Tạp chí Tin học và Điều khiển học, T. 19, S. 4 (2003), 347–352

NHẬN DẠNG THANH ĐIỆU TIẾNG VIỆT

ĐOÀN PHAN LONG, NGUYỄN NGỌC SAN

Học viện Công nghệ Bưu chính Viễn thông

Abstract. Vietnamese language is a typical monosyllabic tonal one. To which tone identification is undoubtedly an essential component in regard the speech recognition of unlimited words. This paper presents an efficient method for tone classification of isolate Vietnamese syllables. Several suprasegmental feature parameters for tone identification are extracted from the voiced portion and then fed into a neuron network classifier.

Tóm tắt. Tiếng Việt là loại hình ngôn ngữ đơn lập (ngôn ngữ không biến hình - amorphous) và có thanh điệu. Việc xác định dấu trong ngôn ngữ tiếng Việt là công việc quan trọng và cơ bản để giải quyết việc nhận dạng tiếng nói tiếng Việt với số lượng từ không hạn chế. Bài báo trình bày phương pháp nhận dạng dấu trong các âm tiết tiếng Việt phát âm rời rạc với việc áp dụng các thuật toán để xác định các thông số đặc trưng siêu đoạn tính từ phần hữu thanh. Thử nghiệm áp dụng mạng Nơron để xác định thanh điệu trong tiếng Việt với 5 đầu vào là các thông số đặc trưng và 6 đầu ra tương ứng với 6 thanh điệu trong tiếng Việt.

1. MỞ ĐẦU

Việc nghiên cứu để nhận dạng thanh điệu trong tiếng Việt là cần thiết để giúp quá trình xây dựng một hệ thống nhận dạng tiếng Việt hoàn chỉnh với số lượng từ vựng lớn. Cho đến nay, rất ít các công trình được nghiên cứu và công bố trong lĩnh vực nhận dạng tiếng Việt nói chung và nhận dạng thanh điệu trong tiếng Việt nói riêng. Một số bài báo đã đề cập đến vấn đề nhận dạng từ có thanh điệu khác nhau ([5]), song các bài báo này không đề cập đến các phương pháp và thuật toán để xác định các thông số đặc trưng, từ đó có thể áp dụng các phương pháp tìm kiếm khác nhau để xác định thanh điệu trong các âm tiết tiếng Việt.

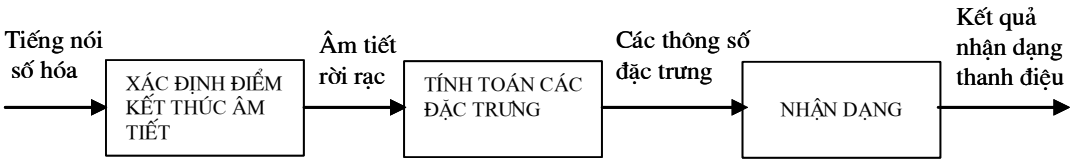
Cũng giống như tiếng Trung quốc, Thái..., tiếng Việt là ngôn ngữ đơn lập không biến hình và có thanh điệu. Mỗi âm tiết đều mang trong nó một thanh điệu đóng vai trò là một âm vị mang tính siêu đoạn. Đó là loại âm vị không có âm đoạn, không độc lập tồn tại nhưng cũng có chức năng phân biệt nghĩa, nhận diện từ ([5]). Đây là một đặc điểm riêng của tiếng Việt so với các ngôn ngữ Châu Âu. Tiếng Việt có 16 nguyên âm, 21 phụ âm và 2 bán nguyên âm (semi-vowel). Mỗi âm tiết có thể kết hợp với một trong sáu thanh và tạo thành 6 từ khác nhau. Tất nhiên trong số đó có nhiều kết hợp không tồn tại ([3]).

