ỨNG DỤNG MẠNG NƠ-RON NHÂN TẠO ĐỂ ĐIỀU KHIỂN THIẾT BỊ BẰNG GIỌNG NÓI TIẾNG VIỆT

Nguyễn Chí Ngôn[1], Trần Thanh Hùng[1]
Trương Thị Thanh Tuyền[2] và Nguyễn Thái Nghe[2]

ABSTRACT

This paper presents a neural networks-based method for a robot control system using Vietnamese voice commands. A STFT-based method for formant estimation is used to extract important features of recorded waveforms to generate the training data. A multi-layer feed-forward neural network is trained to recognize four words of any speakers, which are 'Trái', 'Phải', 'Tới', and 'Lui'. Testing our system to control a wireless car shows the stability, accuracy of approximately 90% and ability to extend the system.

Keywords: ***Artificial neural networks, Speech recognition, Pitch period, Formant detection, Micro-controller, Control system***

Title: A method of applying neural networks to control system by Vietnamese speech

TÓM TẮT

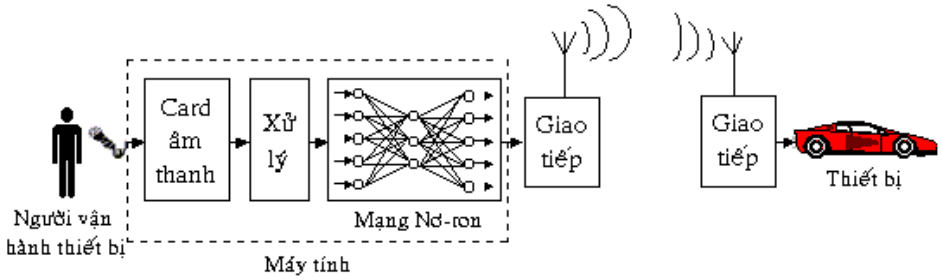
Bài báo đề cập đến một giải pháp ứng dụng mạng nơ-ron nhân tạo (Artificial neural networks) để điều khiển thiết bị bằng giọng nói tiếng Việt. Phép biến đổi Fourier thời gian ngắn - STFT (Short time Fourier Transform) được áp dụng để trích các đặc trưng cơ bản của tín hiệu tiếng nói. Một mạng nơ-ron nhân tạo được huấn luyện để nhận dạng giọng nói tiếng Việt của bất kỳ người nào, khi họ đọc một trong bốn từ lệnh 'Trái', 'Phải', 'Tới' và 'Lui' (áp dụng để điều khiển robot). Kết quả nghiên cứu được kiểm chứng thông qua việc điều khiển từ xa một xe vô tuyến. Độ chính xác được ước lượng xấp xỉ 90% và khả năng mở rộng tập lệnh điều khiển là rất cao.

Từ khóa: ***Mạng nơ-ron nhân tạo, nhận dạng tiếng nói, chu kì cao độ, trích các formant, vi điều khiển, hệ thống điều khiển***

1 GIỚI THIỆU

Ứng dụng nhận dạng tiếng nói để điều khiển thiết bị là một lĩnh vực thiết thực trong cuộc sống. Có nhiều phương pháp tiếp cận đến nhận dạng tiếng nói, song do tính phức tạp vốn có của mỗi ngôn ngữ và mỗi chất giọng của từng dân tộc, mà lĩnh vực này luôn là một thách thức to lớn đối với những người đam mê.

Bài viết này chúng tôi mong muốn tìm kiếm một giải pháp ứng dụng trí tuệ nhân tạo trong cuộc sống. Cụ thể là áp dụng mạng nơ-ron nhân tạo (artificial neural networks, gọi tắt là mạng nơ-ron) để nhận dạng một số từ cơ bản của tiếng Việt, đủ để điều khiển một mini-robot. Với mục tiêu mong muốn là bất kì người sử dụng nào cũng có thể vận hành tốt thiết bị bằng cách đọc các lệnh vào micro của máy tính.Mạng Nơ-ron sẽ nhận dạng từ điều khiển vừa đọc, và gởi đến mạch giao tiếp thiết bị byte điều khiển tương ứng (Hình 1).



Hình 1: Sơ đồ khối hệ thống

[1] Bộ môn Viễn thông và Kỹ thuật Điều khiển, Khoa Công nghệ Thông tin.
[2] Bộ môn Hệ thống thông tin và Toán ứng dụng, Khoa Công nghệ Thông tin.

