược và có điểm nhấn nên cấu trúc ngữ pháp ngôn ngữ ký hiệu nhiều khi không thống nhất, cùng một câu có thể sắp xếp nhiều cách khác nhau (thường thì điểm nhấn được đưa lên đầu câu để gây hiệu quả chú ý) [7]

- Ví dụ 2:

Bình thường: Hôm qua, tôi gặp lại người bạn thân ở công viên. (Trong câu này, điểm nhấn là GẶP, và BẠN THÂN)

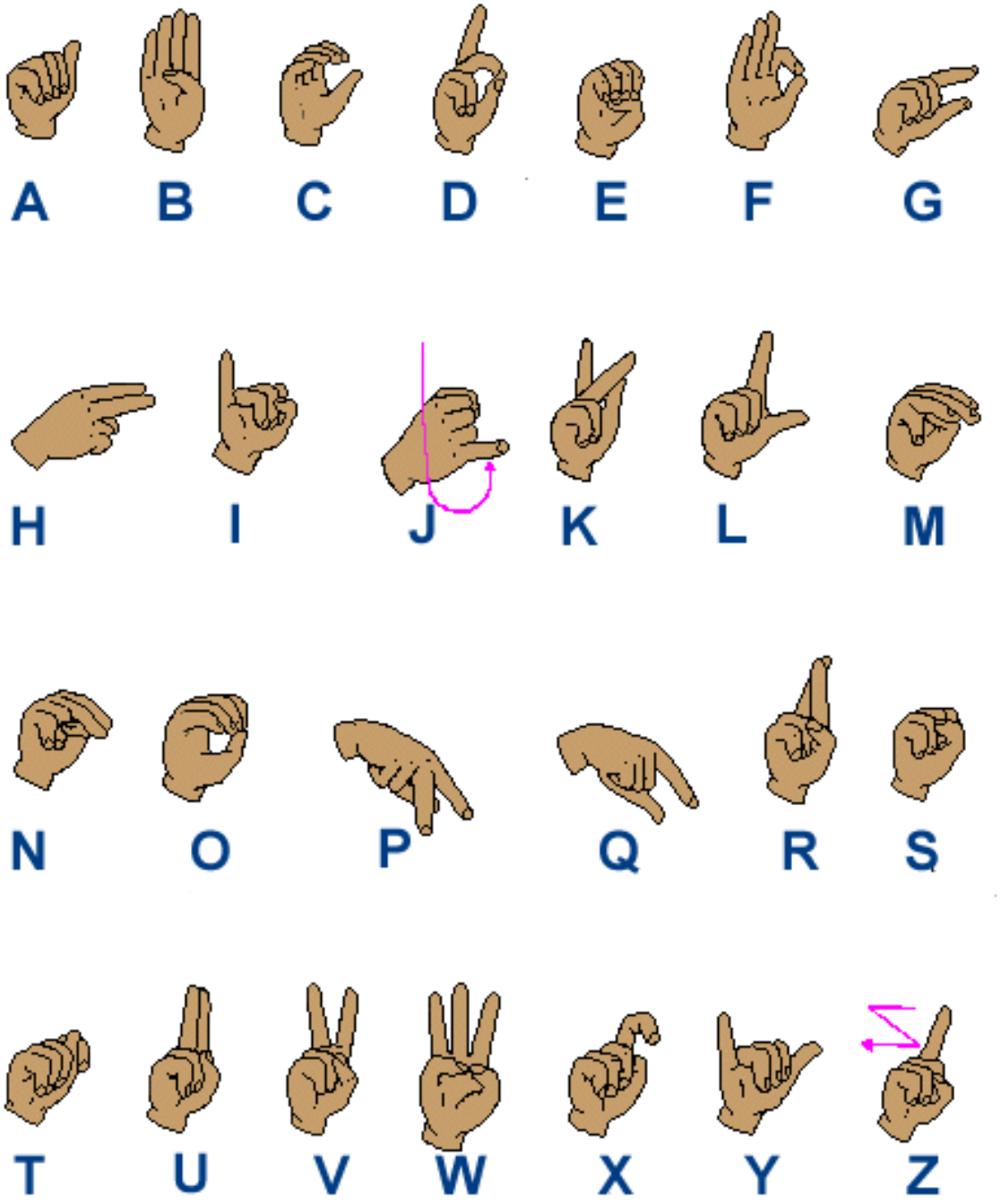
Ngôn ngữ kí hiệu: Bạn thân GẶP ở công viên hôm qua

Vậy đề tài phải rút gọn các từ thừa trong câu trước, sau đó mới đánh vần câu.

Quy định Bảng kí hiệu tay quốc tế được thể hiện như hình 2.2

Bảng kí hiệu tay theo chuẩn tiếng Việt được thể hiện như hình 2.3

Để đánh vần một chữ, người ta sẽ đưa lần lượt các kí tự để tạo thành một chữ. Ví dụ như từ TÔI sẽ được đánh vần theo thứ tự 19 + 14 + 26 + 9 trên hình 2.3 [8]



Hình 2.2 Bảng chữ cái theo ngôn ngữ kí hiệu

(Nguồn: https://pro.edu.vn)

Hình 2.3 Bảng chữ cái Việt Nam theo chuẩn ngôn ngữ kí hiệu Việt Nam

(Nguồn: https://pro.edu.vn)

Dự án sẽ sử dụng dữ liệu ở hình 2.3 làm dữ liệu cho việc đánh vần. Các từ cần đánh vần sẽ được tạo thành một danh sách các kí hiệu cần thực thi.

Một số từ trong ngôn ngữ kí hiệu vẫn được dùng nhanh, ví dụ như hình 2.4 và hình 2.5

Hình 2.4 Một số các từ thông dụng trong ngôn ngữ kí hiệu 1

(Nguồn: Giao tiếp với trẻ em giảm thính lực

-TS. Nguyễn Thị Xuyên - Thứ trưởng Bộ Y tế)

Hình 2.5 Một số các từ thông dụng trong ngôn ngữ kí hiệu 2

(Nguồn: Wikihow.vn)

Đề tài kết hợp với trung tâm giáo dục cho người khiếm thính trên địa bàn Gò Vấp để xây dựng bộ data các từ thông dụng này theo chuẩn ngôn ngữ kí hiệu Việt Nam

2.2 Tổng quan công nghệ Nhận dạng giọng nói

2.2.1 Giới thiệu về công nghệ Nhận dạng giọng nói

Nhận dạng tiếng nói là một quá trình nhận dạng mẫu, với mục đích là phân lớp (classify) thông tin đầu vào là tín hiệu tiếng nói thành một dãy tuần tự các mẫu đã được học trước đó và lưu trữ trong bộ nhớ. Các mẫu là các đơn vị nhận dạng, chúng có thể là các từ, hoặc các âm vị. Nếu các mẫu này là bất biến và không thay đổi thì công việc nhận dạng tiếng nói trở nên đơn giản bằng cách so sánh dữ liệu tiếng nói cần nhận dạng với các mẫu đã được học và lưu trữ trong bộ nhớ. Khó khăn cơ bản của nhận dạng tiếng nói đó là tiếng nói luôn biến thiên theo thời gian và có sự khác biệt lớn giữa tiếng nói của những người nói khác nhau, tốc độ nói, ngữ cảnh và môi trường âm học khác nhau.

Các nghiên cứu về nhận dạng tiếng nói dựa trên ba nguyên tắc cơ bản:

- Tín hiệu tiếng nói được biểu diễn chính xác bởi các giá trị phổ trong một khung thời gian ngắn (short-term amplitude spectrum). Nhờ vậy ta có thể trích ra các đặc điểm tiếng nói từ những khoảng thời gian ngắn và dùng các đặc điểm này làm dữ liệu để nhận dạng tiếng nói.
- Nội dung của tiếng nói được biểu diễn dưới dạng chữ viết, là một dãy các ký hiệu ngữ âm. Do đó ý nghĩa của một phát âm được bảo toàn khi chúng ta phiên âm phát âm thành dãy các ký hiệu ngữ âm.
- Nhận dạng tiếng nói là một quá trình nhận thức. Thông tin về ngữ nghĩa (semantics) và suy đoán (pragmatics) có giá trị trong quá trình nhận dạng tiếng nói, nhất là khi thông tin về âm học là không rõ ràng.

Cách tiếp cận nhận dạng tiếng nói bằng thống kê bao gồm: sử dụng mô hình Markov ẩn, mạng nơ-ron, sử dụng cơ sở tri thức, v.v..

2.2.2 Dữ liệu mở của google

Hình 2.6 Google Cloud speech API

(Nguồn: Google)

Để đáp ứng nhu cầu sử dụng dữ liệu hiện nay, Google đưa ra một gói dữ liệu, gọi là Google Cloud speech API. Dữ liệu này được áp dụng thuật toán mạng thần kinh học sâu (deep learning neural network) để nhận dạng giọng nói tự động (ASR).

Google Cloud speech API có dữ liệu ngôn ngữ của 125 quốc gia và biến thể.

Name	BCP-47	Model	Profanity filter	Automatic punctuation	Diarization (Beta)	Boost (Beta)	Word level confidence
Urdu (Pakistan)	ur-PK	Default	✓			✓	
Uzbek (Uzbekistan)	uz-UZ	Command and search	✓				
Uzbek (Uzbekistan)	uz-UZ	Default	✓				
Vietnamese (Vietnam)	vi-VN	Command and search	✓			✓	
Vietnamese (Vietnam)	vi-VN	Default	✓			✓	
Zulu (South Africa)	zu-ZA	Command and search	✓			✓	
Zulu (South Africa)	zu-ZA	Default	✓			✓	

Hình 2.7 Danh sách ngôn ngữ được hỗ trợ trong dữ liệu của google

(Nguồn: Google Cloud)

Google khuyến nghị kết hợp dữ liệu này với các công nghệ xử lí ngôn ngữ tự nhiên để đưa ra những ứng dụng tốt nhất, trong đó có hỗ trợ voice bots và phân tích cảm xúc cho lời nói.

Các tính năng chính của bộ dữ liệu Google Cloud speech API[9]:

- Thích ứng lời nói: tùy chỉnh nhận dạng giọng nói để phiên âm các thuật ngữ cụ thể theo miền và các từ hiếm bằng cách cung cấp gợi ý và tăng độ chính xác phiên âm của các từ hoặc cụm từ cụ thể. Tự động chuyển đổi số nói thành địa chỉ, năm, tiền tệ và nhiều hơn nữa bằng cách sử dụng các lớp.
- Thích ứng môi trường: Chọn từ một loạt các mô hình được đào tạo để điều khiển bằng giọng nói và gọi điện thoại và sao chép video được tối ưu hóa cho các yêu cầu chất lượng cụ thể của miền. Ví dụ như trường hợp là âm thanh

cuộc gọi điện thoại thì có thể chọn miền là cuộc gọi điện thoại, kết quả cho sẽ chính xác hơn (ví dụ như cuộc điện thoại được ghi ở tốc độ lấy mẫu 8kHz)

- Truyền nhận dạng giọng nói: Nhận kết quả nhận dạng giọng nói theo thời gian thực khi API xử lý đầu vào âm thanh được truyền phát từ micrô của ứng dụng hoặc được gửi từ tệp âm thanh được ghi trước (nội tuyến hoặc qua Lưu trữ đám mây).
- Nhận dạng đa kênh: Speech-to-Text có thể nhận ra các kênh riêng biệt trong các tình huống đa kênh (ví dụ: hội nghị video) và chú thích các bản ghi để giữ trật tự.
- Xử lí nhiễu: Speech-to-Text có thể xử lý âm thanh ồn từ nhiều môi trường mà không yêu cầu loại bỏ tiếng ồn bổ sung.
- Lọc nội dung: Có thể tuỳ chọn lọc từ thô tục trong kết quả văn bản

  Các tính năng đang được phát triển, dùng thử [9]
- Tự động phát hiện câu thoại thuộc ngôn ngữ nước nào
- Tự động điền dấu câu (dấu chấm, dấu phẩy)
- Xác định người nói

Đề tài sử dụng API của google để phát triển chức năng chuyển văn bản thành giọng nói, nhờ đó tăng khả năng chính xác cao hơn so với các API khác.

2.3 Tổng quan công nghệ Xử lí ngôn ngữ tự nhiên

2.3.1 Giới thiệu về công nghệ xử lí ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên (natural language processing - NLP) là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ-công cụ hoàn hảo nhất của tư duy và giao tiếp.

