ỨNG DỤNG NHẬN DẠNG TIẾNG NÓI TIẾNG VIỆT BẰNG MÔ HÌNH MARKOV ẨN ĐỂ ĐIỀU KHIỂN MOBILE ROBOT

USING HIDDEN MARKOV MODEL TO RECOGNIZE VIETNAMESE VOICE

Trần Thu Hà,
Trần Tiến Đức
ĐH Sư Phạm Kỹ Thuật, TP. HCM

TÓM TẮT

Khối mã hóa được thiết kế bằng phần mềm nhận dạng tiếng Việt bằng mô hình Markov ẩn, mô hình âm vị và dùng thuật toán Vertebi để nhận dạng từ hoặc câu lệnh điều khiển. Chọn tối ưu từ lệnh điều khiển với sai số nhỏ nhất khi nhận dạng từ lệnh, giọng nói điều khiển có thể là giọng nam hoặc giọng nữ. Đối tượng điều khiển là mobile robot được thiết kế trên cơ sở sử dụng IC STM32F103RCT6.

ABSTRACT

The system consists of an encoder transferring the vietnamese voice into commands and execution unit - mobile robot. The encoder is designed by a Vietnamese recognition software using phoneme model Hidden Markov Model (HMM) and algorithms of Vertebi for recognizing the words or sentence of control command. Optimization and selection of the command words with minimum recognition error, control voice can be male or female. Object as mobile robot controller is designed on the basis of IC STM32F103RCT6.

1 Giới thiệu:

Trong giai đoạn công nghiệp hoá hiện đại hoá hiện nay của nước nhà, tự động hoá quá trình sản xuất hết sức có ý nghĩa trong việc nâng cao năng suất sản xuất. Trong đó robot đóng vai trò quan trọng trong việc thay thế con người làm việc. Điều khiển robot có thể điều khiển bằng các phương pháp khác nhau, trong đó điều khiển bằng tiếng nói là công nghệ điều khiển linh hoạt. Đã có nhiều công trình nghiên cứu về lĩnh vực nhận dạng tiếng nói (Speech recognition) trên cơ sở lý thuyết các hệ thống thông minh nhân tạo, nhiều kết quả đã trở thành sản phẩm thương mại như Via Voice của IBM, Dragon, Spoken Toolkit của CSLU (Central of Spoken Laguage Understanding)..., các hệ thống bảo mật thông qua nhận dạng tiếng nói các hệ quay số điện thoại bằng giọng nói... Triển khai những công trình nghiên cứu và đưa vào thực tế ứng dụng để điều khiển sự hoạt động của robot là một việc làm hết sức có ý nghĩa đặc biệt trong giai đoạn công nghiệp hoá hiện đại hoá hiện nay. Tuy nhiên ở Việt Nam công nghệ này còn đang ở quá trình nghiên cứu và thử nghiệm.

Do đặc thù tiếng Việt là một ngôn ngữ đơn âm và có thanh điệu. do đó âm vị của từng từ tiếng việt sẽ khác nhau và khác nhau cả mức độ của giọng nam giọng nữ và đặc trưng vùng miền. Việc nghiên cứu ứng dụng điều khiển bằng tiếng nói tiếng Việt vẫn là vấn đề đang mở cho nhiều nhóm đề tài nghiên cứu và ứng dụng.

