Sự Phát Triển Của Nhận Dạng Giọng Nói

NGUYỄN Đình-Nguyên[1];NGÔ-HỒ Anh-Khôi[2]
Khoa Kỹ Thuật- Công Nghệ
Trường Đại học Nam Cần Thơ
Việt Nam
[1,]Sinh Viên Nghiên Cứu; [2,]Giảng Viên Hướng Dẫn

***Tóm tắt*—Trong thời đại công nghệ phát triển mạnh mẽ, cùng với cuộc cách mạng công nghiệp 4.0 trên lĩnh vực công nghệ kỹ thuật số. Trong đó ngành trí tuệ nhân tạo có bước tiến mới và đạt những thành tựu nổi bật trên nhiều hướng nghiên cứu khác nhau, sự tương tác giữa con người và máy tính trở nên dễ dàng hơn thông qua các thiết bị đầu vào như chuột, bàn phím, camera, microphone…Và có nhiều cách khác để con người giao tiếp với máy tính, trong đó có giao tiếp bằng nhận dạng giọng nói. Nhu cầu giao tiếp với máy tính bằng tiếng nói đã và đang trở nên cần thiết trong cuộc sống hiên tại và cũng là lĩnh vực giao tiếp tự nhiên và mới mẻ nhất ở thời điểm hiện tại.**

Từ khóa—trí tuệ nhân tạo, sự phát triển giọng nói, nhận dạng, nhận dạng giọng nói.

I. GIỚI THIỆU

Công nghệ thông tin (CNTT), (tên tiếng Anh: Information Technology hay còn gọi là IT) nó là một nhánh ngành kỹ thuật sử dụng máy tính và phần mềm trên máy tính để thực hiện việc lưu trữ, bảo vệ, xử lỹ, xây dựng và thiết kế phần mềm,v.v.. trên nhiều lĩnh vực khác nhau.

Công nghệ thông tin là ngành xoay quanh lĩnh vực công nghệ và hoạt động trên nhiều lĩnh vực khác nhau như phần mềm máy tính, phần cứng máy tính, an toàn thông tin, lập trình,v.v.... Nói cách khác, bất cứ thứ gì được biểu diễn dưới dạng thông tin dữ liệu, thông tin tri thức thông qua các phương tiện truyền thông khác thì được xem là một phần của lĩnh vực công nghệ thông tin.

Trong lĩnh vực CNTT, trí tuệ nhân tạo hay AI (tiếng anh: Artificial Intelligence), là một lĩnh vực đang được áp dụng rông rãi vài đi đôi với cuộc cách mạng công nghiệp 4.0. Trí tuệ nhân tạo được hiểu một cách ngắn gọn là trí thông minh do con người tạo nên và áp đặt lên các thiết bị vô tri như robot, máy móc,v.v…nó trái ngược hoàn toàn với trí tuệ của con người vì nó không có cảm xúc.

Đặc biệt, lĩnh vực có nguồn gốc ra đời trước trí tuệ nhân tạo là nhận dạng giọng nói. Cụ thể, nhận dạng giọng nói trải qua một thời kì gần 3 thế kỷ và phát triển đi đôi với trí tuệ nhân tạo, nó đã và đang là một trong các lĩnh vực tiên phong trong cuộc cách mạng công nghiệp 4.0 hiện nay.

Về phương diện kỹ thuật, nhận diện giọng nói ra đời từ năm 1877 khi nhà bác học Thomas Edison phát minh ra máy ghi âm, thiết bị đầu tiên có thể ghi và tái tạo lại âm thanh.

Nhưng công nghệ nhận dạng giọng nói này chỉ hiểu được số. Sau đóm hệ thống Audrey do Bell Labs chế tạo vào năm 1952 và được coi là thiết bị nhân dạng giọng nói đầu tiên, chỉ nhận ra 10 chữ số được nói bởi 1 giọng duy nhất.

Năm 1962 máy Shoe Box được IBM phát triển, nó có thể nhận ra 16 từ tiếng anh, 10 chữ số và 6 lệnh số học. Đây là bản nâng cấp thiết của hệ thống Audrey.

Từ những năm 1971- 1976, Bộ Quốc phòng Mỹ đã tài trợ cho chương trình DARPA SUR (Nghiên cứu hiểu về lời nói), dẫn đến sự ra đời của Harpy bởi vì Carnegie Mellon có thể hiểu được 1011 từ.

Đến năm 1992, Apple cũng sản xuất hệ thống nhận dạng giọng nói liên tục theo thời gian thực hiện có thể nhận ra tới 20.000 từ.

Đến năm 2008, Google nổi lên ứng dụng Google Voice Search dành cho iPhone.

Vào năm 2010, Google đã giới thiệu ứng dụng nhận dạng được cá nhân hóa trên các thiết bị Android sẽ ghi lại các truy vấn giọng nói của người dùng khác nhau để phát triển một mô hình giọng nói nâng cao, nó bao gồm 230 tỷ từ tiếng anh.

Cuối cùng năm 2011, Siri của Apple đã được triển khai trong iPhone 4S cũng dựa trên điện toán đám mây.

II. SỰ PHÁT TRIỂN CỦA NHẬN DẠNG GIỌNG NÓI TRƯỚC NĂM 2010

A. Dynamic Time Warping - DTW

1) Khái niệm:

Dynamic Time Warping (còn gọi là độ vênh thời gian động) là một cách để so sánh hai trình tự thời gian thường không đồng bộ với nhau một cách hoàn hảo. Nó là một phương pháp để tính toán sự phù hợp tối ưu giữa hai chuỗi. DTW hữu ích trong nhiều lĩnh vực như nhận dạng giọng nói, khai thác dữ liệu, thị trường tài chính,v.v... Nó thường được sử dụng trong khai thác dữ liệu để đo khoảng cách giữa hai chuỗi thời gian.

Nói chung, DTW là một phương pháp tính toán sự phù hợp tối ưu giữa hai chuỗi nhất định với các quy tắc và giới hạn nhất định:

- Mọi chỉ mục từ chuỗi đầu tiên phải được khớp với một hoặc nhiều chỉ số từ chuỗi khác và ngược lại.
- Chỉ mục đầu tiên từ chuỗi đầu tiên phải được khớp với chỉ mục đầu tiên từ chuỗi khác (nhưng nó không phải là khớp duy nhất của nó).
- Chỉ mục cuối cùng từ chuỗi đầu tiên phải được khớp với chỉ mục cuối cùng từ chuỗi khác (nhưng nó không phải là khớp duy nhất của nó).
- Ánh xạ của các chỉ số từ dãy đầu tiên sang các chỉ số từ dãy khác phải tăng đơn điệu và ngược lại, tức là nếu j>I là các chỉ số từ chuỗi đầu tiên, sau đó không được có hai chỉ số l>k trong chuỗi khác, chỉ mục đó I được so khớp với chỉ mục l và chỉ mục j được so khớp với chỉ mục k, và ngược lại.

