KỸ THUẬT NHẬN DẠNG TIẾNG NÓI và ỨNG DỤNG TRONG ĐIỀU KHIỂN

TS. Nguyễn Văn Giáp
KS. Trần Việt Hồng
Bộ môn Cơ điện tử - Khoa Cơ khí – Đại học Bách Khoa TPHCM
nvgiap@dme.hcmut.edu.vn; tvhong@dme.hcmut.edu.vn

TÓM TẮT

Vấn đề nghiên cứu các phương pháp nhận dạng tiếng nói đã và đang thu hút rất nhiều sự đầu tư và nghiên cứu của các nhà khoa học trên khắp thế giới. Tuy nhiên cho đến nay kết quả mang lại vẫn chưa hoàn toàn làm hài lòng những người nghiên cứu do tính chất quá phức tạp và không cố định của đối tượng nhận dạng là tiếng nói con người. Đặc biệt, đối với tiếng Việt thì kết quả càng còn nhiều hạn chế. Bài báo trình bày một hướng nhận dạng tiếng nói tiếng Việt, dựa trên việc trích đặc trưng tiếng nói bằng phương pháp MFCC và bộ nhận dạng dùng mạng HMM. Kết quả được kiểm nghiệm thực tế bằng mô hình xe điều khiển từ xa.

ABSTRACT

Researching and inventing speech recognition methods have been paid much considerations by many scientists over the world. However, the achievements don't satisfy researchers' demands because of the complexity and unstability of speech until now. Especially with Vietnamese speech, the results are more unsatisfied. The paper suggests a synthetic method for recogniting Vietnamese speech: extract speech's particularities by MFCC method and recognize by HMM network. The results are experimented through a model of RF controlled car.

1 ĐẶT VẤN ĐỀ

1.1 Giới thiệu

Ngày nay, cùng với sự phát triển của ngành điện tử và tin học, các hệ thống máy tự động đã dần thay thế con người trong nhiều công đoạn của công việc. Máy có khả năng làm việc hiệu quả và năng suất cao hơn con người rất nhiều. Song cho đến nay, vấn đề giao tiếp người – máy tuy đã được cải thiện nhiều nhưng vẫn còn rất thủ công: thông qua bàn phím và các thiết bị nhập dữ liệu khác. Giao tiếp với thiết bị máy bằng tiếng nói sẽ là phương thức giao tiếp văn minh và tự nhiên nhất, dấu ấn giao tiếp người – máy sẽ mất đi mà thay vào đó là cảm nhận của sự giao tiếp giữa người với người, nếu hoàn thiện thì đây sẽ là một phương thức giao tiếp tiện lợi và hiệu quả nhất.

Do có sự khác biệt về mặt ngữ âm giữa các ngôn ngữ nên ta không thể áp dụng các chương trình nhận dạng khác để nhận dạng tiếng Việt. Một hệ thống nhận dạng tiếng nói ở nước ta phải được xây dựng trên nền tảng của tiếng nói tiếng Việt.

1.2 Tình hình nghiên cứu trong và ngoài nước

Vấn đề nhận dạng tiếng nói tiếng Việt chỉ mới được quan tâm nghiên cứu trong những năm gần đây và chưa có một chương trình nhận dạng hoàn chỉnh nào được công bố.

Trên thế giới đã có rất nhiều hệ thống nhận dạng tiếng nói (tiếng Anh) đã và đang được ứng dụng rất hiệu quả như: Via Voice của IBM, Spoken Toolkit của CSLU (Central of Spoken Laguage Under-standing)… nhưng trong tiếng Việt thì còn rất nhiều hạn chế.