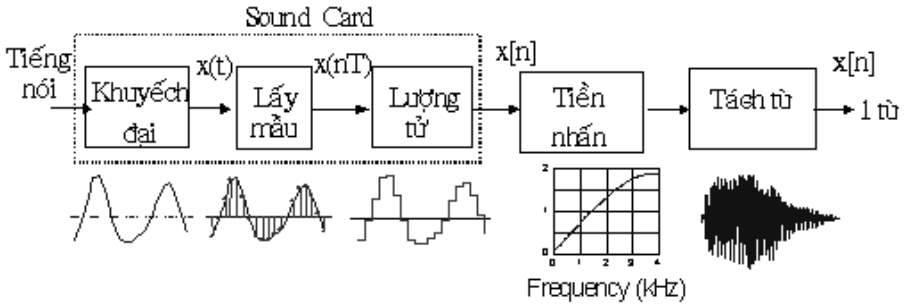
Chúng tôi tiến hành xây dựng hệ thống có thể điều khiển thiết bị bằng 4 từ: 'Tới', 'Lui', 'Trái' và 'Phải'. Kết quả thử nghiệm rất khả quan và cho thấy khả năng mở rộng tập lệnh điều khiển là rất cao (Nguyễn Chí Ngôn và Trịnh Hữu Phúc, 2002; Nguyen Chi Ngon, Tran Thanh Hung, Truong Thi Thanh Tuyen and Nguyen Thai Nghe 2005).

2 XÂY DỰNG CƠ SỞ DỮ LIỆU DÙNG CHO VIỆC HUẤN LUYỆN

Trước tiên, chúng tôi cần xây dựng một cơ sở dữ liệu dùng để huấn luyện mạng (gọi là tập mẫu). Tập mẫu này có được thông qua việc thu thập dữ liệu của nhiều giọng đọc khác nhau và xử lý để chỉ giữ lại những đặc trưng cơ bản của nó. Sau đó, quá trình huấn luyện mạng được thực hiện. Kết thúc quá trình này, mạng nơ-ron có thể phân loại các từ khác nhau, từ đó có thể nhận dạng được các từ đã học mà không cần đến không gian dữ liệu mẫu nữa. Tương ứng với 4 từ lệnh dùng để điều khiển robot, "tới", "lui", "trái" và "phải", chúng tôi đánh dấu các dữ liệu đặc trưng đã phân tích được thành 4 nhóm. Quá trình nhận dạng, thực chất là phân loại (classification) từ cần kiểm tra thuộc nhóm nào trong 4 nhóm dữ liệu trên (Ngôn *et al*, 2002; 2005).

2.1 Tiền xử lý dữ liệu

Sau khi ghi âm, tín hiệu tiếng nói cần được xử lý để hạn chế nhiễu. Đồng thời, một giải thuật tách từ được áp dụng để xác định thời điểm bắt đầu và kết thúc của tín hiệu (bởi vì thời gian cho phép soundcard ghi âm thường dài hơn tín hiệu thực tế). Hình 2 trình bày nguyên tắc tiền xử lý dữ liệu. Hình 3 minh họa dạng tín hiệu trước và sau khi xử lý.



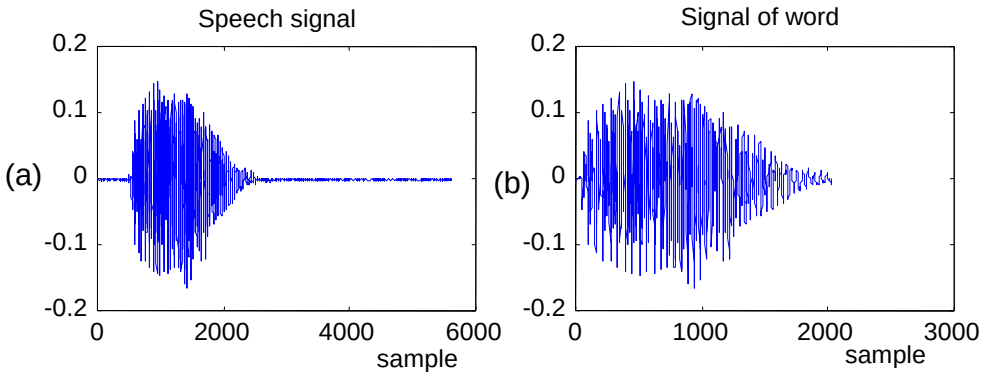
Hình 2: Nguyên tắc tiền xử lý dữ liệu tiếng nói