2) Giải thuật:

a) Công thức.

Giả sử chúng ta có hai chuỗi như sau:

$X = x[1], x[2],\ldots,x[i],\ldots,x[n]$

$Y = y[1], y[2],\ldots,y[j],\ldots,y[m]$

Các chuỗi X và Y có thể được sắp xếp để tạo thành một lưới n-by-m, trong đó mỗi điểm (i, j) là sự liên kết giữa x [i] và y [j].

Một đường cong W ánh xạ các phần tử của X và Y để giảm thiểu khoảng cách giữa chúng. W là một dãy các điểm lưới (i, j). Chúng ta sẽ xem một ví dụ về đường cong vênh sau.

b) Đường công vênh và khoảng cách DTW.

Đường dẫn tối ưu đến (i_k, j_k) có thể được tính bằng:

$$D_{min}(i_k, j_k) = \min_{i_{k-1}, j_{k-1}} D_{min}(i_{k-1}, j_{k-1}) + d(i_k, j_k | i_{k-1}, j_{k-1})$$

trong đó d là khoảng cách Euclide. Sau đó, chi phí đường dẫn tổng thể có thể được tính như:

$$D = \sum_k d(i_k, j_k)$$

c) Hạn chế đối với chức năng Warping.

Đường cong được tìm thấy bằng cách sử dụng phương pháp lập trình động để căn chỉnh hai chuỗi. Đi qua tất cả các con đường có thể là "bùng nổ tổ hợp". Do đó, vì mục đích hiệu quả, điều quan trọng là phải hạn chế số lượng đường dẫn cong vênh có thể xảy ra và do đó các hạn chế sau được nêu ra:

d) Nhận dạng từ bằng giọng nói.

- Điều kiện ranh giới : Ràng buộc này đảm bảo rằng đường cong bắt đầu bằng điểm bắt đầu của cả hai tín hiệu và kết thúc bằng điểm cuối của chúng.

$$i_1 = 1, i_k = n \quad \text{and} \quad j_1 = 1, j_k = m$$

- Điều kiện đơn điệu : Ràng buộc này bảo toàn thứ tự thời gian của điểm (không quay ngược thời gian).

$$i_{t-1} \leq i_t \quad \text{and} \quad j_{t-1} \leq j_t$$

- Điều kiện liên tục (kích thước bước) : Ràng buộc này hạn chế chuyển tiếp đường dẫn đến các điểm liền kề trong thời gian (không nhảy trong thời gian).

$$i_t - i_{t-1} \leq 1 \quad \text{and} \quad j_t - j_{t-1} \leq 1$$

- Điều kiện cửa sổ cong vênh: Có thể hạn chế các điểm cho phép nằm trong một cửa sổ cong vênh nhất định có chiều rộng ω (một số nguyên dương).

$$|i_t - j_t| \leq \omega$$

- Điều kiện độ dốc : Đường cong có thể bị hạn chế bằng cách hạn chế độ dốc, và do đó tránh các chuyển động cực đoan theo một hướng.

  Di chuyển ngang: $(i, j) \rightarrow (i, j + 1)$

  Di chuyển dọc: $(i, j) \rightarrow (i + 1, j)$

  Di chuyển theo đường chéo: $(i, j) \rightarrow (i + 1, j + 1)$

3) Ứng dụng:

a) Nhận dạng từ bằng giọng nói.

Do tốc độ nói của mỗi người có sự khác nhau, dao động phi tuyến tính sẽ xảy ra trong mẫu giọng nói so với trục thời gian, cần được loại bỏ. Đối sánh DP là một thuật toán dựa trên lập trình động (DP), nó sử dụng hiệu ứng chuẩn hóa theo thời gian, trong đó các dao động trong trục thời gian được mô hình hóa bằng cách sử dụng hàm làm cong thời gian phi tuyến tính.

Kiểm tra hai mẫu giọng nói bất kỳ, chúng ta có thể loại bỏ sự khác biệt về thời gian của chúng bằng cách làm cong trục thời gian của một mẫu để đạt được sự trùng hợp với trục còn lại. Mặc khác, nếu hàm cong vênh được phép nhận bất kỳ giá trị nào, có thể rất ít phân biệt giữa các từ thuộc các loại khác nhau. Vì vậy, để tăng cường sự phân biệt giữa các từ thuộc các loại khác nhau, các hạn chế đã được áp dụng đối với độ dốc hàm cong vênh.

b) Phân tích sức mạnh tương quan.

Đồng hồ không ổn định được sử dụng để đánh bại phân tích công suất ngây thơ. Một số kỹ thuật được sử dụng để chống lại sự phòng thủ này, một trong số đó là sự cong vênh thời gian động.

B. Chuỗi Markov - Markov

1) Khái niệm:

Một xích Markov hay chuỗi Markov là một quá trình ngẫu nhiên mô tả một dãy các biến cố khả dĩ trong đó xác suất của mỗi biến cố chỉ phụ thuộc vào trạng thái của biến cố trước đó. Một dãy vô hạn đếm được, trong đó xích thay đổi trạng thái theo từng khoảng thời gian rời rạc, cho ta một xích Markov thời gian rời rạc (DTMC). Một quá trình diễn ra trong thời gian liên tục được gọi là xích Markov thời gian liên tục (CTMC). Chúng được đặt tên theo nhà toán học người Nga Andrey Markov.

Xích Markov có được ứng dụng rộng rãi làm mô hình thống kê của nhiều quá trình đời thực, như là nghiên cứu hệ thống điều khiển hành trình trong các xe motor, hàng đợi hay hàng người đến sân bay, tỉ giá hối đoái tiền tệ và sự biến đổi của dân số quần thể.

Quá trình Markove là cơ sở cho phương pháp mô phỏng ngẫu nhiên xích Markov Monte Carlo, được dùng để mô phỏng việc lấy mẫu từ một phân bố xác suất phức tạp, và có ứng dụng trong thống kê Bayes, nhiệt động lực học, cơ học thống kê, vật lý, hóa học, kinh tế, tài chính, xử lý tín hiệu, lý thuyết thông tin và trí tuệ nhân tạo.

2) Giải thuật:

a) Xích Markov thời gian rời rạc.

Một xích Markov thời gian rời rạc là một dãy các biến ngẫu nhiên X_1, X_2, X_3, ... với tính chất Markov, tức xác suất chuyển sang trạng thái tiếp theo chỉ phụ thuộc vào trạng thái hiện tại chứ không phụ thuộc vào những trạng thái trước đó:

$$\Pr(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \Pr(X_{n+1} = x \mid X_n = x_n)$$

nếu cả hai xác suất có điều kiện này có nghĩa, tức là nếu:

$$\Pr(X_1 = x_1, \ldots, X_n = x_n) > 0.$$

Những giá trị khả dĩ của X_i tạo thành một tập đếm được S gọi là không gian trạng thái của xích.

b) Xích Markov thời gian liên tục.

