



(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) Cộng hòa xã hội chủ nghĩa Việt Nam (VN)
CỤC SỞ HỮU TRÍ TUỆ

(11)



1-0037671

(51)⁷ G10L 15/22; G10L 15/02; G10L 15/17 (13) B

(21) 1-2019-04135

(22) 26/06/2018

(86) PCT/CN2018/092899 26/06/2018

(87) WO2019/001428 03/01/2019

(30) 201710514348.6 29/06/2017 CN

(45) 27/11/2023 428

(43) 25/09/2019 378A

(73) ADVANCED NEW TECHNOLOGIES CO., LTD. (KY)

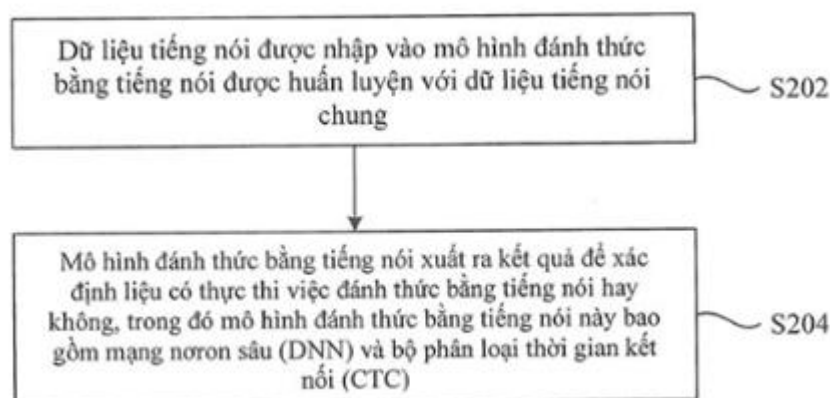
Cayman Corporate Centre, 27 Hospital Road, George Town, Grand Cayman KY1-9008, Cayman Islands

(72) WANG, Zhiming (CN); ZHOU, Jun (CN); LI, Xiaolong (US).

(74) Công ty Luật TNHH T&G (TGVN)

(54) PHƯƠNG PHÁP ĐÁNH THỨC BẰNG TIẾNG NÓI, THIẾT BỊ ĐIỆN TỬ VÀ
PHƯƠNG TIỆN LƯU TRỮ PHI CHUYÊN TIẾP ĐỌC ĐƯỢC BẰNG MÁY TÍNH

(57) Sáng chế đề cập đến phương pháp, thiết bị đánh thức bằng tiếng nói, và thiết bị điện tử. Phương pháp này bao gồm các bước: thi hành việc đánh thức bằng tiếng nói nhờ sử dụng mô hình đánh thức bằng tiếng nói mà nó bao gồm mạng nơron sâu (Deep Neural Network, DNN) và bộ phân loại thời gian kết nối (Connectionist Temporal Classifier, CTC). Mô hình đánh thức bằng tiếng nói này có thể được thu bằng cách huấn luyện với dữ liệu tiếng nói chung.



Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến lĩnh vực các công nghệ phần mềm máy tính, và cụ thể là đến phương pháp, đề cập đến thiết bị, phương pháp đánh thức bằng tiếng nói, và thiết bị điện tử.

Tình trạng kỹ thuật của sáng chế

Cùng với sự phát triển nhanh chóng của các thiết bị di động, các công nghệ liên quan đến tiếng nói cũng đang trở nên ngày càng phổ biến. Ví dụ, việc nhận dạng tiếng nói được sử dụng trong các công cụ trợ lý hội thoại ngày càng phổ biến như Siri của Apple, Cortana của Microsoft, và Alexa của Amazon để nâng cao trải nghiệm của người dùng và mức độ tương tác con người-máy tính.

Một công nghệ tương tác tiếng nói quan trọng là phát hiện từ khóa (Keyword Spotting, KWS), mà nó cũng có thể được gọi chung là đánh thức bằng tiếng nói. Dựa trên giải pháp kỹ thuật đã biết, cần có giải pháp đánh thức bằng tiếng nói không dựa vào dữ liệu tiếng nói theo từ khóa cụ thể.

Bản chất kỹ thuật của sáng chế

Phương pháp, thiết bị đánh thức bằng tiếng nói, và thiết bị điện tử được đề xuất trong các phương án của sáng chế này, để giải quyết vấn đề kỹ thuật sau đây: cần có giải pháp đánh thức bằng tiếng nói mà không dựa vào dữ liệu tiếng nói theo từ khóa cụ thể.

Để giải quyết vấn đề kỹ thuật nêu trên, các phương án của sáng chế này được thi hành như sau.

Phương pháp đánh thức bằng tiếng nói được đề xuất trong một phương án của sáng chế này bao gồm các bước:

nhập dữ liệu tiếng nói vào mô hình đánh thức bằng tiếng nói được huấn luyện với dữ liệu tiếng nói chung, và xuất ra, bởi mô hình đánh thức bằng tiếng nói này, kết quả để xác định liệu có thực thi việc đánh thức bằng tiếng nói hay không, trong đó mô hình

đánh thức bằng tiếng nói này bao gồm mạng nơ-ron sâu (Deep Neural Network, DNN) và bộ phân loại thời gian kết nối (Connectionist Temporal Classifier, CTC).

Thiết bị đánh thức bằng tiếng nói được đề xuất trong một phương án của sáng chế này bao gồm mô-đun đầu vào và mô hình đánh thức bằng tiếng nói;

trong đó dữ liệu tiếng nói được nhập bởi mô-đun đầu vào vào mô hình đánh thức bằng tiếng nói được huấn luyện với dữ liệu tiếng nói chung, và mô hình đánh thức bằng tiếng nói xuất ra kết quả để xác định liệu có thực thi việc đánh thức bằng tiếng nói hay không, trong đó mô hình đánh thức bằng tiếng nói này bao gồm DNN và CTC.

Thiết bị điện tử được đề xuất trong một phương án của sáng chế này bao gồm:

ít nhất một bộ xử lý; và

bộ nhớ được kết nối theo cách truyền thông được với ít nhất một bộ xử lý này; trong đó

bộ nhớ lưu trữ lệnh có thể thực thi được bởi ít nhất một bộ xử lý này, và lệnh này được thực thi bởi ít nhất một bộ xử lý này để cho phép ít nhất một bộ xử lý này:

nhập dữ liệu tiếng nói vào mô hình đánh thức bằng tiếng nói được huấn luyện với dữ liệu tiếng nói chung, và xuất ra, bởi mô hình đánh thức bằng tiếng nói này, kết quả để xác định liệu có thực thi việc đánh thức bằng tiếng nói hay không, trong đó mô hình đánh thức bằng tiếng nói này bao gồm DNN và CTC.

Ít nhất một trong số các giải pháp kỹ thuật nêu trên, được chấp nhận trong các phương án của sáng chế này, có thể đạt được các hiệu quả có lợi sau đây: thay vì dựa trên dữ liệu tiếng nói theo từ khóa cụ thể, mô hình đánh thức bằng tiếng nói này có thể được huấn luyện với dữ liệu tiếng nói chung tiếp cận được, và sau đó mô hình đánh thức bằng tiếng nói được huấn luyện này có thể được sử dụng cho việc đánh thức bằng tiếng nói, điều này có ích cho việc cải thiện độ chính xác của việc đánh thức bằng tiếng nói.

Mô tả vắn tắt các hình vẽ

Để mô tả các giải pháp kỹ thuật trong các phương án của sáng chế này hoặc trong giải pháp kỹ thuật đã biết một cách rõ ràng hơn, các hình vẽ kèm theo cần được sử dụng trong phần mô tả về các phương án hoặc giải pháp kỹ thuật đã biết này sẽ được giới thiệu vắn tắt trong phần sau đây. Rõ ràng, các hình vẽ kèm theo được mô tả trong phần sau đây chỉ là một số phương án được bộc lộ trong bản mô tả này. Những người có trình

độ trung bình trong lĩnh vực kỹ thuật còn có thể thu được các hình vẽ kèm theo khác theo các hình vẽ kèm theo này mà không cần nỗ lực sáng tạo.

Fig.1 là sơ đồ giản lược của kiến trúc tổng thể liên quan đến giải pháp của sáng chế này trong kịch bản ứng dụng thực tế;

Fig.2 là lưu đồ giản lược của phương pháp đánh thức bằng tiếng nói theo một phương án của sáng chế này;

Fig.3 là sơ đồ giản lược của khung của mô hình đánh thức bằng tiếng nói trên Fig.2 theo một phương án của sáng chế này;

Fig.4 là sơ đồ giản lược của việc trích xuất đặc điểm của mô đun trích xuất đặc điểm trên Fig.3 trong kịch bản ứng dụng thực tế theo một phương án của sáng chế này;

Fig.5 là sơ đồ cấu trúc giản lược của mạng nơron sâu (DNN) trên Fig.3 trong kịch bản ứng dụng thực tế theo một phương án của sáng chế này;

Fig.6 là sơ đồ cấu trúc giản lược của bộ phân loại thời gian kết nối (CTC) trên Fig.3 trong kịch bản ứng dụng thực tế theo một phương án của sáng chế này; và

Fig.7 là sơ đồ cấu trúc giản lược của thiết bị đánh thức bằng tiếng nói tương ứng với Fig.2 theo một phương án của sáng chế này.

Mô tả chi tiết sáng chế

Phương pháp, thiết bị đánh thức bằng tiếng nói, và thiết bị điện tử được đề xuất trong các phương án của sáng chế.