1.3 Mục tiêu của đề tài

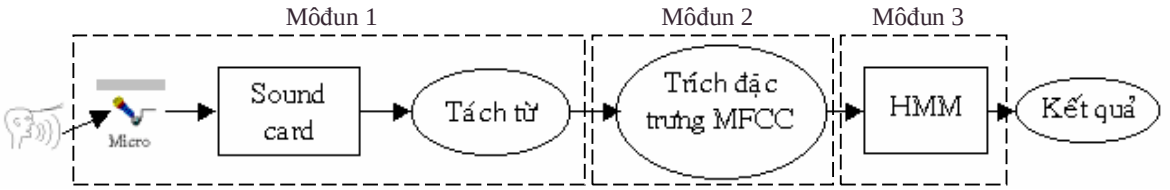
Đề tài này nghiên cứu thử nghiệm một hướng nhận dạng tiếng nói - tiếng Việt dựa trên việc trích đặc trưng của tiếng nói bằng phương pháp MFCC (Mel-Frequency Ceptrums Coefficients), và nhận dạng bằng mô hình HMM (Hidden Markov Models). Đồng thời, một mô hình điều khiển bằng tiếng nói – tiếng Việt được xây dựng với bộ từ vựng nhỏ, thiết lập hệ thống điều khiển bằng tiếng nói với một tập lệnh cố định. Tập lệnh này dùng để điều khiển Robot, và mô hình điều khiển xe bằng tiếng nói hoàn chỉnh là một ứng dụng thực tế mang tính thử nghiệm của đề tài.

2 XÂY DỰNG HỆ THỐNG NHẬN DẠNG TIẾNG NÓI

Một hệ thống nhận dạng nói chung thường bao gồm hai phần: phần huấn luyện (training phase) và phần nhận dạng (recognition phase). "Huấn luyện" là quá trình hệ thống "học" những mẫu chuẩn được cung cấp bởi những tiếng khác nhau (từ hoặc âm), để từ đó hình thành bộ từ vựng của hệ thống. "Nhận dạng" là quá trình quyết định xem từ nào được đọc căn cứ vào bộ từ vựng đã được huấn luyện. Sơ đồ tổng quát của hệ thống nhận dạng tiếng nói được thể hiện trên hình 1.

Để thuận tiện cho việc kiểm tra và đánh giá kết quả, từ sơ đồ trên chúng tôi chia chương trình nhận dạng thành ba mô-đun riêng biệt:

- *Mô-đun 1*: Thực hiện việc ghi âm tín hiệu tiếng nói, tách tiếng nói khỏi nền nhiễu và lưu vào cơ sở dữ liệu.
- *Mô-đun 2*: Trích đặc trưng tín hiệu tiếng nói đã thu ở mô-đun 1 bằng phương pháp MFCC, đồng thời thực hiện ước lượng vector các vector đặc trưng này.
- *Mô-đun 3*: Xây dựng mô hình Markov ẩn với 6 trạng thái, tối ưu hóa các hệ số của HMM tương ứng với từng từ trong bộ từ vựng, tiến hành nhận dạng một từ được đọc vào micro.



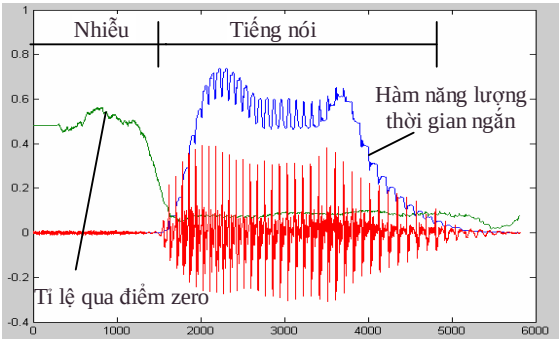
Hình 1 *Sơ đồ tổng quát hệ thống nhận dạng tiếng nói.*

2.1 Thực hiện mô-đun 1

Nhiệm vụ của mô-đun này là thu tín hiệu từ micro, dùng kỹ thuật xử lý đầu cuối để phát hiện phần tín hiệu tiếng nói và phần tín hiệu nhiễu. Từ đó ta có thể tách tiếng nói ra khỏi nền nhiễu (chỉ thu tín hiệu tiếng nói mà không thu tín hiệu nhiễu nền).

Tuy có nhiều phương pháp tách tiếng nói khác nhau, nhưng qua quá trình nghiên cứu và thử nghiệm các tác giả nhận thấy sự kết hợp giữa phương pháp hàm năng lượng thời gian ngắn và tỉ lệ qua điểm zero cho kết quả tốt hơn.

Phương pháp này dựa vào tính chất năng lượng của tín hiệu tiếng nói thường lớn hơn năng lượng của tín hiệu nhiễu và tỉ lệ qua điểm zero của nhiễu sẽ lớn hơn tín hiệu tiếng nói. Hình 2 cho thấy mối quan hệ giữa tín hiệu thu được, giá trị của hàm năng lượng thời gian ngắn và tỉ lệ qua điểm zero.