2. Mục tiêu bài báo:

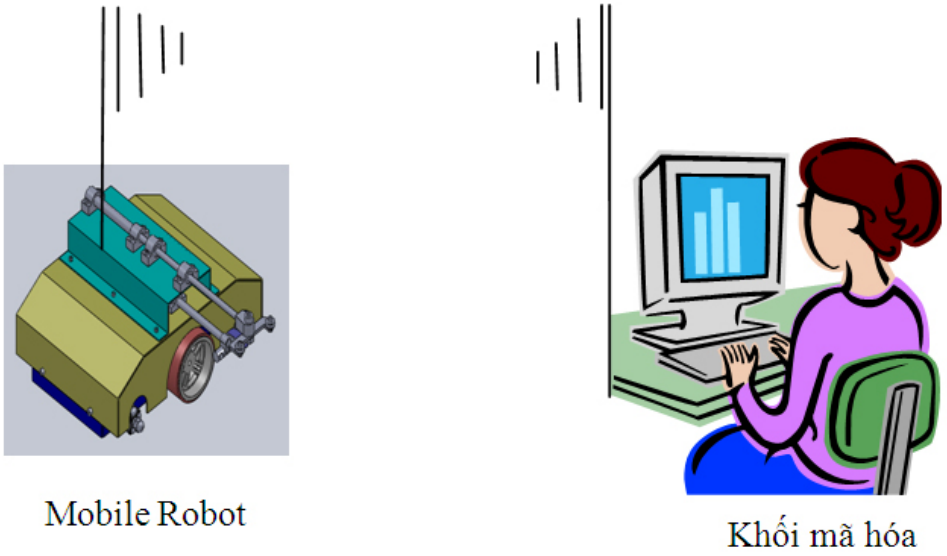
Phân tích tiếng việt để xây dựng hệ nhận dạng tiếng nói dựa trên phương pháp thống kê theo mô hình Markov ẩn nhận dạng âm vị và dùng giải thuật Vertebi nhận dạng từ lệnh điều khiển.

Xây dựng phần mềm nhận dạng từ lệnh tiếng Việt để điều khiển Robot. Chọn lựa từ lệnh điều khiển với xác suất nhận dạng với sai số nhỏ nhất.

Thiết kế thi công mobile robot điều khiển

bằng giọng nói.

3. Hệ thống điều khiển robot bằng tiếng nói:



Hình 1: Mô hình điều khiển robot bằng tiếng nói.

Mô hình điều khiển robot bằng tiếng nói có cấu trúc gồm 02 khối trên hình 1 bao gồm khối phát tín hiệu điều khiển là khối mã hóa tiếng nói và khối thu tín hiệu để thực thi lệnh điều khiển là Mobile robot.Tín hiệu đã mã hóa và gửi tới mobile robot qua bộ thu phát RF.

Khối mã hóa tiếng nói có sơ đồ khối như hình 2 là khối nhận lệnh tiếng nói điều khiển tại micro qua sound card và tín hiệu được được nhận dạng âm vị bằng mô hình HMM và sử dụng giải thuật Vertebi để nhận dạng từ lệnh và hiển thị trên giao diện điều khiển các lệnh và tín hiệu sẽ được mã hóa chuyển đến RF module [5,6,7,8].

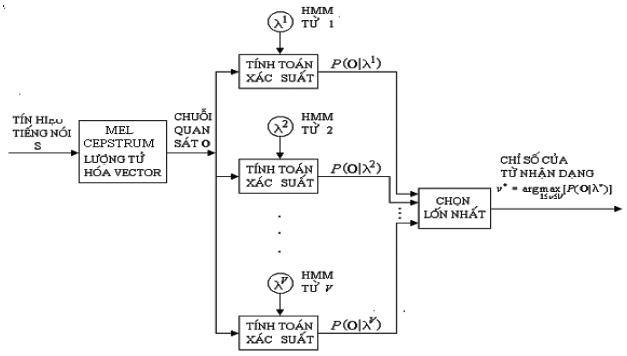
a- Phân tích tiếng việt để xây dựng hệ thống nhận dạng;

Bài báo đề xuất chương trình nhận dạng tiếng nói tiếng Việt sử dụng mô hình Markov ẩn theo mô hình âm vị để trích tham số của tín hiệu tiếng nói phân tích cepstrum thông qua dãy bộ lọc theo thang tần số Mel – Mel Frequency Cepstral Coefficients (MFCC). Sử dụng kỹ thuật lượng tử hóa vector - vector quantization - dùng để lấy trung bình đặc tính của các frame cũng như đánh nhãn các vector và được ứng dụng trong nhận dạng tiếng nói bằng mô hình Markov ẩn. Ta chọn và phân mỗi từ thành nhiều frame, mỗi frame có "*N*" mẫu. Các frame của tiếng nói sẽ biểu diễn qua hàm năng lượng ngắn hạn [1,2] và xử lý tiếng nói bằng thuật toán phát hiện điểm đầu và cuối của một từ căn cứ vào hàm năng lượng ngắn hạn[5,7,8]:

a. Với mỗi frame, tính năng lượng ngắn hạn *E[5.8]*, nếu *E* lớn hơn giá trị ngưỡng cho trước thì đánh dấu frame đó là bắt đầu của một từ, ký hiệu là frame *B*, ngược lại thì xét frame kế cho đến khi xác định được frame *B*. Nếu không xác định được frame *B* thì tín hiệu đó không phải là tiếng nói.