Để cho phép người có hiểu biết trung bình trong lĩnh vực tương ứng hiểu rõ hơn về các giải pháp kỹ thuật trong sáng chế này, các giải pháp kỹ thuật trong các phương án của sáng chế sẽ được mô tả rõ ràng và đầy đủ trong phần sau đây với tham chiếu đến các hình vẽ kèm theo trong các phương án của sáng chế. Rõ ràng, các phương án được mô tả chỉ đơn thuần là một số, chứ không phải là tất cả, các phương án của sáng chế. Tất cả các phương án khác thu được bởi những người có trình độ trung bình trong lĩnh vực kỹ thuật dựa trên các phương án của sáng chế mà không cần nỗ lực sáng tạo sẽ đều nằm trong phạm vi bảo hộ của sáng chế.

Để dễ hiểu, nguyên lý của các giải pháp của sáng chế được giải thích như sau. Trong sáng chế này, mô hình đánh thức bằng tiếng nói, bao gồm mạng nơron sâu (DNN) và bộ phân loại thời gian kết nối (CTC), được huấn luyện với dữ liệu tiếng nói chung. Mô hình đánh thức bằng tiếng nói được huấn luyện có thể được sử dụng cho việc đánh

thức bằng tiếng nói và hỗ trợ các từ khóa được định nghĩa bởi người dùng được kích hoạt bởi việc đánh thức bằng tiếng nói. Hơn nữa, mô hình đánh thức bằng tiếng nói có thể được sử dụng trong các thiết bị công suất thấp chẳng hạn như các điện thoại di động và thiết bị gia dụng, bởi vì DNN được chứa trong mô hình đánh thức bằng tiếng nói này có thể không quá phức tạp, và ví dụ, có thể chỉ có ba hoặc bốn lớp với hai hoặc ba trăm nút trong mỗi lớp. Mô hình đánh thức bằng tiếng nói này có thể được gọi là CTC-KWS, và KWS ở đây là việc phát hiện từ khóa được nêu trong phần tình trạng kỹ thuật.

DNN là bộ cảm nhận đa lớp, mà nó có lớp ẩn giữa lớp đầu vào và lớp đầu ra và có thể mô phỏng các mối liên hệ phi tuyến phức hợp. CTC là bộ phân loại được tạo cấu trúc để thực hiện tác vụ gắn thẻ nhãn, và không đòi hỏi việc căn chỉnh cưỡng bức (forced alignment) giữa đầu vào và đầu ra.

Fig.1 là sơ đồ giản lược của kiến trúc tổng thể liên quan đến giải pháp của sáng chế này trong kịch bản ứng dụng thực tế. Hai phần có liên quan chủ yếu đến kiến trúc tổng thể: dữ liệu tiếng nói và mô hình đánh thức bằng tiếng nói. Mô hình đánh thức bằng tiếng nói bao gồm DNN và CTC. Việc đánh thức bằng tiếng nói có thể được thi hành bằng cách nhập dữ liệu tiếng nói vào mô hình đánh thức bằng tiếng nói để xử lý.

Giải pháp của sáng chế được mô tả chi tiết trong phần sau đây dựa trên nguyên lý và kiến trúc tổng thể nêu trên.

Fig.2 là lưu đồ giản lược của phương pháp đánh thức bằng tiếng nói theo một phương án của sáng chế này. Từ góc nhìn của các chương trình, phần thân thực thi của lưu trình này có thể là chương trình trên máy chủ hoặc thiết bị đầu cuối, ví dụ, chương trình huấn luyện mô hình, chương trình nhận dạng tiếng nói, ứng dụng đánh thức bằng tiếng nói, v.v. Từ góc nhìn của các thiết bị, phần thân thực thi khả dĩ của lưu trình này là, nhưng không giới hạn ở, ít nhất một trong số các thiết bị sau đây mà chúng có thể đóng vai trò là các máy chủ hoặc các thiết bị đầu cuối: điện thoại di động, máy tính bảng, thiết bị đeo được thông minh, máy ô tô, máy tính cá nhân, máy tính cỡ vừa, cụm máy tính, v.v.

Lưu trình trên Fig.2 có thể bao gồm các bước sau đây.

Trong S202, dữ liệu tiếng nói được nhập vào mô hình đánh thức bằng tiếng nói được huấn luyện với dữ liệu tiếng nói chung.

Trong phương án này của sáng chế, tiếng nói có thể được giám sát bởi phần thân thực thi hoặc phần thân khác để thu được dữ liệu tiếng nói. Khi tiếng nói được giám sát, người dùng có thể nói ra từ khóa định trước để kích hoạt mô hình đánh thức bằng tiếng nói để thực thi việc đánh thức bằng tiếng nói.

Trong S204, mô hình đánh thức bằng tiếng nói xuất ra kết quả để xác định liệu có thực thi việc đánh thức bằng tiếng nói hay không, trong đó mô hình đánh thức bằng tiếng nói này bao gồm DNN và CTC.

Trong phương án này của sáng chế, so với dữ liệu tiếng nói theo từ khóa cụ thể được nêu trong phần tình trạng kỹ thuật, dữ liệu tiếng nói chung được mô tả trong bước S202 ít bị hạn chế hơn và, do đó, có thể tiếp cận được một cách dễ dàng. Ví dụ, nó có thể là tuyển tập nhận dạng tiếng nói liên tục từ vựng lớn (Large Vocabulary Continuous Speech Recognition, LVCSR) hoặc loại tương tự.

Trong phương án này của sáng chế, DNN được chứa trong mô hình nhận dạng tiếng nói có thể dự đoán sự phân bố xác suất hậu nghiệm của chuỗi âm vị phát âm tương ứng với các đặc điểm tiếng nói đầu vào. DNN có thể được theo sau bởi CTC để cho ra điểm số độ tin cậy tương ứng với chuỗi âm vị phát âm được dự đoán. Kết quả để xác định liệu có thực hiện việc đánh thức bằng tiếng nói hay không có thể được xuất ra dựa trên điểm số độ tin cậy này.

Với phương pháp trên Fig.2, thay vì dựa vào dữ liệu tiếng nói theo từ khóa cụ thể, mô hình đánh thức bằng tiếng nói có thể được huấn luyện với dữ liệu tiếng nói chung tiếp cận được, và hơn nữa, mô hình đánh thức bằng tiếng nói được huấn luyện có thể được sử dụng cho việc đánh thức bằng tiếng nói, điều này là có ích cho việc cải thiện độ chính xác của việc đánh thức bằng tiếng nói.

Mô hình đánh thức bằng tiếng nói này cũng vượt qua các hạn chế của dữ liệu tiếng nói theo từ khóa cụ thể và hỗ trợ các từ khóa được kích hoạt được định nghĩa bởi người dùng. Do đó, điều này sẽ thuận tiện và linh hoạt hơn trong các ứng dụng thực tế và hữu ích cho việc cải thiện trải nghiệm của người dùng.

Dựa trên phương pháp trên Fig.2, một số giải pháp thi hành cụ thể và các giải pháp mở rộng của phương pháp này được đề xuất thêm trong các phương án của sáng chế, các giải pháp này được mô tả trong phần sau đây.

Để dễ hiểu, sơ đồ giản lược của khung của mô hình đánh thức bằng tiếng nói trên Fig.2 được đề xuất trong một phương án của sáng chế, như được thể hiện trên Fig.3.

Khung trên Fig.3 lần lượt bao gồm môđun trích xuất đặc điểm, DNN, và CTC. Trong bước S204, việc xuất ra, bởi mô hình đánh thức bằng tiếng nói, kết quả để xác định liệu có thực hiện việc đánh thức bằng tiếng nói hay không cụ thể có thể bao gồm việc:

trích xuất các đặc điểm âm thanh từ dữ liệu tiếng nói đầu vào;

nhập các đặc điểm âm thanh này vào DNN để xử lý để lần lượt thu được xác suất lớp của các đặc điểm âm thanh tương ứng với mỗi âm vị phát âm;

nhập xác suất lớp này vào CTC để xử lý để thu được điểm số độ tin cậy của thuật ngữ đánh thức bằng tiếng nói tương ứng với chuỗi âm vị phát âm; và

xác định liệu có thực thi việc đánh thức hay không tùy theo điểm số độ tin cậy này, và xuất ra kết quả xác định.

Các phần khác nhau của khung trên Fig.3 được mô tả chi tiết kết hợp thêm với Fig.4, Fig.5, và Fig.6 theo lưu trình đánh thức bằng tiếng nói nêu trên.

Fig.4 là sơ đồ giản lược của việc trích xuất đặc điểm của môđun trích xuất đặc điểm trên Fig.3 trong kịch bản ứng dụng thực tế theo một phương án của sáng chế này.

Trên Fig.4, khi phân đoạn của tiếng nói "Zhi Ma Kai Men" được giám sát lúc đó, chuỗi nhãn mục tiêu tương ứng với nó là chuỗi âm vị phát âm, mà nó có thể được biểu diễn là: "zhi1ma2kai1men2", trong đó các con số đại diện cho các tiếng. Ngoài các âm vị chẳng hạn như các âm vị ban đầu và các âm vị cuối cùng, các âm vị tiếng cũng được tính đến như đơn vị tạo mô hình. Trong ứng dụng thực tế, tất cả các âm vị độc lập với ngữ cảnh hoặc phụ thuộc vào ngữ cảnh đều có thể được tính đến, trong số đó các âm vị phụ thuộc vào ngữ cảnh là nhiều hơn. Tuy nhiên, khi xem xét đến việc làm giảm gánh nặng tính toán sau đó của DNN, ưu tiên là chỉ xem xét các âm vị độc lập với ngữ cảnh, cụ thể là 72 đơn vị âm vị tiếng Trung Quốc độc lập với ngữ cảnh, bao gồm đơn vị khoảng trống.

