Tạp chí Tin học và Điều khiển học, T.21, S.1 (2005), 69–74

NHẬN DẠNG THANH ĐIỆU TIẾNG NÓI TIẾNG VIỆT BẰNG MẠNG NƠRON PHÂN TẦNG

LÊ TIẾN THƯỜNG, TRẦN TIẾN ĐỨC

Trường Đại học Bách khoa
Đại học Quốc gia Tp. Hồ chí Minh

Abstract. Vietnamese is a monosyllabic and tonal language. Tone recognition is important because tone affects the lexical identification of words. The fundamental frequency (F0) contour of tone is fitted by the quadric curve and normalized. A hierarchical neural networks with three neural networks is used to recognize the tones. More than 300 lines of poetry from the Truyện Kiều written by Nguyễn Du are used as training data in the networks and other 200 lines of poetry another are used to test, where each line is recorded only one time. The experimental result has average accuracy 92.6%.

Tóm tắt. Tiếng Việt là ngôn ngữ đơn lập có thanh điệu. Nhận dạng thanh điệu là quan trọng vì thanh điệu có chức năng phân biệt nghĩa của từ. Đường nét tần số cơ bản F0 của thanh điệu được làm mịn bằng đa thức bậc hai và chuẩn hóa. Sử dụng mạng nơron phân tầng gồm ba mạng nơron để nhận dạng thanh điệu. Hơn 300 câu thơ trong Truyện Kiều của Nguyễn Du làm dữ liệu học và 200 câu thơ khác cũng trong Truyện Kiều làm dữ liệu thử, trong đó mỗi câu thơ chỉ ghi một lần. Kết quả thử nghiệm có độ chính xác trung bình 92,6% cho phép kết luận hướng nghiên cứu này là thích hợp.

1. MỞ ĐẦU

Tiếng Việt là loại ngôn ngữ đơn lập, có thanh điệu. Bất kỳ âm tiết nào cũng có một thanh điệu nhất định và thanh điệu bao giờ cũng tồn tại trong âm tiết. Thanh điệu là sự thay đổi tần số cơ bản F0 hay cao độ của giọng nói, có chức năng phân biệt nghĩa của từ. Tiếng Việt có tất cả 6 thanh điệu theo thứ tự: ngang, huyền, ngã, hỏi, sắc và nặng, vì vậy việc nhận dạng thanh điệu sẽ làm giảm đáng kể số lượng các âm tiết khi nhận dạng tiếng Việt. Nhiều ngôn ngữ khác ở Phương Đông như tiếng Hán, tiếng Thái và ở Tây Phi như tiếng Zulu, tiếng Hausa cũng có hiện tượng này.

Cho đến nay vấn đề nhận dạng thanh điệu tiếng Việt còn chưa được hoàn chỉnh. Hiện tại, chúng tôi chỉ tìm được hai nghiên cứu về vấn đề này [1, 4], nhưng cả hai đều không theo hướng trích hay làm mịn đường nét F0 (F0 contour) của thanh điệu, trong khi đó tình hình nhận dạng tiếng Hán và tiếng Thái có phần phong phú hơn [5 - 7]. Bài báo này được tổ chức như sau: Phần 2 nêu phương pháp trích đặc điểm thanh điệu tiếng Việt. Phần 3 trình bày thử nghiệm dùng mạng nơron phân tầng gồm ba mạng nơron để đánh giá độ thành công của hệ nhận dạng dựa trên dữ liệu học và thử là các câu thơ trong Truyện Kiều của Nguyễn Du. Cuối cùng là nhận xét và kết luận.

2. TRÍCH ĐẶC ĐIỂM

2.1. Trích tần số cơ bản FO

Bước 1. Tần số cơ bản F0 chỉ được xác định trên khung (frame) hữu thanh của âm tiết. Trong tiếng nói tiếng Việt, các âm vị f, s, ş có số lần đổi dấu (zero-crossing rate) trong khung khá lớn nên dễ dàng kết luận khung đó là vô thanh và bỏ qua khung đó, đối với khung có năng lượng quá nhỏ ta cũng bỏ qua.

Bước 2. Tần số cơ bản F0 của người nam trong khoảng 80 - 200 Hz và của người nữ trong khoảng 120 - 240 Hz nên tiếng nói được tiền nhấn bằng bộ lọc

$$y[n] = s[n] - as[n-1] \tag{1}$$

với $a = -0{,}93$ để làm nổi tần số thấp. Bước 2 có thể có hay không trong hệ thống.