b. Tính năng lượng ngắn hạn *E* của frame kế cho đến khi *E* nhỏ hơn giá trị ngưỡng thì đánh dấu frame đó là kết thúc của một từ, ký hiệu là frame *K*.

c. Hiệu chỉnh điểm bắt đầu bằng cách tính hàm năng lượng ngắn hạn *e* của các frame con xung quanh frame *B*, rồi chọn frame con thích hợp khi so sánh với , mỗi frame con phải có độ dài nhỏ hơn frame.



Hình 3: Lưu đồ quy trình nhận dạng âm vị bằng HMM.

d. Tương tự hiệu chỉnh điểm kết thúc ở frame *K*.Thực hiện đề xuất mô hình lọc nguồn tạo tiếng nói phổ tiếng nói hữu thanh có khuynh hướng suy giảm toàn bộ -6dB/octave khi tần số tăng lên.

e. Xây dựng chương trình nhận dạng tiếng nói tiếng Việt liên tục sử dụng mô hình âm vị phụ thuộc sử dụng giải thuật Vertebi để xác định các ma trận trạng thái, ma trận quan sát và ma trận vị trí ban đầu. Trên hình 3, trình bày lưu đồ nhận dạng tiếng việt bằng mô hình âm vị của HMM và giải thuật Vertebi để nhận dạng từ. Ta cần nhận dạng bộ từ vựng có *V* từ, mỗi từ đều có mô hình Markov riêng và được nói *K* lần (có thể một hay nhiều người nói), ta thực hiện các bước sau đây:

1. Với mỗi từ *v* trong bộ từ vựng, ta phải xây dựng một mô hình Markov ẩn , tức là ta phải ước lượng các tham số của mô hình dựa trên tập dữ liệu huấn luyện.

2. Với mỗi từ chưa biết, ta xây dựng mô hình nhận dạng như trên Hình 3. Tín hiệu tiếng nói được trích đặc điểm bằng phương pháp Mel-Cepstrum hay LPC-Cepstrum, thông qua bộ lượng tử hóa vector ta có được chuỗi quan sát . Tiếp theo ta tính xác suất cho tất cả mô hình , , và chọn từ có xác suất lớn nhất, tức là (1)

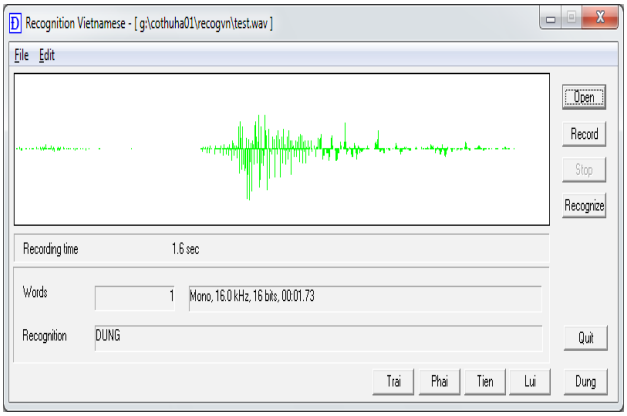
Bước tính xác suất dùng thuật toán Viterbi, và cần phép tính. Với bộ từ vựng từ, mô hình phương trình 1 trạng thái và quan sát cho mỗi từ chưa biết, tổng cộng hết phép tính (mỗi phép tính gồm một phép nhân, một phép cộng, và một phép tính mật độ quan sát).

Cần phải tính đến thời gian trích đặc điểm và lượng tử hóa vector [4,5].

b- Giao diện điều khiển bằng tiếng nói:

Giao diện điều khiển robot bằng tiếng nói như hình 4. Trên giao diện ta có thể nhận thấy các nút điều khiển bằng tay ở chế độ sử dụng lệnh cài đặt sẵn.

Khi điều khiển bằng giọng nói, lệnh sẽ hiển thị sóng của từ lệnh điều khiển và hiển thị từ bằng tiếng việt không dấu.