2.2 Trích đặc trưng tín hiệu tiếng nói

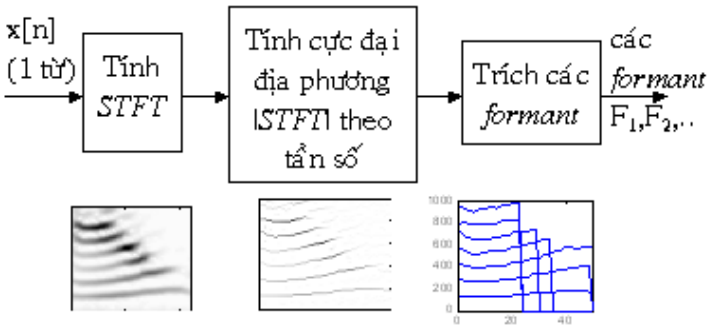
Tách được các đặc trưng cơ bản của tín hiệu tiếng nói có ý nghĩa rất quan trọng vì đó là cơ sở để nhận dạng. Các nghiên cứu cho thấy, hai thành phần đặc trưng quan trọng nhất, đó là *chu kì cao độ (pitch period)* và *các formant* (Tran Thanh Hung, Q.P. Ha, G. Dissanayake, 2004; 2005). *Chu kì cao độ* chính là dạng sóng một chu kì của phần gần tuần hoàn (ứng với nguyên âm) trong tiếng nói, do đó thường được xử lý ở miền thời gian. Trong khi đó các *formant* liên quan đến phổ tần số của tín hiệu. Đối với tiếng nói, các *formant* không cố định mà thay đổi chậm theo thời gian. Do đó chỉ có thể thu được các *formant* bằng cách phân tích và biểu diễn tín hiệu tiếng nói ở miền thời gian-tần số.

Qua thực nghiệm chúng tôi nhận thấy, với cùng một người, nếu người đó đọc các từ khác nhau thì *formant* tương ứng cũng khác nhau. Nếu nhiều người cùng đọc một từ, thì *formant* tương ứng có sự khác biệt không nhiều. Do đó, chúng tôi quyết định trích các *formant* này và dùng nó để làm dữ liệu huấn luyện mạng nơ-ron.

Hình 4 mô tả nguyên tắc trích *formant* của tín hiệu tiếng nói dùng phép biến đổi *STFT* (Short Time Fourier Transform).



Hình 3: Kết quả xử lý, (a) trước khi xử lý, (b) sau khi xử lý – gọi là 1 từ tín hiệu



Hình 4: Nguyên tắc trích formant của tín hiệu tiếng nói

Do tín hiệu tiếng nói là tín hiệu không dừng, nên không thể áp dụng phép phân tích Fourier thông thường. Song, nếu chúng ta chia tín hiệu tiếng nói ra thành từng đoạn đủ nhỏ theo thời gian, thì tín hiệu tiếng nói trong mỗi đoạn có thể xem là tín hiệu dừng, và do đó có thể lấy biến đổi Fourier trên từng đoạn tín hiệu này. Đây là nguyên lý của phép biến đổi Fourier thời gian ngắn, còn gọi là biến đổi Fourier cửa sổ hóa.

Trong STFT, tín hiệu cần phân tích f(t) đầu tiên được nhân với một hàm cửa sổ w(t-τ) để lấy được tín hiệu trong một khoảng thời gian ngắn xung quanh thời điểm τ. Sau đó phép biến đổi Fourier bình thường được tính trên đoạn tín hiệu này. Kết quả ta được một hàm theo tần số và thời gian STFT f(ω,τ) xác định bởi (dấu (*) ký hiệu cho thành phần liên hợp phức):

$$STFT_f(\omega,\tau) = \int_{-\infty}^{\infty} w^*(t-\tau) f(t) e^{-j\omega t} dt \qquad (1)$$

STFT tại thời điểm τ được xem là phổ cục bộ của f(t) xung quanh thời điểm τ, do cửa sổ tương đối ngắn làm triệt tiêu tín hiệu ngoài vùng lân cận. Vì vậy STFT có tính định vị theo thời gian. Cửa sổ phân tích càng hẹp thì sự định vị này càng tốt (còn được gọi là độ phân giải theo thời gian).