2. MÔ HÌNH HỆ THỐNG NHẬN DẠNG THANH ĐIỆU

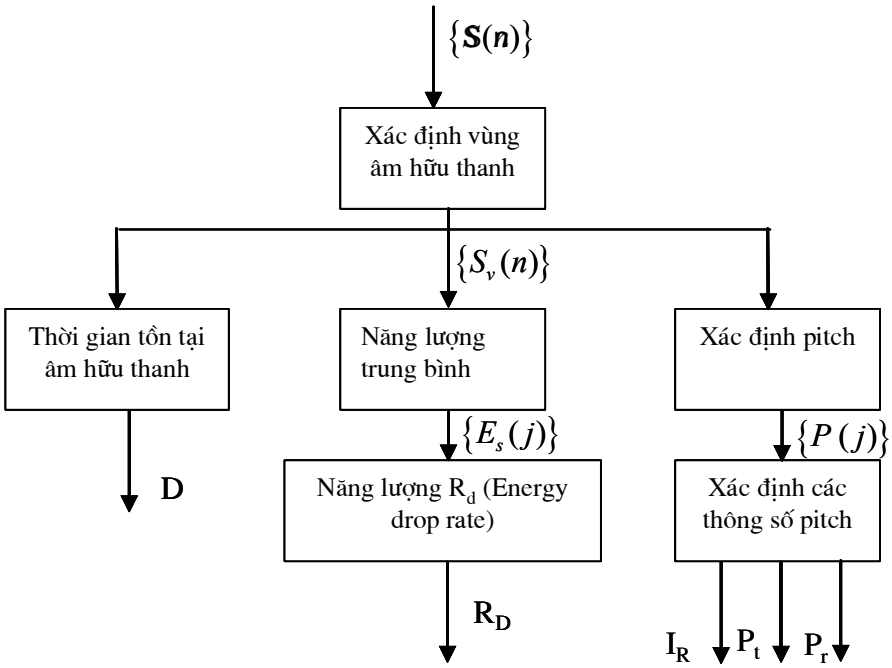
Bài báo mô tả quá trình nhận dạng thanh điệu trong tiếng Việt với cách phát âm rời rạc (Hình 1). Với cách phát âm liên tục, do tiếng Việt là tiếng đơn âm, bằng một số phương pháp sử dụng phân tích phổ và năng lượng ta có thể nhận dạng điểm bắt đầu, điểm kết thúc của mỗi âm tiết. Khi đã xác định được các điểm này, ta hoàn toàn có thể áp dụng phương pháp nhận dạng thanh điệu như phần dưới sẽ trình bày.

Phương pháp tiếp cận sử dụng các thông số pitch và các đặc trưng siêu đoạn tính. Để đơn giản hóa quá trình và tăng thời gian tính toán, ta chỉ xác định 5 thông số cơ bản để xác

định thanh điệu trong tiếng Việt (hình 2). Các đặc trưng này được dùng để làm các thông số cho các phương pháp nhận dạng có thể được lựa chọn để xác định thanh điệu của tiếng nói tiếng Việt, ví dụ như HMM, Neural network... Trong tiếng Việt, các thanh chỉ có thanh tính như thanh bằng, thanh huyền, để xác định ta chỉ cần các đặc trưng cơ bản là các thông số pitch của âm thanh, trong khi các thanh còn lại là các thanh ngã, sắc, hỏi, nặng, do đặc trưng phi điệu tính như hiện tượng yết hầu hoá, thanh hầu hoá, ta cần tính toán các thông số đặc trưng siêu đoạn tính là các thông số về năng lượng.