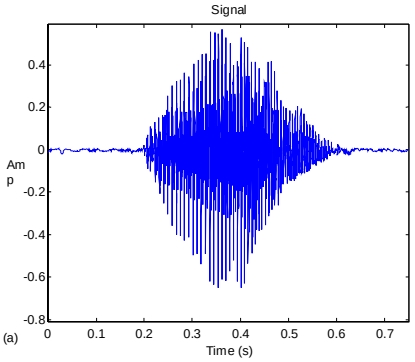


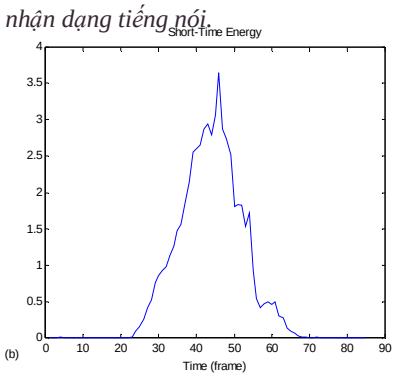
Hình 2 *Sự tương quan giữa tín hiệu tiếng nói và nền nhiễu.*

Với một cửa sổ kết thúc tại mẫu thứ m, hàm năng lượng thời gian ngắn E(m) được xác định bởi:

$$E(m) = \sum_{n=-\infty}^{\infty} [s(n)w(m-n)]^2 \qquad \textbf{(2.1)}$$

Đồ thị của hàm năng lượng thời gian ngắn của một đoạn tín hiệu được thể hiện trên hình 3.





Hình 3 *Tín hiệu (a) và năng lượng thời gian ngắn (b)*

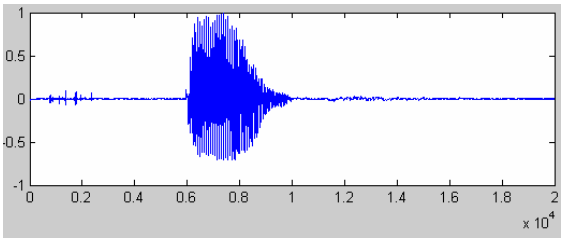
Tỷ lệ qua điểm zero (zero crossing rate) là một thông số cho biết số lần mà biên độ tín hiệu đi qua điểm zero trong một khoảng thời gian cho trước được xác định bởi:

$$Z_s(m) = \frac{1}{N} \sum_{n=m-N+1}^{m} \frac{|\text{sgn}\{s(n)\} - \text{sgn}\{s(n-1)\}|}{2} w(m-n) \qquad \textbf{(2.2)}$$

trong đó, N là chiều dài của cửa sổ w(m-n).

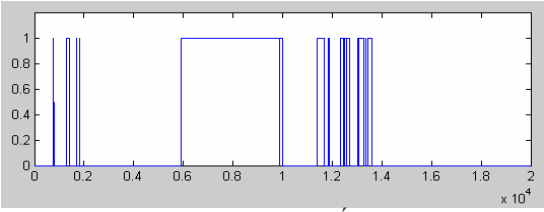
Nhiều thuật toán phát hiện đầu cuối được dựa trên độ lớn của tín hiệu năng lượng thời gian ngắn và tỉ lệ qua điểm zero để cố gắng phát hiện chính xác đến mức có thể. Quá trình cơ bản của thuật toán như sau: một mẫu tín hiệu nhỏ của nền nhiễu được lấy trong suốt khoảng "lặng" (silence) cho đến trước điểm bắt đầu của tín hiệu tiếng nói. Từ đây ngưỡng tiếng nói được xác định dựa trên năng lượng khoảng lặng và năng lượng đỉnh. Ban đầu, những điểm kết thúc được xác định ở những nơi năng lượng tín hiệu vượt qua ngưỡng này, sau đó ta tính khoảng cách giữa hai điểm xem có thoả mãn độ dài của một từ hay không. Tương tự ta áp dụng cho tỉ lệ qua điểm zero. **[4-6]**

Ví dụ: tín hiệu thu vào từ micro bao gồm nhiễu nền và tiếng nói có đồ thị như sau:



Hình 4 *Tín hiệu của từ "tới".*

Qua quá trình xử lý theo chu trình trên ta có được đồ thị dạng xung như sau:



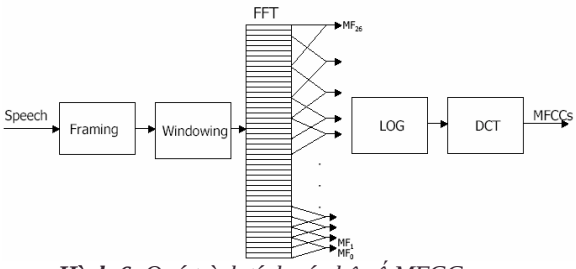
Hình 5 *Dạng xung sau xử lý kết hợp hàm năng lượng thời gian ngắn và tỉ lệ qua điểm zero*

Từ hình 5 ta thấy chỉ cần xác định độ dài tối thiểu của một từ là ta có thể tách từ ra khỏi nền nhiễu. Đến đây mô-đun 1 đã hoàn thành nhiệm vụ. Đây là một phần rất quan trọng trong một hệ thống nhận dạng tiếng nói, nó ảnh hưởng rất lớn đến kết quả nhận dạng.

2.2 Thực hiện mô-đun 2

Đến đây chúng ta đã có được các mẫu tiếng nói đã được khử nhiễu. Mô-đun 2 thực hiện việc trích đặc trưng các mẫu tiếng nói đã thu ở mô-đun 1. Có nhiều phương pháp trích đặc trưng khác nhau như: wavelets, LPC, MFCC… Ở đây chọn phương pháp MFCC (trích đặc trưng theo thang tần số Mel) do tốc độ tính toán cao, độ tin cậy lớn và đã được sử dụng rất hiệu quả trong các chương trình nhận dạng tiếng nói trên thế giới.

Sơ đồ giải thuật phương pháp MFCC như sau:



Hình 6 *Quá trình tính các hệ số MFCC.*

➢ Cửa sổ hoá tín hiệu (Windowing)

Những phương pháp đánh giá phổ cổ điển chỉ đáng tin cậy trong trường hợp tín hiệu dừng (stationary signal), ví dụ một tín hiệu mà những đặc trưng là bất biến đối với thời gian. Đối với tín hiệu tiếng nói thì điều này chỉ có được trong một khoảng thời gian ngắn, việc này có thể thực hiện được bằng cách "cửa sổ hoá" một tín hiệu x'(n) thành một chuỗi liên tục những cửa sổ tuần tự $x_t(n)$, t=1,2,……,T, gọi là những frame.

Trong hệ thống nhận dạng tự động thì dạng cửa sổ thường dùng nhất là Hamming window, đáp ứng xung của nó là một hàm cosin tăng:

$$w(n)=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{N-1}\right) & n=0,...,N-1\\ 0 & n\quad khac\end{cases}$$

➢ Phân tích phổ

Nếu những giá trị có khoảng cách đều nhau, tức là xem $w=\dfrac{2\pi k}{N}$, thì biến đổi Fourier rời rạc (DFT) của tất cả các frame của tín hiệu là:

$$X_t(k)=X_t(e^{j2\pi k/N}) \qquad k=0,...,N-1.$$

Bên cạnh đó nếu số mẫu N là bội số của 2 (N=2p, p là số nguyên) thì độ phức tạp tính toán sẽ giảm đáng kể khi dùng phương pháp FFT (Fast Fourier Transform).