Một xích Markov thời gian liên tục $(X_t)_{t \geq 0}$ được định nghĩa bởi một không gian trạng thái hữu hạn hoặc đếm được S, một ma trận tốc độ chuyển đổi Q với các chiều bằng với chiều của không gian trạng thái, và một phân bố xác suất ban đầu trên không gian trạng thái đó. Với $i \neq j$, phần tử q_{ij} không âm và diễn tả tốc độ của quá trình khi chuyển từ trạng thái i

sang trạng thái j. Các phần tử q_{ij} được chọn sao cho mỗi hàng của ma trận tốc độ chuyển đổi có tổng bằng 0.

3) Ứng dụng:

Ứng dụng chuỗi Markov trong các lĩnh vực như: vật lý, hoá học, sinh học, âm nhạc, thể thao, công nghệ,…
Một số ứng dụng trong các lĩnh vực điển hình như:

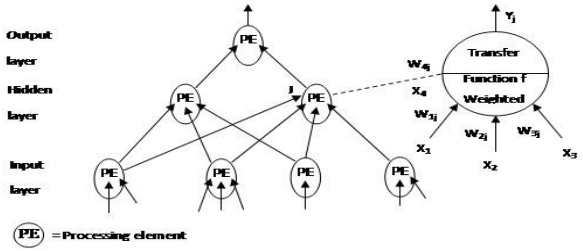
- Vật lý: Hệ thống Markov xuất hiện rộng rãi trong nhiệt động lực học, cơ học thống kê, cơ học lượng tử…
- Sinh học: Động lực học quần thể, sinh học hệ thống, các mô hình khoang, Phylogenetics và tin sinh học…
- Nhận dạng giọng nói: Mô hình Markov là cơ sở cho hầu hết các hệ thống nhận dạng giọng nói hiện đại nhất hiện nay.
- Âm nhạc: Markov được sử dụng trong sáng tác theo nhạc thuật toán.

C. Mạng Nơ-ron nhân tạo - Artificial Neural Network

1) Khái niệm:

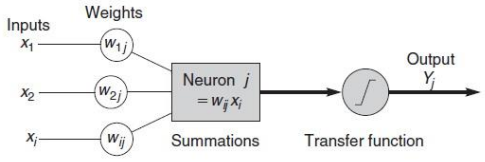
Mạng Nơron nhân tạo (Artificial Neural Network-ANN) là mô hình xử lý thông tin được mô phỏng dựa trên hoạt động của hệ thống thần kinh của sinh vật, bao gồm số lượng lớn các Nơron được gắn kết để xử lý thông tin. ANN giống như bộ não con người, được học bởi kinh nghiệm (thông qua huấn luyện), có khả năng lưu giữ những kinh nghiệm hiểu biết (tri thức) và sử dụng những tri thức đó trong việc dự đoán các dữ liệu chưa biết (unseen data).

2) Giải thuật:



Trong đó các Processing Elements (PE) của ANN gọi là Nơron, mỗi Nơron nhận các dữ liệu vào (Inputs) xử lý chúng và cho ra một kết quả (Output) duy nhất. Kết quả xử lý của một Nơron có thể làm Input cho các Nơron khác.

Quá trình xử lý thông tin của một ANN:



- Inputs (dữ liệu vào): Mỗi Input tương ứng với 1 thuộc tính (attribute) của dữ liệu (patterns).
- Output (kết quả): Kết quả của một ANN là một giải pháp cho một vấn đề.
- Connection Weights (Trọng số liên kết) : Đây là thành phần rất quan trọng của một ANN, nó thể hiện mức độ quan trọng (độ mạnh) của dữ liệu đầu vào đối với quá trình xử lý thông tin (quá trình chuyển đổi dữ liệu từ Layer này sang layer khác). Quá trình học (Learning Processing) của ANN thực ra là quá trình điều chỉnh các trọng số (Weight) của các input data để có được kết quả mong muốn.
- Summation Function (Hàm tổng): Tính tổng trọng số của tất cả các input được đưa vào mỗi Nơron (phần tử xử lý PE). Hàm tổng của một Nơron đối với n input được tính theo công thức sau:

$$Y = \sum_{i=1}^{n} X_i W_i$$

- Transfer Function (Hàm chuyển đổi): Hàm tổng (Summation Function) của một Nơron cho biết khả năng kích hoạt (Activation) của Nơron đó còn gọi là kích hoạt bên trong (internal activation). Các Nơron này có thể sinh ra một output hoặc không trong ANN (nói cách khác rằng có thể output của 1 Nơron có thể được chuyển đến layer tiếp trong mạng Nơron hoặc không). Mối quan hệ giữa Internal Activation và kết quả (output) được thể hiện bằng hàm chuyển đổi (Transfer Function).
- Việc lựa chọn Transfer Function có tác động lớn đến kết quả của ANN. Hàm chuyển đổi phi tuyến được sử dụng phổ biến trong ANN là sigmoid (logical activation) function.

$$Y_T = 1/(1 + e^{-Y})$$

- Trong đó:

  Y_T: Hàm chuyển đổi

  Y: Hàm tổng

- Kết quả của Sigmoid Function thuộc khoảng [0,1] nên còn gọi là hàm chuẩn hóa (Normalized Function).

3) Ứng dụng:

Các ứng dụng lĩnh vực bao gồm hệ thống nhận dạng và điều khiển (điều khiển phương tiện, dự án quỹ đạo, điều khiển quá trình, quản lý tài nguyên tự nhiên), hóa lượng tử, chơi game nói chung , nhận mẫu, nhận dạng, chẩn đoán y tế, tài chính (ví dụ: hệ thống giao dịch tự động), khai thác dữ liệu,...

ANN đã được sử dụng để tăng độ tin cậy phân tích tốc độ của các cơ sở hạ tầng chịu ảnh hưởng của thiên tai và dự án lún nền. ANN cũng được sử dụng để xây dựng hộp đen đen trong khoa học địa chỉ: thủy văn, mô hình đại dương và kỹ thuật ven biển and địa chỉ. ANN sử dụng trong an ninh mạng, với phân biệt mục giữa các hoạt động hợp pháp và hoạt động độc hại. Ví dụ: máy học đã được sử dụng để phân loại phần mềm độc hại Android, để xác định các thuộc tính miền về các kẻ đe dọa tác động và để phát hành các URL có cơ chế bảo mật. Nghiên cứu được tiến hành trên ANN hệ thống được thiết kế để kiểm tra thâm nhập, nhằm mục đích phát hiện botnet, gian thẻ tín dụng và mạng phân loại.

