KẾT HỢP PHƯƠNG PHÁP LƯỢNG TỬ HÓA VECTOR VÀ MÔ HÌNH MARKOV ẨN TRONG NHẬN DẠNG TIẾNG NÓI TIẾNG VIỆT - ỨNG DỤNG TRONG ĐIỀU KHIỂN TIVI

NGUYỄN TÚ HÀ
Khoa Vật lý, Trường Đại học Sư phạm, Đại học Huế
Email: nguyentuha82@gmail.com

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 các nhà nghiên cứu do tính phức tạp và không ổn định của tiếng nói. Đặc biệt, đối với nhận dạng tiếng nói tiếng Việt thì kết quả còn nhiều hạn chế. Bài báo trình bày một hướng nhận dạng tiếng nói tiếng Việt, sử dụng mô hình Markov ẩn (Hidden Markov Model - HMM) kết hợp với phương pháp lượng tử hóa vector (Vector Quantization - VQ) để nhận dạng tiếng nói. Kết quả được kiểm nghiệm thực tế bằng mô hình điều khiển tivi.

Từ khóa: nhận dạng tiếng nói; lượng tử hóa vector; mô hình Markov ẩn.

1. ĐẶT VẤN ĐỀ

Hiện nay, vấn đề tìm hiểu và thực hiện một hệ thống nhận dạng tiếng nói đã được đưa vào nghiên cứu trong các viện nghiên cứu trên khắp thế giới [1], [3]. Những ứng dụng mà hệ thống này mang lại là vô cùng to lớn và có ý nghĩa như: xe lăn cho người tàn tật được điều khiển bằng tiếng nói; điều khiển máy tính hoặc các hệ thống tự động bằng tiếng nói.

Trên thế giới đã có nhiều hệ thống nhận dạng tiếng nói đã và đang được ứng dụng rất hiệu quả như: ViaVoice, Dragon Naturally Speaking, Spoken Toolkit, Google… Các hệ thống nhận dạng này áp dụng cho ngôn ngữ tiếng Anh, vì vậy, không thể áp dụng hệ thống này cho nhận dạng tiếng Việt. Do đó, một hệ thống nhận dạng tiếng nói tiếng Việt cần phải được xây dựng để có thể ứng dụng cho người Việt Nam.

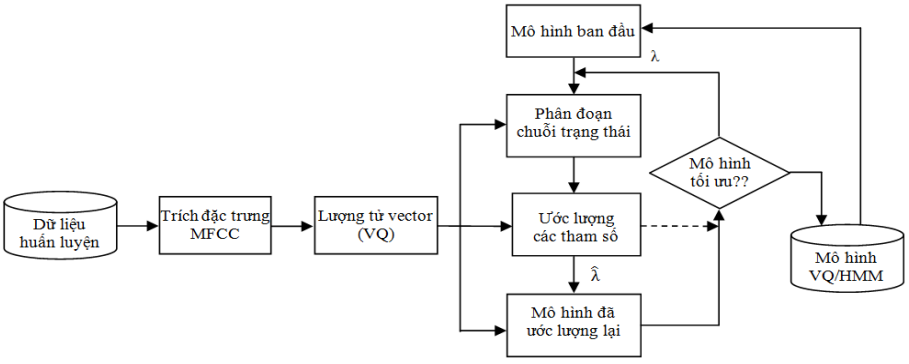
Một đề xuất mới của chúng tôi về một phương pháp nhận dạng tiếng nói tiếng Việt, sử dụng mô hình Markov ẩn rời rạc để nhận dạng tiếng nói kết hợp với phương pháp lượng tử hóa vector. Hệ thống được kiểm nghiệm thực tế bằng việc xây dựng mô hình nhận dạng tiếng nói tiếng Việt gồm các nhóm lệnh điều khiển tivi.

2. HỆ THỐNG NHẬN DẠNG TIẾNG NÓI

Một hệ thống nhận dạng nói chung thường bao gồm hai phần: phần huấn luyện và phần nhận dạng. “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.

Tạp chí Khoa học và Giáo dục, Trường Đại học Sư phạm Huế
ISSN 1859-1612, Số 01(45)/2018: tr. 113-118
Ngày nhận bài: 28/12/2016; Hoàn thành phản biện: 19/4/2017; Ngày nhận đăng: 06/7/2017