Mục tiêu của lĩnh vực này là giúp máy tính hiểu và thực hiện hiệu quả những nhiệm vụ liên quan đến ngôn ngữ của con người như: tương tác giữa người và máy, cải thiện hiệu quả giao tiếp giữa con người với con người, hoặc đơn giản là nâng cao hiệu quả xử lý văn bản và lời nói.

Xử lý ngôn ngữ tự nhiên bao gồm hiểu ngôn ngữ tự nhiên (Natural Language Understanding – NLU) và sinh ngôn ngữ tự nhiên (Natural Language Generation – NLG). Trong đó, hiểu ngôn ngữ tự nhiên (NLU)bao gồm 4 bước chính sau đây[10]:

- Phân tích hình vị: là sự nhận biết, phân tích, và miêu tả cấu trúc của những hình vị trong một ngôn ngữ cho trước và các đơn vị ngôn ngữ khác, như từ gốc, biên từ, phụ tố, từ loại,… Có hai loại bài toán điển hình trong phần này, bao gồm bài toán tách từ (word segmentation) và gán nhãn từ loại (POS).
- Phân tích cú pháp: là quy trình phân tích một chuỗi các biểu tượng, ở dạng ngôn ngữ tự nhiên hoặc ngôn ngữ máy tính, tuân theo văn phạm hình thức. Văn phạm hình thức thường dùng trong phân tích cú pháp của ngôn ngữ tự nhiên bao gồm Văn phạm phi ngữ cảnh (Context-free grammar – CFG), Văn phạm danh mục kết nối (Combinatory categorial grammar – CCG), và Văn phạm phụ thuộc (Dependency grammar – DG). Đầu vào của quá trình phân tích là một câu gồm một chuỗi từ và nhãn từ loại của chúng, và đầu ra là một cây phân tích thể hiện cấu trúc cú pháp của câu đó. Các thuật toán phân tích cú pháp phổ biến bao gồm CKY, Earley, Chart, và GLR.
- Phân tích ngữ nghĩa: là quá trình liên hệ cấu trúc ngữ nghĩa, từ cấp độ cụm từ, mệnh đề, câu và đoạn đến cấp độ toàn bài viết, với ý nghĩa độc lập của chúng. Nói cách khác, việc này nhằm tìm ra ngữ nghĩa của đầu vào ngôn từ. Phân tích ngữ nghĩa bao gồm hai mức độ: Ngữ nghĩa từ vựng biểu hiện các ý nghĩa của những từ thành phần, và phân biệt nghĩa của từ; Ngữ nghĩa thành phần liên quan đến cách thức các từ liên kết để hình thành những nghĩa rộng hơn.
- Phân tích diễn ngôn: Ngữ dụng học là môn nghiên cứu về mối quan hệ giữa ngôn ngữ và ngữ cảnh sử dụng (context-of-use). Ngữ cảnh sử dụng bao gồm danh tính của người hoặc vật, và vì thế ngữ dụng học bao gồm những nghiên cứu về cách ngôn ngữ được dùng để đề cập (hoặc tái đề cập) tới người hoặc vật. Ngữ cảnh sử dụng bao gồm ngữ cảnh diễn ngôn, vì vậy ngữ dụng học cũng bao gồm những nghiên cứu về cách thức cấu tạo nên diễn ngôn, và cách người nghe hiểu người đang đối thoại với mình.

Một số ứng dụng của xử lí ngôn ngữ tự nhiên[11]:

- Truy xuất thông tin (Information Retrieval – IR) có nhiệm vụ tìm các tài liệudưới dạng không có cấu trúc (thường là văn bản) đáp ứng nhu cầu về thông

tin từ những nguồn tổng hợp lớn. Những hệ thống truy xuất thông tin phổ biến nhất bao gồm các công cụ tìm kiếm như Google, Yahoo, hoặc Bing search. Những công cụ này cho phép tiếp nhận một câu truy vấn dưới dạng ngôn ngữ tự nhiên làm đầu vào và cho ra một danh sách các tài liệu được sắp xếp theo mức độ phù hợp.

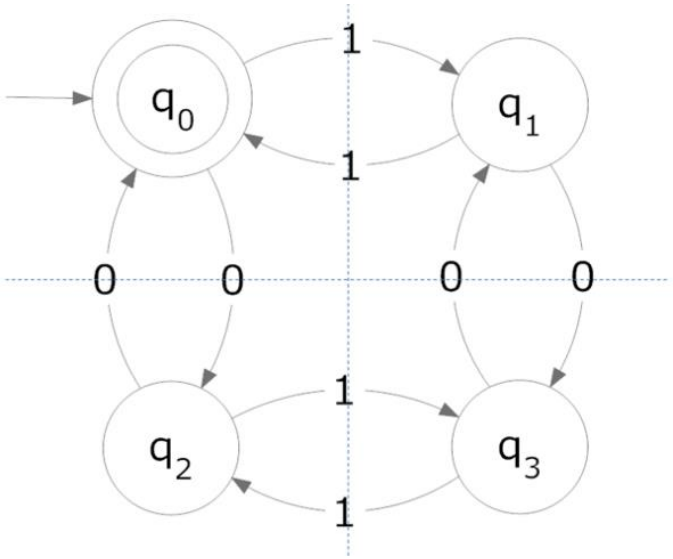
- Trích chọn thông tin (Information Extraction) nhận diện một số loại thực thể được xác định trước, mối quan hệ giữa các thực thể và các sự kiện trong văn bản ngôn ngữ tự nhiên. Khác với truy xuất thông tin trả về một danh sách các văn bản hợp lệ thì trích chọn thông tin trả về chính xác thông tin mà người dùng cần. Những thông tin này có thể là về con người, địa điểm, tổ chức, ngày tháng, hoặc thậm chí tên công ty, mẫu sản phẩm hay giá cả.
- Trả lời câu hỏi (QA) có khả năng tự động trả lời câu hỏi của con người ở dạng ngôn ngữ tự nhiên bằng cách truy xuất thông tin từ một tập hợp tài liệu. Một hệ thống QA đặc trưng thường bao gồm ba mô đun: Mô đun xử lý truy vấn (Query Processing Module) – tiến hành phân loại câu hỏi và mở rộng truy vấn; Mô đun xử lý tài liệu (Document Processing Module) – tiến hành truy xuất thông tin để tìm ra tài liệu thích hợp; và Mô hình xử lý câu trả lời (Answer Processing Module) – trích chọn câu trả lời từ tài liệu đã được truy xuất.
- Tóm tắt văn bản tự động là bài toán thu gọn văn bản đầu vào để cho ra một bản tóm tắt ngắn gọn với những nội dung quan trọng nhất của văn bản gốc. Có hai phương pháp chính trong tóm tắt, là phương pháp trích xuất (extractive) và phương pháp tóm lược ý (abstractive). Những bản tóm tắt trích xuất được hình thành bằng cách ghép một số câu được lấy y nguyên từ văn bản cần thu gọn. Những bản tóm lược ý thường truyền đạt những thông tin chính của đầu vào và có thể sử dụng lại những cụm từ hay mệnh đề trong đó, nhưng nhìn chung được thể hiện ở ngôn ngữ của người tóm tắt.

2.3.2 Xử lí ngôn ngữ tiếng Việt

Tiếng Việt được xếp vào loại đơn lập – tức phi hình thái, không biến hình. Cùng với đó, tiếng Việt được viết theo trật tự S – V – O. (subject (S), verb (V) and object (O)).

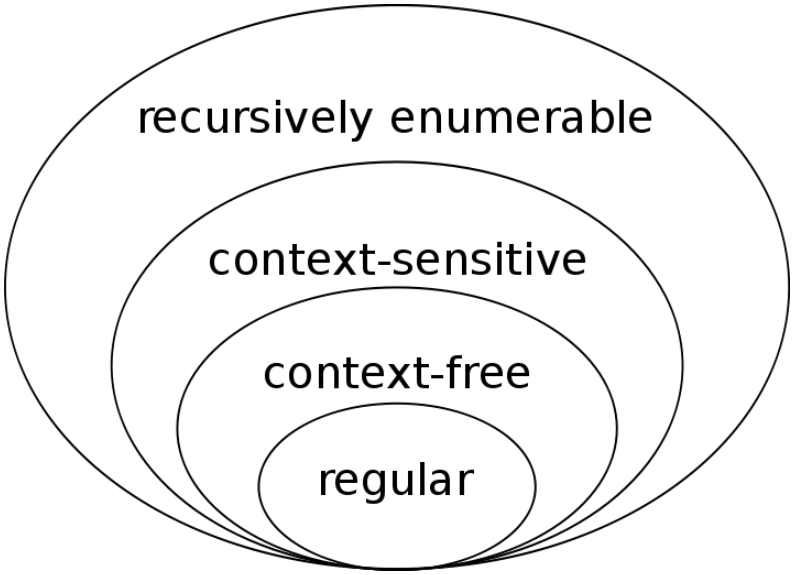
Ngôn ngữ hình thức – Formal Language

Ngôn ngữ hình thức (Formal Language) là một tập các chuỗi (string) được xây dựng dựa trên một bảng chữ cái (alphabet), được ràng buộc bởi các luật (rule) hoặc văn phạm (grammar) đã được định nghĩa trước. Alphabet có thể là tập các ký tự trong ngôn ngữ tự nhiên (Natural Language) hoặc tập tự định nghĩa các ký tự. Mô hình ngôn ngữ tự nhiên tuân theo quy luật của chuỗi Markov và được hình thức hóa đầu tiên bởi Noam Chomsky được gọi là 'Mô hình phân cấp Chomsky'. Sau này những mô hình này được dùng để tạo ra ngôn ngữ lập trình hoặc các ứng dụng trong các nghiên cứu dịch tự động.



Hình 2.8 Tiền đề trong việc xây dựng lý thuyết Automata là ngôn ngữ hình thức

(Nguồn: Đỗ Bá Lâm - Đại học Bách khoa Hà Nội)



Hình 2.9 Mô hình phân cấp Chomsky

(Nguồn: Lê Thanh Hương - Đại học Bách khoa Hà Nội)

Các khái niệm cơ bản về xử lí ngôn ngữ tự nhiên

- Bộ chữ (Alphabet Set): tập các ký hiệu (vô hạn hoặc hữu hạn).
  Ví dụ: Tập 26 chữ Roman alphabet, Tập $\sum$ ={0,1}, …
- Chuỗi (String) hoặc từ (Word): là một chuỗi các chữ cái trên Alphabet nào đó
  Ví dụ ‘abc ‘; ‘0101110’ ; …
  Chuỗi rỗng (không chứa ký tự nào trong Alphabet). (ký hiệu ε , |ε| = 0).
- Ngôn ngữ rỗng (Empty Language): một ngôn ngữ không chứa bất kì câu nào được gọi là ngôn ngữ rỗng (ký hiệu: ∅).
- Một ngôn ngữ trên một bộ chữ Σ là tập các chuỗi trên Σ . Σ* là tập chứa tất cả các chuỗi trên Σ bao gồm cả ε. Ví dụ với Σ = {0,1} thì: Σ* = { ε, 0, 1,00, 01, 10, 11, 000, 001,…}
- Ngôn ngữ L là tập những chuỗi có chiều dài hữu hạn trên một bộ chữ hữu hạn Σ nào đó. Nếu ngôn ngữ L hữu hạn ta chỉ cần liệt kê tất cả các chuỗi để biểu diễn các trường hợp và xét ngữ nghĩa cho từng trường hợp, nhưng vì ngôn ngữ tự nhiên vô hạn nên ta cần văn phạm để xét nghĩa.

Văn Phạm – Grammar : G = { N, Σ, P, S}

- N: tập các từ vựng phụ trợ, như các phạm trù ngữ pháp, kí hiệu không kết thúc (non-terminal).
- S: tập các từ của ngôn ngữ, gọi là ký hiệu kết thúc (terminal).
- P: tập các luật văn phạm, gọi là luật sản sinh (Production), $N \cap \Sigma = \emptyset$
- S : là yếu tố nguyên thủy của ngữ pháp, $S \in N$
- Một luật P có dạng : $a \rightarrow b$ ($a, b \in N \cup \Sigma$)
- X là tập các phần tử của chuỗi .
- Xi là tập của những chuỗi có chiều dài i.
- Nếu P trong văn phạm đều có dạng: $X \rightarrow a$ ($X \in N, a \in N \cup \Sigma$), văn phạm đó gọi là phi ngữ cảnh (Context-Free Grammar: CFG).