Để thấy rõ STFT cũng định vị trong miền tần số, ta có thể áp dụng định lý Parserval:

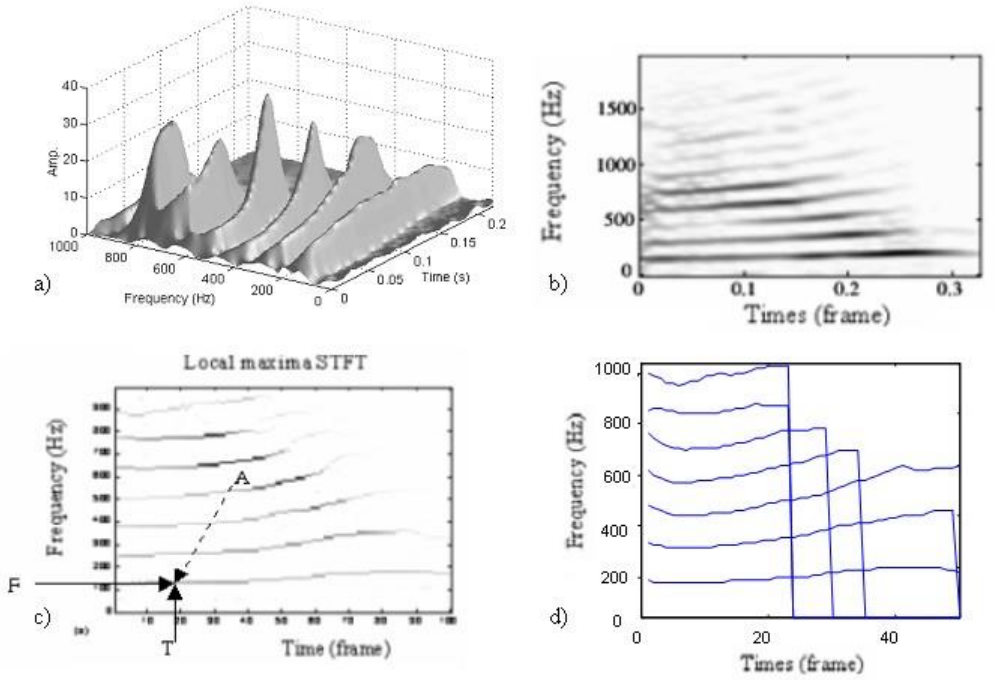
$$\int_{-\infty}^{\infty} f(t)\, g^*(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) G^*(\omega) d\omega \qquad (2)$$

Biểu thức (1) có thể viết lại như sau:

$$STFT_f(\omega,\tau) = \frac{e^{-\omega\tau}}{2\pi}\int_{-\infty}^{\infty} W^*(\omega'-\omega)F(\omega')e^{j\omega'\tau}d\omega' \quad (3)$$

với W(ω') và F(ω') lần lượt là phổ của cửa sổ w(t) và của tín hiệu f(t).

Trong biểu thức (3), W(ω'-ω) có tác dụng như một lọc dãy thông tập trung quanh tần số đang phân tích ω và có băng thông bằng với băng thông của w(t), làm giới hạn phổ của tín hiệu F(ω') xung quanh ω. Rõ ràng STFT có tính định vị theo tần số. Tính định vị này (còn gọi là độ phân giải tần số) càng tốt nếu băng thông của cửa sổ phân tích càng hẹp.



Hình 5: Kết quả trích formant bằng STFT của từ "Tới"
a) Spectralgram; b) Các cực đại của STFT – được xác định là các ridge của spectralgram;
c) Cực đại địa phương của STFT; d) trích các formant

Hàm cửa sổ thường dùng trong STFT là cửa sổ Kaiser, hàm này được định nghĩa từ hàm Bessel bậc 0.

$$w[n] = \begin{cases} \dfrac{I_0\left[\beta\sqrt{1-[(n-\alpha)/\alpha]^2}\right]}{I_0(\beta)}, & 0 \le n \le M \\ 0, & n \notin [0,M] \end{cases} \quad (4)$$

với α=M/2 và I0(β) là hàm cải biên của hàm Bessel bậc 0 (modified zero-order Bessel function), được định nghĩa là:

$$I_0(\beta) = \frac{1}{2\pi}\int_0^{2\pi} e^{\beta\cos\theta}d\theta \quad (5)$$

Hàm Kaiser có thể thay đổi linh hoạt nhờ vào thông số hình dạng (shape parameter) β. Với các giá trị β khác nhau, cửa sổ Kaiser sẽ có hình dạng khác nhau.