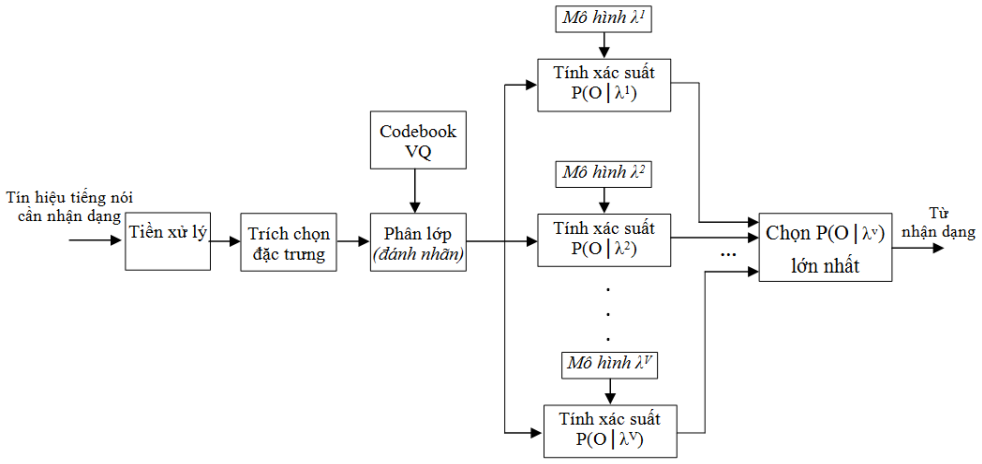
Quá trình huấn luyện được thực hiện như sau: Giả sử hệ thống cần nhận dạng bộ từ vựng có V từ. Đầu tiên chúng ta phải huấn luyện để xây dựng mô hình Markov ẩn λ_v của các từ trong bộ từ vựng bằng cách áp dụng bài toán 3 của mô hình HMM (bài toán huấn luyện) [4]. Trong quá trình huấn luyện, mỗi từ sẽ được nói nhiều lần (có thể do một hay nhiều người nói), sau đó chúng ta sẽ tiến hành trích đặc trưng bằng phương pháp đường bao phổ (Mel frequency cepstral coefficient - MFCC). Các vector đặc trưng này sẽ được lượng tử hóa vector để phân lớp và đưa vào mô hình HMM để ước lượng các tham số của mô hình một cách tối ưu cho từng từ. Như vậy kết quả được một tập gồm V codebook kích thước M, và V mô hình HMM.



Hình 1. *Quá trình huấn luyện mô hình VQ/HMM*

Để nhận dạng, chúng ta sẽ áp dụng bài toán 1 (bài toán ước lượng) [4]. Từ cần nhận dạng sẽ được trích đặc trưng bằng phương pháp MFCC và phân lớp bằng lượng tử hóa vector để có được tập quan sát $O=\{o_1 o_2 \ldots o_T\}$. Tiếp theo, ta sẽ tính xác suất $P(O|\lambda_v)$ cho tất cả các mô hình ($1 \le v \le V$) và chọn từ v có xác suất lớn nhất, tức là:

$$v^* = \arg\max_{0 \le v \le V} \left[P\left(O \mid \lambda_v\right) \right]$$



Hình 2. *Quá trình nhận dạng từ rời rạc bằng mô hình VQ/HMM*

2.1. Tiền xử lý

Tín hiệu tiếng nói sau khi được thu và trước khi trích chọn đặc trưng, phải tiến hành tiền xử lý tín hiệu tiếng nói. Mục đích của việc tiền xử lý tín hiệu tiếng nói để loại bỏ nhiễu, chuẩn hóa biên độ, làm rõ tín hiệu, xác định các lệnh điều khiển, tách từ.

2.2. Trích chọn đặc trưng

Trích chọn đặc trưng là quá trình thực hiện các phân tích nhằm xác định các thông tin quan trọng, đặc trưng, ổn định của tín hiệu tiếng nói. Đối với một hệ nhận dạng tiếng nói, việc trích chọn đặc trưng của tiếng nói là cần thiết. Điều này giúp giảm thiểu số lượng dữ liệu trong việc huấn luyện và nhận dạng, dẫn đến số lượng công việc tính toán trong hệ thống giảm đáng kể. Bên cạnh đó, việc trích chọn đặc trưng còn làm rõ sự khác biệt của tiếng này so với tiếng khác, làm mờ đi sự khác biệt của cùng hai lần phát âm khác nhau của cùng một tiếng. Từ quá trình này, chúng ta sẽ có được chuỗi vector quan sát O.

Có nhiều phương pháp trích chọn đặc trưng khác nhau như: FBA, LPC, MFCC, PLP… Mỗi phương pháp có những ưu điểm và nhược điểm riêng. Tuy nhiên, phương pháp dựa trên việc tính hệ số MFCC (Mel-scale Frequency Cepstral Coefficient) được sử dụng vì nó phổ biến và hiệu quả nhất. Vì vậy trong nghiên cứu này sử dụng phương pháp MFCC làm công cụ để trích chọn đặc trưng cho hệ thống nhận dạng.