Giải thuật phân tích cú pháp Earley

Earley biểu diễn luật P thông qua dấu chấm "•". Dấu chấm "•" là một siêu ký hiệu (metasymbol) không thuộc về N hay Σ. Vị trí dấu thay đổi theo trạng thái đang xét.

Ví dụ một luật sản sinh P ở trạng thái S(j) : (A → α • β, i).

Ví dụ cụ thể

Phân tích câu "tôi ăn quả cam."

Cho tập luật P:

S → N VP	1
S → P VP	2
VP → V N	3
VP → V NP	4
NP → N N	5
NP → N A	6
AP → R A	7

Bảng 2.1 Bảng luật P của ví dụ

Non-terminal: S, NP, VP, AP.

Terminal: P, N, V, A, R.

S	Câu	N	Danh từ
VP	Cụm động từ	V	Động từ
NP	Cụm danh từ	A	Tính từ
AP	Cụm tính từ	R	Phụ từ
P	Đại từ	M	Số

Bảng 2.2 Phân thích Non-Terminal và Terminal

Áp dụng giải thuật Earley ta được bảng

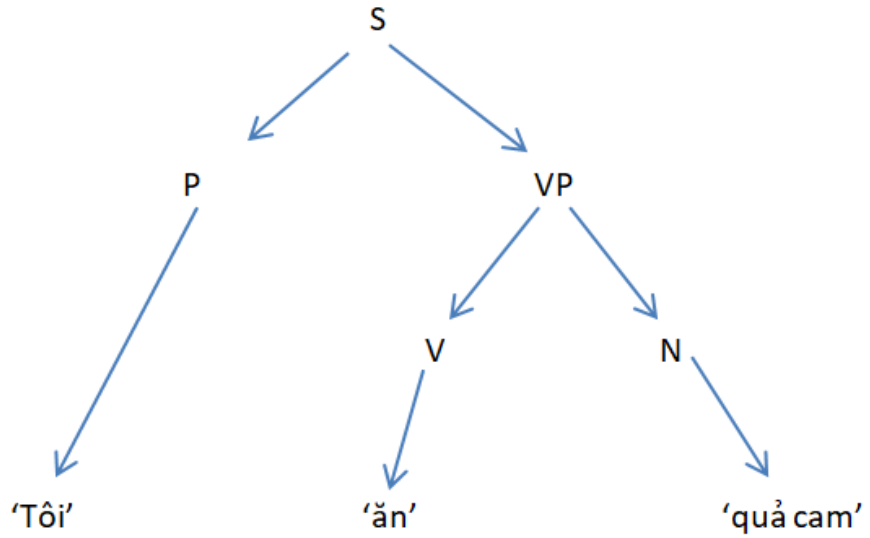
0	1	2	3	4
	‘tôi’ là đại từ P	‘ăn’ là động từ V	‘quả cam’ là danh từ N	
1: S → • N VP	2 : S → P • VP	3: VP → V • N	3: VP → V N • *	2: S → P VP • **
2: S → • P VP	3: VP → • V N	4: VP → V • NP	5: NP → N • N	
	4: VP → • V NP	5: NP → • N N	6: NP → N • A	
		6: NP → • N A		

Bảng 2.3 Kết quả quá trình xử lí ví dụ

- Bước 0: Ta xét từ gốc ROOT ký hiệu là S, lấy tất cả các luật của S và các non-terminal đầu tiên được suy diễn từ S nếu có. Dấu • được để ngay đầu, có ý nghĩ tiếp theo ta sẽ xét phần tử kế tiếp dấu chấm •.
- Bước 1: Xét từ đầu tiên ‘tôi’ là đại từ nhân xưng, đáp ứng với dòng 2 của bảng 0. Ta dịch chuyển dấu • để xác nhận phần tử đầu thành công và sẽ xét tiếp phần tử kế.
- Bước 2: Xét phần tử kế ‘ăn’ là động từ, cả hai luật 3,4 đều đáp ứng nên ta xét cùng lúc cả hai trường hợp.
- Bước 3: Xét từ cuối ‘quả cam’ là danh từ thỏa luật 3 và kết thúc.

Nếu trong quá trình xét ta gặp non-terminal thì liệt kê tại cùng bảng và duyệt dựa vào đó cho đến khi dấu chấm • ở phía cuối suy diễn và độ dài câu tương ứng với các phần tử đã xét thành công thì kết thúc.

Từ các bước trên ta có được kết quả được cây suy dẫn:



Hình 2.10 Cây cấu trúc của ví dụ

Nhập nhằng trong xử lí ngôn ngữ Tiếng Việt

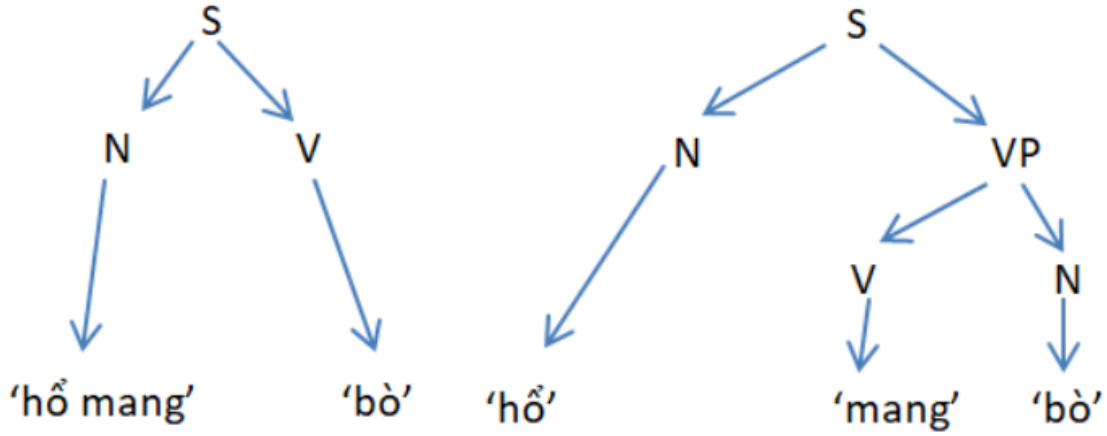
Đối với xử lý ngôn ngữ khái niệm "Nhập nhằng" là hiện tượng khi câu hoặc từ có nhiều nghĩa dẫn tới việc một câu có thể có nhiều cây suy dẫn. Với tiếng Việt – loại ngôn ngữ đơn lập, nhập nhằng còn xảy ra khi ta có hệ thống từ ghép, từ láy, …[12]

Ví dụ:

"quần áo" – N N , hoặc "quần áo" – N

"nóng lòng" – A N , hoặc "nóng lòng" – A

Trong phân tích câu 'hổ mang bò', ta được hai cây suy dẫn:



Hình 2.11 Hai trường hợp cây cấu trúc từ một câu giống nhau

(Nguồn: Đỗ Bá Lâm - Đại học Bách khoa Hà Nội)

Tiếng Anh và tiếng Việt có nhiều điểm khác biệt (do loại hình ngôn ngữ, do nền văn hoá,…).

Khác về ngữ âm học, hình vị, ranh giới từ, sự từ vựng hoá (như: ox – bò đực, anh – elder brother ,…); từ loại; trật tự từ, kết cấu câu. Do đó việc áp dụng thuật giải Earley cho tiếng Việt còn gặp nhiều khó khăn.

Cái bài toán giải quyết vấn đề nhập nhằng: Tiền xử lý (Pre-Processing), Phân tích hình thái (Morphology), Phân đoạn từ (Word Segmentation), Phân tích ngữ pháp (Parser), Gán nhãn ngữ nghĩa (Semantics), …

Hiện này các ứng dụng tiêu biểu như sửa lỗi chính tả, lỗi cú pháp; dịch tự động; phát hiện vi phạm bản quyền, spam ; tóm tắt rút trích nội dung văn bản, … đều sử dụng công nghệ Natural Language Processing – NLP. [12]

2.3.3 Thư viện Underthesea

Underthesea là thư viện mã nguồn mỡ dành cho ngôn ngữ Python, nhằm hỗ trợ việc nghiên cứu, hướng dẫn và phát triển công nghệ xử lí ngôn ngữ tự nhiên Việt Nam.
Để cài đặt thư viện underthesea, ta có thể sử dụng lệnh sau trên Anaconda powershell:

```
Pip install underthesea
```

Các chức năng hiện tại của thư viện

- Phân đoạn câu
- Phân đoạn từ
- Phân loại từ
- Chunking
- Nhận dạng thực thể
- Phân loại văn bản
- Phân tích cảm xúc

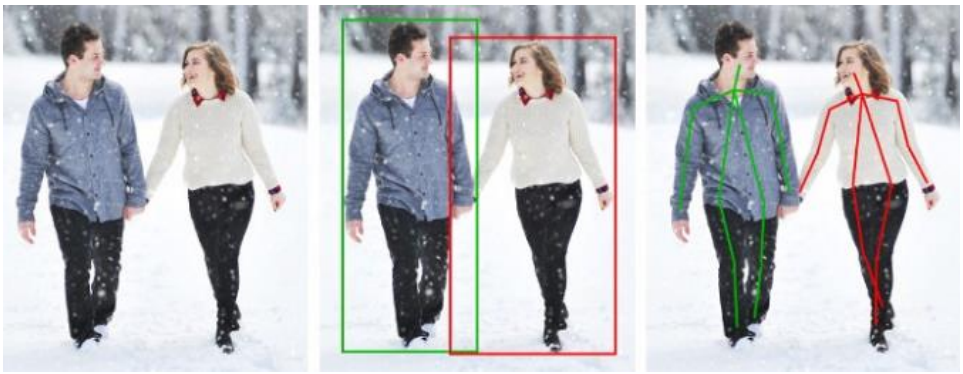
Đề tài sử dụng thư viện underthesea cho việc xử lí văn bản thu được từ bước trên. Vì là thư viên riêng cho xử lí ngôn ngữ tự nhiên của Việt Nam nên việc chọn thư viện underthesea làm cho kết quả cao hơn so với thư viện nước ngoài.

2.4 Tổng quan công nghệ HandTracking

2.4.1 Giới thiệu về phương pháp OpenPose [13]

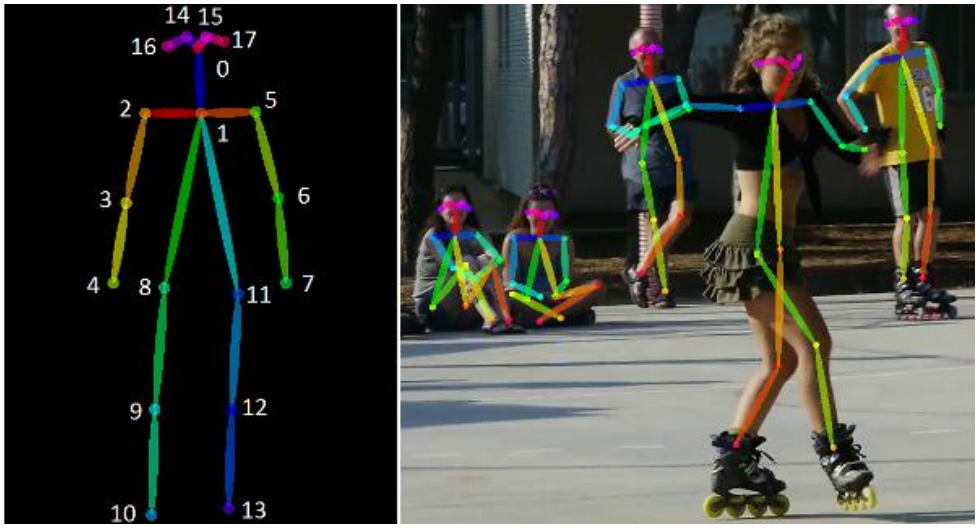
Bộ xương người đại diện cho định hướng của một người trong một định dạng đồ họa. Về cơ bản, nó là một tập hợp các tọa độ có thể được kết nối để mô tả tư thế của người đó. Mỗi phối hợp trong bộ xương được gọi là một phần (hoặc khớp, hoặc một điểm then chốt). Một kết nối hợp lệ giữa hai phần gọi là một cặp (hoặc một chi).