Hình 4: Giao diện điều khiển robot bằng tiếng nói

Để nhận dạng từ lệnh điều khiển ta phải thực hiện hai chế độ: Chế độ 01: dạy - tập huấn từ lệnh điều khiển; Chế độ 02: chế độ điều khiển robot bằng tiếng nói

* ***Chế độ 01- Chế độ dạy - tập huấn từ lệnh điều khiển***: là chế độ mà các câu lệnh được thu âm trong môi trường trong nhà, do một giọng nam hoặc nữ đọc, sử dụng micro thông thường gắn với máy tính.

Trong chế độ dạy - tập huấn trên hình 5 sẽ tiến hành theo trình tự các khối: Giọng nói thu từ micro âm thanh được lấy mẫu theo thời gian và được lượng tử hóa theo biên độ, xác định các vector để trích đặc trưng âm thanh bằng phương pháp MFCC (Mel-Frequency Cepstrums Coefficients) với các bộ lọc Mel [3,4,7]. Xác định âm vị để huấn luyện các hệ số ma trận xác suất trạng thái vị trí, ma trận xác suất quan sát và ma trận khởi tạo trạng thái ban đầu để xác định âm vị bằng cách chọn lựa xác suất lớn nhất, sau đó dùng giải thuật Vertebi để nhận dạng từ. Từ lệnh tiếng Việt được nhận dạng sau 100 ms sẽ được hiển thị trên giao diện điều khiển đồng thời được lưu trữ âm vị trong bộ nhớ [1,2,4,6].

Ví dụ: Để điều khiển robot rẽ trái ta dạy cho

phần mềm từ "Trái", ta nói vào micro từ lệnh "Trái". Tín hiệu tiếng nói này được trích đặc điểm theo MFCC giả sử có số lượng là 6 frame và lượng tử hóa dùng Markov ẩn để huấn luyện các tham số cho các âm vị <tr>, <a> và <i ngắn> và lệnh "Tiến" bao gồm <t>, <ie >, < n> [5].

Ta có ma trận xác suất của lệnh "Trái " và lệnh "Tiến".

tr	0.1	0.1	0.2	0.1	0.3	0.1
a	0.2	0.3	0.3	0.4	0.2	0.2
i	0.2	0.3	0.4	0.8	0.6	0.2
t	0.3	0.2	0.1	0.1	0.9	0.8
ie	0.4	0.3	0.2	0.1	0.2	0.2
n	0.8	0.9	0.4	0.2	0.1	0.1

Hình 6: Bảng ma trận xác suất âm vị

Bài toán nhận dạng bằng giải thuật Viterbi câu lệnh là: "Trái" <tr - a - i ngắn>và "tiến <t - ie - n> gồm 6 âm vị độc lập theo thứ tự âm vị 1 là <tr>, âm vị 2 là <a>, âm vị 3 là <i>, âm vị 4 là < t>, âm vị 5 là <ie>, âm vị 6 là<n>. Trên hình 5 hiển thị tiếng nói được nhận dạng theo âm vị <tr> với xác xuất là 0.1 sau đó âm vị tiếp theo là <tr> hoặc <a> nhưng ta thấy là

xác xuất âm vị <tr> là 01 và xác xuất <a> là 0.3 lớn hơn xác xuất của <tr>, do đó Vertebi sẽ chuyển đến âm vị <a> có xác xuất là 0.3. Tiếp theo là <I ngắn> - với xác xuất 0.4 do đó Vertebi sẽ tiến tới 0.4. Như vậy câu lệnh "trái" sẽ được nhận dạng với tổng ma trận xác xuất là 2,2.

Trên hình 7 bảng thử nghiệm nhận dạng các từ đơn tiếng việt ta thấy các từ nhận dạng càng có nhiều âm ghép sai số nhận dạng càng lớn.

Kết quả thử nghiệm sử dụng phần mềm nhận dạng từ cho thấy giọng nữ có sai số nhận dạng cao hơn giọng nam và trên cơ sở thực nghiệm ta có số lần dạy khoảng 90 - 100 lần, do đó ta chọn điểm tối ưu để dạy nhận dạng giọng nói là 100 lần cho giọng nam và cho giọng nữ.