Các đặc điểm âm thanh có thể được trích xuất bởi môđun trích xuất đặc điểm từ dữ liệu tiếng nói đầu vào, mà nó cụ thể có thể bao gồm việc trích xuất các khung đặc điểm âm thanh của dữ liệu tiếng nói đầu vào từ cửa sổ theo khoảng thời gian được chỉ rõ, trong đó mỗi khung trong số các khung đặc điểm âm thanh này là các năng lượng

giàn lọc lôgarit nhiều chiều; việc xếp chồng lần lượt nhiều khung đặc điểm âm thanh liền kề; việc lấy lần lượt các khung đặc điểm âm thanh được xếp chồng làm các đặc điểm âm thanh được trích xuất từ dữ liệu tiếng nói chung; và thêm vào đó, các khung đặc điểm âm thanh được xếp chồng có thể được sử dụng làm các đầu vào của DNN một cách tương ứng.

Các năng lượng giàn lọc lôgarit chỉ các tín hiệu năng lượng được trích xuất bởi giàn lọc lôgarit, mà chúng có thể được biểu diễn dưới dạng vector trong giải pháp của sáng chế để tạo thuận lợi cho việc xử lý mô hình. Nhiều chiều trong phần nêu trên thể hiện nhiều chiều của vector này.

Ví dụ, độ dài được chỉ rõ của cửa sổ thời gian có thể là 25 mili giây, mỗi cửa sổ thời gian có thể di chuyển trong 10 mili giây, và nhiều chiều này ví dụ có thể là 40 chiều. Trên trục thời gian của dữ liệu tiếng nói, các mili giây từ 0 đến 25 có thể được sử dụng làm cửa sổ, và các năng lượng giàn lọc lôgarit 40 chiều được trích xuất tương ứng từ dữ liệu tiếng nói này để đóng vai trò là khung đặc điểm âm thanh thứ nhất; các mili giây từ 10 đến 35 có thể được sử dụng làm cửa sổ, và các năng lượng giàn lọc lôgarit 40 chiều được trích xuất tương ứng từ dữ liệu tiếng nói này để đóng vai trò là khung đặc điểm âm thanh thứ hai; và nhiều khung đặc điểm âm thanh có thể được trích xuất theo cách giống như vậy.

Hơn nữa, mục đích của việc xếp chồng nhiều khung đặc điểm âm thanh liền kề là để cho phép nhiều thông tin hơn từ ngữ cảnh của khung hiện tại, điều này có ích cho việc cải thiện độ chính xác của các kết quả dự đoán kế tiếp. Tiếp theo ví dụ trên đây, khung hiện tại, mười khung liên tiếp liền kề trước khung hiện tại, và năm khung liên tiếp liền kề sau khung hiện tại ví dụ có thể được xếp chồng để thu được đặc điểm xếp chồng 640 chiều để được nhập vào DNN sau đó. Thêm vào đó, việc chuẩn hóa trung bình và phương sai cepstral có thể được tiến hành trên các chiều của đặc điểm xếp chồng, và sau đó việc nhập ngược vào có thể được tiến hành.

Cần lưu ý rằng cách trích xuất đặc điểm và các thông số được tiếp nhận trong ví dụ nêu trên chỉ mang tính ví dụ, và chúng có thể được điều chỉnh khi cần trong các ứng dụng thực tế.

Fig.5 là sơ đồ cấu trúc giản lược của DNN trên Fig.3 trong kịch bản ứng dụng thực tế theo một phương án của sáng chế.

Trên Fig.5, các nơron khác nhau trong DNN được kết nối hoàn toàn. Các đặc điểm âm thanh được trích xuất bởi mô đun trích xuất đặc điểm trên Fig.3 được nhập vào DNN. DNN có thể mô tả mối quan hệ giữa đặc điểm âm thanh được nhập $x_0 \in R^{n_0}$ và đơn vị tạo mô hình j trong lớp đầu ra theo sự ánh xạ hàm sau đây:

$$z_i = x_{i-1} W_i^T + B_i, 1 \leq i \leq N+1; \text{ (công thức I)}$$

$$x_i = \sigma(z_i), 1 \leq i \leq N; \text{ (công thức II)}$$

$$y_j = \frac{\exp(z_{N+1,j})}{\sum_k \exp(z_{N+1,k})}; \text{ (công thức III)}$$

trong đó $x_{i,i>0} \in R^{n_i}$ là đầu ra của lớp ẩn, $W_i \in R^{n_i \times n_{i-1}}$ và $B_i \in R^{n_i}$ lần lượt là các trọng số và các thông số độ dịch (offset), n_i là số lượng nút trên lớp thứ i , $\theta = \{W_i, B_i\}$, " T " biểu thị chuyển vị của ma trận, N là số lượng lớp ẩn, và σ là hàm kích hoạt phi tuyến, ví dụ, hàm đơn vị tuyến tính được chỉnh lưu (Rectified Linear Unit, ReLU) $\sigma(z) = \max(z, 0)$. Công thức III là hàm mũ được chuẩn hóa (softmax), đại diện cho hậu nghiệm được ước lượng của đơn vị nhãn j .

Trong ứng dụng thực tế, mạng nơron hồi quy (Recurrent Neural Network, RNN) cũng có thể được sử dụng cùng với CTC. Tuy nhiên, trong kịch bản mà dữ liệu huấn luyện bị hạn chế như trong phần tình trạng kỹ thuật, các yêu cầu về tiêu thụ năng lượng và tính toán tối thiểu đối với các thiết bị di động có thể dễ dàng được đáp ứng nhờ sử dụng DNN cùng với CTC. Để làm giảm độ phức tạp tính toán, DNN với khoảng hàng trăm nút trong lớp ẩn sẽ thích hợp hơn.

Fig.6 là sơ đồ cấu trúc giản lược của CTC trên Fig.3 trong trường hợp ứng dụng thực tế theo một phương án của sáng chế.

CTC được thiết kế cụ thể cho các tác vụ nhãn tuần tự. Không giống như tiêu chí entropy chéo (cross-entropy) cho việc căn chỉnh mức khung giữa các đặc điểm đầu vào và các nhãn mục tiêu, CTC nhằm tự động học việc căn chỉnh giữa dữ liệu tiếng nói và các chuỗi nhãn (ví dụ, các âm vị hoặc các ký tự, v.v.), qua đó loại bỏ nhu cầu đối với việc căn chỉnh dữ liệu cưỡng bức, và đầu vào không nhất thiết phải giống như độ dài nhãn.

Trên Fig.6, đơn vị tạo mô hình được chỉ rõ được trích xuất từ L , và CTC được đặt trên lớp softmax của DNN. DNN gồm đơn vị $|L|$ và đơn vị khoảng trống. Việc đưa vào

đơn vị khoảng trống làm giảm bớt gánh nặng của việc dự đoán nhãn, do các ký hiệu không được xuất ra tương ứng trong khi không chắn chắn.

$y_j^t (j \in [0, |L|], t \in [0, T])$ được định nghĩa là xác suất mà các đầu ra DNN j trong bước nhảy thời gian t . Chuỗi đầu vào x^T có độ dài khung T và nhãn mục tiêu l^T được cho trước, và $l_i \in L$. Đường CTC $\pi = (\pi_0, \dots, \pi_{T-1})$ là chuỗi nhãn mức khung, mà nó khác với l ở chỗ đường CTC này cho phép sự xuất hiện của các nhãn không trống lặp lại và các đơn vị khoảng trống.

Đường CTC π có thể được ánh xạ đến chuỗi nhãn tương ứng l bằng cách xóa các nhãn lặp lại và các đơn vị khoảng trống. Ví dụ, $\pi("aa-b-c") = \pi("abb-cc-") = "abc"$. Hàm ánh xạ nhiều-một được định nghĩa là τ , và "-" đại diện cho khoảng trống. Nếu x^T được cho trước, và điều kiện xác suất đầu ra của mỗi bước thời gian được giả định là độc lập, thì xác suất của đường π là:

$$p(\pi|x; \theta) = \prod_{t=0}^{T-1} y_{\pi_t}^t; \text{ (công thức IV)}$$

Sau đó, xác suất của l có thể được tính toán dựa trên τ bằng cách tính tổng các xác suất của tất cả các đường được ánh xạ đến l . Trong các ứng dụng thực tế, việc tính tổng tất cả các đường trên CTC theo xét về việc tính toán là khá phiền hà. Đối với vấn đề này, thuật toán lập trình động xuôi và ngược có thể được chấp nhận. Tất cả các đường CTC khả dĩ được biểu diễn một cách súc tích dưới dạng lưới dựa trên thuật toán này, đúng như được thể hiện trên Fig.6.

Vào lúc huấn luyện, CTC nhằm cho phép $\theta^* = \argmin_{\theta} \sum_{(x,l) \in S} -\log(p(l|x; \theta))$, trong đó S thể hiện dữ liệu huấn luyện được sử dụng. Vào lúc giải mã, khi điểm số độ tin cậy được xuất ra bởi CTC lớn hơn ngưỡng được thiết lập, cơ cấu phát hiện có thể quyết định khẳng định theo đó, và có thể xem là các từ khóa tương ứng đã được phát hiện. Ngưỡng được thiết lập này có thể được tinh chỉnh dựa trên tập dữ liệu xác minh.