Hình 1. Mô hình nhận dạng các thanh điệu trong tiếng Việt với cách phát âm rời



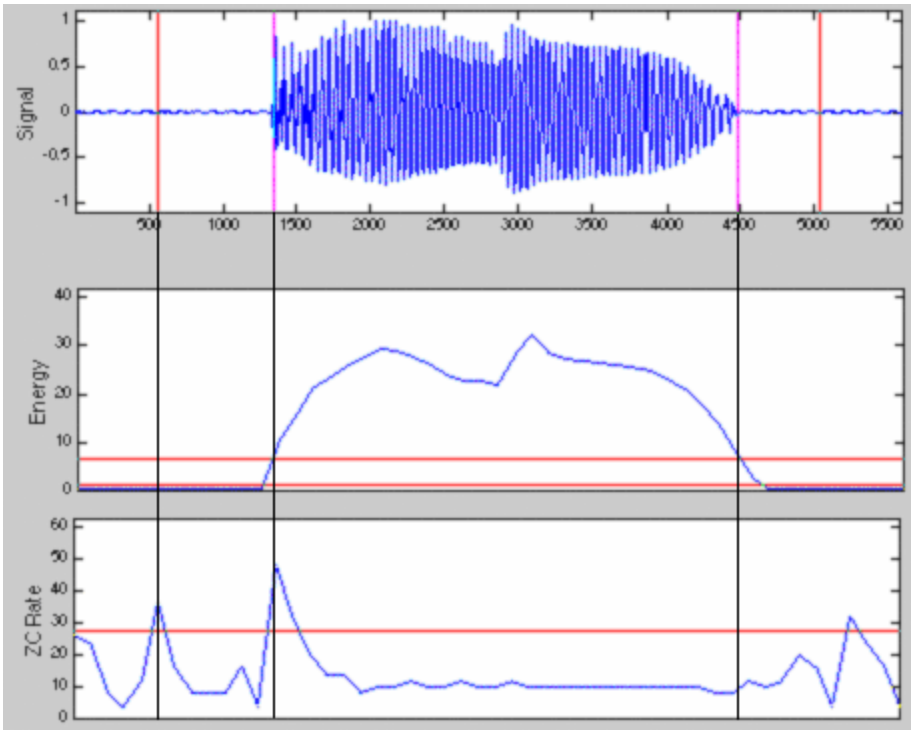
Hình 2. Xác định các thông số đặc trưng siêu đoạn tính

3. XÁC ĐỊNH VÙNG THANH

Trong các ngôn ngữ có thanh điệu nói chung, tiếng Việt nói riêng, tất cả các thông tin về thanh điệu đều nằm trong vùng thanh. Do vậy, vấn đề đầu tiên của quá trình nhận dạng thanh âm là xác định chính xác vùng thanh của tiếng nói ([4]). Các thông số đặc trưng để xác định thanh âm đều nằm ở trong vùng thanh này. Để phân biệt chính xác vùng thanh, người ta sử dụng các thông số đặc trưng về năng lượng và tỷ lệ trung bình vượt không thời gian ngắn (Zero-crossing rate).

Tín hiệu âm thanh đầu vào $\{S(n)\}$ được chia thành các khung 10 ms với tỷ lệ chồng lên nhau là 50%. Trong mỗi khung tín hiệu, các thông số năng lượng $\{E(j)\}$ và tỷ lệ trung bình vượt không thời gian ngắn $\{Z(j)\}$ được tính toán, trong đó $E(j)$ và $Z(j)$ là các thông số năng

lượng và tỷ lệ điểm cắt không tại khung tín hiệu j. Điểm bắt đầu và điểm kết thúc của phần âm thanh (voiced portion) được xác định bằng cách tìm xuôi và ngược từ khung tín hiệu có năng lượng lớn nhất. Điểm bắt đầu là điểm mà tại đó khung tín hiệu có năng lượng nhỏ hơn mức năng lượng được định trước E_{it} hoặc tỷ lệ trung bình vượt không thời gian ngắn lớn hơn giá trị định trước Z_i. Điểm kết thúc được xác định nhờ thông số năng lượng của khung tín hiệu (Hình 3).