Bước 3. Tín hiệu được lọc thông thấp có tần số cắt bằng 500 Hz với đáp ứng tần số bằng phẳng nhất để giảm bớt ảnh hưởng của các formant cao hơn và thành phần tần số cao nhưng vẫn đảm bảo tồn tại tần số cơ bản lớn nhất.

Bước 4. Để giữ lại các đỉnh lớn trên tín hiệu, biên độ tín hiệu được giữ nguyên nếu trị tuyệt đối biên độ lớn hơn ngưỡng và gán bằng zero nếu nhỏ hơn ngưỡng. Chia khung đang khảo sát thành bốn phần, xác định trị tuyệt đối biên độ lớn nhất trên mỗi phần, sắp tăng dần bốn giá trị này và gọi là max1, max2, max3, max4. Nếu max2 $>$ 0,9 max4 thì max = max4, ngược lại max = max2, ngưỡng được chọn bằng 0,7 max. Ngưỡng này thích hợp cho những khung trong miền chuyển tiếp giữa hai âm vị vì nó vẫn giữ được những đỉnh nhỏ khi trong khung vừa có những đỉnh nhỏ của âm vị trước, vừa có những đỉnh lớn của âm vị sau trong một âm tiết.

Bước 5. Tiếp theo tín hiệu được đưa đến hàm hiệu biên độ trung bình (average magnitude difference function - AMDF) [3, 8]

$$d(p) = \sum_{n=0}^{N-1} |s(n) - s(n+p)| \tag{2}$$

với $s[n]$ là tín hiệu sau xử lý ngưỡng, N là độ dài của khung và p được lấy trong khoảng pitch từ 80 đến 200 tương ứng với tần số cơ bản 80 - 200 Hz. Chọn điểm cực tiểu d(P0) rồi suy ra P0 là chu kì pitch hay tần số cơ bản F0 = 16000/P0, ở đó tần số lấy mẫu là 16 kHz. Đối với các khung có d(P0) $> 0{,}7\, d_{\max}(p)$ được phân loại là vô thanh và gán F0 = 0.

Bước 6. Sau khi đã xác định F0 của các khung trong toàn bộ âm tiết, ta cần xử lí các khung có F0 = 0. Nếu các khung là vô thanh ở đầu hay cuối âm tiết thì thay F0 của các khung đó bằng F0 của khung hữu thanh kế cận. Nếu các khung là vô thanh ở giữa âm tiết thì thay F0 của các khung đó bằng giá trị cách đều F0 của hai khung biên hữu thanh.

Bước 7. Cuối cùng đường nét tần số cơ bản được làm trơn bằng lọc trung vị (median filter) k với $k = 5$ hay lọc trung bình di chuyển có trọng (weighted moving average filter) với đáp ứng xung h = [0,1 0,2 0,4 0,2 0,1] cho âm tiết có độ dài lớn hơn 10 khung.

2.2. Làm mịn và chuẩn hóa đường nét F0

Do việc thử nghiệm được tiến hành trên mạng nơron có số nút ở lớp đầu vào cố định nên

ta cần chuẩn hóa độ dài đường nét F0 của âm tiết thành cố định. Gọi L là độ dài cố định, l là độ dài của đường nét F0, khi đó

$$x[i] = \frac{L-1}{l-1} i\,, \quad i = 0, \ldots, l-1. \tag{3}$$

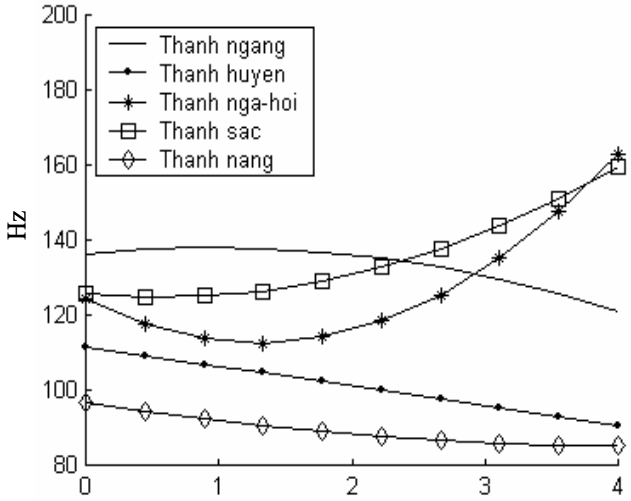
Làm mịn đường nét F0 bằng đa thức bậc hai $ax^2 + bx + c$ là phù hợp nhất vì khuynh hướng vận động của F0 là ngang, xuống, lên và xuống rồi lên. Tối thiểu hóa sai số trung bình bình phương