Trong phương án này của sáng chế, mô hình có thể được huấn luyện bởi phương pháp giảm gradient, ưu tiên là bởi phương pháp giảm gradient ngẫu nhiên không đồng bộ, để tối ưu hóa theo cách lặp lại các thông số trong mô hình đánh thức bằng tiếng nói cho đến khi việc huấn luyện hội tụ.

Ví dụ, DNN và CTC có thể được huấn luyện trên máy chủ có đơn vị xử lý đồ họa (Graphics Processing Unit, GPU). Các thông số mạng được khởi tạo ngẫu nhiên để được phân bố đều trong khoảng $(-0,02; 0,02)$, tỷ lệ học ban đầu là 0,008, và động lượng là 0,9.

Để dễ hiểu, tỷ lệ học và chức năng của nó được mô tả như sau.

Tỷ lệ học là thông số được sử dụng trong phương pháp giảm gradient. Trong phương pháp giảm gradient, nghiệm có thể được khởi tạo đầu tiên, và trên cơ sở nghiệm này, chiều di chuyển và kích thước bước di chuyển được xác định, sao cho sau khi giải pháp ban đầu này di chuyển theo chiều này và kích thước bước này, đầu ra của hàm mục tiêu có thể được làm giảm xuống. Sau đó, nó được cập nhật thành nghiệm mới, chiều di chuyển tiếp theo và kích thước bước tiếp theo được tìm kiếm liên tục, và sau khi quy trình này được thực hiện lặp lại, hàm mục tiêu liên tục được làm giảm xuống, để cuối cùng tìm ra nghiệm, sao cho hàm mục tiêu là tương đối nhỏ. Trong quy trình tìm nghiệm, nếu kích thước bước quá lớn và việc tìm kiếm không đủ cẩn thận thì nghiệm tốt có thể bị bỏ qua, và nếu kích thước bước quá nhỏ thì quy trình tìm nghiệm sẽ tiến hành quá chậm. Do đó, điều quan trọng là thiết lập kích thước bước một cách phù hợp. Tỷ lệ học được sử dụng để điều chỉnh kích thước bước ban đầu. Trong phương pháp giảm gradient, kích thước bước trong mỗi lần điều chỉnh là bằng tỷ lệ học nhân với gradient.

Trong việc huấn luyện nêu trên, tập dữ liệu xác minh cũng có thể được sử dụng để xác minh chéo mô hình đánh thức bằng tiếng nói để xác định liệu việc huấn luyện có hội tụ hay không.

Để cải thiện hiệu năng và độ mạnh của mô hình đánh thức bằng tiếng nói, thêm nhiều biện pháp được đề xuất trong các giải pháp của sáng chế.

Một biện pháp là huấn luyện thích ứng. Cụ thể, mô hình chung có thể được tinh chỉnh với dữ liệu tiếng nói của một số từ khóa cụ thể và ở tỷ lệ học tương đối thấp. Dựa trên sự xem xét này, khi mô hình đánh thức bằng tiếng nói được huấn luyện, dữ liệu tiếng nói theo từ khóa cụ thể cũng có thể được thu nhận, và mô hình đánh thức bằng tiếng nói có thể được huấn luyện với dữ liệu tiếng nói theo từ khóa cụ thể này. Tỷ lệ học được sử dụng trong việc huấn luyện nhỏ hơn tỷ lệ học được sử dụng trong việc huấn luyện của mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói chung.

Một biện pháp khác là học chuyển giao (transfer learning). Cụ thể, các thông số mạng có thể không được khởi tạo một cách ngẫu nhiên, mà tham chiếu đến một mạng

tương ứng sẵn có mà nó có cấu trúc liên kết giống như mạng mục tiêu ngoại trừ các đơn vị được tạo hạt mịn (fine-grained) trong lớp đầu ra, và có thể sử dụng tiêu chí entropy chéo. Học chuyển giao có thể được xem xét đến đặc biệt là khi dữ liệu huấn luyện có quy mô lớn.

Các biện pháp khác bao gồm, ví dụ, việc sử dụng các lệnh vector liên quan đến kiến trúc (ví dụ, NEON của ARM) để tăng tốc hơn nữa phép tính nhân, v.v.

Như được nêu trước đây, giải pháp của sáng chế có thể dễ dàng hỗ trợ các từ khóa đánh thức bằng tiếng nói được định nghĩa bởi người dùng. Chuỗi nhãn mục tiêu tương ứng với các từ khóa được định nghĩa bởi người dùng như vậy có thể được xác định qua từ điển.

Phương pháp đánh thức bằng tiếng nói được đề xuất trong phương án này của sáng chế là như được mô tả trong phần nêu trên. Dựa trên cùng nguyên lý mô tả, thiết bị tương ứng được đề xuất thêm trong một phương án của sáng chế, như được thể hiện trên Fig.7.

Fig.7 là sơ đồ cấu trúc giản lược của thiết bị đánh thức bằng tiếng nói tương ứng với Fig.2 theo một phương án của sáng chế này. Các hộp đứt nét thể hiện các mô đun tùy chọn. Thiết bị này có thể được đặt trên phần thân thực thi của lưu trình trên Fig.2, bao gồm mô đun đầu vào 701 và mô hình đánh thức bằng tiếng nói 702.

Dữ liệu tiếng nói được nhập bởi mô đun đầu vào 701 vào mô hình đánh thức bằng tiếng nói 702 được huấn luyện với dữ liệu tiếng nói chung, và mô hình đánh thức bằng tiếng nói 702 xuất ra kết quả để xác định liệu có thực thi việc đánh thức bằng tiếng nói hay không, trong đó mô hình đánh thức bằng tiếng nói bao gồm DNN và CTC.

Tùy chọn, dữ liệu tiếng nói chung bao gồm tuyển tập LVCSR.

Tùy chọn, thiết bị này còn bao gồm mô đun huấn luyện 703; và việc huấn luyện, bởi mô đun huấn luyện 703, mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói chung bao gồm việc:

tối ưu hóa theo cách lặp lại, bởi mô đun huấn luyện 703, các thông số trong mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói chung bởi phương pháp giảm gradien ngẫu nhiên không đồng bộ cho đến khi việc huấn luyện hội tụ.

Tùy chọn, mô đun huấn luyện 703 còn thu nhận dữ liệu tiếng nói theo từ khóa cụ thể; và

huấn luyện mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói theo từ khóa cụ thể này, trong đó tỷ lệ học được sử dụng trong việc huấn luyện này nhỏ hơn tỷ lệ học được sử dụng trong việc huấn luyện của mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói chung.

Tùy chọn, mô đun huấn luyện 703 xác minh chéo mô hình đánh thức bằng tiếng nói với tập dữ liệu xác minh trong việc huấn luyện để xác định liệu việc huấn luyện có hội tụ hay không.

Tùy chọn, việc xuất ra, bởi mô hình đánh thức bằng tiếng nói 702, kết quả để xác định liệu có thực thi việc đánh thức bằng tiếng nói hay không cụ thể bao gồm việc:

trích xuất, bởi mô hình đánh thức bằng tiếng nói 702, các đặc điểm âm thanh từ dữ liệu tiếng nói đầu vào;

nhập các đặc điểm âm thanh này vào DNN được chứa trong mô hình đánh thức bằng tiếng nói 702 để xử lý để lần lượt thu xác suất lớp của các đặc điểm âm thanh tương ứng với mỗi âm vị phát âm;

nhập xác suất lớp này vào CTC được chứa trong mô hình đánh thức bằng tiếng nói 702 để xử lý để thu điểm số độ tin cậy của thuật ngữ đánh thức bằng tiếng nói tương ứng với chuỗi âm vị phát âm; và

xác định liệu có thực thi việc đánh thức hay không tùy theo điểm số độ tin cậy này, và xuất ra kết quả xác định.

Tùy chọn, việc trích xuất, bởi mô hình đánh thức bằng tiếng nói 702, các đặc điểm âm thanh từ dữ liệu tiếng nói đầu vào cụ thể bao gồm việc:

trích xuất, bởi mô hình đánh thức bằng tiếng nói 702, các khung đặc điểm âm thanh của dữ liệu tiếng nói đầu vào từ cửa sổ theo khoảng thời gian được chỉ rõ, trong đó mỗi khung trong số các khung đặc điểm âm thanh này là các năng lượng giàn lọc lôgarit nhiều chiều;

xếp chồng lần lượt nhiều khung đặc điểm âm thanh liên kề; và

lấy lần lượt các khung đặc điểm âm thanh được xếp chồng làm các đặc điểm âm thanh được trích xuất từ tiếng nói được giám sát.

Dựa trên cùng nguyên lý mô tả, thiết bị điện tử tương ứng còn được đề xuất trong một phương án của sáng chế, thiết bị này bao gồm:

ít nhất một bộ xử lý; và

bộ nhớ được kết nối theo cách truyền thông được với ít nhất một bộ xử lý này; trong đó

bộ nhớ lưu trữ lệnh có thể thực thi được bởi ít nhất một bộ xử lý này, và lệnh này được thực thi bởi ít nhất một bộ xử lý này để cho phép ít nhất một bộ xử lý này:

nhập dữ liệu tiếng nói vào mô hình đánh thức bằng tiếng nói được huấn luyện với dữ liệu tiếng nói chung, và xuất ra, bởi mô hình đánh thức bằng tiếng nói này, kết quả để xác định liệu có thực thi việc đánh thức bằng tiếng nói hay không, trong đó mô hình đánh thức bằng tiếng nói này bao gồm DNN và CTC.