Một số phương pháp tiếp cận Ước tính Pose của con người đã được giới thiệu trong những năm qua. Các phương pháp sớm nhất (và chậm nhất) thường ước tính tư thế của một người trong một hình ảnh chỉ có một người bắt đầu. Các phương pháp này thường xác định các phần riêng lẻ trước, sau đó hình thành các kết nối giữa chúng để tạo tư thế.



Hình 2.12 Kết quả phương pháp OpenPose

(Nguồn: Alain Pham)



Hình 2.13: Định dạng keypoint COCO cho bộ xương người (trái)

Và Kết xuất bộ xương người (phải)

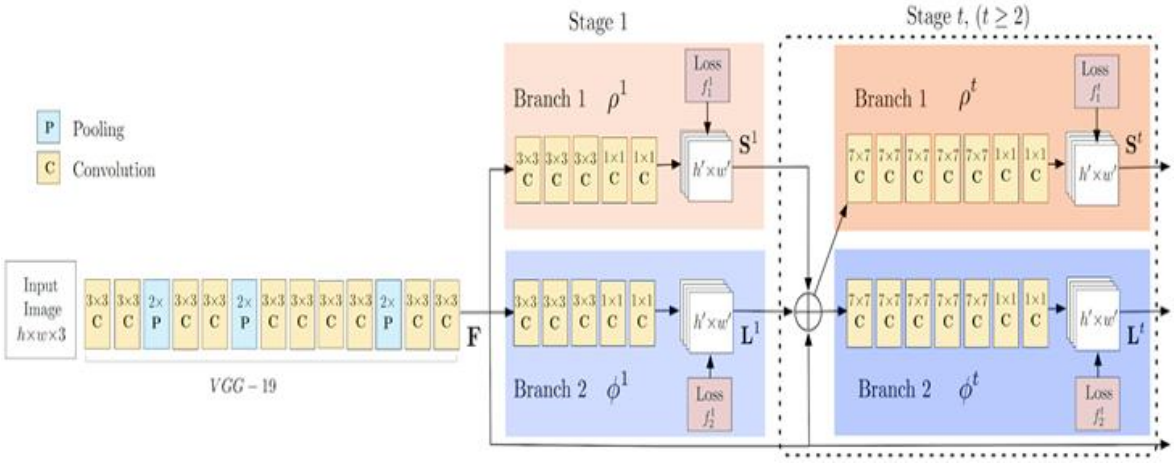
(Nguồn: Gines Hidalgo)

Ước tính tư thế nhiều người khó hơn trường hợp một người vì không xác định được vị trí và số người trong một hình ảnh. Thông thường, chúng ta có thể giải quyết vấn đề trên bằng một trong hai cách tiếp cận:

- Cách tiếp cận đơn giản là kết hợp máy dò người trước, sau đó là ước tính các bộ phận và sau đó tính toán tư thế cho mỗi người. Phương pháp này được gọi là phương pháp từ trên xuống.
- Một cách tiếp cận khác là phát hiện tất cả các bộ phận trong hình ảnh (tức là các bộ phận của mỗi người), tiếp theo là liên kết / nhóm các bộ phận thuộc về những người khác biệt. Phương pháp này gọi là phương pháp từ dưới lên.

OpenPose là một trong những cách tiếp cận từ dưới lên phổ biến nhất để ước tính tư thế của nhiều người.

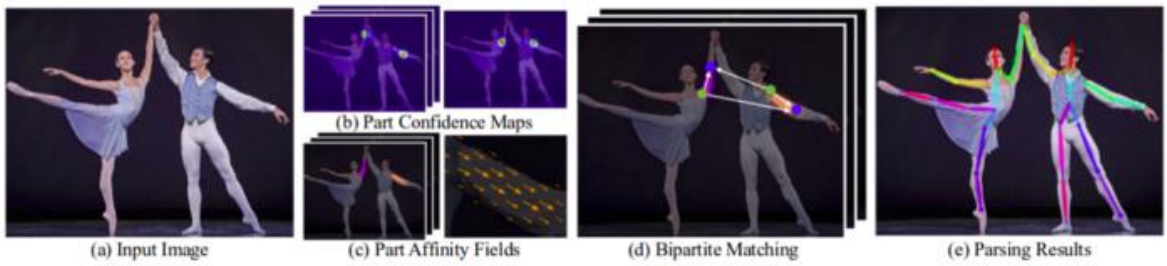
Cũng như nhiều cách tiếp cận từ dưới lên, trước tiên OpenPose phát hiện các phần (điểm chính) thuộc về mỗi người trong ảnh, sau đó là gán các phần cho các cá nhân riêng biệt. Hình 2.14 là kiến trúc của mô hình OpenPose.



Hình 2.14: Sơ đồ khối của kiến trúc OpenPose

(Nguồn: Tomas Simon)

Mạng OpenPose lần đầu tiên trích xuất các tính năng từ một hình ảnh bằng cách sử dụng một vài lớp đầu tiên (VGG-19 trong hình 2.14) Các tính năng sau đó được đưa vào hai nhánh song song của các lớp chập. Nhánh đầu tiên dự đoán một bộ gồm 18 bản đồ độ tin cậy, với mỗi bản đồ đại diện cho một phần cụ thể của bộ xương người. Nhánh thứ hai dự đoán một bộ gồm 38 trường có ái lực phần (PAF) đại diện cho mức độ liên kết giữa các phần.



Hình 1.15 Các bước ước tính tư thế con người bằng phương pháp OpenPose

(Nguồn: Yaser Sheikh)

Các giai đoạn kế tiếp được sử dụng để tinh chỉnh các dự đoán được thực hiện bởi mỗi chi nhánh. Sử dụng bản đồ độ tin cậy của phần, đồ thị lưỡng cực được hình thành giữa các cặp phần. Sử dụng các giá trị PAF, các liên kết yếu hơn trong các biểu đồ lưỡng cực được cắt tỉa. Thông qua các bước trên, bộ xương của con người có thể được ước tính và chỉ định cho mỗi người trong ảnh.

2.4.2 Module OpenMMD

OpenMMD đại diện cho dự án Deep-Learning dựa trên OpenPose có thể chuyển đổi trực tiếp video của người thật sang chuyển động của các mô hình hoạt hình (ví dụ Miku, Anmicius). OpenMMD có thể được gọi là OpenPose + MikuMikuDance (MMD).

Hình 2.16 Ví dụ mô hình 3D: Anmicius

(Nguồn: Zhang Xinyi)

MMD là một chương trình hoạt hình miễn phí cho phép người dùng tạo hiệu ứng và tạo phim hoạt hình 3D bằng các mô hình 3D như Miku và Anmicius.

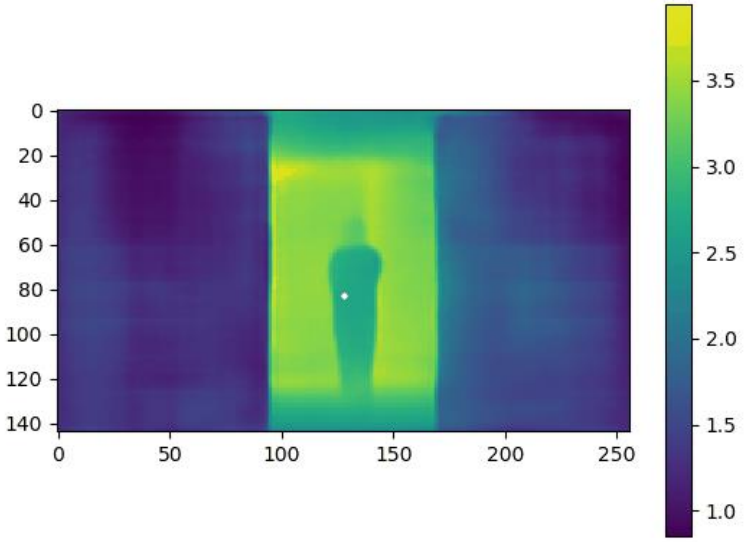
OpenPose và MMD chỉ là "lối vào" và "lối ra" của hộp ứng dụng. Có ba Mô hình Deep Learning được đào tạo trước trong hộp để xử lý và chuyển đổi dữ liệu được định dạng.

Đặc trưng của OpenMMD

- Sử dụng mô hình học không giám sát từ dữ liệu 2D và 3D
- Xác định độ sâu trường ảnh
- Xác định các điểm chính trên cơ thể người
- Có thể sử dụng để tạo hiệu ứng đồ hoạ thủ công

Ví dụ cụ thể của module được thể hiện theo thứ tự hình 2.17 đến hình 2.20

Hình 2.17
Video nguyên bản
(Nguồn: Zhang Xinyi)



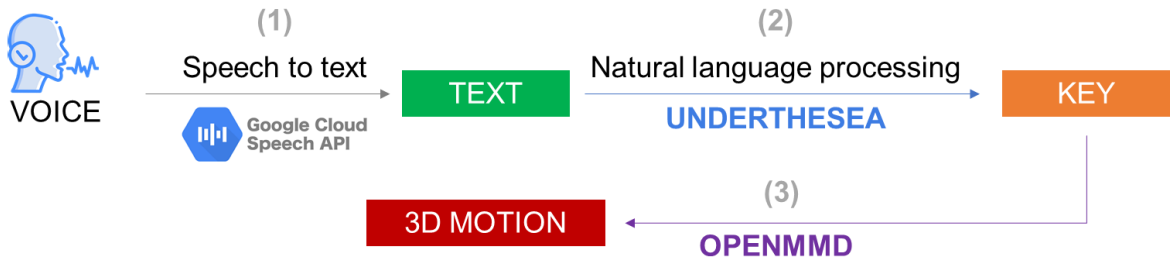
Hình 2.18
Tính độ sâu trường ảnh
(Nguồn: Zhang Xinyi)

Hình 2.19
Xác định điểm chính cơ thể
(Nguồn: Zhang Xinyi)

Hình 2.20
Kết quả của quá trình OpenPose
(Nguồn: Zhang Xinyi)

CHƯƠNG 3
NỘI DUNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

3.1 Tổng quan hệ thống



Hình 3.1 Sơ đồ khối tổng quan hệ thống

Đề tài được chia thành 3 bước chính (hình 3.1):

Bước 1: Chuyển lời nói tiếng Việt thành văn bản dựa trên API của Google

Phần mềm có thể nhận dữ liệu từ bàn phím, với trường hợp này, phần mềm có thể bỏ qua bước 1.

Bước 2: Sử dụng thư viện Underthesea để xử lí văn bản vừa thu được và chuyển thành các lệnh điều khiển 3D

Bước 3: Sử dụng Module OpenMMD để mô phỏng nội dung dưới dạng 3D

3.2 Dữ liệu tương đương giữa ngôn ngữ tiếng Việt và ngôn ngữ kí hiệu

Đề tài đưa ra ba bộ dữ liệu cho 3 chức năng khác nhau, bao gồm:

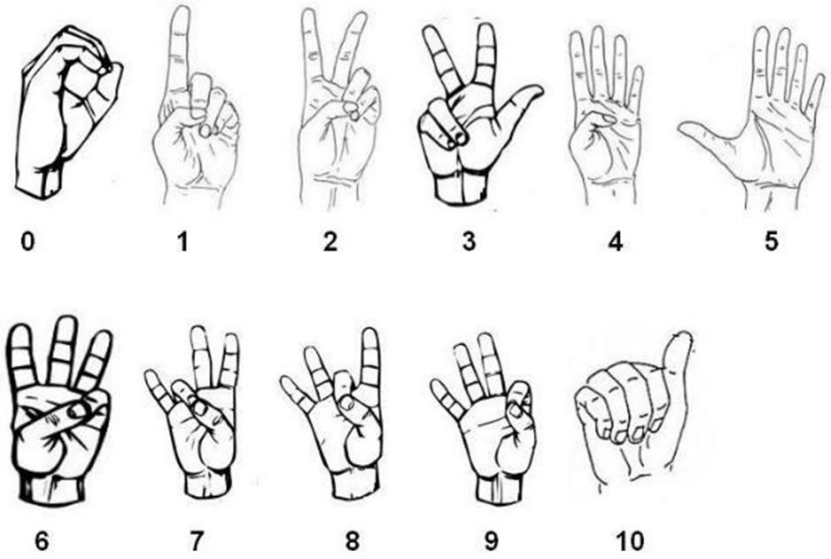
- Dữ liệu số: Number_data
- Dữ liệu bản chữ cái: Spell_data
- Dữ liệu các từ thông dụng: Quick_data

Lấy ví dụ câu sau: “2 con chó to quá”

Nguyên bản	2	Con chó	To quá
Rút gọn	**2**	**chó**	**to**
Dạng	Số	Danh từ	Tính từ
Tập dữ liệu sử dụng	**Number_data**	**Spell_data**	**Quick_data**

Bảng 3.1 Dữ liệu tương ứng cho các từ khác nhau

Đối với số liệu, phần mềm sẽ tự động sử dụng tập Number_data.