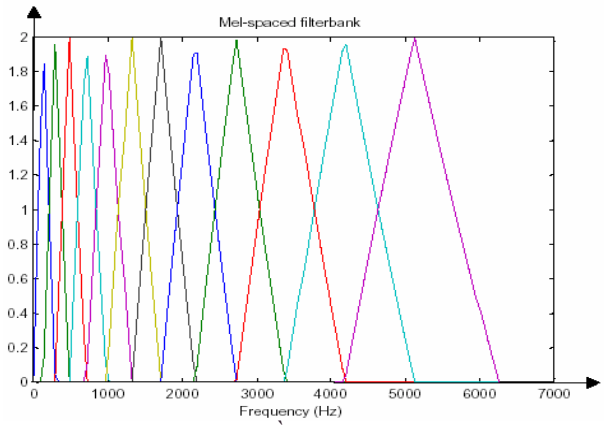
➢ Lọc xử lý

Những nghiên cứu về sinh lý học chứng tỏ rằng mức độ cảm nhận đối với tần số tín hiệu tiếng nói của con người không theo một tỉ lệ tuyến tính. Ứng với mỗi tone là có một tần số f, được đo bằng đơn vị Hz. Để mô tả chính xác sự tiếp nhận tần số của hệ thống thính giác, người ta đã xây dựng một thang khác – thang Mel. Thang tần số mel tuyến tính ở tần số dưới 1000 Hz và logarit ở tần số trên 1000 Hz. Một quan hệ ánh xạ tương ứng giữa thang tần số thực (vật lý, Hz) và thang tần số sinh lý Mel được cho bởi công thức sau:

$$F_{mel}=\frac{1000}{\log_{10}2}\left(1+\frac{F_{Hz}}{1000}\right)$$

hay $$F_{mel}=2595.\log_{10}\left(1+\frac{F_{Hz}}{1000}\right) \qquad \textbf{(2.3)}$$

Việc phân tích phổ sẽ thể hiện những đặc trưng tín hiệu tiếng nói mà do chính hình dạng của vùng phát âm tạo ra. Những đặc trưng phổ của tín hiệu tiếng nói sẽ có được sau khi cho qua những bộ lọc. Đối với thang tần số Mel thì một lọc cho mỗi thành phần tần số mong muốn (hình 7). Bộ lọc này có đáp ứng tần số dạng tam giác, và khoảng cách hay băng thông được xác định bởi một hằng số Mel.



Hình 7 *Một ví dụ về bộ lọc thang Mel*

➢ Tính năng lượng logarit (LOG)

Các bước trước đóng vai trò làm phẳng phổ, thực hiện một xử lý giống như tai của con người. Đến

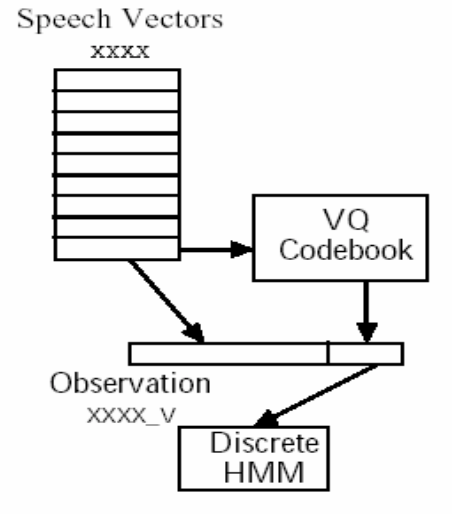
bước này tính toán logarit của bình phương độ lớn những hệ số tại ngõ ra bộ lọc. Chú ý rằng tai người thực hiện rất tốt việc xử lý độ lớn và logarit. Hơn thế nữa, xử lý độ lớn thì loại bỏ những thông tin không cần thiết trong khi xử lý logarit thực hiện một nén động, trích đặc trưng ít nhạy đối với những biến đổi động.

- Tính phổ tần số mel

Bước cuối cùng trong việc tính phổ tần số mel (MFCC) bao gồm thực hiện biến đổi ngược DFT trên độ lớn logarit của ngõ ra của bộ lọc.

Chú ý rằng do năng lượng phổ log là thực và đối xứng nên biến đổi DFT ngược được nói gọn là chuyển đổi cosine rời rạc (Discrete Cosine Transform – DCT). Tính chất của DCT là tạo ra những đặc trưng rất khác nhau. DCT cũng có tác dụng làm phẳng phổ nếu chỉ có những hệ số đầu tiên được giữ lại. Trong nhận dạng tiếng nói thì số hệ số MFCC thường nhỏ hơn 15. [6]

Sau khi tín hiệu tiếng nói được trích đặc trưng thì mỗi từ được được đặc trưng bởi một ma trận hệ số thực. Do mô hình HMM rời rạc được ứng dụng để nhận dạng nên những vector đặc trưng này phải được ước lượng vector (VQ) thành một chỉ số codebook rời rạc. Thuật toán phổ biến dùng để thiết kế codebook là LBG (Linde, Buzo và Gray).