Hình 3. Xác định vùng thanh

4. XÁC ĐỊNH THÔNG SỐ ĐẶC TRƯNG PITCH

Vùng thanh được chia thành 16 khung tín hiệu có độ dài bằng nhau. Từ các mẫu tín hiệu, các giá trị pitch trong từng khung tín hiệu được tính toán ([2]). Ta áp dụng thuật toán của Sondhi ([1]). Các giá trị Pitch được xác định bởi vectơ đặc trưng $\{P(1), P(2), ..., P(i), ..., P(16)\}$ trong đó $\{P(i)\}$ là giá trị Pitch tại khung tín hiệu i. Giá trị Pitch khởi đầu P_I và giá trị Pitch kết thúc P_F được định nghĩa như sau

$$P_I = (P(3) + P(4))/2 \tag{1}$$

$$P_F = (P(13) + P(14))/2 \tag{2}$$

Thông số quan hệ Pitch (Pitch-related Parameter) I_R được gọi là chỉ số tăng được xác định bởi tập các thông số vectơ đặc trưng $\{P(i)\}$ như sau

$$I_R = k\frac{\max\{P(i)\} - \min\{P(i)\}}{\max\{P(i)\} + \min\{P(i)\}} \quad \text{với} \quad 3 \le i \le 14$$

trong đó

$$k = \begin{cases} 1 & \arg\max\{P(i)\} > \arg\min\{P(i)\} \\ -1 & \arg\max\{P(i)\} \leq \arg\min\{P(i)\} \end{cases} \tag{3}$$

Dấu của I_R thể hiện Pitch tại thời điểm đó tăng lên hay giảm xuống trong khi giá trị tuyệt đối của I_R thể hiện giá trị pitch.

5. THỜI GIAN VÀ TỶ LỆ SUY GIẢM NĂNG LƯỢNG

Trong ngôn ngữ đơn âm nói chung và trong tiếng Việt nói riêng, thời gian D được định nghĩa là khoảng thời gian của vùng có thanh của một âm (Voice portion). Nếu khoảng thanh được phát hiện chứa N_v khung liên tiếp với độ dài mỗi khung là 10 ms (và tỷ lệ chồng lên nhau là 50%) thì D được xác định như sau

$$D = (N_v + 1) \times 5ms \tag{4}$$

Để tính tỷ lệ suy giảm năng lượng ở cuối mỗi âm, trước tiên giá trị năng lượng của các khung tín hiệu $\{E_v(j)\}$ của vùng thanh được làm trơn tính theo công thức gần đúng với mỗi vùng cửa sổ gồm 5 khung tín hiệu. Giá trị năng lượng được làm trơn $\{E_{sv}(j)\}$ được tính như sau

$$E_{sv}(j) = \frac{1}{5} \sum_{m=-2}^{2} E_v(j+m) \tag{5}$$

Coi $j_{\max}$ là khung tín hiệu có mức năng lượng được làm trơn lớn nhất, t_d là thời gian mà tại đó giá trị $\{E_{sv}(j)\}$ suy giảm từ 90% đến 10% của giá trị $E_{sv}(j_{\max})$. Khi đó tỷ lệ suy giảm năng lượng R_D được tính như sau

$$R_D = \frac{1}{t_d} \tag{6}$$

6. TIÊU CHUẨN HÓA CÁC THÔNG SỐ ĐẶC TRƯNG

Do tín hiệu âm thanh được tạo ra bởi cùng một người nói tại các thời điểm khác nhau không bao giờ giống nhau hoàn toàn. Nó luôn có sự sai lệch do các yếu tố về trọng âm, ngữ điệu, tốc độ,... Giá trị thông số Pitch được tính như trên bị phụ thuộc rất nhiều vào người nói. Những người có độ tuổi khác nhau cho giá trị Pitch khác nhau rất xa. Giọng nữ thường có giá trị Pitch cao hơn giọng nam giới. Ngay cả với cùng người nói giá trị Pitch tại các thời điểm khác nhau cũng rất khác nhau. Điều này dẫn đến vấn đề, với các thanh âm phụ thuộc nhiều vào giá trị Pitch như dấu sắc, huyền, không dấu rất khó được xác định hoặc cho sai số rất lớn. Điều tương tự cũng xảy ra đối với các giá trị thời gian D và tỷ lệ suy giảm năng lượng R_D.

