KỸ THUẬT NHẬN DẠNG GIỌNG NÓI SỬ DỤNG MÔ HÌNH MARKOV ẨN

SPEECH'S REGCONITION USING MARKOV'S MODEL

Nguyễn Thế Xuân Long[1], Mai Lam[2], Dương Quốc Hoàng Tú[3]

[123]Trường Cao đẳng Công nghệ thông tin – Đại học Đà Nẵng;

Email: long.ntx@gmail.com, mlam.udn@gmail.com, citdqhtu@gmail.com

Tóm tắt - Bài toán nhận dạng giọng nói đã và đang được nhiều nhà nghiên cứu quan tâm và có nhiều phương pháp được đề xuất để giải quyết bài toán này. Tuy nhiên cho đến nay kết quả mang lại vẫn chưa làm hài lòng các nhà nghiên cứu do tính chất phức tạp và không cố định của đối tượng nhận dạng là tiếng nói con người. Đặc biệt với tiếng Việt thì kết quả còn nhiều hạn chế. Bài báo trình bày một hướng nhận dạng tiếng nói tiếng Việt dựa trên cơ sở của phương pháp nhận dạng mẫu dựa theo mô hình Markov ẩn (HMM).

Từ khóa -nhận dạng giọng nói; nhận dạng tiếng Việt, cải thiện chất lượng nhận dạng giọng nói; mô hình Markov ẩn; phương pháp nhận dạng giọng nói.

Abstract-Nowadays, speech recognition is familiar and have been interested by many scientists; there are many methods, directions proposed to solve this problem. However, the results of those researchs have not yet satisfied the scientists due to complexity of human voices; especially in Vietnamese's voices. In this article, we will present one direction to recognize Vietnamese's voice base on Markov's model (HMM).

Key words -speech recognition; Vietnamese's voice recognition; improve quality of speech recognition; hidden Markov's model; speech recognition's methods

1. Đặt vấn đề

Cùng với sự phát triển của ngành công nghệ thông tin, các hệ thống tự động đã dần thay thế các công đoạn của công việc. Nhận dạng tiếng nói là một kỹ thuật có thể được ứng dụng trong rất nhiều lĩnh vực. Ở Việt Nam, từ những năm 90 đã có rất nhiều bài báo đề cập vấn đề xử lý nhận dạng tiếng Việt. Tuy nhiên, các kết quả này vẫn còn nhiều hạn chế, đó là do sự khác biệt về ngôn ngữ văn bản, văn phạm câu, cấu trúc âm vị, cách phát âm và ngôn điệu… Đó là chưa nói đến chúng ta không có sẵn một cơ sở dữ liệu tiếng Việt đủ phong phú để thực nghiệm.

Đã có rất nhiều mô hình được đề xuất để thực hiện như : mô hình Bayes, Maximum Likelihood Estimation (MLE), mô hình hỗn hợp phân bố Guass (Gausse Markov Model), Gausse Classifier (GC)... Trong khuôn khổ bài báo này, chúng tôi trình bày một thử nghiệm áp dụng mô hình Markov ẩn (Hidden Markov Model-HMM)trong việc nhận diện giọng nói.

Phần còn lại của bái báo được cấu trúc như sau. Phần 2 nếu một số kiến thức cơ bản về mô hình Markov ẩn được dùng trong các thử nghiệm. Phần 3 trình bày mô hình thử nghiệm nhận dạng âm thanh tiếng Việt dựa trên mô hình Markov ẩn. Các kết quả được trình bày tiếp sau. Phần cuối là kết luận.

2. Cơ sở lý thuyết

2.1. Mô hình Markov ẩn

Mô hình Markov ẩn (HMM) là mô hình thống kê trong đó hệ thống được mô hình hóa được cho là một quá trình Markov với các tham số không biết trước và nhiệm vụ là xác định các tham số ẩn từ các tham số quan sát được, dựa trên sự thừa nhận này. Các tham số của mô hình được rút ra sau đó có thể sử dụng để thực hiện các phân tích kế tiếp, ví dụ cho các ứng dụng nhận dạng mẫu.