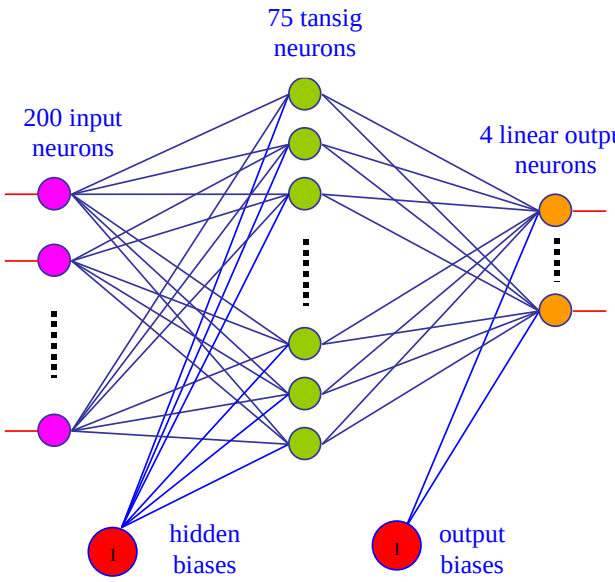
Kết quả phân trích đặc trưng của từ "Tới" được minh họa trên hình 5. Sau khi rời rạc hóa các formant tại 20 thời điểm trên trục thời gian và 10 vị trí trên trục tần số, đặc trưng của tiếng nói được qui về dạng 1 ma trận dữ liệu (10x20), tương ứng với 200 nút vào của

mạng nơ-ron (xem phần 3.1). Tập hợp tất cả các ma trận dữ liệu này, chính là tập mẫu dùng để huấn luyện mạng.

3 XÂY DỰNG MẠNG NƠ-RON

3.1 Cấu trúc mạng

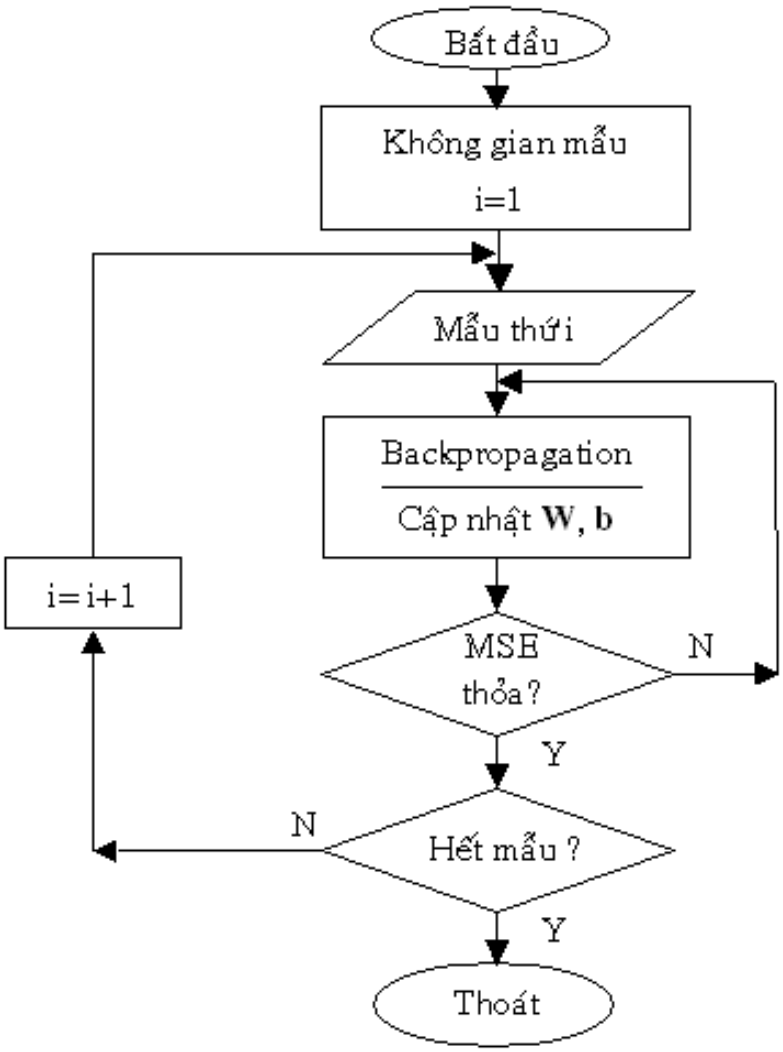
Việc xác định cấu trúc tối ưu cho một mạng nơ-ron tuỳ thuộc vào lĩnh vực ứng dụng của nó. Các công bố cho thấy rằng, một mạng truyền thẳng nhiều lớp (Multilayer Feed Forward Neural Networks), với lớp giữa phi tuyến và đủ lớn, có khả năng xấp xỉ một hàm phi tuyến bất kỳ (Rich, E. and Knight, K. 1991. Nguyễn Hoàng Phương, Bùi Công Cường, Nguyễn Doãn Phước, Phan Xuân Minh, Chu Văn Hỉ, 1998). Qua quá trình thử nghiệm, chúng tôi chọn được mạng Nơ-ron dùng nhận dạng là mạng truyền thẳng nhiều lớp với các thông số của mạng như sau: lớp vào (input layer) gồm 200 nút (tương ứng với 200 điểm đặc trưng của mỗi mẫu dữ liệu đã phân tích); lớp ẩn (hidden layer) gồm 75 nút, với hàm kích hoạt phi tuyến 'tansig' (được xác định bằng phương pháp leo đồi – Hill climbing method [Phương et al. 1998]); và lớp ra (output layer) gồm 4 nút, với hàm kích hoạt tuyến tính 'purelin' (Hình 6).