Dựa trên cùng nguyên lý mô tả, phương tiện lưu trữ máy tính phi khả biến tương ứng với lệnh có thể thực thi được bằng máy tính được lưu trữ trên đó còn được đề xuất trong một phương án của sáng chế, trong đó lệnh có thể thực thi được bằng máy tính này được tạo cấu hình để:

nhập dữ liệu tiếng nói vào mô hình đánh thức bằng tiếng nói được huấn luyện với dữ liệu tiếng nói chung, và xuất ra, bởi mô hình đánh thức bằng tiếng nói này, kết quả để xác định liệu có thực thi việc đánh thức bằng tiếng nói hay không, trong đó mô hình đánh thức bằng tiếng nói này bao gồm DNN và CTC.

Các phương án cụ thể của sáng chế đã được mô tả trong phần nêu trên. Các phương án khác nằm trong phạm vi của yêu cầu bảo hộ kèm theo. Trong một số hoàn cảnh, các hành động hoặc các bước được mô tả trong các điểm yêu cầu bảo hộ có thể được thực hiện theo trình tự khác với thứ tự trong các phương án này và vẫn có thể thu được kết quả mong muốn. Thêm vào đó, các quy trình được thể hiện trên các hình vẽ kèm theo không nhất thiết phải thu được kết quả mong muốn theo trình tự cụ thể hoặc trình tự liên tiếp được thể hiện. Việc xử lý đa nhiệm và việc xử lý song song cũng có thể là khả dĩ hoặc có thể là ưu điểm theo một số cách thi hành.

Các phương án trong bản mô tả này được mô tả tăng dần, các phần giống hoặc tương tự của các phương án có thể được thu khi tham chiếu lẫn nhau, và mỗi phương án nhấn mạnh một phần khác với các phương án khác. Đặc biệt, các phương án về thiết bị, thiết bị điện tử, và phương tiện lưu trữ máy tính phi khả biến về cơ bản là tương tự với các phương án về phương pháp, vì vậy được mô tả một cách đơn giản. Đối với các phần liên quan, hãy tham khảo các phần mô tả của các phần trong các phương án về phương pháp.

Thiết bị, thiết bị điện tử, và phương tiện lưu trữ máy tính phi khả biến được đề xuất trong các phương án của sáng chế là tương ứng với phương pháp. Do đó, thiết bị, thiết bị điện tử, và phương tiện lưu trữ máy tính phi khả biến này cũng có các hiệu quả kỹ thuật hữu ích tương tự như các hiệu quả kỹ thuật của phương pháp tương ứng. Do các hiệu quả kỹ thuật hữu ích của phương pháp này đã được mô tả chi tiết trên đây, các hiệu quả kỹ thuật hữu ích của thiết bị, thiết bị điện tử, và phương tiện lưu trữ máy tính phi khả biến này sẽ không được nêu chi tiết ở đây.

Trong những năm 1990, sự cải tiến về công nghệ có thể được phân biệt rõ ràng là sự cải tiến về phần cứng (ví dụ, sự cải tiến về cấu trúc mạch chẳng hạn như điôt, tranzito, và bộ chuyển mạch) hay sự cải tiến về phần mềm (sự cải tiến về thủ tục của phương pháp). Tuy nhiên, cùng với sự phát triển của các công nghệ, các sự cải tiến về nhiều thủ tục của phương pháp hiện nay có thể được coi là các cải tiến trực tiếp về các cấu trúc mạch phần cứng. Hầu hết tất cả các nhà thiết kế lập trình các thủ tục phương pháp được cải thiện vào trong các mạch phần cứng để thu được các cấu trúc mạch phần cứng tương ứng. Do đó, sẽ là không phù hợp nếu giả định rằng sự cải tiến về thủ tục của phương pháp không thể được thi hành nhờ sử dụng môđun thực thể phần cứng. Ví dụ, thiết bị logic lập trình được (Programmable Logic Device, PLD) (ví dụ, mảng cổng lập trình được dạng trường (Field Programmable Gate Array, FPGA)) là mạch tích hợp như vậy, và các chức năng logic của nó được xác định bởi thiết bị lập trình bởi người dùng. Các nhà thiết kế tự lập trình để “tích hợp” hệ thống kỹ thuật số vào PLD, mà không cần nhờ các nhà sản xuất chip thiết kế và sản xuất chip mạch tích hợp chuyên dụng. Thêm vào đó, hiện nay, việc lập trình hầu hết được thi hành nhờ sử dụng phần mềm trình biên dịch logic thay vì việc sản xuất thủ công chip mạch tích hợp. Phần mềm trình biên dịch logic là tương tự như trình biên dịch phần mềm được sử dụng để phát triển và viết chương trình, và các mã gốc trước khi biên dịch cũng cần được viết nhờ sử dụng ngôn ngữ lập trình cụ thể, được gọi là ngôn ngữ mô tả phần cứng (Hardware Description Language, HDL). Có nhiều loại HDL, chẳng hạn như ngôn ngữ biểu thức luận lý cải tiến (Advanced Boolean Expression Language, ABEL), ngôn ngữ mô tả phần cứng của Altera (Altera Hardware Description Language, AHDL), Hợp lưu (Confluence), ngôn ngữ lập trình của đại học Cornell (Cornell University Programming Language, CUPL), HDCal, ngôn ngữ mô tả phần cứng Java (Java Hardware Description Language, JHDL), Lava, Lola,

MyHDL, PALASM, và ngôn ngữ mô tả phần cứng Ruby (Ruby Hardware Description Language, RHDL), trong số đó, ngôn ngữ mô tả phần cứng mạch tích hợp tốc độ rất cao (Very-High-Speed Integrated Circuit Hardware Description Language, VHDL) và Verilog hiện đang được sử dụng phổ biến nhất. Người có hiểu biết trung bình trong lĩnh vực tương ứng cũng sẽ biết rằng mạch phần cứng để thi hành thủ tục của phương pháp logic có thể thu được dễ dàng bằng cách lập trình logic ở một mức độ không đáng kể thủ tục của phương pháp nhờ sử dụng một số ngôn ngữ mô tả phần cứng trên đây và lập trình nó vào trong mạch tích hợp.

Bộ điều khiển có thể được thi hành theo cách thích hợp bất kỳ. Ví dụ, bộ điều khiển có thể sử dụng hình thức bộ vi xử lý hoặc bộ xử lý và phương tiện đọc được bằng máy tính mà nó lưu trữ các mã chương trình đọc được bằng máy tính (chẳng hạn như phần mềm hoặc phần sụn) thực thi được bởi bộ vi xử lý hoặc bộ xử lý, cổng logic, bộ chuyển mạch, mạch tích hợp chuyên dụng (Application Specific Integrated Circuit, ASIC), bộ điều khiển logic lập trình được, và bộ vi điều khiển được nhúng. Các ví dụ về bộ điều khiển này bao gồm, nhưng không giới hạn ở, các bộ vi điều khiển sau đây: ARC 625D, Atmel AT91SAM, Microchip PIC18F26K20, và Silicone Labs C8051F320. Bộ điều khiển của bộ nhớ còn có thể được thi hành dưới dạng một phần của logic điều khiển của bộ nhớ. Người có hiểu biết trung bình trong lĩnh vực tương ứng cũng biết rằng ngoài việc thi hành bộ điều khiển nhờ sử dụng các mã chương trình đọc được bằng máy tính thuần túy, việc lập trình logic các bước của phương pháp để cho phép bộ điều khiển thi hành cùng một chức năng dưới dạng cổng logic, bộ chuyển mạch, ASIC, bộ điều khiển logic lập trình được, và bộ vi điều khiển được nhúng là hoàn toàn khả thi. Do đó, bộ điều khiển như vậy có thể được xem là thành phần phần cứng, và các thiết bị được chứa trong bộ điều khiển này và được tạo cấu hình để thi hành các chức năng khác nhau cũng có thể được xem là các cấu trúc bên trong thành phần phần cứng này. Hoặc, các thiết bị được tạo cấu hình để thi hành các chức năng khác nhau này thậm chí có thể được xem như là cả các môđun phần mềm được tạo cấu hình để thi hành phương pháp lẫn các cấu trúc bên trong thành phần phần cứng.

Hệ thống, thiết bị, các môđun, hoặc các đơn vị được minh họa cụ thể trong các phương án nêu trên có thể được thi hành bởi chip máy tính hoặc một thực thể, hoặc được thi hành bởi sản phẩm có chức năng cụ thể. Thiết bị thi hành điển hình là máy tính. Cụ

thể, ví dụ, máy tính này có thể là máy tính cá nhân, máy tính xách tay, điện thoại tế bào, điện thoại camera, điện thoại thông minh, thiết bị trợ giúp số cá nhân, thiết bị trình diễn phương tiện, thiết bị định vị, thiết bị thư điện tử, bảng điều khiển trò chơi, máy tính bảng, thiết bị đeo được, hoặc sự kết hợp của các thiết bị bất kỳ trong số các thiết bị này.