HMM là một tiến trình ngẫu nhiên kép, bao gồm một tiến trình ẩn chuyển trạng thái theo chuỗi Markov rời rạc và thuần nhất, xen kẽ với một tiến trình phát sinh dãy quan sát. Các ký hiệu được sử dụng trong mô hình Markov ẩn là:

N: số trạng thái trong mô hình

M: số ký hiệu quan sát có thể

T: độ dài của dãy quan sát (hay số ký hiệu trong dãy quan sát)

{1, 2, .., N}: tập các trạng thái

q_t: trạng thái của mô hình tại thời điểm t

V = {v_1, v_2, …, v_M}: tập rời rạc các ký hiệu quan sát

Π = {π_1, π_2, …, π_N}: tập các phân bố xác suất cho trạng thái khởi đầu, π_i là xác suất để trạng thái i được chọn tại thời điểm khởi đầu t = 1; $\pi_i = P(q_1 = i)$;

$$\begin{cases} \sum_{i=1}^{N} \pi_i = 1 \\ \pi_i \geq 0; i = 1, 2, \dots, N \end{cases}$$

A = {a_{ij}}: ma trận xác suất chuyển với a_{ij} là xác suất để trạng thái j xuất hiện tại thời điểm t+1 khi trạng thái i đã xuất hiện tại thời điểm t. Giả thuyết rằng a_{ij} là độc lập với thời gian t: $a_{ij} = P(q_{t+1} = j/q_t = i)$,

$$\begin{cases} \sum_{j-1}^{N} a_{ij} = 1; i = 1,2, \dots, N \\ a_{ij} \geq 0; j = 1,2, \dots, N \end{cases}$$

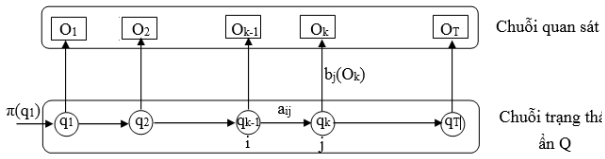
$B = \{b_j(v_k)\}$*:* các hàm đo xác suất phát xạ mẫu, $b_j(v_k) = P(v_k$ được phát sinh khi mô hình ở trạng thái j)

$$\begin{cases} \sum_{k=1}^{M} b_j(v_k) = 1; j = 1,2, \dots, N \\ b_j(v_k) \geq 0; j = 1,2, \dots, N; k = 1,2, \dots, M \end{cases}$$

$O_t$biểu thị ký hiệu quan sát tại thời điểm t.

Bộ ba $\lambda = (A, B, \pi)$ được coi là ký pháp gọn của một mô hình Markov ẩn. A, B và π được gọi là bộ tham số (parameters) của mô hình λ. Hoạt động của HMM có thể được mô tả như sau: tại thời điểm t = 1, mô hình ở trạng thái q1 nào đó và phát sinh ra một ký hiệu quan sát nhất

định O1, sau đó, tại thời điểm t = 2, mô hình chuyển sang trạng thái q2 và phát sinh ký hiệu quan sát O2. Cứ tiếp tục như vậy cho đến thời điểm t = T, mô hình phát sinh được dãy quan sát O = (O1, O2, …, OT) bằng dãy trạng thái Q = (q1, q2, …, qT). Dãy trạng thái Q phụ thuộc vào xác suất chọn trạng thái khởi đầu πi và xác suất chuyển aij. Dãy ký hiệu quan sát {Ot} được HMM phát sinh ra phụ thuộc vào dãy trạng thái Q và các hàm đo xác suất phát xạ mẫu bj(.). Trong trường hợp tập V các ký hiệu quan sát là không gian mẫu không đếm được, các hàm bj(.) có thể cho bằng hàm mật độ của một phân phối xác suất nào đó.