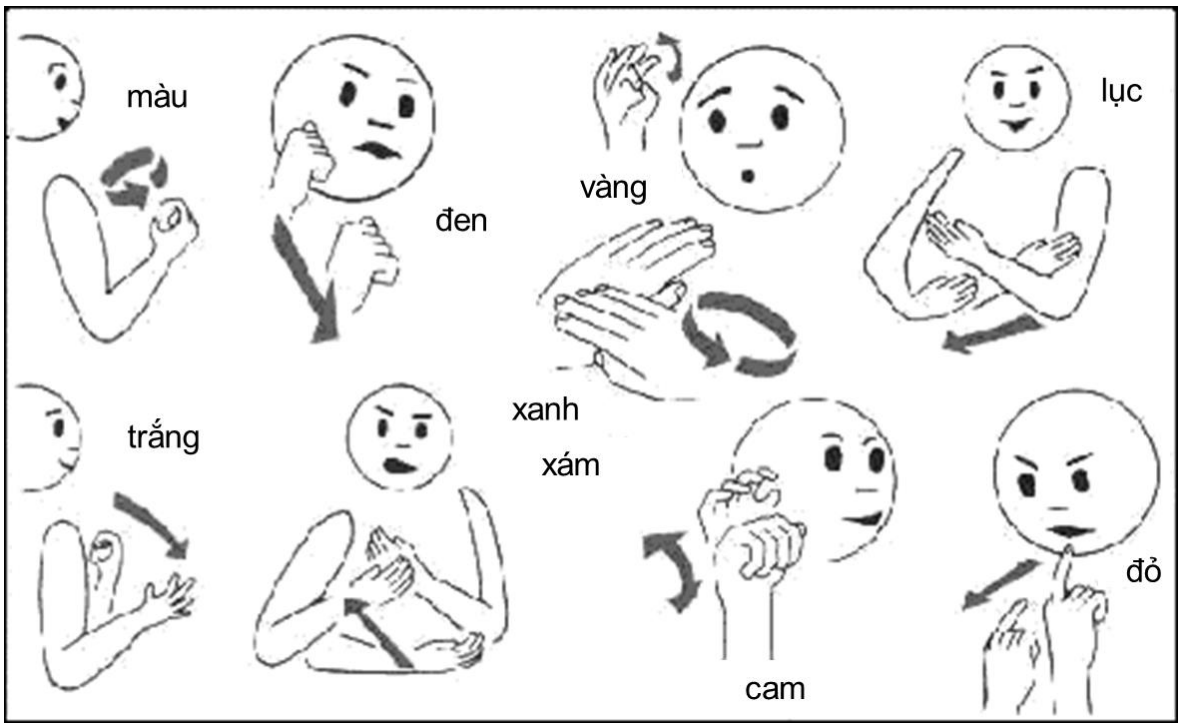
Hình 3.2 Dữ liệu số - Number_data

Đối những từ không phải là số, phần mềm sẽ kiểm tra xem từ đó có thuộc danh sách những từ trong dữ liệu từ thông dụng không. Nếu không thuộc từ thông dụng, phần mềm sẽ đánh vần từ đó dựa trên tập dữ liệu Spell_data.

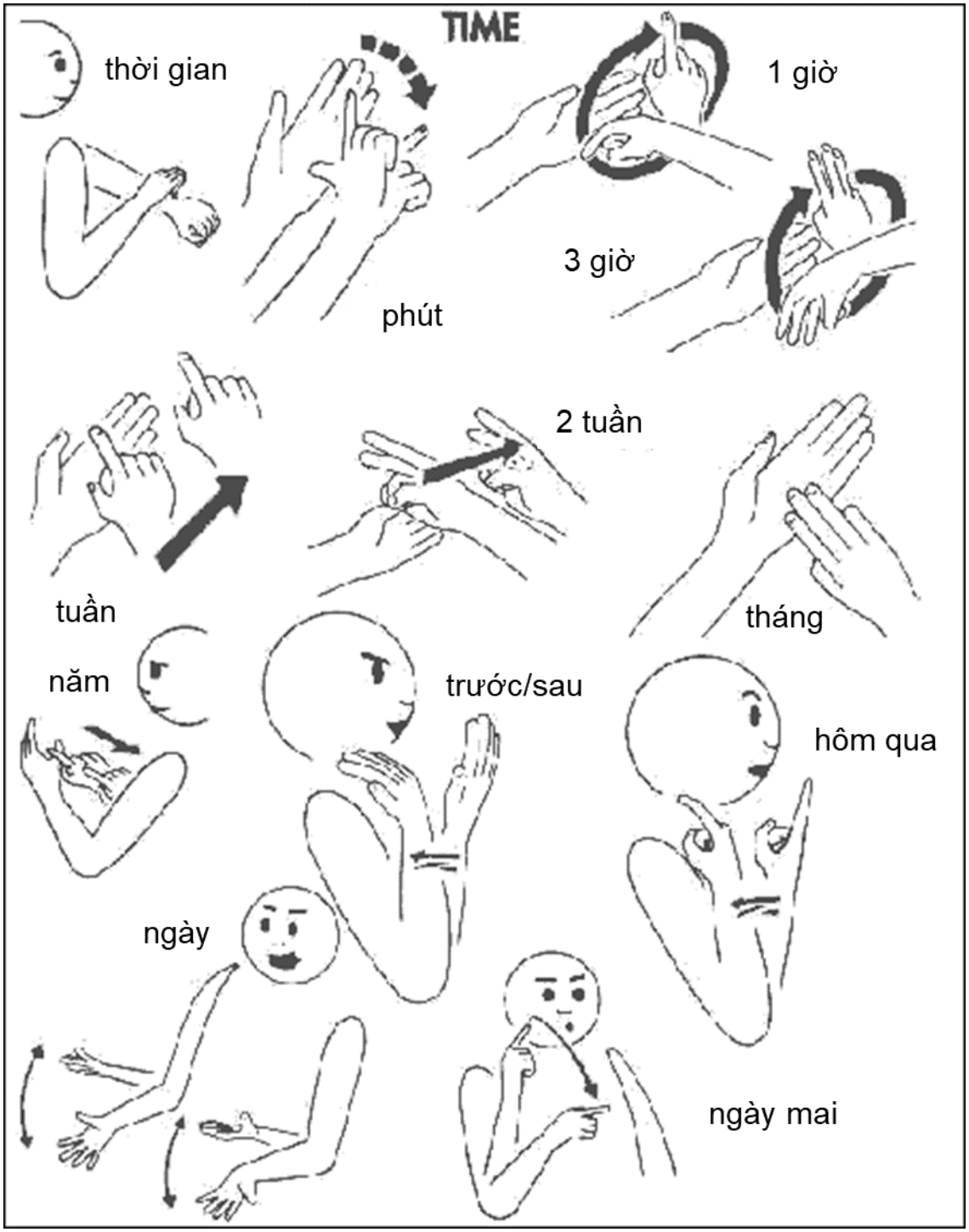
Hình 3.3 Dữ liệu bảng chữ cái – Spell_data

Những từ đã được liệt kê trong danh sách các từ thông dụng, phần mềm sẽ thực hiện nó trong tập dữ liệu Quick_data. Quick_data được chia thành nhiều phần: Danh từ, động từ, tính từ.

Hình 3.4 Một số dữ liệu trong tập các từ thông dụng – Quick_data 1



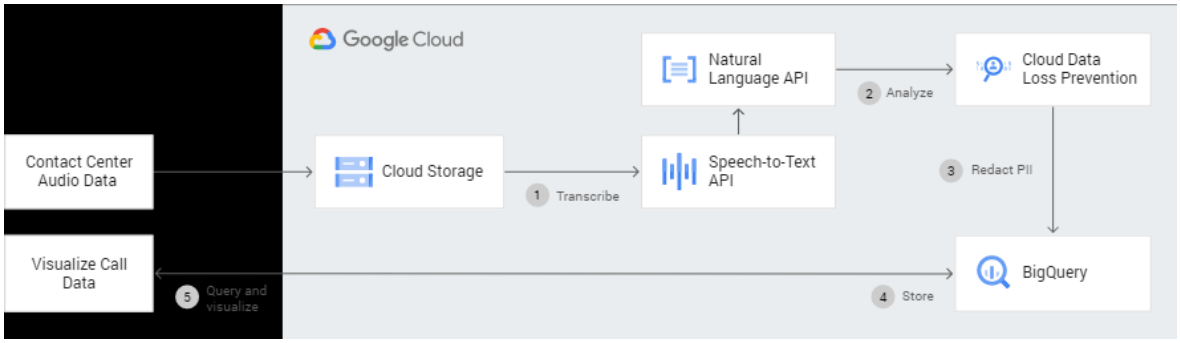
Hình 3.5 Một số dữ liệu trong tập các từ thông dụng – Quick_data 2



Hình 3.6 Một số dữ liệu trong tập các từ thông dụng – Quick_data 3

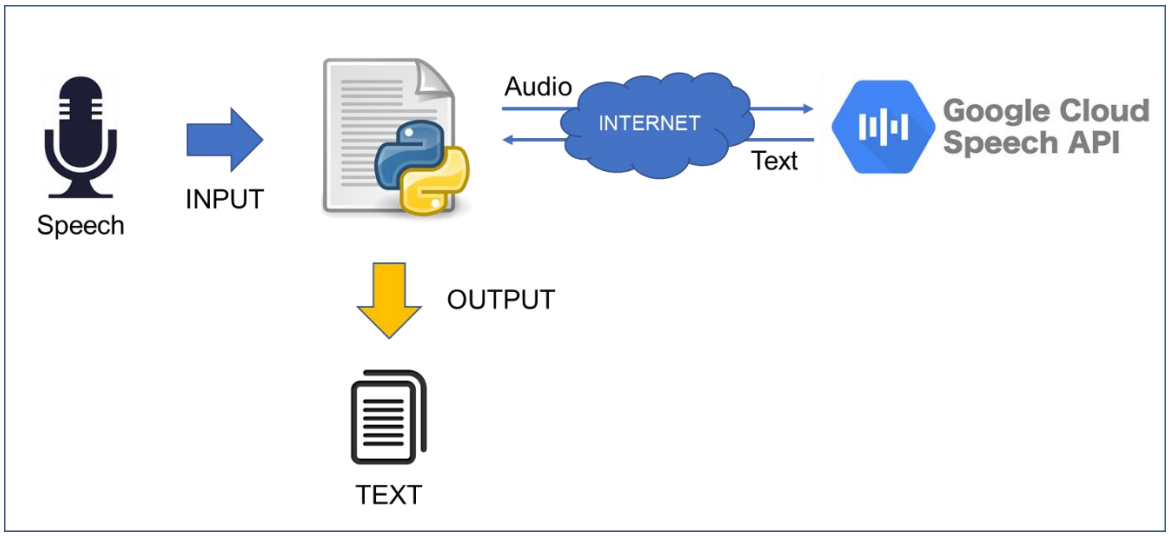
Tập dữ liệu này sẽ được cập nhập liên tục. Tập Quick_data càng lớn, số từ đánh vần sẽ càng ít, tốc độ phiên dịch cũng càng nhanh và khiến người nhìn dễ hiểu hơn.