Thử nghiệm nhận dạng hiển thị các từ trong tập lệnh từ đơn cho thấy rõ đối với các từ có các nguyên âm ghép hoặc từ có dấu sắc có sai số lớn hơn các từ lệnh "Lùi", "Dừng". Nhận dạng chính xác nhất là từ lệnh "Lùi" có các đơn âm <l, u, i ngắn>, từ được nhận giá trị sai số là 2,5 %. Từ lệnh "Dừng" có âm vị "Ng" là từ ghép do đó sai số cao hơn là 4% (tập huấn 100 lần)[8].

Tiến hành thử nghiệm nhận dạng từ lệnh quay trái 90 độ và quay trái một góc 45 độ và 60 độ. Chọn từ lệnh điều khiển nhận dạng sau 5 lượt tập huấn với số lần dạy 120, 100, 90, 60 và 50 lần, ta có kết quả chọn lệnh tối ưu so sánh các câu lệnh với mục đích robot quay trái 45 độ ta có các phương án lệnh điều khiển là "Trái bốn"sai số cho 100 lần dạy là 17,78 %; " Trái bốn lăm"sai số cho 100 lần dạy là 21,2 %; "Trái một " sai số nhận dạng của 100 lần dạy là 15 %. Từ kết quả ta chọn lệnh điều khiển là " trái một " cho ý muốn điều khiển robot rẽ trái 45 độ[8].

**** Chế độ 02- Chế độ điều khiển robot***:

Tín hiệu điều khiển sẽ được nói qua Micro, âm thanh được lấy mẫu theo thời gian và được lượng tử hóa theo biên độ, xác định các vector để trích đặc trưng âm thanh bằng phương pháp MFCC(Mel-Frequency Cepstrums Coefficients) với các bộ lọc Mel. Xác định âm vị lệnh điều khiển nhận dạng từ lệnh sau thời gian là 100 ms. Từ lệnh điều khiển được nhận và so sánh với các từ đã được dạy tập huấn sẽ hiển thị trên dao diện điều khiển và được mã hóa chuyển đến module RF thông qua cổng giao tiếp nối tiếp COM của máy tính và module APC200A để phát lệnh điều khiển [8].

Khối thu có sơ đồ khối hình 9 trên robot sẽ nhận lệnh để điều khiển và chuyển đổi mã lệnh đến bộ điều khiển trung tâm điều khiển động cơ thực hiện các lệnh tiến, lùi, trái, phải, trái một, phải một…

Robot hoạt động ổn định khi nhận lệnh điều khiển. Hệ thống nhận dạng tiếng nói là từ lệnh của tiếng việt chính xác và robot có khả năng hoạt động trong bán kính 1000 m.

Robot có thể điều khiển với từ lệnh "trái" hoặc "phải" và sẽ quay tròn về phía trái hoặc phía phải, mãi cho đến khi có lệnh dừng. Robot này có thể ứng dụng với các mục đích biểu diễn hoặc ứng dụng thực hiện nhiệm vụ quét, rửa máy móc công nghiệp hoặc dùng làm giá đỡ cho việc mài một sản phẩm tròn...

Nếu việc nhận dạng thực sảy ra sai số sẽ có trường hợp từ lệnh nhầm thì việc thực thi của robot sẽ sai lệnh và nếu âm vị của từ nhận dạng sai, từ nhận dạng được không có trong bộ lưu trữ thì lệnh sẽ không được mã hóa và robot sẽ đứng yên. Đây là một trong những giới hạn của vấn đề nghiên cứu.

4- Kết luận:

Điều khiển bằng tiếng nói là vấn đề lý thú đang được nhiều nhóm nghiên cứu thực hiện và quan tâm hiện nay, bài báo là một trong những ứng dụng của việc phân tích để nhận dạng tiếng nói, phân tích chọn lựa các thông số tập huấn số lượt tập huấn và đánh giá sai số cho các lượt tập huấn để chọn tối ưu từ lệnh điều khiển. Từ lệnh điều khiển ngắn gọn dễ nhớ và có xác suất nhận dạng chính xác, là các từ "trái", "phải", " tiến", "lùi", "dừng", "Phải một", "Phải hai", "Trái một", "Trái hai".