Hình 1. *Mô hình Markov ẩn*

2.2. Huấn luyện mô hình Markov ẩn

Bài toán: Với dãy huấn luyện O cần hiệu chỉnh các tham số của mô hình λ để cực đại hóa P(O/λ). Ta có:

$$P(O,Q/\lambda) = \pi_{q1}.b_{q1}(O_1).a_{q1q2}.b_{q2}(O_2).a_{q2q3}\dots a_{qT-1qT}.b_{qT}(O_T)$$

Và

$$P(Q/\lambda) = \sum_{Q} P(O,Q/\lambda) = \sum_{Q} \pi_{q1}b_{q1}(O_1)a_{q1q2}b_{q2}(O_2)\dots a_{qT-1qT}(O_T)$$

Đặt $\alpha_t(i) = P(O_1, O_2, \dots, O_t, q_t = i/\lambda)$ và $\beta_t(i) = P(O_{t+1}, O_{t+2}, \dots, O_T/q_t = i, \lambda)$, 1 ≤ t ≤ T với giá trị khởi tạo $\alpha_1(i) = \pi_i b_i(O_1)$ và $\beta_T(i) = 1, 1 \le i \le N$

Định nghĩa công thức truy hồi$a_{t+1}(j)$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} a_t(i)a_{ij}\right] b_j(O_{t+1}) \text{ với } t = 1,2,\dots,T-1$$

Tương tự, định nghĩa công thức $\beta_t(i)$ cho tính toán ngược như sau:

$$\beta_t(i) = \left[\sum_{j=1}^{N} a_{ij}b_j(O_{t+1})\right]\beta_{t+1}(j) \text{ với } t = T-1, T-2, \dots, 1$$

Thuật toán tiến lùi Baum-Welch (Forward-Backward Baum-Welch algorithm):

Bước 1. Xác định:

$$\gamma_t(i) = P(q_t = i/O,\lambda) = \frac{P(q_t = i, O|\lambda)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)}$$

Bước 2. Xác định:

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j/O,\lambda) = \frac{P(q_t = i, q_{t+1} = j, O|\lambda)}{P(O|\lambda)} = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)}$$

Bước 3. Chỉnh tham số:

$$\overline{\pi_i} = \gamma_1(i);\ \overline{a_{ij}} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)};\ \overline{b_j}(v_k) = \frac{\sum_{t=1,O_t=v_k}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)}$$

Bước 4. Nếu $P(O/\lambda_{mới}) \le P(O/\lambda_{cũ})$ thì kết thúc. Ngược lại quay lại bước 1.

2.3. Nhận dạng mô hình Markov ẩn

Bài toán: Cho mô hình $\lambda = (A, B, \pi)$ và một dãy quan sát $O = (O_1, O_2, \dots, O_T)$. Cần tìm dãy trạng thái $Q = (q_1, q_2, \dots, q_T)$ để xác suất $P(O, Q/\lambda)$.

Thuật toán Viterbi:

Bước 1. Gọi

$$f(k,j) = \max_{\{q_t\}_{t=1}^{k}, q_k = j} P(O_1, O_2, \dots, O_k, q_1, q_2, \dots, q_k|\lambda)$$

Bước 2. Khởi tạo cơ sở quy hoạch động: $f(1,j) = \pi_j b_j(O_1)$.

Bước 3. Tính bảng phương án f bằng công thức truy hồi:

$$f(k,j) = \max_{1\le i\le N}(f(k-1,i).a_{ij}.b_j(O_k))$$

Lưu vết:

$$Trace(k,j) = arg\max_{1\le i\le N}\left(f(k-1,i).a_{ij}.b_j(O_k)\right), (k \ge 2)$$