Để dễ mô tả, thiết bị được chia thành các đơn vị khác nhau dựa trên các chức năng, và các đơn vị này được mô tả riêng biệt. Theo một cách thi hành của sáng chế, các chức năng của các đơn vị khác nhau cũng có thể được thi hành trong một hoặc nhiều mảnh phần mềm và/hoặc phần cứng.

Người có hiểu biết trung bình trong lĩnh vực tương ứng sẽ hiểu rằng các phương án của sáng chế có thể được đề xuất dưới dạng phương pháp, hệ thống, hoặc sản phẩm chương trình máy tính. Do đó, các phương án của sáng chế có thể được thi hành dưới dạng phương án hoàn toàn là phần cứng, phương án hoàn toàn là phần mềm, hoặc phương án kết hợp phần mềm và phần cứng. Hơn nữa, các phương án của sáng chế có thể ở dạng sản phẩm chương trình máy tính được thi hành trên một hoặc nhiều phương tiện lưu trữ sử dụng được bằng máy tính (bao gồm, nhưng không bị giới hạn ở, bộ nhớ đĩa từ, CD-ROM, bộ nhớ quang và dạng tương tự) bao gồm các mã chương trình sử dụng được bằng máy tính.

Sáng chế được mô tả với việc tham chiếu đến các lưu đồ và/hoặc các sơ đồ khối của phương pháp, thiết bị (hệ thống) và sản phẩm chương trình máy tính theo các phương án của sáng chế. Cần hiểu rằng các lệnh chương trình máy tính có thể được sử dụng để thi hành mỗi quy trình và/hoặc khối trong các lưu đồ và/hoặc các sơ đồ khối và các sự kết hợp của các quy trình và/hoặc các khối trong các lưu đồ và/hoặc các sơ đồ khối này. Các lệnh chương trình máy tính có thể được cung cấp cho máy tính đa năng, máy tính chuyên dụng, bộ xử lý được nhúng hoặc bộ xử lý của một thiết bị xử lý dữ liệu lập trình được khác để tạo ra máy, sao cho máy tính hoặc bộ xử lý của một thiết bị xử lý dữ liệu lập trình được khác này thực thi lệnh để tạo ra thiết bị được tạo cấu hình để thi hành các chức năng được chỉ định trong một hoặc nhiều quy trình trong lưu đồ và/hoặc một hoặc nhiều khối trong sơ đồ khối.

Các lệnh chương trình máy tính cũng có thể được lưu trữ trong bộ nhớ đọc được bằng máy tính mà có thể hướng dẫn cho máy tính hoặc một thiết bị xử lý dữ liệu lập trình được khác này làm việc theo cách cụ thể, sao cho lệnh được lưu trữ trong bộ nhớ

đọc được bằng máy tính tạo ra sản phẩm bao gồm thiết bị ra lệnh, và thiết bị ra lệnh này thi hành các chức năng được chỉ định bởi một hoặc nhiều quy trình trong lưu đồ và/hoặc một hoặc nhiều khối trong sơ đồ khối.

Các lệnh chương trình máy tính này cũng có thể được nạp vào máy tính hoặc một thiết bị xử lý dữ liệu lập trình được khác này, sao cho một loạt bước hoạt động được thực thi trên máy tính hoặc một thiết bị lập trình được khác này để tạo ra quy trình được thi hành bằng máy tính, và do đó, lệnh được thực thi trong máy tính hoặc một thiết bị lập trình được khác này cung cấp các bước để thi hành các chức năng được chỉ định trong một hoặc nhiều quy trình trong lưu đồ và/hoặc một hoặc nhiều khối trong sơ đồ khối.

Theo cấu hình điển hình, thiết bị tính toán bao gồm một hoặc nhiều đơn vị xử lý trung tâm (CPU), giao diện đầu vào/đầu ra, giao diện mạng, và bộ nhớ:

Bộ nhớ này có thể bao gồm phương tiện đọc được bằng máy tính chẳng hạn như bộ nhớ khả biến, bộ nhớ truy cập ngẫu nhiên (RAM), và/hoặc bộ nhớ phi khả biến, ví dụ, bộ nhớ chỉ đọc (ROM) hoặc RAM tác động nhanh (flash RAM). Bộ nhớ này là một ví dụ về phương tiện đọc được bằng máy tính.

Phương tiện đọc được bằng máy tính này bao gồm phương tiện phi khả biến và khả biến cũng như phương tiện di chuyển được và không di chuyển được, và có thể thi hành việc lưu trữ thông tin thông qua phương pháp hoặc công nghệ bất kỳ. Thông tin này có thể là lệnh đọc được bằng máy tính, cấu trúc dữ liệu, và môđun của chương trình hoặc dữ liệu khác. Một ví dụ về phương tiện lưu trữ của máy tính bao gồm, nhưng không bị giới hạn ở, bộ nhớ thay đổi pha (phase change memory, PRAM), bộ nhớ truy cập ngẫu nhiên tĩnh (static random access memory, SRAM), bộ nhớ truy cập ngẫu nhiên động (dynamic random access memory, DRAM), các loại RAM khác, ROM, bộ nhớ chỉ đọc lập trình được xóa được bằng điện (electrically erasable programmable read-only memory, EEPROM), bộ nhớ tác động nhanh (flash memory) hoặc các công nghệ bộ nhớ khác, bộ nhớ chỉ đọc dạng đĩa compac (compact disk read-only memory, CD-ROM), đĩa đa năng số (digital versatile disc, DVD) hoặc các bộ lưu trữ quang học khác, băng cát xét, bộ lưu trữ băng từ/đĩa từ hoặc các thiết bị lưu trữ từ tính khác, hoặc phương tiện không truyền dẫn khác bất kỳ, và có thể được sử dụng để lưu trữ thông tin truy cập được vào thiết bị tính toán. Theo định nghĩa trong bản mô tả này, phương tiện đọc được bằng

máy tính không bao gồm phương tiện chuyển tiếp, chẳng hạn như tín hiệu dữ liệu được điều biến và sóng mang.

Cần lưu ý thêm rằng các thuật ngữ "bao gồm", "gồm" hoặc các biến thể khác bất kỳ của chúng được dự định để bao trùm sự bao gồm không loại trừ, sao cho quy trình, phương pháp, vật phẩm hoặc thiết bị bao gồm một chuỗi phần tử không chỉ bao gồm các phần tử đó, mà còn bao gồm các phần tử khác không được liệt kê rõ ràng, hoặc còn bao gồm các phần tử cố hữu đối với quy trình, phương pháp, vật phẩm hoặc thiết bị này. Khi không có thêm các giới hạn, một phần tử được định ra bởi "bao gồm một ..." không loại trừ việc quy trình, phương pháp, vật phẩm hoặc thiết bị bao gồm phần tử đó còn có các phần tử giống hệt khác.

Sáng chế có thể được mô tả trong ngữ cảnh thông thường của lệnh có thể thực thi được bằng máy tính được thực thi bằng máy tính, ví dụ, môđun chương trình. Nhìn chung, môđun chương trình này bao gồm một đoạn chương trình, một chương trình, một đối tượng, một tổ hợp, một cấu trúc dữ liệu, và loại tương tự được sử dụng để thực thi một tác vụ cụ thể hoặc thi hành một loại dữ liệu trừu tượng cụ thể. Sáng chế cũng có thể được thi hành trong các môi trường tính toán phân tán. Trong các môi trường tính toán phân tán này, tác vụ được thực thi nhờ sử dụng các thiết bị xử lý từ xa được kết nối qua mạng truyền thông. Trong các môi trường tính toán phân tán này, môđun chương trình có thể được đặt trong phương tiện lưu trữ máy tính cục bộ và từ xa bao gồm thiết bị lưu trữ.

Các phương án trong bản mô tả này được mô tả tăng dần, các phần giống hoặc tương tự của các phương án có thể được thu khi tham chiếu lẫn nhau, và mỗi phương án nhấn mạnh một phần khác với các phương án khác. Đặc biệt, phương án về hệ thống về cơ bản là tương tự như phương án về phương pháp, vì vậy được mô tả một cách đơn giản. Đối với các phần liên quan, hãy tham khảo các phần mô tả của các phần trong phương án về phương pháp.

Phần mô tả nêu trên chỉ đơn thuần là các phương án của sáng chế, và không được sử dụng để giới hạn sáng chế. Đối với người có hiểu biết trung bình trong lĩnh vực tương ứng, sáng chế có thể có các thay đổi và cải biến khác nhau. Sự cải biến, thay thế tương đương, cải tiến và dạng tương tự bất kỳ được thực hiện mà không nằm ngoài mục đích và nguyên lý của sáng chế sẽ nằm trong phạm vi bảo hộ của sáng chế.

YÊU CẦU BẢO HỘ

1. Phương pháp đánh thức bằng tiếng nói, phương pháp này bao gồm các bước:

huấn luyện mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói chung, trong đó bước huấn luyện này bao gồm bước: tối ưu hóa theo cách lặp lại các thông số trong mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói chung này thông qua phương pháp giảm gradient ngẫu nhiên không đồng bộ cho đến khi việc huấn luyện hội tụ;

thu nhận dữ liệu tiếng nói theo từ khóa cụ thể, và huấn luyện mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói theo từ khóa cụ thể này, trong đó tỷ lệ học được sử dụng trong việc huấn luyện với dữ liệu tiếng nói theo từ khóa cụ thể này là nhỏ hơn tỷ lệ học được sử dụng trong việc huấn luyện của mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói chung;

nhập dữ liệu tiếng nói vào mô hình đánh thức bằng tiếng nói được huấn luyện với dữ liệu tiếng nói chung và dữ liệu tiếng nói theo từ khóa cụ thể này; và

xuất ra, bởi mô hình đánh thức bằng tiếng nói này, kết quả để xác định liệu có thực thi việc đánh thức bằng tiếng nói hay không, trong đó mô hình đánh thức bằng tiếng nói này bao gồm mạng nơ-ron sâu (Deep Neural Network, DNN) và bộ phân loại thời gian kết nối (Connectionist Temporal Classifier, CTC).

