Tạp chí Khoa học - Đại học Cần Thơ
website: sj.ctu.edu.vn

ỨNG DỤNG MÔ HÌNH MARKOV ẨN ĐỂ NHẬN DẠNG TIẾNG NÓI TRÊN FPGA

Nguyễn Cao Quí[1]
[1] *Bộ môn Điện tử Viễn thông, Khoa Công nghệ, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 11/10/2011
Ngày chấp nhận: 25/03/2013

Title:
Speech recognition using hidden Markov model on FPGA

Từ khóa:
Nhận dạng, tiếng nói, mô hình Markov ẩn, FPGA

Keywords:
Speech recognition, hidden Markov model, FPGA

ABSTRACT

Hidden Markov Model (HMM) is a statistical model, well suited for pattern recognition: speech, image, handwriting,... HMM has widely used in the last several years because of the two reasons. First it can perform with high accuracy in a wide range of application, second the model structure can be changed easily to fit practical applications.

This paper focused on studying of HMM for speech recognition and installing it on FPGA. HMM has many parameters, so choosing appropriate parameters of the model for the FPGA is included in the project. The selection is very important and must balance between time and accuracy.

TÓM TẮT

Mô hình Markov ẩn (HMM) là một mô hình thống kê, thích hợp ứng dụng trong việc nhận dạng mẫu: tiếng nói, hình ảnh và chữ viết… HMM được ứng dụng rộng rãi trong những năm gần đây vi hai lý do. Thứ nhất, mô hình có độ chính xác cao trong nhiều ứng dụng; Thứ hai, cấu trúc mô hình có thể thay đổi dễ dàng cho phù hợp với từng ứng dụng cụ thể.

Bài báo này tập trung nghiên cứu mô hình Markov ẩn theo hướng ứng dụng nhận dạng tiếng nói và cài đặt mô hình này lên chip FPGA. HMM có nhiều tham số, vì vậy việc lựa chọn tham số sao cho tốt nhất cũng được thực hiện trong đề tài. Việc lựa chọn này rất quan trọng, nó phải đạt được sự cân bằng giữa tốc độ xử lý và độ chính xác.
Hệ thống nhận dạng này được cài đặt trên FPGA để nhận dạng các từ đơn, số lượng từ trong bộ từ vựng có thể thay đổi nhờ khả năng có thể huấn luyện của HMM.

Do hệ thống nhận dạng này được cài đặt trên FPGA nên nó chiếm khoảng không nhỏ, thích hợp ứng dụng trong giao tiếp người-máy, robot, điều khiển bằng tiếng nói hay hỗ trợ người khuyết tật…

1 GIỚI THIỆU

HMM được bắt đầu xây dựng và công bố từ những năm 1960, đây là mô hình toán học về thống kê. Nhiều năm sau đó (từ 1980), mô hình này được bắt đầu nghiên cứu để ứng dụng trong lĩnh vực nhận dạng. Do đạt được độ chính xác cao và có khả năng thay đổi cấu trúc dễ dàng nên mô hình này ngày càng được sử dụng rộng rãi trong nhiều lĩnh vực, đặc biệt là trong lĩnh vực nhận dạng tiếng nói.

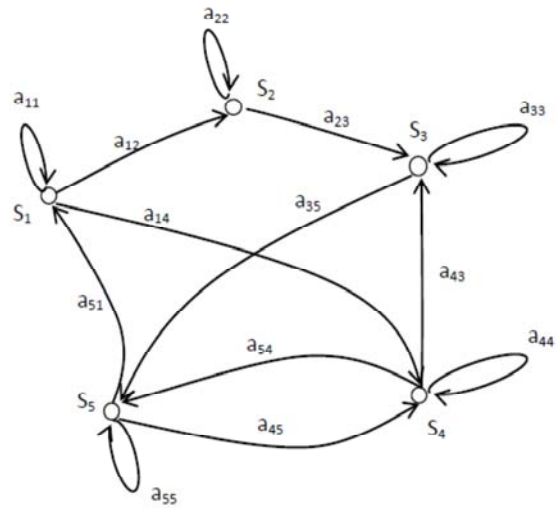
Năm 1952, phòng thí nghiệm Bell phát triển máy nhận dạng tiếng nói đơn với các từ vựng là các số. Hệ thống này dựa trên sự cộng hưởng phổ tần của các số nên có độ chính xác rất thấp. Đến những năm 1970, việc nghiên cứu máy nhận dạng tiếng nói đạt được một bước tiến đáng kể nhờ vào sự phát triển của lý thuyết nhận dạng mẫu và thuật toán tiên đoán tuyến tính LPC (linear predictive coding) để rút trích đặc trưng của tín hiệu tiếng nói. Từ năm 1980, các phương pháp thống kê bắt đầu được sử dụng mạnh mẽ trong kỹ thuật nhận dạng, đặc biệt là mô hình Markov do nó có độ chính xác cao.