Bước 4. Tìm dãy trạng thái tối ưu: $q_T = arg\max_j f(T,j)$

$$q_t = Trace(t+1, q_t + 1), t = T-1, T-2, \dots, 1$$

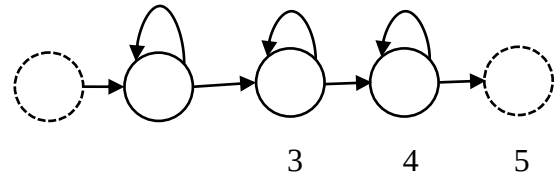
3. Thử nghiệm nhận dạng âm thanh Tiếng việt

3.1. Môi trường thực nghiệm

Cơ sở dữ liệu dùng cho thực nghiệm bao gồm 120 câu. Các câu được thu âm trong môi trường kín, do một giọng nam đọc, sử dụng micro tiêu chuẩn gắn với máy tính, card âm thanh sử dụng Sound Blaster 5.1, tốc độ lấy mẫu 8000Hz, PCM 8 bit môn 8kBps. Thử nghiệm dùng bộ thư viện của Trung tâm nghiên cứu nhận dạng tiếng nói thuộc Viện sau đại học Oregon Hoa Kỳ phát triển để xây dựng hệ thống nhận dạng dựa mô hình Markov, cũng như kết hợp mạng nơ-ron với mô hình Markov.

3.2. Thử nghiệm với mô hình Markov ẩn

Mô hình Markov được xây dựng dựa trên bộ thư viện CSLU Toolkit bao gồm 5 trạng thái như hình 2. Trong đó có ba trạng thái quan sát, một trạng thái khởi đầu và 1 trạng thái kết thúc.



Hình 2. *Mô hình Markov ẩn dùng trong thử nghiệm*

Ma trận xác suất (5x5) chuyển trạng thái trong mô hình được khởi tạo như sau:

0.0 1.0 0.0 0.0 0.0

0.0 0.6 0.4 0.0 0.0

0.0 0.0 0.5 0.5 0.0

0.0 0.0 0.0 0.6 0.4

0.0 0.0 0.0 0.0 0.0

Các quan sát O_j chính là vector đặc tính gồm 30 thành phần của từng khung tín hiệu. Với mỗi khung tín hiệu 10ms, tính hệ số cepstral MEL cùng với đạo hàm bậc một, bậc hai của từng hệ số và giá trị của từng hệ số trừ giá trị trung bình. Mô hình HMM monophone độc lập được áp dụng cho từng đơn vị nhận dạng. Khởi tạo mô hình sử dụng phương pháp lượng tử hóa vector (VQ). Mô hình được huấn luyện dựa trên thuật toán EM (expectation/maximization). Trong huấn luyện, mô hình nhúng dùng để kết hợp các mô hình độc lập nhằm đánh giá lại các tham số dựa trên thuật toán lùi Baum-Welch đã được trình bày ở phần 2.2. Mô hình được huấn luyện bằng 120 câu được gán nhãn bằng tay. Sau khi huấn luyện, sử dụng mô hình để nhận dạng trên một tập tử gồm 50 câu được chọn ngẫu nhiên từ cơ sở dữ liệu, các câu dùng để kiểm tra này khác với câu được dùng trong huấn luyện để đảm bảo khách quan. Sau đây là một số kết quả nhận dạng dùng mô hình Markov ẩn. Độ chính xác được chia thành hai mức từ và mức câu.

***Bảng 1**. Độ chính xác của mô hình Markov ẩn*

Số câu dùng để huấn luyện	Độ chính xác	
	Từ	Câu
120	86%	51%

Mô hình Markov ẩn HMM đã được ứng dụng thành công trong các hệ thống nhận dạng tiếng nói. Điểm mạnh của HMM là rất phù hợp cho việc biểu diễn một chuỗi đơn vị tiếng nói theo thời gian. Tuy nhiên, HMM có đặc điểm là mạnh về mô hình hóa từng loại mẫu nhưng yếu về khả năng phân biệt giữa các loại mẫu. Do đó, kết quả nhận dạng của HMM đối với các từ có độ khác biệt ít có độ chính xác không cao (bảng 1). Tỷ lệ nhận dạng đối với mức câu khá thấp là do tỷ lệ lỗi chèn, xóa nhiều khá cao (34%, 1.08%).