ANN đã được đề xuất như một công cụ để mô phỏng các thuộc tính của hệ thống mở rộng nhiều phần thân. Trong bộ não nghiên cứu, ANN nghiên cứu hành vi ngắn hạn của tế bào thần kinh riêng lẻ, động lực của mạch thần kinh phát sinh từ sự tương tác giữa các tế bào thần kinh riêng lẻ và cách hành vi có thể phát sinh từ đại diện mô-đun thần kinh hỗ trợ cho các hệ thống hoàn chỉnh. Các nghiên cứu được xem xét tính toán thời gian dài và thời gian ngắn của hệ thống thần kinh và mối

quan hệ của chúng tôi với tập tin học và ghi nhớ từ lẻ nơ-ron đến hệ thống cấp độ.

III. Sự Phát Triển Của Nhận Dạng Giọng Nói Sau Năm 2010

Trước những năm 2010, nhận dạng giọng nói được biết đến một cách

A. *Mạng neuron sâu (DNN-Deep neural Network)*

1) Khái niệm:

Mạng neuron sâu (DNN-Deep neural Network) là một mạng neuron nhân tạo (ANN) với nhiều đơn vị lớp ẩn giữa lớp đầu vào và đầu ra. Tương tự như các ANN nông, các DNN nông có thể mô hình mối quan hệ phi tuyến phức tạp. Các kiến trúc DNN, ví dụ như để phát hiện và phân tích đối tượng tạo ra các mô hình hỗn hợp trong đó đối tượng này được thể hiện như một thành phần được xếp lớp của các hình ảnh nguyên thủy. Các lớp phụ cho phép các thành phần của các đặc điểm từ các lớp thấp hơn, đem lại tiềm năng của mô hình hóa dữ liệu phức tạp với các đơn vị ít hơn so với một mạng lưới nông thực hiện tương tự như vậy.

Các DNN thường được thiết kế như các mạng nuôi tiến, nhưng nghiên cứu gần đây đã áp dụng thành công kiến trúc học sâu đối với các mạng nơ ron tái phát cho các ứng dụng chẳng hạn như mô hình hóa ngôn ngữ. Các mạng neuron sâu tích chập (CNN) được sử dụng trong thị giác máy tính nơi thành công của chúng đã được ghi nhận. Gần đây hơn, các CNN đã được áp dụng để mô hình hóa âm thanh cho nhận dạng giọng nói tự động (ASR), nơi chúng đã cho thấy sự thành công trong các mô hình trước đó. Để đơn giản, ta hãy nhìn vào việc huấn luyện các DNN được đưa ra ở đây.

2) Giải thuật:

$$p_j = \frac{\exp(x_j)}{\sum_k \exp(x_k)}$$

3) Ứng dụng:

Deep Neural Network được ứng dụng trong lĩnh vực Deep Learning để xử lý và giải quyết vấn đề.

Dịch thuật: Phương pháp học sâu (deep learning algorithms) sẽ dịch các ngôn ngữ, giúp khách du lịch hay những người có nhu cầu về công việc có thể dễ dàng trao đổi, tương tác với nhau mà không cần phải chật vật như xưa.

Phương tiện không người lái: Nghe có vẻ viễn tưởng nhưng trên thực tế những chiếc máy bay không người lái và những xe hơi tự hành giờ đây không còn xa lạ với đời sống của chúng ta. Nhờ deep learning, sau khi các thuật toán tiếp nhận và xử lý các bộ dữ liệu sẽ có khả năng hành động giống như con người, "nhìn" được thực tế đường đi, di chuyển, dừng lại hay tránh các xe khác.

Trợ lý ảo & các dịch vụ chatbots/bots: Các nhà cung cấp dịch vụ trực tuyến giờ đây thường dùng các trợ lý ảo hay các dịch vụ chatbots, bots để chăm sóc khách hàng. Việc này không chỉ mang lại hiệu quả về chi phí mà còn nâng cao được năng suất nhờ vào số lượng câu hỏi được giải đáp cũng như tốc độ phản hồi tăng lên nhanh chóng.

Y học:Thông qua phương pháp deep learning và công nghệ deep neural network, bệnh nhân giờ đây được chẩn đoán, kê đơn một cách nhanh chóng và chính xác hơn. Các công ty, tổ chức trong lĩnh vực y tế ngày càng đầu tư nhiều vào loại hình công nghệ này nhằm giảm tải cho các bác sĩ và giải phóng áp lực cho bệnh viện hay các cơ sở khám chữa.

Mua sắm & Giải trí: Sáng nay bạn đăng một bức hình đạp phải bãi "mìn" của chú cún cưng khi vừa tỉnh giấc lên Facebook thì lập tức, tối (thậm chí là chỉ vài tiếng sau) bạn sẽ thấy quảng cáo thảm chùi chân, thuốc xịt dạy cún đi vệ sinh đúng chỗ và những thứ tương tự hiện liên tục trên News Feed của mình. Đây chính là kết quả có được nhờ Deep Learning và công nghệ Deep Neural Network mà ta đang đề cập.

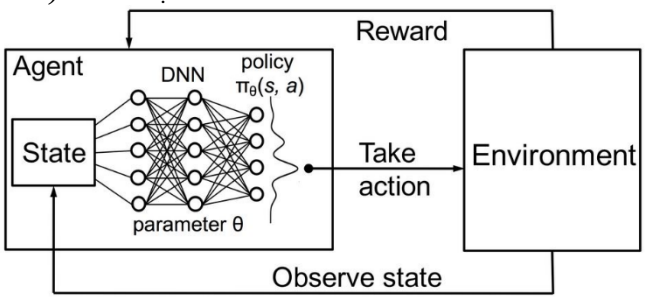
Nhận dạng tiếng nói tự động
Nhận dạng hình ảnh
Xử lý ngôn ngữ tự nhiên
Khám phá dược phẩm và độc chất học
Quản lý quan hệ khách hàng (CRM)
Tin sinh học

B. *Hệ thống action-state-reward*

1) Khái niệm:

Đây là một kết hợp mạnh mẽ giữa Deep Learning và Reinforcement Learning, trong đó Neural Network tham gia vào quá trình quyết định hành động của Reinforcement Learning, từ đó dẫn đến một quá trình học tối ưu hơn!

2) Giải thuật:



3) Ứng dụng:

- Tô màu cho tranh trắng đen Thông thường, cần có một một chuyên gia chỉnh ảnh để tô màu cho các bức ảnh trắng đen và công việc này tốn rất nhiều công sức và thời gian. Tuy nhiên, hiện nay, sử dụng một Convolutional Network lớn để học màu sắc, thuật toán của Machine Learning có khả năng tô màu cho các ảnh trắng đen như cách một chuyên gia sẽ làm.
- Lồng tiếng cho phim câm Ứng dụng này sử dụng cả Convolutional Neural Network và Recurrent Neural Network, trong đó Convolutional Neural Network được dùng để nhận dạng cảnh phim, sau đó Recurrent Neural Network được sử dụng để tạo ra chuỗi âm thanh (tiếng trống) phù hợp với cảnh phim vừa được nhận diện. Khi một chiếc máy tính sử dụng cả hai thuật toán trên và một chuyên gia con người cùng lồng tiếng cho một cảnh phim, không ai có thể phân biệt được âm thanh được tạo ra là do máy hay do người (bài kiểm tra này gọi là Turing test, bài kiểm tra độ thông minh của trí tuệ nhân tạo).
- Máy phiên dịchSử dụng Recurrent Neural Network, Machine Learning có khả năng phiên dịch giữa hai thứ tiếng. Hơn nữa, thuật toán không chỉ có khả năng phiên dịch các câu cú được nhập vào mà còn có khả năng phiên dịch hình ảnh nhờ sử dụng Convolution

Neural Network để nhận diện các từ xuất hiện trong ảnh.