Hình 6: Cấu trúc mạng nơ-ron ứng dụng

3.2 Huấn luyện mạng

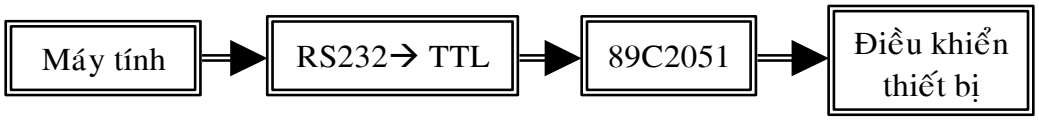
Nói chung giải thuật huấn luyện mạng nơ-ron tương đối phức tạp. Tuy nhiên, điều thuận lợi là phần mềm Matlab đã hỗ trợ rất nhiều công cụ. Chúng tôi đã áp dụng giải thuật huấn luyện Levenberg-Marquardt. Đây là một giải thuật có độ hội tụ bậc hai và là giải thuật nhanh nhất của Matlab (Demuth, H. and M. Beale, 2005). Quá trình huấn luyện mạng (lưu đồ tổng quát cho trên Hình 7) chúng tôi có kiểm tra bằng dữ liệu của 3 người không có giọng đọc trong tập mẫu để đánh giá khả năng 'nhớ'của mạng. Quá trình huấn luyện được thực hiện khoảng 4 giờ trên máy PC Celeron 1.8GHz, 256MB DRAM. Các giải thuật huấn luyện được trình bày chi tiết trong (Nelson, M. M. and Illingworth, W.T. 1991. Rich *et al.* 1991 and Demuth *et al.* 2005).