Mục tiêu của đề tài này là tạo ra thiết bị nhận dạng tiếng nói nhỏ gọn nhưng có độ chính xác và đạt tốc độ cao. Vì vậy công việc chính trong đề tài này là nghiên cứu lý thuyết HMM và tập trung vào ứng dụng của HMM trong lĩnh vực nhận dạng tiếng nói, lựa chọn các thông số thích hợp của mô hình để có thể cài đặt máy nhận dạng lên một chip FPGA. Máy nhận dạng này được thử nghiệm với các từ nói đơn và thông qua quá trình thực nghiệm sẽ điều chỉnh lại các thông số của mô hình để đạt được độ chính xác cao nhất.

2 MÔ HÌNH MARKOV ẨN

Đây là một mô hình thống kê, thành phần của mô hình bao gồm tập N trạng thái $\{S_i\}$, các trạng dịch chuyển qua lại với nhau với một xác suất nhất định, tập xác suất di chuyển này được gọi là *ma trận dịch chuyển trạng thái* $A=[a_{ij}]$. Mô hình hoạt động khi cho chuỗi dữ liệu đầu vào $O=[o_1,o_2, o_t,..o_T]$ gọi là *chuỗi quan sát*, đây là dữ liệu trích rút từ tiếng nói cần nhận dạng trong ứng dụng nhận dạng tiếng nói. Mỗi quan sát o_t có một xác xuất xuất hiện trên mỗi trạng thái S_i, tập hợp các xác xuất này gọi là *phân phối xác xuất của quan sát* $B = \{b_j(o_t)\}_{j=1}^{N}$.

Ngoài ra còn có tập $\pi = \{\pi_i\}_{i=1}^{N}$ là xác suất quan sát đầu tiên o_1 tại trạng thái i. Tập $\lambda=\{S, A, B, \pi\}$ là các tham số của HMM. Khi có chuỗi quan sát được đưa vào mô hình, từ đầu ra ở các trạng thái sẽ rút ra được các tham số ẩn trong chuỗi quan sát (Juang and L.R. Rabiner, 1991). Hình 1 là một ví dụ HMM 5 trạng thái.



Hình 1: HMM 5 trạng thái

Hai vấn đề chính của HMM:

HMM có hai vấn đề chính cần phải giải quyết để nó có thể ứng dụng trong hệ thống nhận dạng (Jeff Bilmes, 2002):

– Vấn đề 1: Nhận dạng. Cho chuỗi quan sát $O=\{o_1,o_2,..., o_T)$ và một mô hình HMM λ. Tính xác suất $P(O|\lambda)$ của chuỗi O trên mô hình đó

– Vấn đề 2: Huấn luyện. Làm thế nào điều chỉnh các tham số của mô hình λ để $P(O|\lambda)$ cực đại, nghĩa là tối ưu hóa λ.

2.1 Giải quyết vấn đề 1

Để xác định xác suất chuỗi quan sát O trên một mô hình có sẵn λ, chúng ta dùng thuật toán hướng tới (forward algorithm).

1) Khởi tạo:

$$\alpha_t(i) = \pi_i b_i(O_1),\ 1 \le i \le N. \qquad (1)$$

2) Quy nạp:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}),\ 1 \le t \le T-1,\ 1 \le j \le N. \qquad (2)$$

3) Kết thúc:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i). \qquad (3)$$

Chúng ta cũng có thể dùng thuật toán lùi (back algorithm).