2.3. VQ Codebook

Trong mô hình HMM rời rạc, toàn bộ không gian đặc trưng âm thanh được chia làm một số trung bình các vùng, bằng thủ tục phân vùng như lượng tử hóa vectơ (VQ). Trọng tâm của mỗi vùng được tiêu biểu bởi một codeword vốn là một chỉ mục đến codebook. Mỗi mẫu tiếng nói được đổi thành một codeword bằng cách tìm ra vector gần nhất trong codebook. Mỗi codebook có M codeword được gọi là codebook cỡ M. M cũng là số kí hiệu quan sát được của 1 trạng thái trong HMM. Như vậy, trong HMM rời rạc, số quan sát là hữu hạn. Nhược điểm của mô hình dạng này là có sai số trong quá trình lượng tử hoá (nếu kích thước của codebook là nhỏ). Ngược lại nếu kích thước của codebook lớn thì sẽ phải trả giá bằng số lượng tính toán sẽ tăng lên. Trong nghiên cứu này sử dụng thuật toán Split Binary (hay thuật toán LBG) [6].

2.4. Ước lượng các tham số của mô hình HMM

Đối với mỗi từ trong bộ từ vựng, chúng ta xây dựng một mô hình HMM bằng cách ước lượng các thông số của mô hình một cách tối ưu dựa trên chuỗi dữ liệu quan sát trong quá trình huấn luyện. Trong nghiên cứu này sử dụng thuật toán Baum-Welch [6], [8], đây là một trong những phương pháp tối ưu thành công nhất.

2.5. Nhận dạng

Đối với mỗi từ cần nhận dạng, hệ thống tính toán mô hình có khả năng với tất cả mô hình đã huấn luyện và chọn ra mô hình có khả năng nhất.Một phương pháp thông dụng hay được dùng để giải quyết bài toán này là dùng thuật toán tìm kiếm Viterbi [9]. Đây là

thuật toán dựa trên phương pháp lập trình động (Dynamic Programing Method) để tìm ra một dãy các trạng thái tối ưu duy nhất.

3. THỰC NGHIỆM VÀ KẾT QUẢ

3.1. Thực nghiệm

Trong nghiên cứu này, chúng tôi chọn các từ để huấn luyện là:***tắt, bật, tivi, tăng, giảm, chuyển, âm, kênh, một, hai, ba, bốn, năm, sáu, bảy, tám, chín, không;***và các câu lệnh điều khiển tivi có cú pháp:

- $\begin{cases} \text{"Bật"} \\ \text{"Tắt"} \end{cases} + \text{"Tivi"}$
- $\begin{cases} \text{"Tăng"} \\ \text{"Giảm"} \end{cases} + \begin{cases} \text{"Âm"} \\ \text{"Kênh"} \end{cases}$
- "Chuyển"+"kênh" + {"Không","Một", …"Chín"} + {"Không","Một", …"Chín"}

Cơ sở dữ liệu được xây dựng trong nghiên cứu này được thu thập từ 150 người nói gồm 70 nam và 80 nữ, có độ tuổi từ 18 đến 30. Các người nói được hướng dẫn phát âm chuẩn theo một tốc độ nhất định và việc thu âm được thực hiện trong phòng thu ít nhiễu. Các tập tin âm thanh được thu từ chương trình Adobe Audition, sử dụng PCM, lấy mẫu tại tần số 16.000Hz với 16bit và lưu trữ dưới định dạng WAV.

Việc thu âm được thực hiện gồm hai mục đích, thu âm để chuẩn bị cơ sở dữ liệu cho quá trình huấn luyện mô hình và cho quá trình nhận dạng.

3.2. Phương pháp đánh giá

Để đánh giá hệ thống, trong nghiên cứu này chúng tôi sử dụng phương pháp thực nghiệm với thống kê và so sánh kết quả trực tiếp. Mỗi nhóm dữ liệu thực nghiệm được đọc vào một cách ngẫu nhiên và ghi nhận kết quả trả ra từ chương trình, sau đó tính tỉ lệ nhận dạng từ đúng, tỉ lệ nhận dạng lỗi sai.

Đối với quá trình huấn luyện và kiểm tra, kết quả được chia thành 2 nhóm: nhóm 100 người được huấn luyện và nhóm 50 người không được huấn luyện.