Quá trình thực hiện cho việc dạy tập huấn tiếng nói có thể chọn từ 90 đến 100 lần đọc của giọng nam hoặc giọng nữ. Tiếng nói được phân tích theo các đặc trưng âm vị và lưu lại bộ nhớ. Các từ lệnh điều khiển được nhận dạng chính xác đến 96 %[8]. Trong tương lai sẽ phát triển giao diện có thể nhận lệnh điều khiển của các giọng nói của nam hoặc nữ chuẩn (của Hà nội gốc) mà không cần tập huấn lại các lệnh.

Bài báo đã giới thiệu hệ thống điều khiển robot bằng tiếng nói với giao diện điều khiển hiện thị từ

lệnh điều khiển bằng tiếng Việt không dấu. Ứng dụng của hệ thống điều khiển bằng tiếng nói trong nhiều lãnh vực như công nghệ sản xuất trò chơi điện tử với robot thông minh, điều khiển tiếng nói trong công nghiệp, trong công nghệ thông tin, trong công nghệ sản xuất và lãnh vực dịch vụ y tế ,v.v…

TÀI LIỆU THAM KHẢO

[1] Thuong Le-Tien, "A study on the continuous wavelet transform for the Vietnamese speech processing", *Proceedings of the 97 international conference on natural information and intelligent information systems*, Vol. 2, New Zealand, 1997.

[2] Lê Tiến Thường, Trần Tiến Đức, "Nhận dạng tiếng nói tiếng Việt liên tục bằng mạng nơ-ron", *Tạp chí Phát triển khoa học và công nghệ*, Đại học Quốc gia Tp. HCM, Số 10, Tập 5, 2002.

[3] Yiping Wang and Zhefeng Zhao - A Noise – "Robust Speech Recognition System Based on Wavelet Neural Network". Lecture Notes in Computer Science, 2011, Volume 7004, Artificial Intelligence and Computational Intelligence, Pages 392-397.

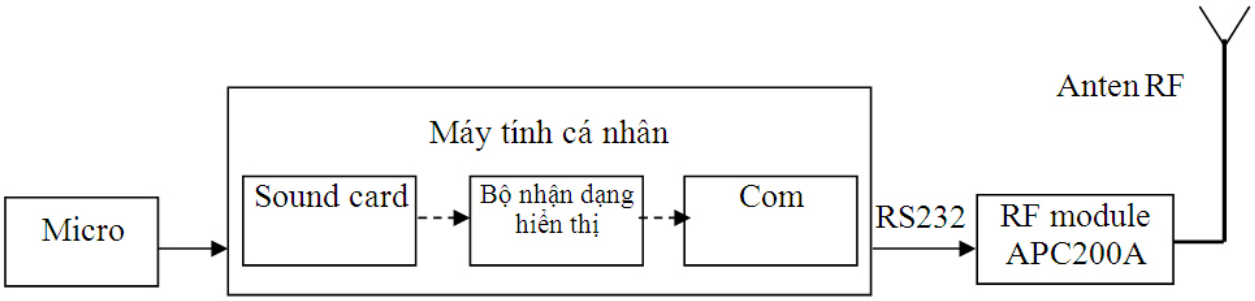
[4] Murakami Ryo, *Voice recognition robot and its control method*. Toyota Motor corp., Japan patent application : JP 2008126329 (A); Publication date 2008-06-05.

[5] Trần Tiến Đức, "Nhận dạng 10 chữ số tiếng Việt phát âm rời bằng mô hình Markov ẩn có mật độ quan sát liên tục"*, Tạp chí Khoa học và công nghệ các trường đại học kỹ thuật*, Số 27+28, 2001.

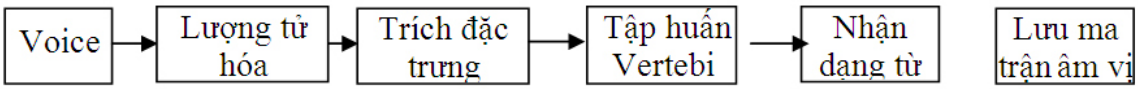
[6] Astrov Sergey; Bauer Josef, *Method of speaker adaptation for a Hidden Markov model based voice recognition system*. Patent application US: US 2006074665 (A1); Applicant(s): Siemens Aktiengesellschaft; Publication date: 2006-04-06 .