1) Khởi tạo:

$$\beta_T(i) = 1, \quad 1 \le i \le N. \tag{4}$$

2) Quy nạp:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad t = T-1, T-2, ..., 1, \ 1 \le i \le N. \tag{5}$$

3) Kết thúc:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \pi_1 \beta_1(i). \tag{6}$$

2.2 Giải quyết vấn đề số học

Trong bài báo này, tác giả sử dụng thuật toán hướng tới để xác định xác suất của một chuỗi quan sát. Nhưng các hệ số $\alpha_t(i)$ là các xác suất có giá trị không âm và nhỏ hơn hoặc bằng 1, nếu chuỗi quan sát lớn thì $P(0|\lambda)$ có thể rất nhỏ và vượt quá khả năng tính toán của hệ xử lý toán học. Vì vậy, tác giả sử dụng công cụ logarit cơ số 10 để xác định lại xác suất của chuỗi quan sát $P(0|\lambda)$ như sau:

$$P(O \mid \lambda) = \sum_{t=1}^{T} \log_{10}\left(\sum_{i=1}^{N} \alpha(i) \right) \tag{7}$$

2.3 Giải quyết vấn đề 2

Nội dung vấn đề 2 là thực hiện quá trình huấn luyện hệ thống để điều chỉnh mô hình λ sao cho đạt được các thông số tối ưu.

$$\bar{\pi}_i = \frac{\sum_{r=1}^{R} \alpha_1(i)\beta_1(i)}{\sum_{r=1}^{R}\sum_{i=1}^{N} \alpha_T} \tag{8}$$

$$\overline{a_{ij}} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{r=1}^{R}\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)} \tag{9}$$

$$\overline{b_j}(k) = \frac{\sum_{r=1}^{R}\sum_{t=1, O_t = v_k}^{T} \alpha_t(j)\beta_t(j)}{\sum_{r=1}^{R}\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)} \tag{10}$$

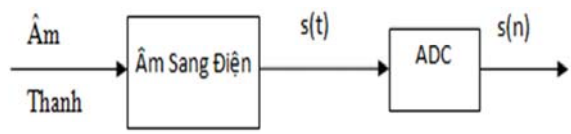
Với R: Số mẫu tiếng nói để luấn luyện của mỗi từ.

V={v1,v2,… vK} là codebook của tập chuỗi quan sát O dùng để huấn luyện.

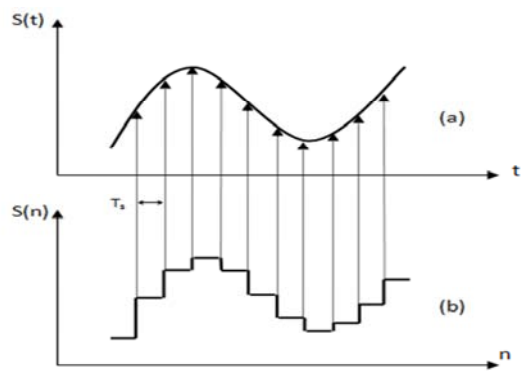
3 TIẾNG NÓI VÀ ĐẶC TRƯNG

3.1 Tiếng nói

Tiếng nói là một dạng sóng âm thanh dùng để giao tiếp của con người. Để có thể lưu trữ, xử lý, phân tích, nhận dạng với sự hỗ trợ của máy tính, tín hiệu tiếng nói cần phải được chuyển thành tín hiệu điện tương tự và qua bộ biến đổi ADC (Analog-to-Digital Converter) để chuyển tín hiệu tương tự thành tín hiệu số như Hình 2. Các dạng tín hiệu được miêu tả trong Hình 3.



Hình 2: Biến đổi tín hiệu âm thanh sang tín hiệu số

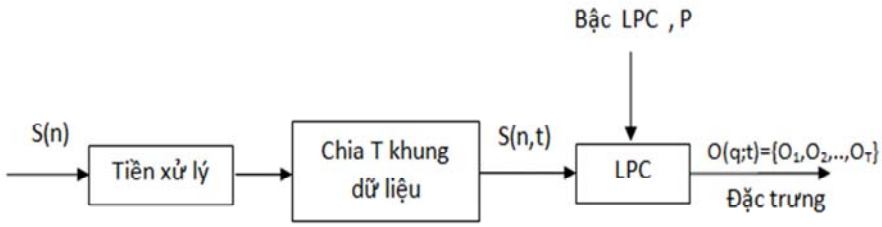


Hình 3: (a). Tín hiệu tương tự, (b). Tín hiệu số

3.2 Trích rút đặc trưng tiếng nói

Tiếng nói hay âm thanh trước khi được phân tích hoặc nhận dạng cần phải được rút trích các đặc trưng của nó. Bởi vì dữ liệu tiếng nói có