- Nhận diện chữ viết taySử dụng Recurrent Neural Network, thuật toán có khả năng học cách viết chữ từ nét chữ của người. Sau đó, thuật toán có thể viết ra bất kỳ dòng chữ nào theo yêu cầu sử dụng nét viết tương tự.
- Viết truyện – làm thơ – viết quảng cáo…

IV. NGOẠI VI VÀ NỘI VI

A. Ngoại vi

Ngày nay, có 3 khái niệm được nhắc tới khi nói đến cách để làm một công cụ thông minh hơn. Đó là: Máy học (Machine Learning), Máy học sâu (Deep Learning), Trí tuệ nhân tạo (Artificial Intelligence). Trí tuệ nhân tạo là trí tuệ máy móc được tạo ra bởi con người. Trí tuệ này có thể tư duy, suy nghĩ, học hỏi,… như con người. Xử lý dữ liệu ở mức độ rộng hơn, quy mô hơn, hệ thống, khoa học và nhanh hơn so với con người. AI có 3 loại khác nhau dựa vào năng lực để phân chia:

- AI hẹp (Artificial Narrow Intelligence - ASI): Trí tuệ nhân tạo được cho là hẹp khi máy có thể thực hiện một nhiệm vụ cụ thể tốt hơn so với con người. Nghiên cứu hiện tại về AI hiện đang ở cấp độ này.
- AI toàn năng (Artificial General Intelligence - AGI): là cấp độ thứ hai mà trí tuệ nhân tạo đạt đến trạng thái chung khi nó có thể thực hiện bất kỳ nhiệm vụ sử dụng trí tuệ nào có cùng độ chính xác như con người
- AI siêu phàm (Artificial Superintelligence - ASI) là một trí tuệ có khả năng tư duy gấp hàng tỷ con người gộp lại, nó liên tục học hỏi, cải tiến, và nó có thể sẽ không đặt "hạnh phúc của con người hay sự tồn tại" là ưu tiên của nó.

1) Machine Learning:

Machine Learning là một hệ thống có thể học từ ví dụ thông qua tự cải thiện và không được lập trình viên mã hóa rõ ràng. Bước đột phá đi kèm với ý tưởng rằng một cỗ máy có thể học hỏi từ dữ liệu (ví dụ) để tạo ra kết quả chính xác.Machine learning theo định nghĩa cơ bản là ứng dụng các thuật toán để phân tích cú pháp dữ liệu, học hỏi từ nó, và sau đó thực hiện một quyết định hoặc dự đoán về các vấn đề có liên quan. Vì vậy, thay vì code phần mềm bằng cách thức thủ công với một bộ hướng dẫn cụ thể để hoàn thành một nhiệm vụ cụ thể, máy được "đào tạo" bằng cách sử dụng một lượng lớn dữ liệu và các thuật toán cho phép nó học cách thực hiện các tác vụ.Machine learning bắt nguồn từ các định nghĩa về AI ban đầu, và các phương pháp tiếp cận thuật toán qua nhiều năm bao gồm:logic programming, clustering, reinforcement learning, and Bayesian networks. Như chúng ta đã biết, không ai đạt được mục tiêu cuối cùng của General AI, và thậm chí cả Narrow AI hầu hết là ngoài tầm với những phương pháp tiếp cận Machine learning sơ khai. Có 2 cách học phổ biến là:

- Học có giám sát (Supervised Learning): Dataset có kết quả thực để kiểm tra.
- Học không giám sát (Unsupervised Learning): Dataset không có kết quả thực để kiểm tra, để máy tính tự nhìn thấy các mối quan hệ tiềm ẩn trong dữ liệu.

Machine Learning được sử dụng trong việc phân loại thư spam, hệ thống đề xuất phim, nhận diện ảnh và giọng nói,... Thêm nữa là một kỹ thuật trong Machine Learning rất được ưa chuộng hiện nay là Neural Network.

2) Deep Learning:

Deep learning là một phần mềm máy tính bắt chước mạng lưới các nơ-ron trong não con người. Nó là một tập hợp con của Machine Learning và được gọi là Deep Learning vì nó sử dụng các deep neurak networks. Có thể nói Deep Learning là kỹ thuật để hiện thực hóa Machine Learning. Một phương pháp tiếp cận thuật toán khác từ cộng đồng machine-learning, Artificial Neural Networks, được nhắc đến nhiều thập kỷ qua. Neural Networks được lấy cảm hứng từ sự hiểu biết về sinh học của bộ não loài người – sự liên kết giữa các nơ-ron. Tuy nhiên, không giống như một bộ não sinh học nơi mà bất kỳ nơ-ron nào cũng có thể liên kết với các nơ-ron khác trong một khoảng cách vật lý nhất định, các mạng thần kinh nhân tạo này có các lớp rời rạc, các kết nối, và các hướng truyền dữ liệu.Chẳng hạn, bạn có thể lấy một hình ảnh, cắt nó thành một nhóm được đặt vào lớp đầu tiên của mạng thần kinh nhân tạo. Trong lớp đầu tiên các nơ-ron cá nhân truyền dữ liệu đến lớp thứ hai. Lớp thứ hai của nơ-ron làm nhiệm vụ của nó, và như vậy, cho đến khi lớp cuối cùng và cho ra sản phẩm cuối cùng.Mỗi nơ-ron đảm nhiệm một chức năng – làm thế nào để biết chính xác liệu rằng nó có liên quan đến nhiệm vụ đang được thực hiện. Vì vậy, suy nghĩ về điểm dừng là một dấu hiệu. Các thuộc tính của một hình ảnh dấu hiệu "dừng" được cắt nhỏ và được "kiểm tra" bởi các nơ-ron – dạng hình trụ, màu đỏ của các động cơ cháy, các chữ cái đặc trưng, kích thước biển báo giao thông, và sự chuyển động hoặc sự thiếu hụt của nó. Nhiệm vụ của mạng thần kinh là để kết luận liệu đây có phải là dấu hiệu dừng hay không. Nó đi kèm với một "vector xác suất". Trong ví dụ của chúng ta, hệ thống có thể xác định chắc chắn đến 86% một dấu hiệu dừng, 7% rằng đó là một dấu hiệu giới hạn tốc độ, và 5% còn lại là một con diều bị mắc kẹt trong cây,(hoặc cái gì đó tương tự) vv … và kiến trúc mạng sau đó sẽ thông báo đến mạng nơron cho dù đó là đúng hay sai.Thậm chí ví dụ này cũng là một sự tiến bộ, bởi vì mạng lưới thần kinh đã có thể làm được tất cả nhưng bị xa lánh bởi cộng đồng nghiên cứu về AI. Nó đã có mặt từ những ngày đầu tiên của AI, và tạo ra rất ít sản phẩm "trí tuệ". Vấn đề là ngay cả những mạng nơ-ron cơ bản nhất cũng có tính toán rất cao, nó không phải là cách tiếp cận thực tiễn. Tuy nhiên, một nhóm nghiên cứu nhỏ do Geoffrey Hinton thuộc trường đại học Toronto đứng đầu, cuối cùng đã parallelizing các thuật toán cho siêu máy tính để chạy và chứng minh khái niệm, nhưng nó không chính xác cho đến khi GPU được triển khai .