***Bảng 2**. Tỷ lệ lỗi giữa các thanh điệu trong nhận dạng bằng mô hình Markov ẩn*

Thanh lỗi	Thanh bị nhận dạng sai						Tổng cộng
	Thanh sắc	Thanh huyền	Thanh hỏi	Thanh ngã	Thanh nặng	Thanh không	
Thanh sắc	-	0	0	1	0	1	2
Huyền	1	-	0	0	0	0	1
Hỏi	1	0	-	0	0	0	1
Ngã	5	1	0	-	1	1	8
Nặng	4	2	1	4	-	1	12
Không	1	0	0	0	0	-	1
Tổng cộng	12	3	1	5	1	3	25

Bảng 2 cho thấy số lượng lỗi nhận dạng nhầm giữa các thanh điệu. Kết quả cho thấy tỷ lệ nhận dạng nhầm ở thanh sắc là cao nhất (12 lỗi chiếm 48%) và thanh hỏi, thanh nặng là thấp nhất (1 lỗi chiếm 4%). Thanh dễ bị nhận dạng nhầm với thanh khác là thanh nặng (12 lỗi chiếm 24%) và thanh ngã (8 lỗi chiếm 32%).

4. Kết Luận

Bài báo này đã trình bày quá trình thực nghiệm nhận dạng một tập các từ tiếng Việt. Các phương pháp sử dụng nhận dạng bao gồm mô hình Markov và mạng nơ-ron ba lớp. Kết quả mô hình Markov ẩn có khả năng ứng dụng trong việc phân biệt các từ, câu. Phân tích tỷ lệ lỗi cho thấy thanh sắc là thanh có tỷ lệ nhận dạng sai nhiều nhất (48% đối với phương pháp HMM). Thanh nặng là thanh có tỷ lệ nhận dạng sai thấp nhất (1%). Thanh không cũng là thanh ít bị nhận dạng sai hơn các thanh khác (12%). Tuy nhiên những kết quả trong bài báo chỉ là những kết quả bước đầu, chúng tôi đang tiến hành thử nghiệm trên cơ sở dữ liệu lớn hơn với các chữ tiếng Việt được phát âm liên tục. Hướng nghiên cứu chính của chúng tôi là xác định được mô hình phiên âm của các âm vị và các từ trong tiếng Việt, cùng với đó là các thử nghiệm trên các mô hình nhận dạng, giữ mô hình mạng nơ-ron và mô hình Markov.

Tài liệu tham khảo

[1] Đặng Ngọc Đức, Lương Chi Mai, *Nhận dạng từ có thanh điệu khác nhau trong tiếng Việt,* Tạp chí Tin học và Điều khiển học, Số 2, trang 131-138, 2003.

[2] J. Schalkwyk, Hosom JP., Ed Kaiser, Khaldom Shobaki, CSLU-HMM: *The CSLU Hidden Markov Modelling Environment*, Center of Spoken Language Understanding (CSLU), Oregon Graduate Institute of Science and Technology, 2000.

[3] B.Yegnanarayana and S. Kishore. AANN: *an alternative to GMM for pattern recognition.* Neural Networks, pages 459–469, 2002.

[4] M. W. Mak K. K. Yiu and S. Y. Kung. *Environment adaptation for robust speaker verification,* In Proc. of Eurospeech, pages 2973–2976, 2003

[5] Shrikanth Narayanan Soonil Kwon, *Speaker change detection using a new weighted distance measure,* In IEEE International Conference on Spoken Language Processing, Denver, USA, volume 4, pages 2537–2540, 2002.

[6] Hong-Jiang Zhang Lie Lu and Hao Jiang, *Content analysis for audio classification and segmentation,* IEEE transactions on speech and audio processing, 10(7):504–516, 2002.