nhiều thông tin nên chỉ rút trích những thông tin cần thiết cần thiết cho việc nhận dạng. Trong nội dung bài báo nào, đặc trưng được rút trích là phổ tần rời rạc và các biến đổi tần số của tín hiệu hiệu tiếng nói. Có một số phương pháp để rút trích các đặc trưng này, tác giả đã chọn phương pháp LPC (L. R. Rabimer and R. W. Schafer, 1979) vì nó được kiểm nghiệm và đánh giá rất hiệu quả trong nhận dạng tiếng nói. Hình 4 mô tả quá trình rút trích đặc trưng của tín hiệu âm thanh hay tiếng nói sử dụng LPC.

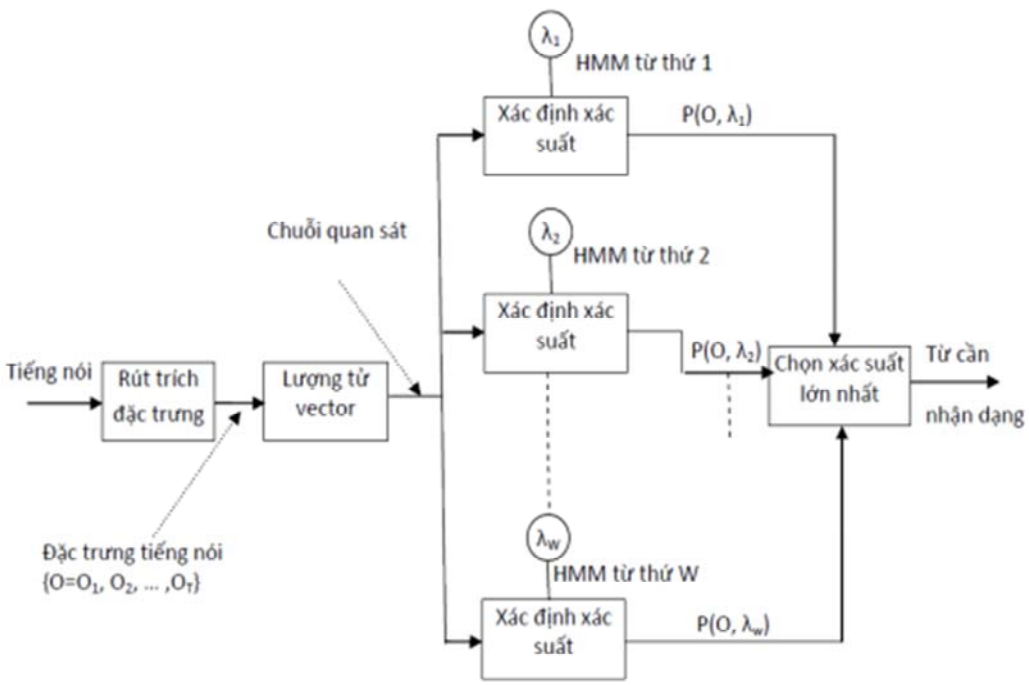


Hình 4: Rút trích đặc trưng của tiếng nói

Mẫu tiếng nói dưới dạng số S(n) trước khi rút trích được đưa qua khối Tiền xử lý để loại bỏ các nhiễu hoặc các tín hiệu tần số cao (vượt qua tần số âm thanh) bằng các mạch lọc số, sau đó sẽ được chia thành T khung dữ liệu. Cuối cùng tín hiệu qua khối LPC với bậc P để đạt được chuỗi đặc trưng $O=[o_1,o_2, o_t,..o_T]$ của mẫu tiếng nói, với mỗi thành phần o_i là một vector có P phần tử dưới dạng số thực.