Hình 7: Lưu đồ thao tác huấn luyện

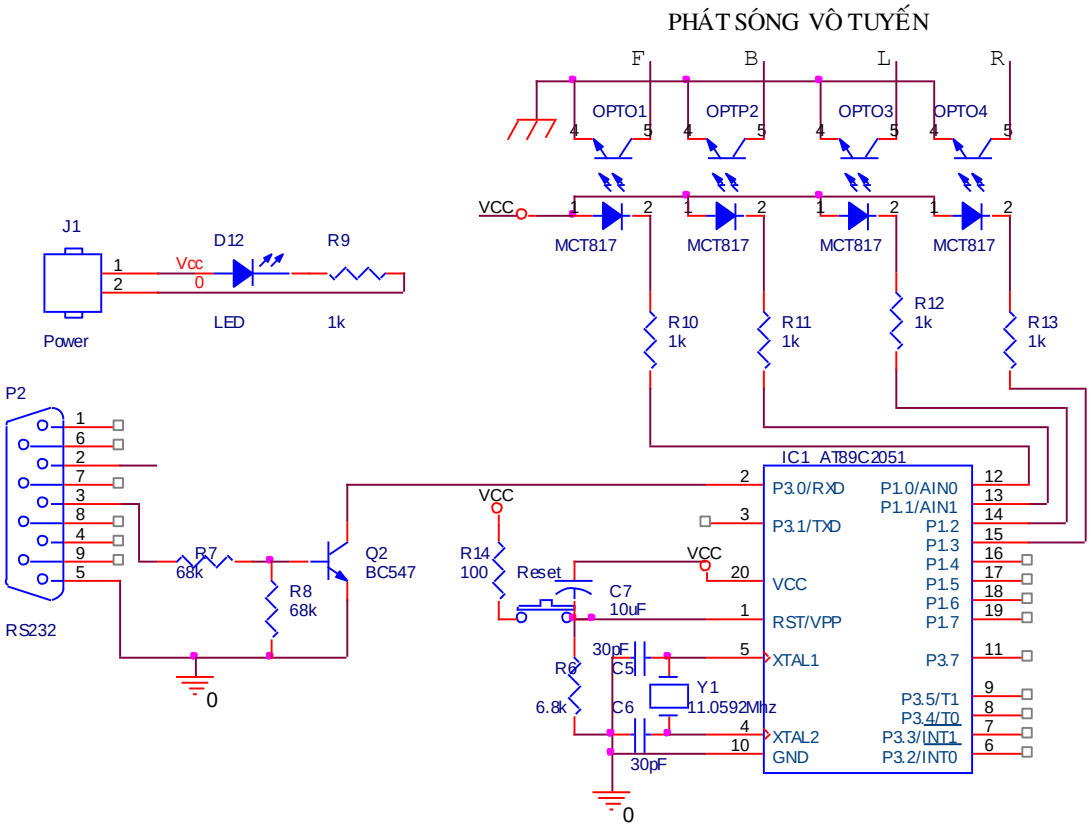
4 MẠCH GIAO TIẾP THIẾT BỊ

Thiết bị được điều khiển thông qua cổng truyền thông nối tiếp RS232. Mạch giao tiếp giữa thiết bị và máy tính được xây dựng dựa trên vi điều khiển 89C2051 (xem Hình 8 và Hình 9).



Hình 8: Nguyên tắc giao tiếp nối tiếp RS232

Hoạt động của thiết bị có thể tóm tắt đơn giản: Sau khi mạng nơ-ron nhận dạng được từ lệnh của người điều khiển (bằng giọng nói), một byte điều khiển được gửi ra cổng RS232; thông qua vi điều khiển 89C2051, byte điều khiển này được gửi tới một mạch phát sóng vô tuyến; thiết bị sẽ chấp hành mệnh lệnh khi bộ thu sóng của nó giải mã được từ điều khiển. Cơ chế này cho phép vận hành thiết bị từ xa thông qua việc đọc lệnh vào máy tính.



Hình 9: Giao tiếp giữa thiết bị và máy tính

5 KẾT QUẢ

Sau khi huấn luyện, mạng được áp dụng để điều khiển từ xa một xe vô tuyến. Một chương trình tự động ghi nhận kết quả nhận dạng (sau khi được xác nhận của người điều khiển), được áp dụng. Thống kê trên 1000 lần đọc các lệnh 'Tới', 'Lui', 'Trái','Phải' của nhiều người, chúng tôi ước lượng được độ chính xác như sau (Bảng 1 và Bảng 2):

- 95% đối với nhóm người có giọng đọc đã được sử dụng để huấn luyện mạng.
- 84% đối với nhóm người có giọng đọc chưa được sử dụng để huấn luyện mạng. Tuy nhiên, chúng ta có thể ghi âm giọng nói của những người này và huấn luyện tiếp để cải thiện tỉ lệ lỗi.

6 KẾT LUẬN

Hiện tại chúng tôi chỉ xây dựng mạng với số lượng từ nhận dạng còn ít (4 từ) dựa trên giọng đọc của 10 người (gồm 9 nam và 1 nữ, chất giọng miền tây nam bộ). Đây cũng là điểm hạn chế của đề tài, vì cơ sở dữ liệu không đủ tổng quát để áp dụng cho những miền khác của Việt nam. Hơn nữa, quá trình huấn luyện, mạng chỉ phân loại một lần tín hiệu đầu vào và chia ra thành 4 nhóm, đặc trưng cho 4 từ: 'Tới', 'Lui', 'Trái' và 'Phải'. Tuy nhiên, kết quả nghiên cứu cho thấy khả năng tăng số lượng từ nhận dạng là khả thi. Trong trường hợp đó, chúng tôi dự kiến sẽ phân lớp dữ liệu trước khi nhận dạng; sao cho mỗi lớp chứa các từ có đặc trưng gần giống nhau. Chẳng hạn, chúng tôi dùng một mạng tổng quát để phân biệt từ vừa đọc thuộc nhóm "thanh bằng " hay "thanh trắc", sau đó đưa vào mạng chuyên biệt để nhận dạng chính xác từ vừa đọc.