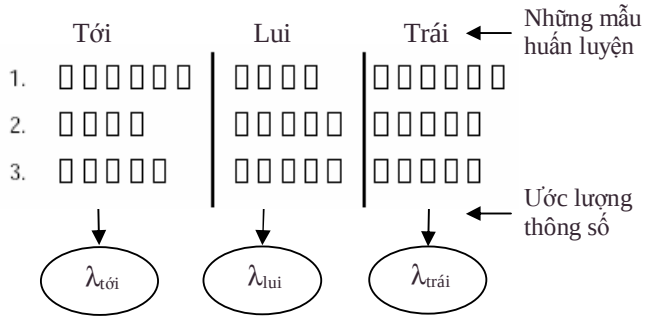
Hình 8 *Ước lượng vector VQ trong nhận dạng.*

Phương pháp được sử dụng để ước lượng vector là phương pháp K-means.

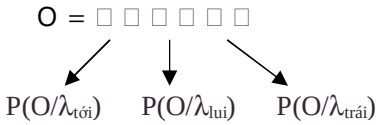
2.3 Thực hiện mô-đun 3

Sau khi đã thực hiện xong 2 mô-đun trên thì chúng ta đã có một cơ sở dữ liệu các vector đặc trưng ứng với từng từ. Trong mô đun này chúng ta sẽ xây dựng một mô hình Markov ẩn với dữ liệu huấn luyện là các vector đặc trưng có được từ mô-đun 2. Sơ đồ huấn luyện và nhận dạng bằng mô hình HMM được thể hiện trên hình 9 với bộ từ vựng gồm 3 từ: tới, lui, trái.

Huấn luyện:



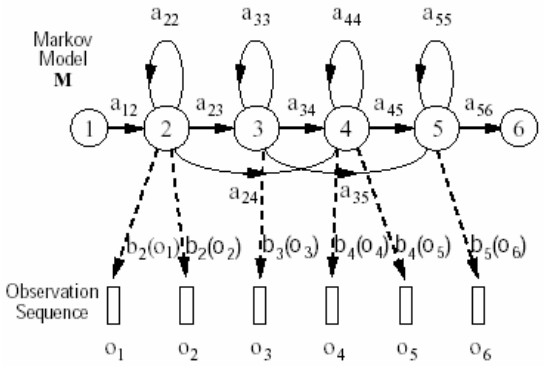
Nhận dạng:



Hình 9 *Sơ đồ mô hình HMM*

Ứng với mỗi từ cần nhận dạng thì chúng ta có một cơ sở dữ liệu các đặc trưng từ các lần đọc khác nhau (như trên sơ đồ là 3 lần lấy mẫu). Sau đó ta sẽ ước lượng các thông số của mô hình $\lambda = (A, B, \pi)$ để xác suất P(O|λ) đạt cực đại, tương ứng với mỗi từ là một λ xác định. Để nhận dạng một từ thì ta chỉ việc tính xác suất chuỗi quan sát của từ đó ứng với các λ đã được huấn luyện, và chọn mẫu nào có xác suất lớn nhất.

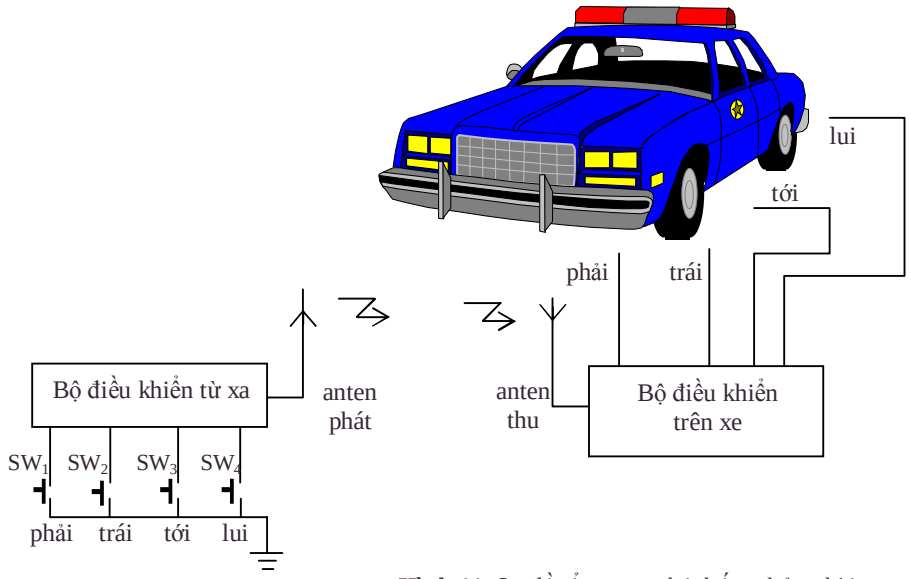
Dựa vào các tài liệu tham khảo và những thông tin về các hệ thống nhận dạng đã xây dựng thành công chúng tôi thấy rằng: đối với nhận dạng tín hiệu tiếng nói thì mô hình HMM thường được chọn là mô hình trái phải (left-right) có từ 5 đến 6 trạng thái. Qua quá trình thử nghiệm, mô hình có 6 trạng thái cho kết quả tốt hơn nên trong chương trình của mình, các tác giả đã xây dựng một HMM với số trạng thái là 6, xem hình 10.