4 MÔ HÌNH MÁY NHẬN DẠNG TIẾNG NÓI ĐƠN DỰA TRÊN HMM

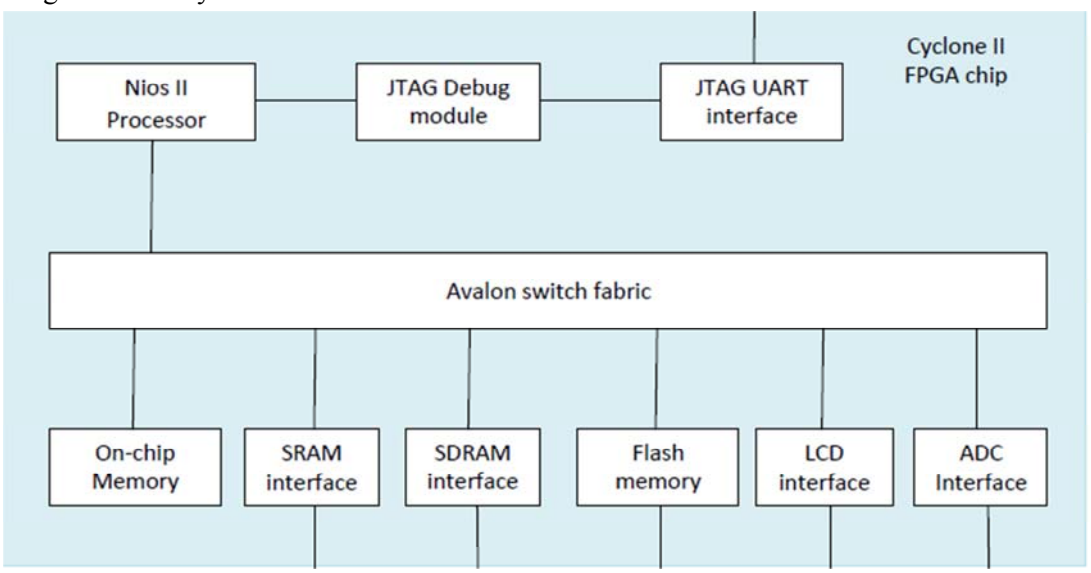
Vấn đề đặt ra là cần xây dựng một máy nhận dạng có thể nhận biết W từ đơn (Lawrence R. Rabiner, 1980). Để làm được điều này trước hết chúng ta phải xây hình tập mô hình cho mỗi từ thông qua quá trình huấn luyện (vấn đề 2). Tiếng nói cần nhận dạng sẽ được rút trích đặc trưng và đưa vào khối lượng tử để rời rạc đặc trưng tiếng nói thành chuỗi quan sát có các phần từ thuộc một codebook hữu hạn (B.H. Juang and L.R. Rabiner, 1991). Chuỗi quan sát sẽ được tính xác suất trên mô hình HMM của mỗi từ (vấn đề 1). Từ được nhận dạng là từ có xác suất cao nhất. Hệ thống được minh họa bằng Hình 5.



Hình 5: Mô hình máy nhận dạng tiếng nói HMM

5 PHƯƠNG PHÁP NGHIÊN CỨU

Mô hình máy nhận dạng tiếng nói cần một phần cứng để cài đặt lên và hoạt động. Phần cứng được lựa chọn trong đề tài này là một chip FPGA vì kỹ thuật này làm cho hệ thống nhỏ gọn với tốc độ xử lý cao và cấu trúc phần cứng bên trong có thể thay đổi khi điều chỉnh các thông số của máy nhận dạng (S J Melnikoff, S F Quigley & M J Russell, 2002). Công cụ hệ thống trên một chip SoPC (system on programmable chip) của Altera được sử dụng để tạo ra các tài nguyên cho hệ thống nhận dạng tiếng nói trên chip FPGA như Hình 6.



Hình 6: Hệ thống phần cứng dùng chíp FPGA Cylone II

Các thành phần chính trên chip FPGA gồm có bộ vi xử lý 32 bit Nios II; JTAG UART để giao tiếp với PC; Avalon switch fabric là thành phần có nhiệm vụ hỗ trợ Nios II giao tiếp với các thành phần khác; SRAM, SDRAM, Flash memory là bộ nhớ của hệ thống; ADC interface giao tiếp bộ biến đổi AD để đọc tín hiệu âm thanh vào hệ thống nhận dạng; LCD interface hiển thị kết quả nhận dạng ra LCD.

6 KẾT QUẢ VÀ THẢO LUẬN

Máy nhận dạng tiếng nói sử dụng mô hình HMM được cài đặt và chạy thử nghiệm trên board FPGA DE2. Các thông số của máy nhận dạng như sau:

Fs= 8kHz (mono) ; Tần số lấy mẫu của bộ biến đổi AD

N=5; Số trạng thái

K=16 ; Kích thước của codebook

D=24 ; Kích thước của vector quan sát

W=5; Số từ vựng: “Một”, “Hai”, ”Ba”, ”Bốn”, “Năm”

Huấn luyện bởi giọng nói của 2 người: 1 nam, 1 nữ. Thực hiện 20 mẫu (tiếng nói) cho một từ vựng.