Với các lý do trên, vấn đề tiêu chuẩn hóa các thông số đặc trưng là rất cần thiết để vấn đề nhận dạng thanh điệu không bị phụ thuộc vào các yếu tố khách quan như trên đã trình bày. Các thông số đặc trưng Pitch được tiêu chuẩn hóa như sau

$$\widehat{P}_I = P_I/P_S; \quad \widehat{P}_F = P_F/P_S \tag{7}$$

trong đó, giá trị P_S là giá trị Pitch thực chất của người nói và được định nghĩa là giá trị pitch khởi đầu trung bình cho tất cả các âm tiết có thanh điệu là dấu sắc, dấu huyền và không dấu.

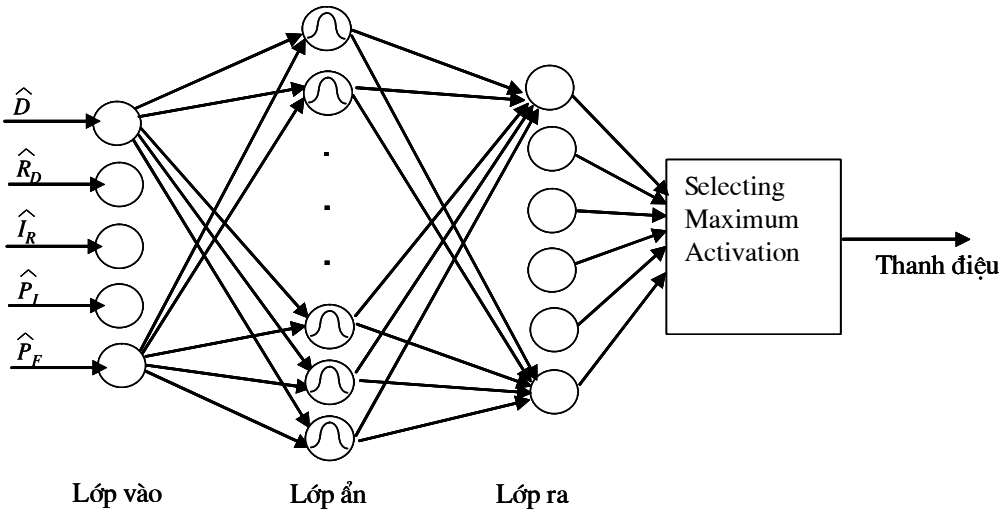
Các thông số đặc trưng về thời gian và tỷ lệ suy giảm năng lượng được tiêu chuẩn hóa như sau

$$\widehat{D} = D/D_S; \quad \widehat{R}_D = R_D/R_{SD} \tag{8}$$

Các thông số D_S và R_{SD} là thời gian trung bình và giá trị ước lượng tỷ lệ suy giảm năng lượng từ cơ sở dữ liệu từ chọn trước.

7. LỰA CHỌN THANH ĐIỆU TRÊN CƠ SỞ MẠNG NƠRON RBF (RADIAL BASIS FUNCTION)

Để xác định và nhận dạng thanh điệu, tác giả sử dụng mạng nơron và sử dụng thuật toán quyết định "Winner-take-all". Mạng nơron này gồm 5 đầu vào tương ứng với 5 thông số đặc trưng của thanh điệu, 6 đầu ra tương ứng với 6 thanh âm là không dấu, sắc, hỏi, ngã, huyền, nặng. Trong bài này, do tác giả áp dụng để nhận dạng thanh điệu phụ thuộc người nói nên mạng nơron sử dụng 25 nút ẩn như hình 4 dưới đây.