3.2. Kết quả thực nghiệm

- Kết quả nhận dạng từ

Bảng 1. *Kết quả nhận dạng từ*

Nhóm dữ liệu		***Tổng***	***Nhận dạng đúng***	***Tỉ lệ***
100 người đã huấn luyện	**Từ**	800	782	97,75%
	Số	1000	981	98.10%
50 người	**Từ**	400	385	96,25%

không huấn luyện	**Số**	500	479	95,80%

- Kết quả nhận dạng câu lệnh

Bảng 2. *Kết quả nhận dạng câu lệnh* (*100 người đã huấn luyện*)

Câu lệnh	***Tổng***	***Nhận dạng đúng***	***Tỉ lệ***
BẬT_TIVI	1000	968	96,8%
TẮT_TIVI	1000	925	92,5%
TĂNG_ÂM	1000	923	92,3%
GIẢM_ÂM	1000	976	97,6%
TĂNG_KÊNH	1000	934	93,4%
GIẢM_KÊNH	1000	979	97,9%
CHUYỂN KÊNH	1000	967	96,7%

Bảng 3. *Kết quả nhận dạng câu lệnh (50 người không huấn luyện)*

Câu lệnh	***Tổng***	***Nhận dạng đúng***	***Tỉ lệ***
BẬT_TIVI	1000	923	92,3%
TẮT_TIVI	1000	879	87,9%
TĂNG_ÂM	1000	824	82,4%
GIẢM_ÂM	1000	935	93.5%
TĂNG_KÊNH	1000	859	85,9%
GIẢM_KÊNH	1000	891	89,1%
CHUYỂN KÊNH	1000	932	93,2%

4. KẾT LUẬN

Dựa trên kết quả thực nghiệm, nghiên cứu đã xây dựng thành công mô hình nhận dạng tiếng nói với tỷ lệ thành công tương đối tốt.

Tuy nhiên, cần nghiên cứu phát triển thêm:

Xây dựng cơ sở dữ liệu lớn hơn để huấn luyện cho các mô hình tốt hơn. Đồng thời phát triển thêm bộ từ vựng để có thể điều khiển thiết bị phong phú hơn.

Tích hợp thêm các giải pháp giảm nhiễu trong khối tiền xử lý để nâng cao hiệu xuất nhận dạng và ứng dụng trong môi trường có nhiễu cao.

Tích hợp hệ thống nhận dạng tiếng nói trên các chip DSP, FPGA ... để có thể ứng dụng thuận tiện hơn và đóng gói thành bộ sản phẩm hoàn thiện.….

TÀI LIỆU THAM KHẢO

[1] Phạm Văn Tuấn (2011). *Bài giảng nhận dạng tiếng nói*, Đại học Bách khoa Đà Nẵng.
[2] Lê Tiến Thường (2002). *Xử lý số tín hiệu và Wavelets-Tập 1*, NXB Đại học Quốc gia TP Hồ Chí Minh.
[3] Gales. M. and S. Young (2007). *The Application of Hidden Markov Models in Speech Recognition*, Foundations and Trends in Signal Processing, Vol.1, No.2, p.p 195-304.
[4] Rabiner, L. R. (1989). *A tutorial on hidden Markov models and selected applications in speech recognition*, Proceedings of IEEE, vol. 77, no. 2, pp. 257–286.
[5] Juang, B. H. and Rabiner, L. R. (1991). *Hidden Markov Models for Speech Recognition*, Technometrics, Vol.33, No.3, pp. 251-272.
[6] Linde, Y., Buzo, A., and Gray, R. M. (1980). *An Algorithm for Vector Quantizer*, IEEE Transactions on Communication, Vol.28, No.1, pp. 84-95.
[7] Segura, J. C., Rubio, A. J., Peinado, A. M., Garcia, P., and Roman, R. (1994). *Multiple VQ Hidden Markov Modeling for Speech Recognition*, Speech Communication, Vol.14, pp. 163-170.
[8] Balwant, A., Sonkamble, D. and Doye, D. (2012). *Speech Recognition Using Vector Quantization through Modified K-means LBG Algorithm*, Computer Engineering and Intelligent Systems, ISSN 2222, Vol.3, No.7, pp.137-144.
[9] Rabiner, L. R. and Juang, B.H. (1993). *Fundamentals of speech recognition*, Prentice-Hall International, Inc.
[10] Le, V.B and Besacierm, L. (2009). *Automatic Speech Recognition for Under-Resourced Languages: Application to Vietnamese Language*, IEEE.

Title: USING THE COMBINATION OF VECTOR QUANTIZATION METHOD AND HIDDEN MARKOV MODELS FOR VIETNAMESE SPEECH RECOGNITION - APLLYING FOR CONTROL THE TELEVISION

Abstract: Researching and inventing speech recognitionmethods have been paid much considerations bymany scientists over the world. However, the achievements don't satisfy researchers' demandsbecause of the complexity and unstability of speech until now. Especially with Vietnamese speech, theresults are more unsatisfied.The paper suggests asynthetic method for recogniting Vietnamesespeech, is based on the combination of Vector Quantization (VQ) method and Hidden Markov Models (HMMs).Theresults are experimented through a model of remote control television.

Keywords: Speech-recognition; Vector Quantization; HMM.