Với các tham số trên, máy nhận dạng đạt độ chính xác trung bình 86% với giọng nói của người có tham gia huấn luyện. Thời gian nhận dạng cho mỗi từ là 1.9 giây. Độ chính xác của máy nhận dạng có thể tăng lên bằng cách tăng các tham số Fs, N, K, D nhưng thời gian nhận dạng cũng tăng theo.

Bảng 1-5 liệt kê kết quả nhận dạng cho mỗi từ, một từ 10 lần thực hiện với giọng nói của người có tham gia huấn luyện. Các giá trị được in đậm là giá trị có xác suất lớn nhất trong một hàng, nó tương ứng với từ được nhận dạng. Thông thường, các xác suất này là những giá trị từ 0 đến 1 nhưng để giải quyết vấn đề tính toán số học nên xác suất này được trình bày dưới dạng logarit cơ số 10 (7) nên ta thấy nó có giá trị âm.

Ví dụ Bảng 1, một người nói lần lượt 10 lần từ "một" trước máy nhận dạng. Hàng thứ nhất là kết quả nhận dạng cho lần nói thứ nhất, máy cho kết quả là các xác suất của từ vừa được nói tương ứng với các từ "một", "hai", "ba", "bốn","năm" lần lượt là: **-48.9618**, -130.581, -147.300, -147.300, -167.300. Vậy, từ mà máy nhận dạng được là từ "một" vì nó có xác suất lớn nhất.

Bảng 1: Xác suất ứng mỗi từ với 10 lần nhận dạng từ "Một"

"Một"	"Hai"	"Ba"	"Bốn"	"Năm"
-48.9618	-130.581	-147.300	-147.300	-167.300
-75.172	-148.501	-165.300	-162.633	-162.408
-34.066	-144.988	-147.300	-147.300	-142.707
-19.941	-143.048	-147.300	-142.173	-119.375
-81.724	-148.161	-161.999	-165.300	-165.300
-171.461	-179.048	-183.300	-178.173	-170.253
-51.609	-147.300	-147.300	-142.433	-137.036
-129.300	**-115.587**	-129.300	-129.300	-129.300
-14.322	-156.407	-165.300	-154.495	-131.899
-17.577	-18.290	-17.999	-21.300	-21.300

Bảng 2: Xác suất ứng mỗi từ với 10 lần nhận dạng từ "Hai"

"Một"	"Hai"	"Ba"	"Bốn"	"Năm"
-147.300	**-12.226**	-141.864	-147.300	-147.300
-145.435	**-25.673**	-135.916	-147.300	-144.619
-158.740	-165.300	-165.300	-165.300	**-157.328**
-147.300	**-37.612**	-135.916	-147.300	-147.300
-147.300	**-13.337**	-127.696	-147.300	-147.300
-183.300	**-70.353**	-169.815	-183.300	-183.300
-165.300	**-61.540**	-153.916	-165.300	-165.300
-255.300	**-132.712**	-243.916	-255.300	-255.300
-129.300	**-75.567**	-124.172	-129.300	-121.649
-255.300	**-217.115**	-243.916	-255.300	-255.300

Bảng 3: Xác suất ứng mỗi từ với 10 lần nhận dạng từ "Ba"

"Một"	"Hai"	"Ba"	"Bốn"	"Năm"
-182.974	-183.30	**-11.118**	-127.257	-130.272
-119.766	-175.532	**-8.772**	-47.582	-59.010
-150.492	-165.300	**-20.744**	-136.880	-138.176
-178.003	-183.300	**-19.570**	-67.580	-73.183
-146.974	-147.300	**-9.596**	-64.152	-68.762
-307.929	-309.300	**-248.855**	-301.255	-302.551
-105.721	-152.449	**-8.276**	-29.010	-40.542
-160.003	-165.300	**-8.8479**	-55.622	-62.108
-142.938	-147.300	-145.594	**-27.723**	-35.822
-124.938	-129.300	-127.594	**-30.478**	-37.866

Bảng 4: Xác suất ứng mỗi từ với 10 lần nhận dạng từ "Bốn"