Nếu chúng ta quay trở lại ví dụ "ký hiệu dừng", rất có thể là khi mạng đang được điều chỉnh hoặc được "đào tạo" thì sẽ có câu trả lời sai – rất nhiều. Những gì nó cần là luyện tập. Nó cần phải nhìn thấy hàng trăm ngàn, thậm chí hàng triệu hình ảnh, cho đến khi trọng lượng của đầu vào nơ-ron được điều chỉnh chính xác đến mức nó có được câu trả lời ngay trong thực tế mọi lúc – sương mù hoặc không có sương mù, nắng hoặc mưa. Vào thời điểm đó mạng thần kinh đã tự dạy cho nó một dấu hiệu dừng như thế nào; Hoặc khuôn mặt của mẹ bạn trong trường hợp của Facebook. Hay một con mèo, đó là điều mà Andrew Ng đã làm trong năm 2012 tại Google.Sự đột phá của Ng là đưa các mạng thần kinh này, và làm cho chúng trở nên to lớn, tăng số layer và các nơ-ron, sau đó chạy một khối lượng lớn dữ liệu thông qua hệ thống để huấn luyện nó. Trong trường hợp của Ng, đó là hình ảnh từ 10 triệu video

trên YouTube. Ng đặt "deep" vào deep learning, mô tả tất cả các lớp trong các mạng nơron này.

B. Nội vi

Nhận dạng giọng nói đã được biết đến hàng thập kỷ, tại sao chỉ đến bây giờ, công nghệ mới thực sự bùng nổ? Theo wikipedia, khó khăn cơ bản của nhận dạng giọng 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. Sự ra đời của Deep Learning đã giúp nhận diện giọng nói chính xác, thậm chí ở ngoài môi trường phòng lab.

Để AI thông minh thì cần phải có dữ liệu để huấn luyện cho nó, cả về nhận diện hình ảnh, văn bản, giọng nói. Google có hàng tỷ người dùng với công cụ tìm kiếm, nó có thể biết được trong khoảng thời gian nào, trong từng thời điểm người dùng quan tâm từ khóa nào, lĩnh vực nào. Đó là một cách người dùng tự tạo dữ liệu cho AI. Cũng còn một cách là người dùng trực tiếp cung cấp dữ liệu cho AI.

1) Cách thức xây dựng công nghệ Nhận dạng, giả lập giọng nói:

Vậy người ta áp dụng công nghệ giọng nói vào phần mềm như thế nào? Thông thường một bộ máy giọng nói sẽ có hai phần. Phần thứ nhất gọi là Speech synthesizer (còn gọi là Text to Speech hay TTS). Đây là một trình tổng hợp giọng nói và thiết bị hoặc ứng dụng xài để tương tác với người dùng, ví dụ: đọc văn bản trên màn hình, thông báo về tiến độ chạy một tác vụ nào đó. Phần thứ hai là một công nghệ nhận dạng cho phép app biết được người dùng đang nói gì, từ đó chuyển thể thành lệnh để thiết bị thực thi hoặc chuyển đổi thành các kí tự nhập liệu. Nói cách khác, đây là thứ thay thế cho bàn phím của chúng ta. Một ứng dụng nhận dạng giọng nói lý tưởng sẽ bao gồm cả hai bộ phận nói trên, nhưng một số app chỉ xài một cái rồi từ từ nâng cấp sau.

Thứ nhất, các nhà phát triển phải xây dựng nên một công nghệ có thể lắng nghe, phân tích và phiên dịch một cách chính xác giọng nói của người dùng. Nếu không thì làm sao app biết bạn đang nói gì, còn nếu độ chính xác không cao thì cũng như không.

Thứ hai, vấn đề bản địa hóa (localization) cũng là một chuyện làm đau đầu các lập trình viên. Mỗi quốc gia sẽ có ngôn ngữ của riêng mình, vấn đề đó là làm thế nào để có thể hỗ trợ càng nhiều ngôn ngữ càng tốt.

Có một kĩ thuật được nhắc đến nhiều trong thời gian gần đây, đó là Xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP). Nó là tập hợp của nhiều thuận toán phức tạp nhằm phân tích mệnh lệnh của người dùng nhưng không bắt buộc họ phải nói theo một cấu trúc câu định sẵn. Nhiều năm trước khi muốn điều khiển bằng giọng nói, bạn chỉ có thể nói những thứ như "Mở bản đồ", "Nhắn tin cho vợ", "Báo thức lúc 5 giờ sáng". Còn bây giờ thì nhờ có NLP, chúng ta có thể nói các câu như "Siri, vui lòng nhắn tin cho vợ của tôi là tôi sẽ về trễ nhé", hay như "Hãy đánh thức tôi lúc 5 giờ sáng ngày mai".

2) Mô hình triển khai công nghệ giọng nói:

Có nhiều cách thức mà các công ty hiện nay đang triển khai voice technology, có thể kể đến 2 phương pháp phổ biến như sau:

Điện toán đám mây: Trong trường hợp này, việc nhận dạng, xử lý ngôn ngữ sẽ diễn ra trên máy chủ của các công ty cung cấp dịch vụ. Phương pháp đám mây giúp việc nhận dạng được chính xác hơn, ứng dụng thì có dung lượng nhỏ, nhưng bù lại thì thiết bị ở phía người dùng phải luôn kết nối với Internet. Độ trễ trong quá trình gửi giọng nói từ máy lên server rồi trả kết quả từ server về lại máy cũng là những thứ đáng cân nhắc.

Tích hợp thẳng vào app: Với phương thức này, quá trình xử lý giọng nói sẽ diễn ra trong nội bộ ứng dụng, không cần giao tiếp với bên ngoài, chính vì thế tốc độ sẽ nhanh hơn. Người dùng cũng không bắt buộc phải kết nối vào mạng thường trực. Tuy nhiên, giải pháp này gặp nhược điểm đó là khi có cập nhật hoặc thay đổi gì đó về bộ máy nhận dạng, nhà sản xuất sẽ phải cập nhật lại cả một app, trong khi với phương thức đám mây thì những thay đổi đó chỉ cần làm ở phía server. Kích thước ứng dụng cũng sẽ tăng lên, có thể lên tới cả vài trăm MB.