3.3 Xây dựng thuật toán "Speech to text"



Hình 3.7 Hệ thống phân tích giọng nói của Google

(Nguồn: Google)

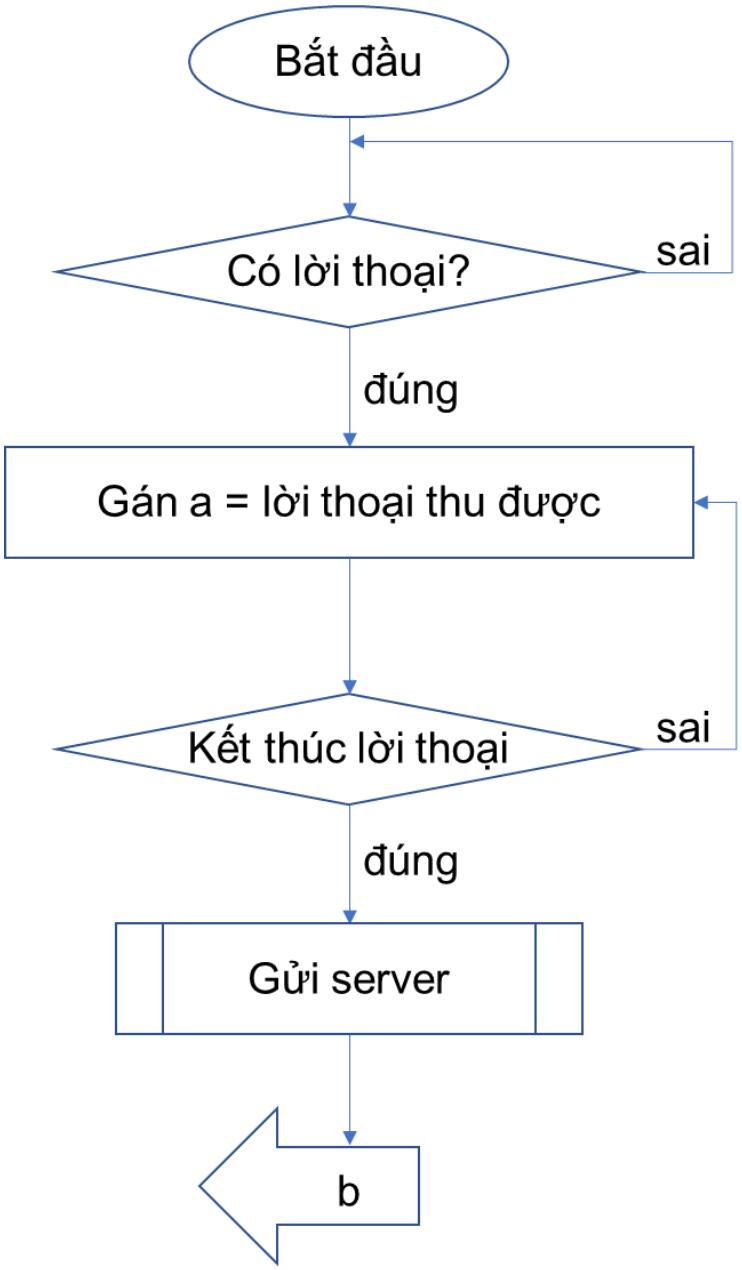


Hình 3.8 Sơ đồ khối thuật toán Speech to text

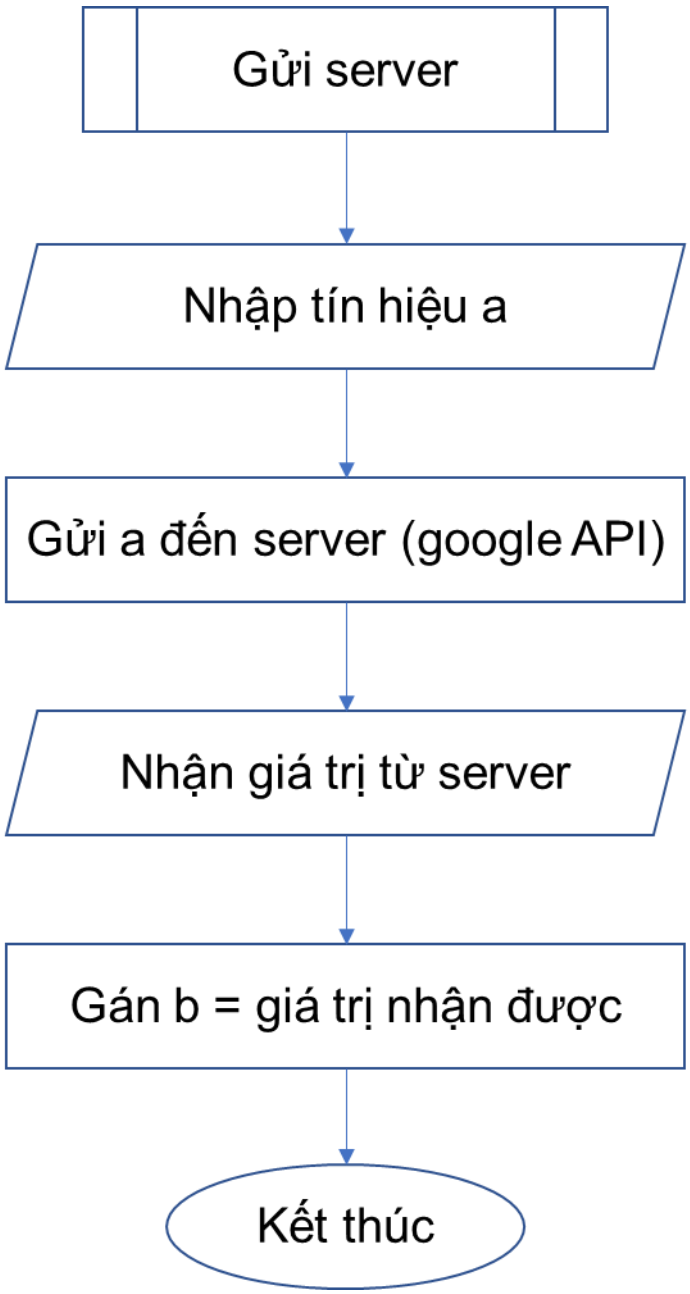
Khi người dùng bắt đầu nói, phần mềm sẽ gắn giá trị nói được là a dưới dạng âm thanh (hình 3.9), phần mềm sẽ luôn kiểm tra liên tục để biết được lúc nào người dùng nói. Khi lời nói dừng lại, phần mềm sẽ hoàn thành việc gán giá trị a và gửi nội dung đến Server dưới dạng Audio (hình 3.8 và hình 3.10).

Server này chính là Google API, nơi sẽ xử lí toàn bộ nội dung âm thanh được gửi đến và trả về giá trị ở dạng văn bản (hình 3.8). Tất cả những bước này được thực hiện ở chương trình con (hình 3.10) và được gọi ở hình 3.8.

Kết thúc bước này, ta có được kết quả là đoạn văn bản thô từ lời nói đầu vào. Kết quả này chỉ và văn bản đơn thuần, không có dấu câu (dấu chấm, phẩy...) nên chương trình không thể hiểu được nội dung của văn bản đó. Để có thể hiểu được ngữ nghĩa của văn bản, ta tiến hành bước tiếp theo là Xử lí ngôn ngữ tự nhiên.



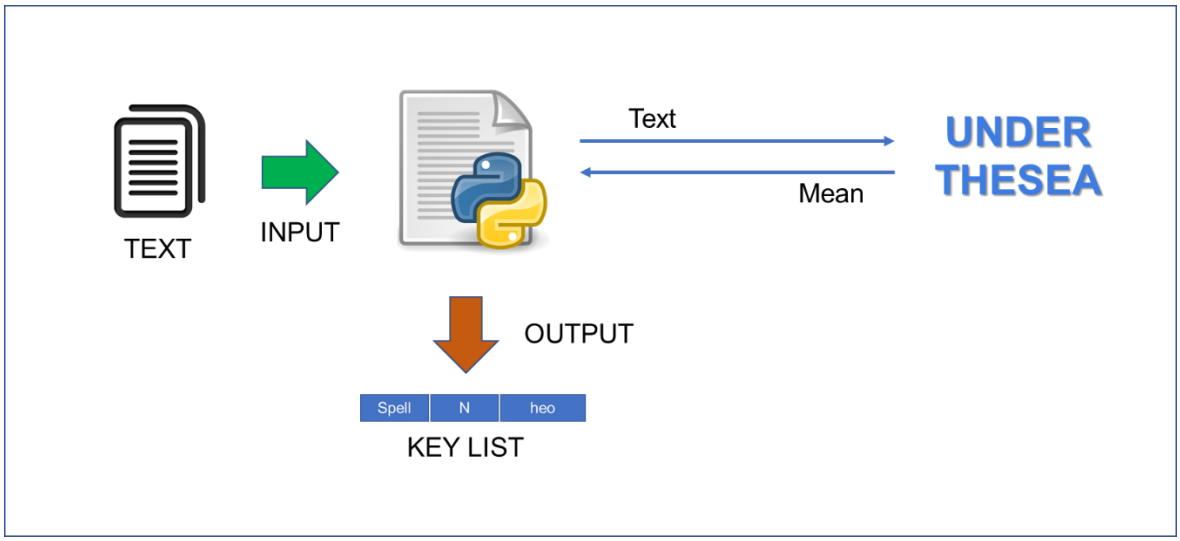
Hình 3.9 Lưu đồ giải thuật chức năng Speech to text



Hình 3.10 Lưu đồ giải thuật chương trình kết nối với Google Cloude

3.4 Xử lí lời nói đầu vào

Ở bước 1, ta có được kết quả là văn bản thô, phần mềm sẽ không thể hiểu được nội dung của đoạn văn bản đó. Để chương trình có thể hiểu được, đề tài áp dụng công nghệ "xử lí ngôn ngữ tự nhiên".



Hình 3.11 Sơ đồ khối xử lí ngôn ngữ đầu vào

Từ gói văn bản thu được, phần mềm sử dụng thư viện Underthesea với bộ dữ liệu có sẵn để thực hiện quá trình xử lí nội dung (Hình 13), bao gồm

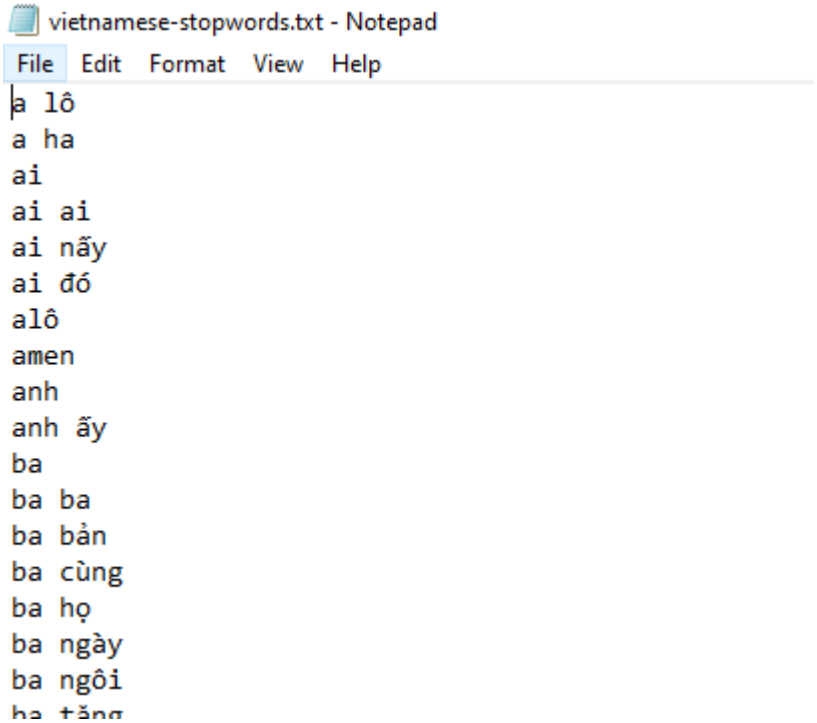
- Tách ý: Văn bản sẽ được tách ra từng câu, thêm dấu chấm để xử lí riêng từng câu, làm tăng độ chính xác. Ở bước này sử dụng gói tokent, với chương trình như sau

Lệnh tách câu
from underthesea import sent_tokenize text = 'Dù anh cứ đi em cũng kệ' token = sent_tokenize(text)
Kết quả thu được: Dù anh cứ đi. Em cũng kệ

- Tách từ: Từ từng câu được tách ở trên, phần mềm sẽ tách từng chữ trong câu ra

Lệnh tách từ trong câu
from underthesea import word_tokenize token2=word_tokenize(token)
[' Dù ', 'anh', 'cứ', 'đi', '.', 'em', 'cũng', 'kệ']