2. Phương pháp theo điểm 1, trong đó dữ liệu tiếng nói chung bao gồm tuyển tập nhận dạng tiếng nói liên tục từ vựng lớn (Large Vocabulary Continuous Speech Recognition, LVCSR).

3. Phương pháp theo điểm 1, còn bao gồm bước:

xác minh chéo mô hình đánh thức bằng tiếng nói với tập dữ liệu xác minh trong việc huấn luyện để xác định liệu việc huấn luyện có hội tụ hay không.

4. Phương pháp theo điểm 1, trong đó bước xuất ra, bởi mô hình đánh thức bằng tiếng nói, kết quả để xác định liệu có thực thi việc đánh thức bằng tiếng nói hay không bao gồm các bước:

trích xuất các đặc điểm âm thanh từ dữ liệu tiếng nói đầu vào;

nhập các đặc điểm âm thanh này vào DNN được chứa trong mô hình đánh thức bằng tiếng nói để xử lý để lần lượt thu xác suất lớp của các đặc điểm âm thanh tương ứng với mỗi âm vị phát âm;

nhập xác suất lớp này vào CTC được chứa trong mô hình đánh thức bằng tiếng nói để xử lý để thu điểm số độ tin cậy của thuật ngữ đánh thức bằng tiếng nói tương ứng với chuỗi âm vị phát âm; và

xác định liệu có thực thi việc đánh thức bằng tiếng nói hay không tùy theo điểm số độ tin cậy, và xuất ra kết quả xác định.

5. Phương pháp theo điểm 4, trong đó bước trích xuất các đặc điểm âm thanh từ dữ liệu tiếng nói được nhập bao gồm các bước:

trích xuất các khung đặc điểm âm thanh của dữ liệu tiếng nói đầu vào từ cửa sổ theo khoảng thời gian được chỉ rõ, trong đó mỗi khung trong số các khung đặc điểm âm thanh này là các năng lượng giàn lọc lôgarit nhiều chiều;

xếp chồng lần lượt nhiều khung đặc điểm âm thanh liên kề; và

lấy lần lượt các khung đặc điểm âm thanh được xếp chồng làm các đặc điểm âm thanh được trích xuất từ dữ liệu tiếng nói đầu vào.

6. Thiết bị điện tử, thiết bị này bao gồm:

ít nhất một bộ xử lý; và

bộ nhớ được kết nối theo cách truyền thông được với ít nhất một bộ xử lý này;

trong đó bộ nhớ lưu trữ lệnh có thể thực thi được bởi ít nhất một bộ xử lý này, và lệnh này được thực thi bởi ít nhất một bộ xử lý này để khiến cho ít thiết bị điện tử này:

huấn luyện mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói chung, trong đó việc huấn luyện này bao gồm việc: tối ưu hóa theo cách lặp lại các thông số trong mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói chung này thông qua phương pháp giảm gradient ngẫu nhiên không đồng bộ cho đến khi việc huấn luyện hội tụ;

thu nhận dữ liệu tiếng nói theo từ khóa cụ thể, và huấn luyện mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói theo từ khóa cụ thể này, trong đó tỷ lệ học được sử dụng trong việc huấn luyện với dữ liệu tiếng nói theo từ khóa cụ thể này là nhỏ hơn tỷ lệ học được sử dụng trong việc huấn luyện của mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói chung;

nhập dữ liệu tiếng nói vào mô hình đánh thức bằng tiếng nói được huấn luyện với dữ liệu tiếng nói chung và dữ liệu tiếng nói theo từ khóa cụ thể này; và

xuất ra, từ mô hình đánh thức bằng tiếng nói này, kết quả để xác định liệu có thực thi việc đánh thức bằng tiếng nói hay không, trong đó mô hình đánh thức bằng tiếng nói này bao gồm mạng nơron sâu (DNN) và bộ phân loại thời gian kết nối (CTC).

7. Thiết bị điện tử theo điểm 6, trong đó dữ liệu tiếng nói chung bao gồm tuyển tập nhận dạng tiếng nói liên tục từ vòm lớn (LVCSR).

8. Thiết bị điện tử theo điểm 6, trong đó lệnh đã nêu còn được thực thi bởi ít nhất một bộ xử lý đã nêu để khiến cho thiết bị điện tử này: xác minh chéo mô hình đánh thức bằng tiếng nói với tập dữ liệu xác minh trong việc huấn luyện để xác định liệu việc huấn luyện có hội tụ hay không.

9. Thiết bị điện tử theo điểm 6, trong đó trong việc xuất ra từ mô hình đánh thức bằng tiếng nói kết quả để xác định liệu có thực thi việc đánh thức bằng tiếng nói hay không, lệnh đã nêu còn được thực thi bởi ít nhất một bộ xử lý đã nêu để khiến cho thiết bị điện tử này:

trích xuất, bởi mô hình đánh thức bằng tiếng nói, các đặc điểm âm thanh từ dữ liệu tiếng nói đầu vào;

nhập các đặc điểm âm thanh vào DNN được chứa trong mô hình đánh thức bằng tiếng nói để xử lý để lần lượt thu xác suất lớp của các đặc điểm âm thanh tương ứng với mỗi âm vị phát âm;

nhập xác suất lớp này vào CTC được chứa trong mô hình đánh thức bằng tiếng nói để xử lý để thu điểm số độ tin cậy của thuật ngữ đánh thức bằng tiếng nói tương ứng với chuỗi âm vị phát âm; và

xác định liệu có thực thi việc đánh thức hay không tùy theo điểm số độ tin cậy này, và xuất ra kết quả xác định.

10. Thiết bị điện tử theo điểm 9, trong đó trong việc trích xuất, bởi mô hình đánh thức bằng tiếng nói, các đặc điểm âm thanh từ dữ liệu tiếng nói đầu vào, lệnh đã nêu còn được thực thi bởi ít nhất một bộ xử lý đã nêu để khiến cho thiết bị điện tử này:

trích xuất, bởi mô hình đánh thức bằng tiếng nói, các khung đặc điểm âm thanh của dữ liệu tiếng nói đầu vào từ cửa sổ theo khoảng thời gian được chỉ rõ, trong đó mỗi khung trong số các khung đặc điểm âm thanh này là các năng lượng giàn lọc lôgarit nhiều chiều;

xếp chồng lần lượt nhiều khung đặc điểm âm thanh liền kề; và

lấy lần lượt các khung đặc điểm âm thanh được xếp chồng làm các đặc điểm âm thanh được trích xuất từ dữ liệu tiếng nói đầu vào.

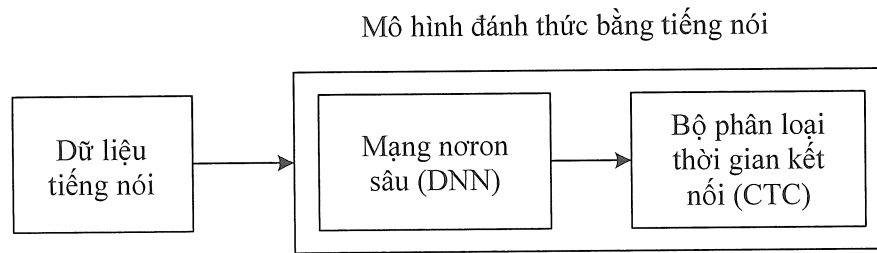
11. Phương tiện lưu trữ phi chuyển tiếp đọc được bằng máy tính có các lệnh được lưu trữ trên đó mà, khi được thực thi bởi bộ xử lý của thiết bị điện tử, khiến cho thiết bị điện tử này thực hiện phương pháp đánh thức bằng tiếng nói, phương pháp này bao gồm các bước:

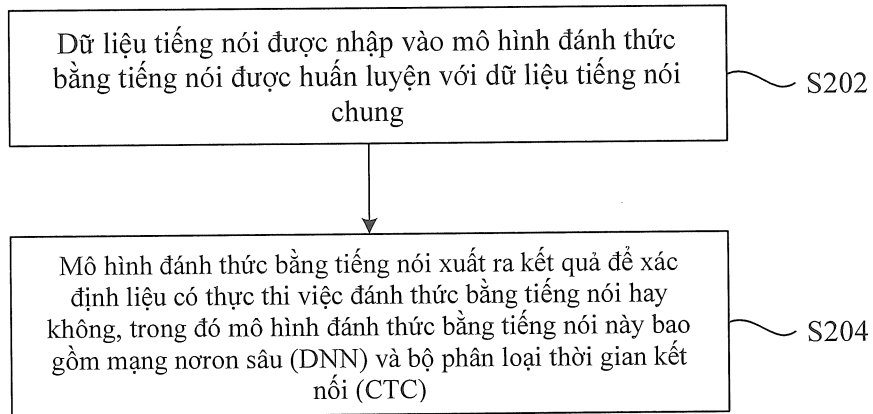
huấn luyện mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói chung, trong đó bước huấn luyện này bao gồm bước: tối ưu hóa theo cách lặp lại các thông số trong mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói chung này thông qua phương pháp giảm gradient ngẫu nhiên không đồng bộ cho đến khi việc huấn luyện hội tụ;