Hình 4. Mạng nơron nhận dạng thanh điệu

8. KẾT QUẢ THỰC NGHIỆM

Môi trường thử nghiệm là cơ sở dữ liệu dùng cho quá trình thực nghiệm bao gồm khoảng 300 câu đọc từ các bài báo trong các tạp chí. Các câu được thu âm trong môi trường trong nhà, do một giọng nam đọc, sử dụng micro thường và card âm thanh của máy tính sách tay IBM-T30, các thông số lấy mẫu như bảng 1 dưới đây. Các câu được gán nhãn bằng tay tới mức âm vị.

Bảng 1. Các thông số lấy mẫu thử nghiệm

STT	Đặc tả thông số	Đơn vị
1	Số lượng người	25 người, 15 giọng nam, 10 giọng nữ
2	Tần số lấy mẫu	11025 Hz
3	Số kênh	1 (Mono)
4	Bít mã hoá	8 bít
5	Tham số	Thời gian âm hữu thanh, năng lượng, Pitch

Hệ thống sau khi được xây dựng, đã cho tiến hành nhận dạng trên cùng một tập dữ liệu

gồm 100 câu. Tập dữ liệu này cũng là tập để thử nghiệm cho mạng trên. Kết quả nhận dạng như sau:

+ Các tham số chính của âm thanh thu được đều phù hợp với cấu trúc tham số xây dựng trong môđun. Các file có nhiễu ở mức thấp và trung bình thì việc tách tín hiệu chính xác hơn.
+ Tính năng nhận dạng thanh điệu cho kết quả ban đầu tương đối tốt với độ chính xác ở mức từ là 83,77% và ở mức câu là 70,15%. Điều này chứng tỏ khả năng phân lớp tốt của mạng nơron RBF (Bảng 2).

Bảng 2. Kết quả nhận dạng

	Sắc	Huyền	Hỏi	Ngã	Nặng	Không	Tổng cộng	Độ chính xác
Sắc	120	2	4	3	0	2	131	92%
Huyền	1	84	3	3	2	0	93	90%
Hỏi	3	4	65	6	3	7	88	74%
Ngã	2	4	3	38	4	6	57	67%
Nặng	4	2	5	6	49	3	69	71%
Không	0	0	2	2	1	93	98	95%
Nhận đúng	120	84	65	38	49	93	536	83.77%

Các thanh "hỏi", "ngã", "nặng" có tỷ lệ nhận dạng nhầm cao hơn các thanh khác.

9. KẾT LUẬN

Bài báo này đã trình bày quá trình thực nghiệm nhận dạng thanh điệu cho tiếng Việt sử dụng kỹ thuật sử lý tiếng nói và mạng nơron. Kết quả nhận dạng cho thấy mạng nơron có khả năng phân biệt các thanh điệu tương đối tốt. Tuy nhiên những kết quả trong bài này chỉ là những kết quả bước đầu. Trên cơ sở thành công của bài báo này, tác giả đang tiếp tục tiến hành nghiên cứu phương pháp nhận dạng tiếng việt trên cơ sở mô hình bán âm tiết, xây dựng các thuật toán để tách và gán nhãn bán âm tiết, phân tích các véc tơ đặc trưng và áp dụng phương pháp tìm kiếm để nhận dạng tiếng Việt có số lượng từ không hạn chế

TÀI LIỆU THAM KHẢO

[1] M. M. Sondhi, New methods of Pitch Extraction, *IEEE trans. on Audio and Electroacoustics, AU* **16** (2) (1968) 262 – 266.

[2] A. Komatsu, A. Ichikawa, Phoneme Recognition in continuous Speech, *Proceedings of ICASSP* (1982) 883 – 886.

[3] Đỗ Xuân Thảo, Lê Hữu Tình, *Giáo trình tiếng Việt 2*, Nhà xuất bản giáo dục, 1997.

[4] L. Rabiner, B. H. Juang, *Fundamental of speech Recognition*, Prentice Hall, 1993.

[5] 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* **19** (2) (2003) 131 – 138.

Nhận bài ngày 15 - 08 - 2003