- Loại từ thừa: StopWords là những từ xuất hiện nhiều trong ngôn ngữ tự nhiên, tuy nhiên lại không mang nhiều ý nghĩa. Ở tiếng việt StopWords là những từ như: để, này, kia... Tiếng anh là những từ như: is, that, this... Có rất nhiều cách để loại bỏ StopWords nhưng có 2 cách chính là: Dùng từ điển và dựa theo tần suất xuất hiện của từ. Đề tài sử dụng phương pháp dùng phương pháp từ điển. Từ điển này được hỗ trợ trên module npm và được công khai trên các diễn đàn "xử lí tiếng Việt", được lưu dưới dạng file.txt. Ta có thể so sánh từng từ trong câu với từ điển, nếu trùng có thể xoá từ đó đi



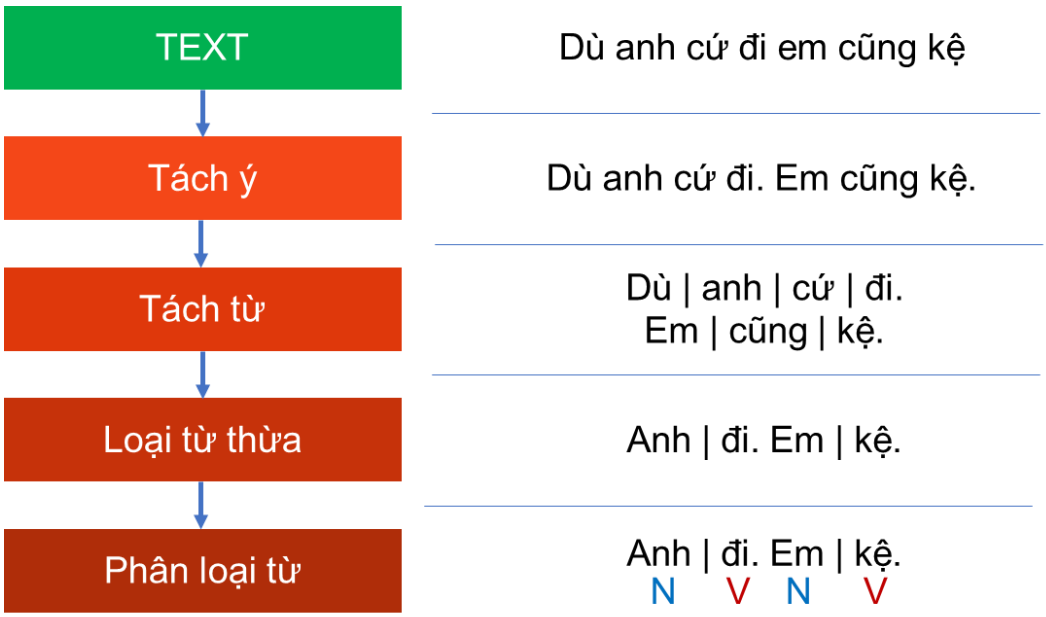
Hình 3.12 Danh sanh Stopword Việt Nam

(Nguồn: Lê Văn Duyệt)

Trong ví dụ, các từ: Dù, cứ, cũng sẽ bị loại

- Phân loại từ: Các từ còn lại sẽ được phân loại từ (danh từ, động từ, tính từ, trạng từ,...)

from underthesea import pos_tag
Pos_tag= pos_tag(token2)
[('Anh', 'N'), ('đi, 'V'), ('em', 'N'), ('kệ', 'V')]



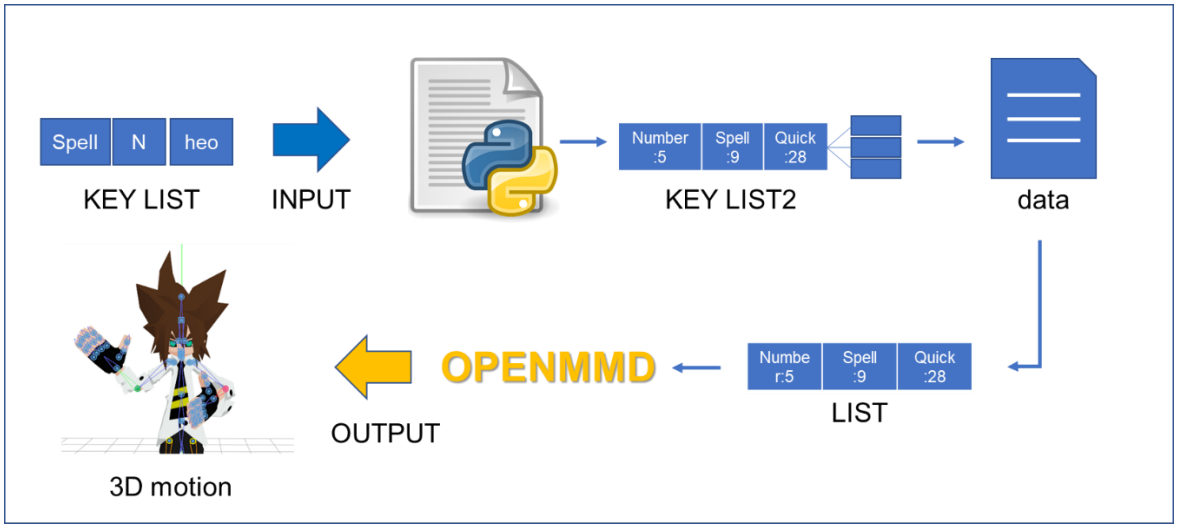
Hình 3.13 Các bước xử lí dữ liệu đầu vào

Sau quá trình phân loại từ, ta được một danh sách gồm cách từ/cụm từ được tách riêng biệt với các thuộc tính tương được. Mỗi từ/cụm từ sẽ được tạo thành 1 mảng gồm 3 phần tử (Hình 3.14). Các phần tử tương đương với:

- Dữ liệu được sử dụng: An[0]
- Từ loại của từ/cụm từ: An[1]
- Nội dung của từ/cụm từ: An[2]

Hình 3.14 Mảng tách từ cụm từ

3.5 Mô hình hoá ngôn ngữ kí hiệu



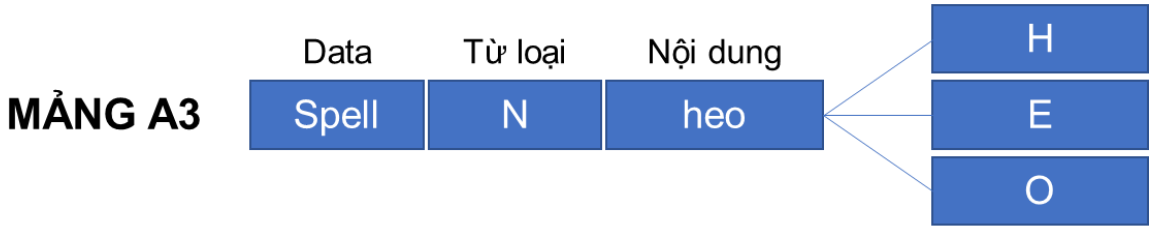
Hình 3.15 Sơ đồ khối chức năng mô phỏng

Có 2 nhiệm vụ cần đạt được trong phần này, đó là tách dữ liệu phần trước thành chuỗi các sự kiện và mô phỏng các sự kiện đó ở dạng 3D

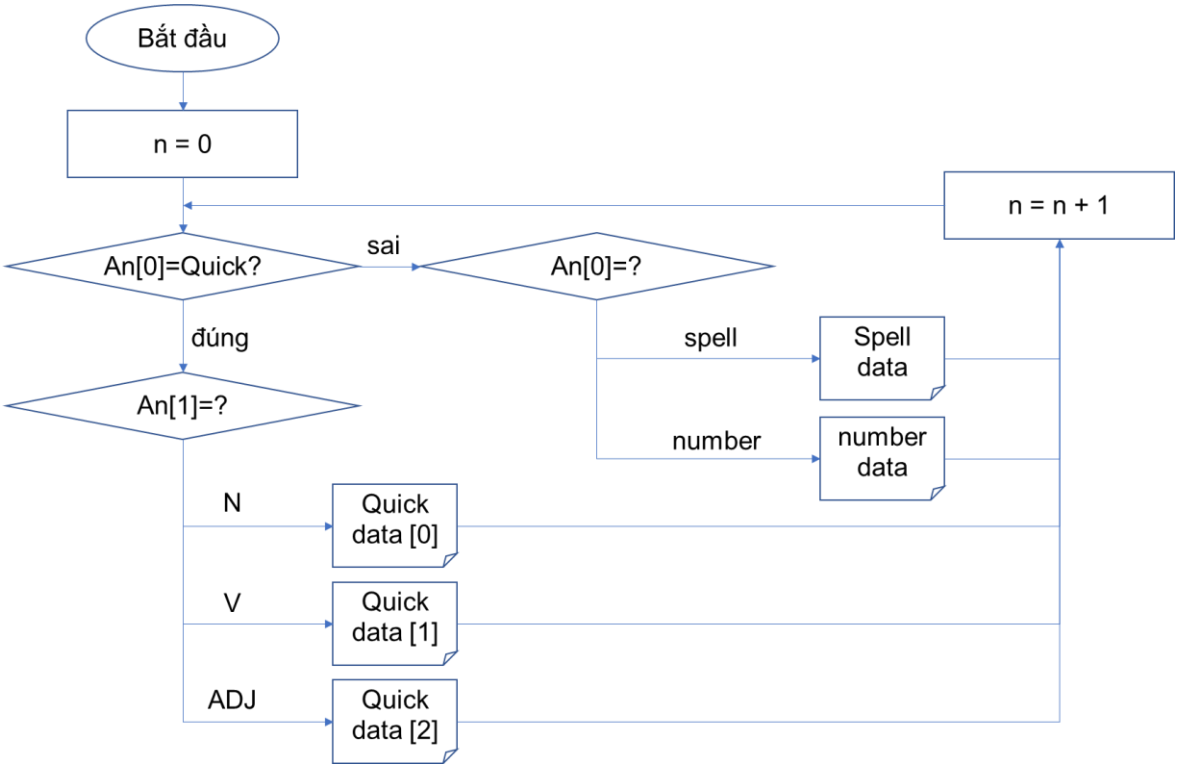
Tách dữ liệu

Từ dữ liệu ở phần trước là các mảng cửa từng từ/cụm từ với 3 phần tử. Chương trình sẽ kiểm tra lần lượt các phần tử đầu tiên của từng mảng, đồng nghĩa với kiểm tra xem từ/cụm từ đó đang cần sử dụng tập dữ liệu nào.

- Nếu là tập số hoặc chữ cái, phần mềm sẽ tách phần tử thứ 3 để chia nội dung của từ/cụm từ đó thành từng chữ như hình 3.16, sau đó lấy dữ liệu tương ứng để so sánh và đưa ra danh sách các sự kiện cần làm.
- Nếu là tập những từ thường gặp, tức là những từ đã được kí hiệu riêng, phần mềm sẽ kiểm tra phần tử thứ 2 của mảng đó, nghĩa là kiểm tra xem từ đó thuộc từ loại nào và so sánh đúng các phần tương ứng thuộc dữ liệu Quick_data. Sơ đồ giải thuật được thể hiện ở hình 3.17.