V. KẾT LUẬN

Sự phát triển của nhận dạng giọng nói là một chủ đề nghiên cứu trong nhiều linh vực như nhận biết ngôn ngữ, nhận biết giọng nói của con người, tương tác với máy tính bằng giọng nói. Bài nghiên cứu này đã nghiên cứu ra sự phát triển của nhận dạng giọng nói qua các giai đoạn và mỗi giai đoạn có sự ra đời của các phương pháp nhận dạng khác nhau. Từ đó, chúng ta dựa vào những nền tảng kiến thức, kinh nghiệm, phương pháp của các thế hệ trước để vận dụng, nghiên cứu và phát triển nó một cách mạnh mẽ. Chúng ta là những con người đang trong cuộc cách mạng công nghiệp 4.0 do đó sự cần thiết của con người trong lĩnh vực mới này rất quan trọng. Sự ra đời của trí tuệ nhân tạo, sự ra đời của máy móc thay cho con người đã và đang chiếm lĩnh vị trí của con người và thay thế bằng sức lao động của máy móc. Chủ đề nghiên cứu này, đã phân tích được sự phát triển của nhận dạng giọng nói và các thuật toán đi cùng qua các thời kỳ khác nhau. Nó ngày càng hoàn thiện hơn, và có thể sẽ được sử dụng rộng rãi trong cuộc sống.

TÀI LIỆU THAM KHẢO

[1] G. Eason, B. Noble, and I. N. Sneddon, "*On certain integrals of Lipschitz-Hankel type involving products of Bessel functions*," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] Ichi.pro, "*Giải thích minh họa cho sự cong vênh thời gian động*" truy cập ngày 30/05/2021 <https://ichi.pro/vi/giai-thich-minh-hoa-cho-su-cong-venh-thoi-gian-dong-60106327800728 >

[3] Jeremy, "*Dynamic Time Warping*" truy cập ngày 30/05/2021 <https://towardsdatascience.com/dynamic-time-warping-3933f25fcdd>

[4] thuvien.vku.udn.vn, "*Kỹ Thuật Nhận Dạng Giọng Nói Sử Dụng Mô Hình Markov Ẩn*" truy cập ngày 30/05/2021 <http://thuvien.vku.udn.vn/bitstream/123456789/292/1/20181209220416.pdf>

[5] vi.wikipedia.org, "*Xích Markov*" truy cập ngày 30/05/2021 <https://vi.wikipedia.org/wiki/X%C3%ADch_Markov >

[6] vi.abadgar-q.com, "*Mạng nơron nhân tạo - Artificial neural network*" truy cập ngày 30/05/2021 <https://vi.abadgar-q.com/wiki/Artificial_neural_network>

[7] vi.abadgar-q.com, "*Chuỗi Markov - Markov chain*" truy cập ngày 30/05/2021 < https://vi.abadgar-q.com/wiki/Markov_chain >

[8] nawapi.gov.vn, "*Giới thiệu tổng quan về Mạng Noron nhân tạo (Artificial Neural Network-ANN)"* truy cập ngày 30/05/2021 <http://nawapi.gov.vn/index.php?option=com_content&view=article&id=3238%3Agii-thiu-tng-quan-v-mng-nron-nhan-to-artificial-neural-network-ann&catid=70%3Anhim-v-chuyen-mon-ang-thc-hin&Itemid=135&lang=vi >

[9] bktt.vn, "*Học Sâu*" truy cập ngày 30/05/2021 <https://bktt.vn/H%E1%BB%8Dc_s%C3%A2u >

[10] duhoctrungquoc.vn, "*Xích Markov*" truy cập ngày 30/05/2021 <https://www.duhoctrungquoc.vn/wiki/vi/X%C3%ADch_Markov>

[11] vi.wikipedia.org, "*Xích Markov*" truy cập ngày 30/05/2021 <https://vi.wikipedia.org/wiki/X%C3%ADch_Markov#C%C3%A1c_lo%E1%BA%A1i_x%C3%ADch_Markov>

[12] Admin UBC, "*Những điều bạn chưa biết về công nghệ nhận diện giọng nói*" truy cập ngày 30/05/2021 <https://ubctv.vn/vi/nhung-dieu-ban-chua-biet-ve-chuc-nang-nhan-dien-giong-noi#:~:text=di%E1%BB%87n%20gi%E1%BB%8Dng%20n%C3%B3i%3A-,%E1%BB%A8ng%20d%E1%BB%A5ng%20gi%E1%BB%8Dng%20n%C3%B3i%20%C4%91%E1%BA%A7u%20ti%C3%AAn%20%C4%91%C6%B0%E1%BB%A3c%20t%E1%BA%A1o%20ra,v%C3%A0%20t%C3%A1i%20t%E1%BA%A1o%20%C3%A2m%20thanh%20 >

[13] nordiccoder.com, "*Deep Neural Network*" truy cập ngày 30/05/2021 <https://nordiccoder.com/blog/deep-neural-network/ >

[14] Thuynt (2019) "*Ứng dụng thực tiễn của Machine Learning – Nhận diện và dự đoán*" truy cập ngày 30/05/2021 <http://hoctructuyen123.net/ung-dung-thuc-tien-cua-machine-learning-nhan-dien-va-du-doan/ >

[15] S J Melnikoff, S F Quigley & M J Russell, 2002, *Implementing a Simple Continuous Speech Recognition System on an FPGA, IEEE*

[16] viblo.asia, "*Machine Learning thật thú vị (6): Nhận diện giọng nói*" truy cập ngày 25/05/2021 < https://viblo.asia/p/machine-learning-that-thu-vi-6-nhan-dien-giong-noi-1Je5E8DylnL >

[17] vncoder.vn, "*Bài 15: Nhận diện giọng nói phần 1 - Lập trình AI bằng Python*", truy cập ngày 25/05/2021, <https://vncoder.vn/bai-hoc/nhan-dien-giong-noi-phan-1-411>

[18] en.wikipedia.org, "*Dynamic Time Warping*" truy cập ngày 25/05/2021 <https://en.wikipedia.org/wiki/Dynamic_time_warping>

[19] laptrinhvagiaithuat.blogspot.com, "*Nhận dạng lời nói trong C#"* truy cập ngày 25/05/2021 <https://laptrinhvagiaithuat.blogspot.com/2016/08/nhan-dang-loi-noi-trong-c.html>

[20] Sharecode.vn, (2017), "*Nhận dạng giọng nói sang text và đọc đoạn text C# có báo cáo pp*" truy cập ngày 25/05/2021,<https://sharecode.vn/source-code/nhan-dang-giong-noi-sang-text-va-doc-doan-text-c-co-bao-cao-pp-10874.htm>