[7] Kawashima Kakahiro, *Robot and voice reproduction method*. Patent application: JP 2005266671 (A); Applicant(s): YAMAHA CORP; Publication date: 2005-09-29.

[8] Trần Thu Hà. Trần Tiến Đức, "Mã hóa tiếng nói thành lệnh điều khiển trong công nghiệp". Đề tài nghiên cứu khoa học cấp bộ năm 2009. Mã số B2009 -22-43.



Hinh 2: Sơ đồ khối của khối mã hóa



Hình 5: Bộ nhận dạng của chế độ dạy - tập huấn

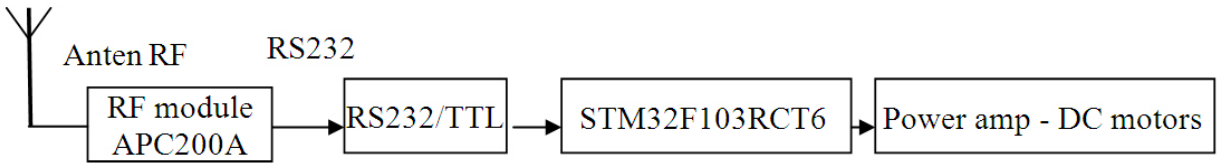
Hinh 5: Sơ đồ khối của khối mã hóa

Số lượt tập huấn	Tiếng nói	Âm vị	Số lần dạy	Hiển thị đúng (Nhận dạng đúng)	Hiển thị sai chữ	Sai số %	Sai số TB	Hiển thị sai chữ	Sai số	Sai số TB
				GIỌNG NAM				GIỌNG NỮ		
1	Tiến	t, ie, n	120	110	11	0.0917	0.0837	12	0.1000	0.0991
2			100	97	8	0.0800		9	0.0900	
3			90	85	6	0.0667		8	0.0889	
4			60	63	6	0.1000		7	0.1167	
5			50	54	4	0.0800		5	0.1000	
1	Tướng	t, ươ, ng	120	109	11	0.0917	0.1074	12	0.1000	0.1262
2			100	90	10	0.1000		11	0.1100	
3			90	82	8	0.0889		10	0.1111	
4			60	53	7	0.1167		9	0.1500	
5			50	43	7	0.1400		8	0.1600	
1	Biến	b, ie, n	120	110	10	0.0833	0.1038	12	0.1000	0.1256
2			100	91	9	0.0900		11	0.1100	
3			90	82	8	0.0889		10	0.1111	
4			60	53	7	0.1167		10	0.1667	
5			50	43	7	0.1400		7	0.1400	
1	Thường	th, ươ,ng	120	110	10	0.0833	0.0882	11	0.0917	0.1014
2			100	92	8	0.0800		9	0.0900	
3			90	83	7	0.0778		8	0.0889	
4			60	54	6	0.1000		7	0.1167	
5			50	45	5	0.1000		6	0.1200	

Hình 7 Bảng thử nghiệm nhận dạng bằng mô hình âm vị của HMM

Từ lệnh điều khiển	Mã số	Thực hiện lệnh	Độ chính xác khi nhận dạng từ lệnh (100 lần dạy)
Lệnh đơn			
Trái	01	Quay trái 90 độ và lệnh tiến sau 100 ms	95 %
Phải	02	Quay phải 90 độ và lệnh tiến sau 100 ms	96%
Lùi	03	Lùi (động cơ quay lui)	97%
Dừng	04	Động cơ đang chạy dừng lại	96%
Tiến	05	Tiến thẳng (động cơ quay tới)	92%
Lệnh đôi			
Trái một	06	Quay trái 45 độ và lệnh tiến	85%
Trái hai	07	Quay trái 60 độ và lệnh tiến	85%
Đi tới (Tiến)	08	Đi tới	86%
Phải một	11	Quay phải 45 độ và tiến thẳng	88 %
Phải hai	12	Quay phải 60 độ và tiến thẳng	88 %

Hình 8 Bảng mã hóa tập lệnh điều khiển robot



Hình 9: Khối thu trên mobile robot

Hình 9: Khối thu trên mobile robot