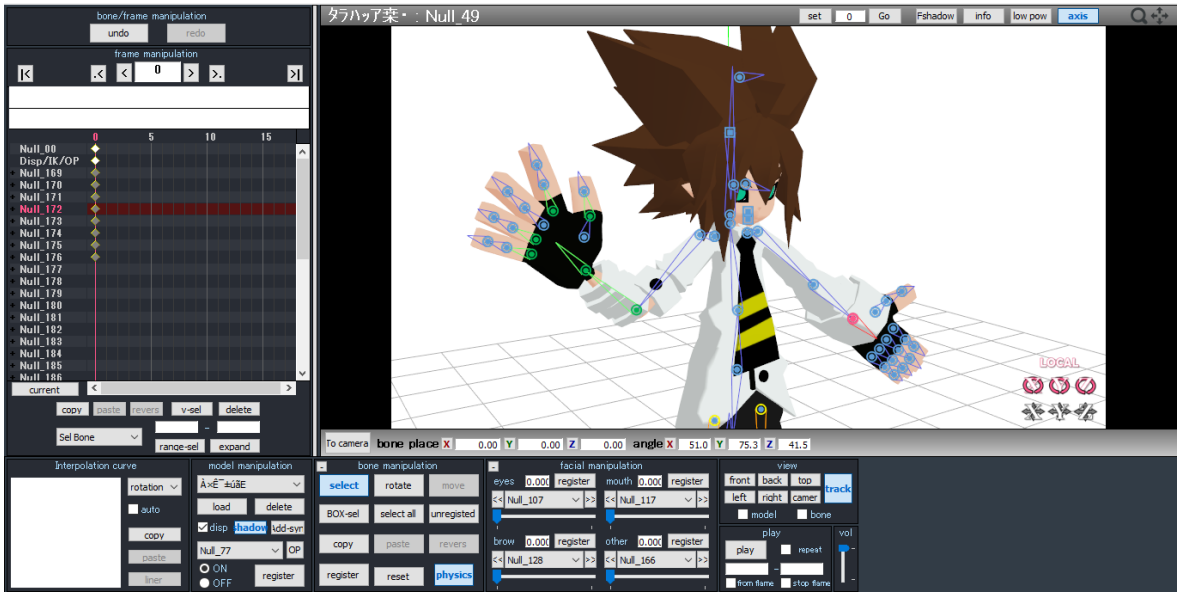
Hình 3.16 Mảng con được tách từ phần tử thứ 3 của mảng chính



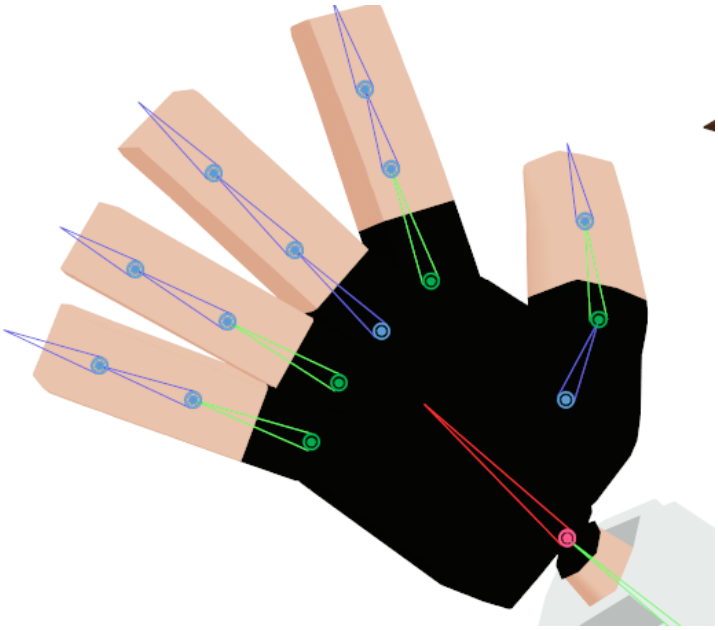
Hình 3.17 Lưu đồ giải thuật chức năng so sánh mảng chính với dữ liệu tương ứng

Mô phỏng

Có 2 cách để tạo dữ liệu mô phỏng. Bao gồm việc sử dụng công nghệ OpenPose, sử dụng video quay lại rồi chuyển nó thành đồ hoạ 3D tự động trên phần mềm OpenMMD. Hoặc thực hiện thủ công trên giao diện chương trình MMD như hình 3.18. Việc thực thi chạy toạ độ thủ công trên giao diện sẽ giúp phần mềm linh hoạt hơn.



Hình 3.18: Giao diện phần mềm



Hình 3.19: Các điểm cố định trên bàn tay

CHƯƠNG 4
KẾT QUẢ VÀ THẢO LUẬN

4.1 Tiến độ thực hiện

Dự án tiến hành nghiên cứu, thực nghiệm tại trung tâm nghiên cứu và giáo dục người khiếm thính (CED)

Về cơ bản, dự án đã hoàn thành xây dựng thuật toán, xây dựng các tập dữ liệu sơ bộ và đang tiến hành hoàn thành đầy đủ các bộ dữ liệu còn thiếu.

4.2 Kết quả thực nghiệm

Sau khi thực nghiệm với một số đối tượng tại Trung tâm nghiên cứu và giáo dục người khiếm thính (CED), đề tài cơ bản có thể lấy được nội dung của người nói trên ngôn ngữ tiếng Việt và tiếng Anh (độ chính xác không cao và thường hay lẫn lộn từ tiếng Anh thành tiếng Việt). Sau khi nhận được nội dung lời nói, phần mềm tiến hành xử lí và đưa ra được các lệnh tương ứng.

Dựa vào kết quả kiểm tra của hệ thống Speech to text dựa trên Google Cloude Speech API, tỷ lệ chính xác chuyển đổi âm thanh thành văn bản lớn hơn 96% đối với môi trường kết nối mạng ổn định [9]. Tốc độ phản hồi cũng giảm trong điều kiện mạng không ổn định. Độ trễ thấp nhất là 50 mili giây [9].

Ở phần xử lí ngôn ngữ tự nhiên, thư viện Underthesea cho độ chính xác cao. Sử dụng bộ dữ liệu VLSP 2016 với 16,858 từ và được chia như bảng 4.1 [14]

Entity Types	**Training Set**	**Test Set**
Location	6,245	1,379
Organization	1,213	274
Person	7,480	1,294
Miscellaneous names	282	49
All	15,222	2,996

Bảng 4.1 Thống kê các thực thể có trong bộ dữ liệu VLSP

Độ chính xác (Precision) được tính theo công thức [4.1], cho kết quả là 90,7% [14]

$$P = \frac{P_{\text{đúng}}}{P_{\text{đúng}} + P_{sai}} \quad [4.1]$$

Recall được gọi là True Positive Rate hay Sensitivity (độ nhạy), được tính theo công thức [4.2], với kết quả là 88.85% [14]

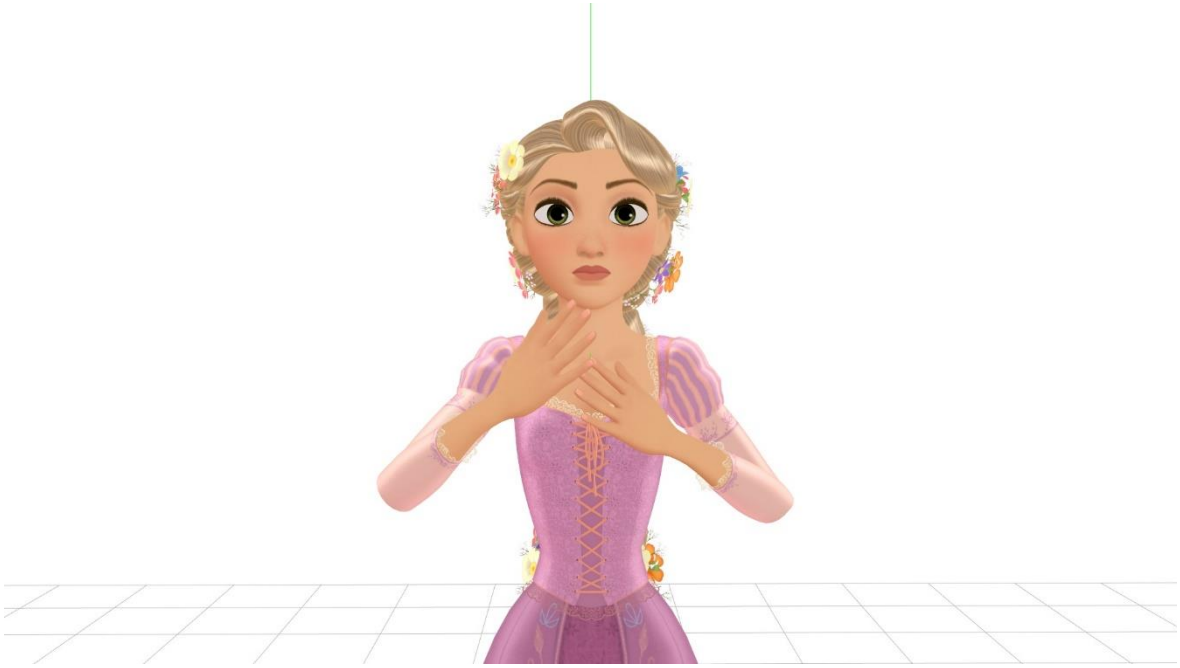
$$R = \frac{R_{\text{đúng}}}{R_{\text{đúng}} + R_{sai}} \quad [4.2]$$

Để bão hoà thông số Precision và Recall, đề tài sử dụng thông số F1-score (trung bình điều hòa - Harmonic mean của các tiêu chí Precision và Recall) với công thức [4.3] và cho ra kết quả cuối cùng là 89.42%

$$F1score = 2\frac{P.S}{P+S} \quad [4.3]$$

Kết quả mô phỏng cho thấy hình ảnh trực quan, có thể hiển thị với nhiều nhân vật khác nhau.

Hình 4.1: Kết quả mô phỏng nhân vật nam



Hình 4.2 Kết quả mô phỏng nhân vật nữ

CHƯƠNG 5
KẾT LUẬN VÀ ĐỀ NGHỊ

5.1 Kết quả khoa học đạt được

Dự án xây dựng thành công thuật toán giúp xử lí tiếng Việt thành ngôn ngữ kí hiệu. Xây dựng một bộ dữ liệu tương ứng giữa ngôn ngữ tiếng Việt và ngôn ngữ kí hiệu theo chuẩn Việt Nam, để các đề tài sau có thể sử dụng cho các mục đích khác nhau.

5.2 Ý nghĩa của dự án

Dự án giúp dịch ngôn ngữ tiếng Việt thành ngôn ngữ kí hiệu, giúp mọi người có thể nói chuyện với người mất khả năng về thính lực. Giúp người mất khả năng thính lực có thể giao tiếp dễ dàng hơn, rút ngắn khoảng cách giữa người điếc, khiếm thị với cộng đồng. Giúp người mất khả năng về thính lực có thể hoà nhập, tiếp thu các kiến thức bên ngoài dễ hơn qua việc dịch nội dung từ các kênh truyền hình thành dạng đồ hoạ 3D.

5.3 Hướng phát triển

Dự án vẫn tiếp tục thu thập dữ liệu về ngôn ngữ kí hiệu với sự hợp tác của trung tâm nghiên cứu và giáo dục người khiếm thính (CED) để làm giàu bộ dữ liệu và tăng độ chính xác của phần mềm.

Dự án phát triển chức năng đọc báo cho người Điếc không biết chữ sau khi hoàn thành bộ dữ liệu “các từ thông dụng”.

TÀI LIỆU THAM KHẢO

Tài liệu tiếng Việt

[1] WHO (2015), *thống kê số người tàn tật trên thế giới*

[2] Bộ Y tế (2017), *thống kê số người mất khả năng thính lực ở Việt Nam*

[5] Nguyễn Văn Ánh (2012), *Đặc điểm của ngôn ngữ ký hiệu Việt Nam*. Truy cập ngày 18 tháng 11 năm 2013

[7] Pearl S. Buck International, Cơ quan Phát triển Quốc tế Hoa Kỳ, Trung tâm tật học Việt Nam (2003), *Ký hiệu của người điếc Việt Nam*. Hà Nội, 2003

[8] TS. Nguyễn Thị Xuyên (2008), *Giao tiếp với trẻ em - Giảm thính lực(khiếm thính) tài liệu số 13*. Nhà xuất bản Y học, Hà Nội, 2008

[12] Đỗ Bá Lâm (2012), *Cải tiến giải thuật earley trong phân tích cú pháp tiếng việt*. Trường Đại học Bách khoa Hà Nội, 2012

Tài liệu tiếng Anh

[3] *Speech and Language Terms and Abbreviations* 2016, Truy cập ngày 2 tháng 12 năm 2006, < http://www.speechlanguage-resources.com>

[4] Encyclopædia Britannica Online. Encyclopædia Britannica Inc (2011), *Deafness*. Truy cập ngày 22 tháng 2 năm 2012

[6] Nguyễn Trần Thủy Tiên (2004). "*Providing higher educational opportunities in Deaf adults in Viet Nam through Vietnamese sign languages: 2000-2003*". Deaf Worlds (bằng tiếng Anh) 20 (3): 232–263.

[9] *Google API,* Truy cập ngày 1 tháng 7 năm 2020, <cloud.google.com>

[10] Daniel Jurafsky, James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition*. Prentice-Hall, 2009.

[11] Christopher Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[13] Z. Cao, T. Simon, S. Wei and Y. Sheikh, "*Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields*," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 1302-1310, doi: 10.1109/CVPR.2017.143.

[14] Pham Quang Nhat Minh (2018), "*A Feature-Rich Vietnamese Named-Entity Recognition Model*" ArXiv, abs/1803.04375.