[21] vncoder.vn, "*Bài 15: Nhận diện giọng nói phần 1 - Lập trình AI bằng Python*" truy cập ngày 25/05/2021, <https://vncoder.vn/bai-hoc/nhan-dien-giong-noi-phan-1-411>

[22] viblo.asia, "*Machine Learning thật thú vị (6): Nhận diện giọng nói*" truy cập ngày 25/05/2021 <https://viblo.asia/p/machine-learning-that-thu-vi-6-nhan-dien-giong-noi-1Je5E8DylnL>

[23] amazon.com.br, "*Trí tuệ nhận tạo nơ rơn*" truy cập ngày 25/05/2021 <https://www.amazon.com.br/Tr%C3%AD-tu%E1%BB%87-nh%C3%A2n-t%E1%BA%A1o-n%C6%A1-ron/dp/1233884247>

[24] sohoatailieu.net.vn, "*Phần mềm nhận dạng giọng nói*" truy cập ngày 25/05/2021 <https://sohoatailieu.net.vn/2020/04/phan-mem-nhan-dang-giong-noi.html>

[25] vi.wikipedia.org, "*Giao diện giọng nói người dùng*" truy cập ngày 25/05/2021 <https://vi.wikipedia.org/wiki/Giao_di%E1%BB%87n_gi%E1%BB%8Dng_n%C3%B3i_ng%C6%B0%E1%BB%9Di_d%C3%B9ng#%E1%BB%A8ng_d%E1%BB%A5ng_trong_th%E1%BB%B1c_t%E1%BA%BF_cu%E1%BB%99c_s%E1%BB%91ng,_ti%E1%BB%81m_n%C4%83ng_trong_t%C6%B0%C6%A1ng_lai_v%C3%A0_nh%E1%BB%AFng_k%E1%BB%B3_v%E1%BB%8Dng_v%E1%BB%81_c%C3%B4ng_ngh%E1%BB%87>

[26] en.wikipedia.org, "*Deep learning*" truy cập ngày 25/05/2021 <https://en.wikipedia.org/wiki/Deep_learning#Automatic_speech_recognition>

[27] en.wikipedia.org, "*Speech recognition*" truy cập ngày 25/05/2021 <https://en.wikipedia.org/wiki/Speech_recognition#End-to-end_automatic_speech_recognition>

[28] Lawrence R. Rabiner, 1980, *A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition, Proceedings of the IEEE*, Vol.77, No.2.

[29] Sharecode.vn, (2017), "Nhận dạng giọng nói sang text và đọc đoạn text C# có báo cáo pp" truy cập ngày 25/05/2021,<https://sharecode.vn/source-code/nhan-dang-giong-noi-sang-text-va-doc-doan-text-c-co-bao-cao-pp-10874.htm>

[30] vncoder.vn, "Bài 15: Nhận diện giọng nói phần 1 - Lập trình AI bằng Python" truy cập ngày 25/05/2021 <https://vncoder.vn/bai-hoc/nhan-dien-giong-noi-phan-1-411 >

[30] 123doc.net , "Nghiên cứu công nghệ nhận dạng giọng nói tiếng Việt sử dụng học máy và ứng dụng vào việc điều khiển thiết bị trong nhà bằng điện thoại Android-đã chuyển đổi" truy cập ngày 29/05/2021 <https://123doc.net/document/5788354-nghien-cuu-cong-nghe-nhan-dang-giong-noi-tieng-viet-su-dung-hoc-may-va-ung-dung-vao-viec-dieu-khien-thiet-bi-trong-nha-bang-dien-thoai-android-da-chuy.htm>

[31] Minh Đào, (2018), "*Ba cấp độ trí tuệ nhân tạo và sự cáo chung của kỷ nguyên con người*" truy cập ngày 25/05/2021 <https://khoahocphattrien.vn/khoa-hoc/doi-net-ve-lang-xuat-ban-an-pham-khoa-hoc-quoc-te/20180301084552450p1c160.htm >

[32] Author (2014), "*Công Nghệ Nhận Dạng Và Giả Lập Giọng Nói*" truy cập ngày 29/05/2021 < https://codelearn.io/sharing/cong-nghe-nhan-dang-va-gia-lap-giong-noi >

[33] Mai Hiền (2017), "*Ứng dụng nhận diện hình ảnh và nhận dạng giọng nói với công nghệ Machine learning*" truy cập ngày 29/05/2021 <https://techinsight.com.vn/ung-dung-nhan-dien-hinh-anh-va-nhan-dang-giong-noi-voi-cong-nghe-machine-learning/ >

[34] Amazingcharts.com, "*Taking Your Clinical Documentation to the Next Level With Voice Recognition*" truy cập ngày 29/05/2021 <https://amazingcharts.com/taking-your-clinical-documentation-to-the-next-level-with-speech-recognition/>

[35] Robert Hoyt, MD, FACP, and Ann Yoshihashi, MD, FACE, "*Lessons Learned from Implementation of Voice Recognition for Documentation*" truy cập ngày 29/05/2021 <https://perspectives.ahima.org/lessons-learned-from-implementation-of-voice-recognition-for-documentation/>

[36] Developer.mozilla.org, "*SpeechRecognition*" truy cập ngày 29/05/2021 <https://developer.mozilla.org/en-US/docs/Web/API/SpeechRecognition>

[37] Sohoatailieu.net.vn (2020), "*Phần mềm nhận dạng giọng nói- nguyên tắc hoạt động và ứng dụng*" truy cập ngày 29/05/2021 <https://sohoatailieu.net.vn/2020/04/phan-mem-nhan-dang-giong-noi.html>

[38] Sestek.com (2014), "*Introduction to Speech Recognition*" truy cập ngày 29/05/2021 <https://www.sestek.com/2014/11/introduction-to-speech-recognition/>

[39] RichieACC, Ryan Lundy,Philipp, Rob,Jorge, Robert, stephbu,(2008). "*C # Nhận dạng giọng nói - Đây có phải là những gì người dùng nói?"* truy cập ngày 29/05/2021 <https://www.it-swarm-vi.com/vi/c%23/c-nhan-dang-giong-noi-day-co-phai-la-nhung-gi-nguoi-dung-noi/958433704/>

[40] Code24h.com (2018), "*Lập trình ứng dụng nhận dạng giọng nói (Text To Speed) bằng VB.NET*" truy cập ngày 29/05/2021 <http://code24h.com/lap-trinh-ung-dung-nhan-dang-giong-noi-text-to-speed-bang-vb-net-d24564.htm>

[41] Little Duck, "*Kỹ thuật nhận dạng giọng nói*" truy cập ngày 29/05/2021 <https://123doc.net/document/3006016-ky-thuat-nhan-dang-giong-noi.html>

[42] Nguyễn Trọng Thảo (2011), "*Ứng dụng nhận dạng giọng nói đỉnh cho Smartphone*" truy cập ngày 29/05/2021 <https://tailieu.vn/doc/ung-dung-nhan-dang-giong-noi-dinh-cho-smartphone-811628.html>