Ngoài ra chúng tôi thấy rằng, kiểm tra trên nhóm người có giọng đọc đã được sử dụng để huấn luyện, mạng nơ-ron làm việc khá hiệu quả. Vì thế, chúng tôi đề xuất 1 phương án

ứng dụng nghiên cứu này vào thực tế, để tạo nên 1 sản phẩm rất có ý nghĩa xã hội. Đó là, chế tạo xe lăn điều khiển bằng giọng nói, dành cho những người khuyết tật bị mất cả 2 tay và 2 chân. Trong trường hợp này, giải pháp huấn luyện là hết sức đơn giản, vì mạng nơ-ron chỉ cần nhận dạng chính giọng chủ của xe lăn mà thôi.

Bảng 1: Kết quả nhận dạng trên nhóm người có giọng nói đã dùng để huấn luyện

Từ lệnh	Số lần đọc	Số lần nhận dạng đúng	Tỉ lệ xấp xỉ
Tới	1000	961	96,10%
Lui	800	755	94,37%
Trái	1150	1037	90,17%
Phải	800	794	99,25%

Bảng 2: Kết quả nhận dạng trên nhóm người có giọng nói chưa dùng để huấn luyện

Từ lệnh	Số lần đọc	Số lần nhận dạng đúng	Tỉ lệ xấp xỉ
Tới	750	652	86,93%
Lui	600	493	82,17%
Trái	800	612	76,65%
Phải	750	681	90,08%

CẢM TẠ

Nghiên cứu này được thực hiện dưới sự hổ trợ của Đại học Cần thơ, trong phạm vi đề tài cấp trường, mang tên "Nghiên cứu ứng dụng mạng nơ-ron nhân tạo để điều khiển thiết bị bằng giọng nói tiếng Việt", thực hiện năm 2003-2004, của nhóm tác giả.

TÀI LIỆU THAM KHẢO

Demuth, H. and M. Beale, 2005. Neural Network Toolbox – User's Guide®. MathWorks, Inc.

Nelson, M. M. and W. T. Illingworth. 1991. A Practical Guide to Neural Nets. Addison-Wesley Publishing Company, ISBN 0-201-52376-0.

Nguyen Chi Ngon ,Tran Thanh Hung , Truong Thi Thanh Tuyen and Nguyen Thai Nghe 2005. A method of control system by Vietnamese speech using Neural Networks. In: Proceedings of Int. Conf. in Computer Science – RIVR'05, February, 21-24, 2005. Can Tho University, Vietnam, pp. 314-317.

Nguyễn Hoàng Phương, Bùi Công Cường, Nguyễn Doãn Phước, Phan Xuân Minh và Chu Văn Hỉ, 1998. Hệ mờ và ứng dụng. Hà nội, Nhà Xuất Bản Khoa Học Kỹ Thuật.

Nguyễn Chí Ngôn, Trịnh Hữu Phúc, 2002. Bước đầu nghiên cứu ứng dụng mạng nơ-ron để điều khiển thiết bị bằng tiếng nói. Trong: Tạp chí Automation Today. Hội KHCN tự động VN, 28:30-32.

Nguyễn Chí Ngôn và Dương Hoài Nghĩa, 2001. Điều khiển dùng mô hình nội mạng Neuron áp dụng vào robot SCARA. Trong: Tạp chí Phát triển KHCN, ĐHQG Tp. HCM, Vol. 4, 8&9:65-71.

Rich, E. and K. Knight. 1991. Artificial Intelligence. Mc-Graw-Hill Inc., 2nd edition, ISBN 0-07-100894-2.

Tran Thanh Hung, Q.P. Ha, G. Dissanayake, 2004. New wavelet-based pitch detection method for human-robot voice interface. accepted by the 2004 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2004, September 28 - October 2, Sendai International Center, Sendai, Japan).

Tran Thanh Hung, Q.P. Ha, G. Dissanayake, 2005. New A wavelet-and neural network -Based voice interface system for wheelchair control, accepted to the Int. J. of Intelligent Systems Technologies and Applications (IJISTA), Special Issue on Biorobotics and Biomechatronics in Australasia.