Hình 10 *Mô hình HMM trái phải với 6 trạng thái.*

3 MÔ HÌNH HỆ THỐNG XE ĐIỀU KHIỂN

Sơ đồ mô hình xe vô tuyến điều khiển bằng tiếng nói từ máy tính được trình bày trên hình 11.



Hình 11 *Sơ đồ tổng quan hệ thống thử nghiệm*

Xe vô tuyến có thể được điều khiển từ xa bằng tiếng nói từ máy tính. Tiếng nói là từ lệnh sẽ được thu vào và nhận dạng trên bộ nhận dạng tiếng nói, và cấp chuỗi từ nhận dạng được cho bộ quyết định để xuất lệnh điều khiển thông qua cổng COM. Một mạch giao tiếp máy tính thông qua cổng nối tiếp (RS232) được thiết kế để điều khiển. Mạch giao tiếp nhận tín hiệu và đóng mở các khoá để chuyển thành tín hiệu của bộ điều khiển từ xa. Mỗi khi có một khoá được đóng hoặc một tổ hợp phím được nhấn, bộ điều khiển từ xa sẽ mã hóa thích hợp và đưa ra anten phát. Tín hiệu điều khiển được điều chế và truyền đến xe bằng sóng vô tuyến với tần số sóng mang F_C = 27MHz. Bộ điều khiển trên xe sẽ tiến hành điều khiển vận hành xe. Mô hình hoạt động tốt với bộ từ vựng gồm 4 từ: phải, trái, tới, lui với kết quả tốt (99%).

4 KẾT LUẬN

Mô hình thử nghiệm nhận dạng tiếng nói tiếng Việt theo hướng kết hợp MFCC và HMM tuy còn nhiều hạn chế nhưng đã đáp ứng được mục tiêu của đề tài. Chương trình được sử dụng để điều khiển robot với bộ từ vựng nhỏ (dưới 16 từ) cho độ chính xác có thể chấp nhận được (trên 90%). Trong thời gian tới nhóm tác giả sẽ tối ưu hóa chương trình nhận dạng để đạt được kết quả cao hơn và tăng tốc độ xử lý.

TÀI LIỆU THAM KHẢO

1. GS. Phạm Văn Ất , *Kỹ thuật lập trình C*, Nhà xuất bản Khoa Học và Kỹ Thuật, 1999.
2. Nguyễn Hoàng Hải – Nguyễn Khắc Kiểm, *Lập trình Matlab*, Nhà xuất bản Khoa Học và Kỹ Thuật, 2003.
3. PGS.TS. Nguyễn Hữu Phương, *Xử lý tín hiệu số*, Nhà xuất bản Giao thông vận tải, 2000.
4. Lê Tiến Thường, *Xử lý tín hiệu số và wavelets*, Nhà xuất bản Đại Học Quốc Gia TP. Hồ Chí Minh, 2002.
5. Claudio Becchetti and Lucio Prina Ricotti, *Speech Recognition Theory and C++ Implementation*, JOHN WILEY & SONS, LTD, 2000.
6. Gordon E.Pelton, *Voice Processing*, McGraw Hill, 1992.
7. John R.Deller & John G.Proakis & John H. L. Hansen, *Discrete – Time Processing of Speech Signals*, Macmillan Publishing Company, 1993.
8. F.J. Owens, *Signal Processing of Speech*, Macmillan, 1993.