"Một"	"Hai"	"Ba"	"Bốn"	"Năm"
-110.149	-152.449	-151.378	**-40.097**	-69.191
-32.066	-180.614	-169.880	**-24.280**	-70.703
-142.007	-147.300	-134.858	-128.226	**-69.061**
-23.840	-147.300	-134.431	-128.332	-58.487
-24.610	-147.300	-127.324	**-17.598**	-25.876
-21.476	-129.300	-115.206	**-10.350**	-49.348
-26.724	-165.300	-152.345	**-24.958**	-65.571
-160.007	-165.300	-146.978	**-51.468**	-70.041
-83.043	-157.532	-151.493	**-56.460**	-88.017
-194.195	-247.532	-241.493	-171.842	**-156.676**

Bảng 5: Xác suất ứng mỗi từ với 10 lần nhận dạng từ "Năm"

"Một"	"Hai"	"Ba"	"Bốn"	"Năm"
-129.300	-129.300	-113.258	-50.510	**-22.537**
-133.471	-183.300	-164.512	-164.488	**-71.715**
-176.862	-183.300	-169.712	-164.614	**-21.520**
-174.202	-180.614	-180.207	-172.414	**-28.575**
-122.862	-129.300	-110.595	-19.751	**-19.152**
-160.007	-165.300	-147.724	-140.776	**-18.269**
-142.007	-147.300	-129.724	-117.076	**-17.485**
-25.249	-180.614	-169.880	-159.546	**-15.750**
-129.300	-129.300	-112.459	-105.717	**-18.299**
-151.0407	-157.532	-146.360	-141.412	**-112.750**

7 KẾT LUẬN VÀ ĐỀ XUẤT

Mô hình Markov ẩn đã chứng tỏ rất thích hợp trong nhận dạng mẫu, đặc biệt là nhận dạng tiếng nói. FPGA là một kỹ thuật hiệu quả để tạo khai phần cứng cho một hệ thống thông minh. Sự kết hợp mô hình Markov và FPGA sẽ tạo ra một hệ thống nhận dạng có độ chính xác cao, nhỏ gọn và dễ dàng thay đổi cấu trúc của hệ thống.

Tuy nhiên, hệ thống trên cần phải được phát triển thêm:

– Khi một từ cần nhận dạng được đưa vào hệ thống, từ này sẽ được tính xác suất trên mô hình của mỗi từ trong bộ từ vựng. Việc tính xác suất này được tiến hành theo tuần tự làm cho tăng thời gian nhận dạng. Để khắc phục chúng ta có thể sử dụng kỹ thuật song song trong FPGA sẽ rút ngắn thời gian nhận dạng W lần.

– Máy nhận dạng chỉ được huấn luyện bởi giọng 2 người nên độ chính xác sẽ thấp khi thực hiện với giọng nói khác. Để giải quyết vấn đề

này chúng ta có thể tiếng hành huấn luyện với số lượng lớn các giọng nói khác nhau trong các môi trường khác nhau.

LỜI CẢM TẠ

Lúc bắt đầu nghiên cứu đề tài này, tôi gặp nhiều khó khăn vì có nhiều vấn đề cần phải giải quyết. Nhờ sự giúp đỡ của thầy Nguyễn Hữu Phương – Đại học Khoa Học Tự Nhiên và sự hỗ trợ về thiết bị, tinh thần của đồng nghiệp trong khoa Công nghệ - Đại học Cần Thơ để tôi có thể từng bước khắc phục khó khăn. Xin được cảm ơn thầy Phương, cảm ơn các đồng nghiệp đã có nhiều đóng góp để tôi có thể hoàn thành đề tài này.

TÀI LIỆU THAM KHẢO

1. B.H. Juang and L.R. Rabiner, 1991, Hidden Markov Models for Speech Recognition, Technometrics, Vol.33, NO.3.
2. Jeff Bilmes, 2002, What HMMs Can Do, Dept of EE, University of Washington.
3. Joseph W. Picone, 1993, Signal Modeling Techniques in Speech Recognition, IEEE.
4. Lawrence R. Rabiner, 1980, A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition, Proceedings of the IEEE, Vol.77, No.2.
5. L. R. Rabimer and R. W. Schafer, 1979, Digital Processing of Speech Signals, Prentice - Hall Inc.
6. S J Melnikoff, S F Quigley & M J Russell, 2002, Implementing a Simple Continuous Speech Recognition System on an FPGA, IEEE.