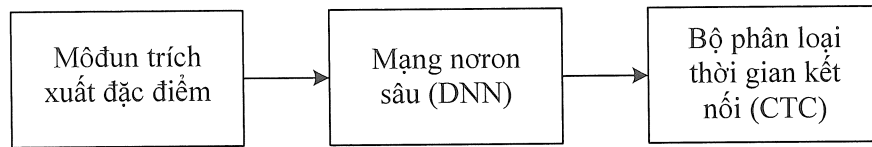
thu nhận dữ liệu tiếng nói theo từ khóa cụ thể, và huấn luyện mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói theo từ khóa cụ thể này, trong đó tỷ lệ học được sử dụng trong việc huấn luyện với dữ liệu tiếng nói theo từ khóa cụ thể này là nhỏ hơn tỷ lệ học được sử dụng trong việc huấn luyện của mô hình đánh thức bằng tiếng nói với dữ liệu tiếng nói chung;

nhập dữ liệu tiếng nói vào mô hình đánh thức bằng tiếng nói được huấn luyện với dữ liệu tiếng nói chung và dữ liệu tiếng nói theo từ khóa cụ thể này; và

xuất ra, bởi mô hình đánh thức bằng tiếng nói này, kết quả để xác định liệu có thực thi việc đánh thức bằng tiếng nói hay không, trong đó mô hình đánh thức bằng tiếng nói này bao gồm mạng nơ-ron sâu (Deep Neural Network, DNN) và bộ phân loại thời gian kết nối (Connectionist Temporal Classifier, CTC).

**FIG. 1**

**FIG. 2**

**FIG. 3**

37671

4/7

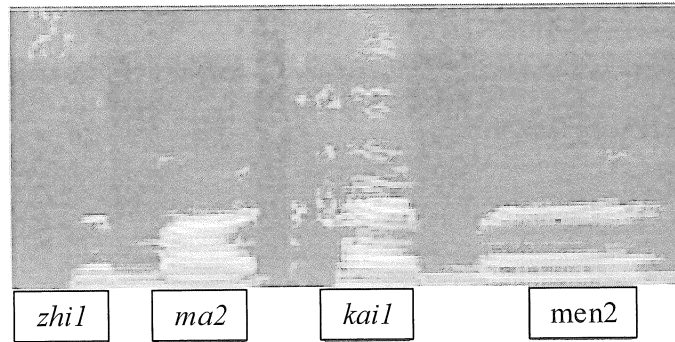


FIG. 4

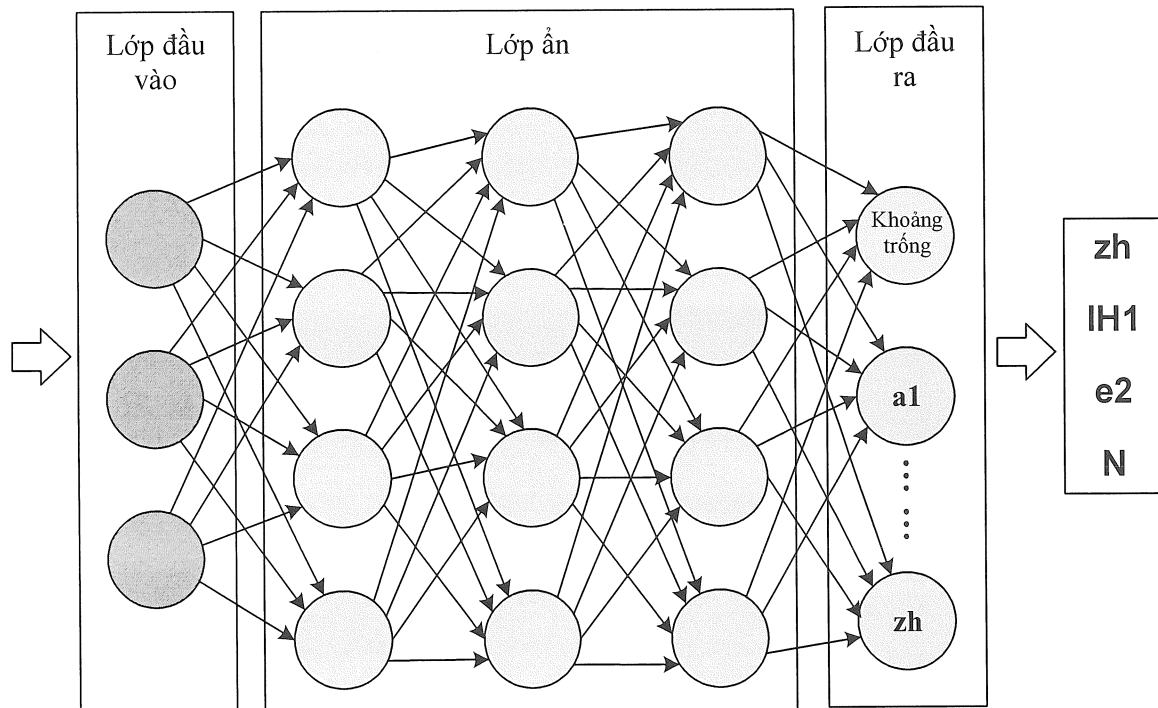


FIG. 5

6/7

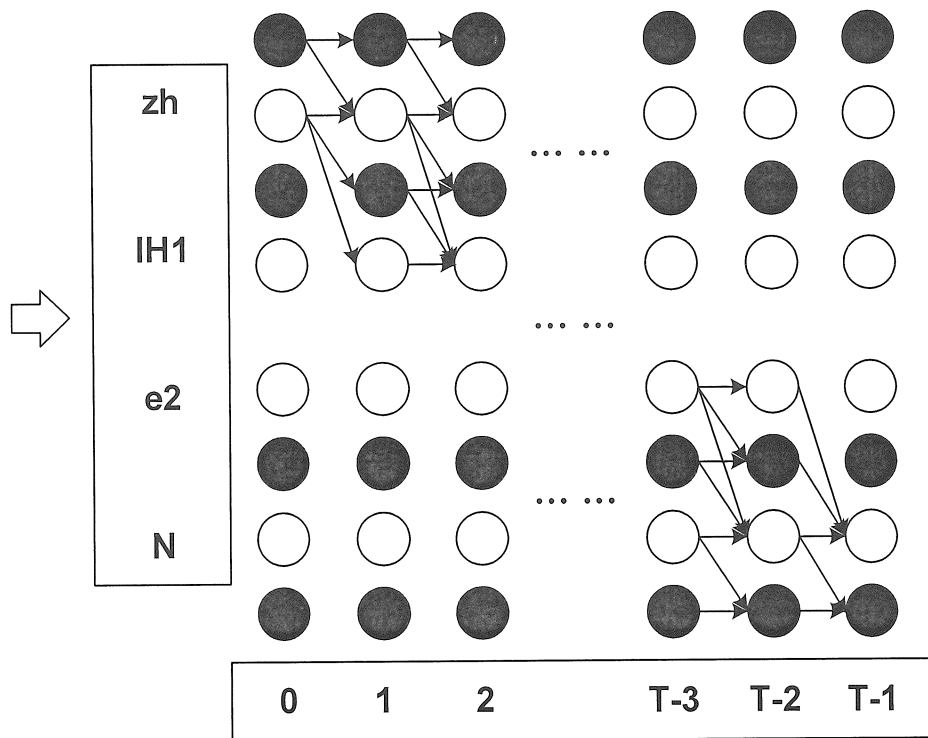


FIG. 6

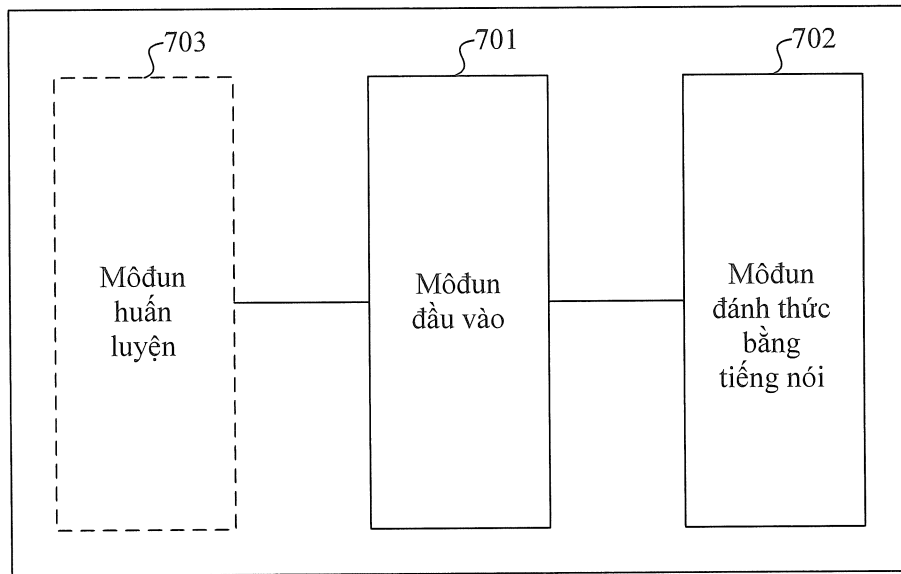


FIG. 7