$$E = \sum_{i=0}^{l-1} [ax^2[i] + bx[i] + c - \mathrm{F0}[i]]^2 \tag{4}$$

sẽ dẫn đến giải hệ phương trình đại số tuyến tính 3 ẩn để xác định a, b và c. Lấy L giá trị của L điểm nội suy cách đều trên đường mịn $\mathrm{F0}_a[i]$ với $i = 0, \ldots, L-1$ và L giá trị đạo hàm tại các điểm đó làm đặc điểm. L giá trị của đường mịn tiếp tục được chuẩn hóa giữa 0 và 1 rồi chuyển sang thang đê-xi-ben để tăng độ phân biệt

$$\mathrm{F0}_n[i] = -20 \log_{10} \Big(\frac{\mathrm{F0}_a[i] - \min + \Delta}{\max - \min} \Big), \quad i = 0, \ldots, L-1, \tag{5}$$

ở đây min và max là giá trị nhỏ nhất và lớn nhất của $\mathrm{F0}_a$ trong toàn bộ dữ liệu, Δ là số dương đủ nhỏ để tránh $\log 0$. L giá trị đạo hàm cũng thực hiện tương tự. Như vậy mỗi thanh điệu được biểu diễn bằng vectơ đặc điểm có $2L$ hệ số. Hình 1 là năm thanh điệu tiếng Việt của người miền Nam không phân biệt ngã hỏi đã làm mịn bằng đường bậc hai.



Hình 1. Năm thanh điệu tiếng Việt của người miền Nam

Ta thấy F0 của thanh ngang cao, tương đối bằng phẳng và hơi đi xuống. Thanh huyền đi xuống đều đặn nhưng thấp hơn thanh ngang. Thanh ngã-hỏi cao hơn thanh huyền, đi xuống rồi đi lên. Thanh sắc cao hơn thanh huyền và đi lên. Thanh nặng thấp nhất và đi xuống nhanh. Đây là nhận xét chung sự vận động đường nét F0, thực tế do biến dạng trong quá trình phát âm nên đường nét F0 của thanh này lẫn với thanh kia là điều bình thường.

3. KẾT QUẢ THỬ NGHIỆM

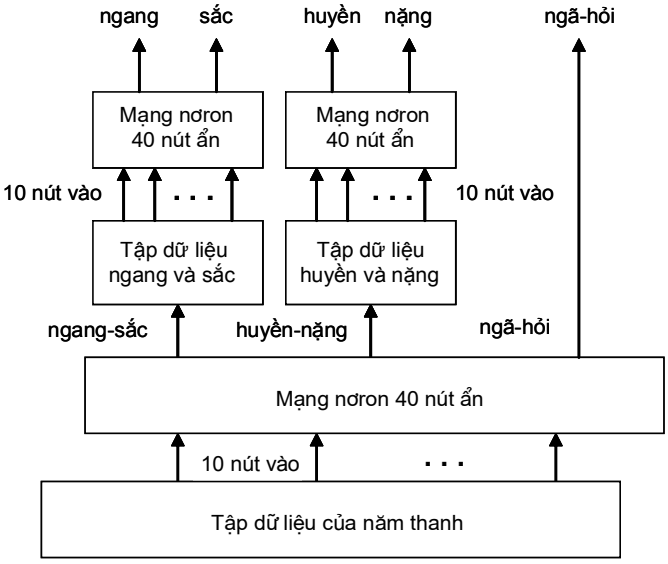
Tiếng nói được phát âm bằng giọng nam của người miền Nam trong môi trường văn phòng có tần số lấy mẫu 16 kHz, độ dài khung 1200 mẫu, độ dời 300 mẫu cho âm tiết dài và 150 mẫu cho âm tiết ngắn. Dữ liệu học gồm 300 câu thơ đầu tiên trong Truyện Kiều của Nguyễn Du và thêm 120 thanh nặng được chọn từ các câu thơ thứ 1000 trở đi, trong đó mỗi câu thơ chỉ ghi một lần. Tổng cộng gồm 708 thanh ngang, 559 thanh huyền, 209 thanh ngã-hỏi, 387 thanh sắc và 339 thanh nặng sau khi đã loại bỏ các thanh sắc và nặng có độ dài quá ngắn do các âm tiết bắt đầu và kết thúc bằng các âm vị như s, ş, f, k, c, t, b, p, chẳng hạn *sắc, bậc, cách, xấp, cập*.... Vectơ đặc điểm của mỗi thanh điệu gồm 10 hệ số logarit với $L = 5$ giá trị của đường mịn và 5 giá trị đạo hàm. Mạng nơron được chọn là mạng perceptron 3 lớp có 10 nút ở lớp đầu vào, 40 nút ở lớp ẩn, cập nhật trọng số khi đã duyệt qua toàn bộ mẫu học, số lần lặp 2000 và được tổ chức như sau:

Bước 1. Chọn 5 nút ở lớp đầu ra tương ứng với 5 thanh điệu. Kết thúc quá trình học, chúng tôi nhận dạng lại chính các mẫu đã học thì thấy tỷ lệ nhận dạng nhầm thanh sắc sang thanh ngang là 15,2% và tỷ lệ nhận dạng nhầm thanh nặng sang thanh huyền là 20,4%, do đó chúng tôi quyết định không phân biệt thanh ngang với sắc, thanh huyền với nặng. Kết quả ta chỉ còn 3 thanh: ngang-sắc, huyền-nặng và ngã-hỏi.

Bước 2. Mạng nơron bây giờ có 3 nút ở lớp đầu ra, sau khi học và nhận dạng lại các mẫu đã học, ta vẫn thấy một số mẫu bị nhận dạng nhầm, đây là những mẫu không tốt do biến dạng khi phát âm và được xóa khỏi tập mẫu học. Lặp lại Bước 2 nhiều lần cho đến khi tất cả các mẫu học được nhận dạng đúng, khoảng 1% mẫu đã bị xóa bỏ.

Bước 3. Tạo lập mạng nơron 2 nút đầu ra để nhận dạng thanh ngang và thanh sắc và cũng loại bỏ các mẫu xấu.

Bước 4. Tương tự, tạo lập mạng nơron 2 nút đầu ra để nhận dạng thanh huyền và thanh nặng và cũng loại bỏ các mẫu xấu.



Hình 2. Mạng nơron phân tầng

Như vậy ta có mạng nơron phân tầng gồm ba mạng nơron. Mạng thứ nhất dùng để nhận dạng 3 thanh: ngang-sắc, huyền-nặng và ngã-hỏi. Mạng thứ hai dùng để nhận dạng 2 thanh:

ngang và sắc sau khi mạng thứ nhất đã nhận dạng mẫu thử là thanh ngang-sắc. Mạng cuối cùng dùng để nhận dạng 2 thanh: huyền và nặng sau khi mạng thứ nhất đã nhận dạng mẫu thử là thanh huyền-nặng. Hình 2 minh họa ba mạng nơron để nhận dạng 5 thanh điệu.

Dữ liệu thử gồm 200 câu thơ từ câu thứ 301 đến 500 trong Truyện Kiều, mỗi câu thơ chỉ ghi một lần, gồm 467 thanh ngang, 372 thanh huyền, 154 thanh ngã-hỏi, 246 thanh sắc và 146 thanh nặng. Bảng 1 là kết quả nhận dạng cho trường hợp sử dụng một mạng nơron và Bảng 2 cho trường hợp mạng nơron phân tầng tính bằng tỷ lệ phần trăm, trong đó kết quả được trình bày theo hàng ngang, chẳng hạn khi thử nghiệm thanh ngang, ta thấy tỉ lệ nhận dạng đúng thanh ngang là 99,4% và nhầm sang thanh sắc là 0,4%, nhầm sang thanh nặng là 0,2%.

Bảng 1. Kết quả nhận dạng dùng một mạng nơron

Thanh	Ngang	Huyền	Ngã-Hỏi	Sắc	Nặng
Ngang	99,4			0,4	0,2
Huyền		92,5	0,5		7,0
Ngã-Hỏi	1,3		94,8	2,6	1,3
Sắc	10,2	0,4	1,2	88,2	
Nặng	1,4	14,4			84,2
Độ chính xác trung bình: 91,8					

Bảng 2. Kết quả nhận dạng dùng mạng nơron phân tầng

Thanh	Ngang	Huyền	Ngã-Hỏi	Sắc	Nặng
Ngang	99,2			0,4	0,4
Huyền		91,6	0,6		7.8
Ngã-Hỏi	1,0		96,1	1,9	1,0
Sắc	8,5	0,1	0,3	91,1	
Nặng	0,6	14,3			85,1
Độ chính xác trung bình: 92,6					

4. NHẬN XÉT VÀ KẾT LUẬN

Độ chính xác của thanh nặng được cải tiến từ 84,2% lên 85,1% và thanh sắc từ 88,2% lên 91,1% do được xử lý riêng. Ngoài ra, khi học thanh nặng bị xóa bỏ nhiều nhất nên khi thử độ chính xác của thanh nặng sẽ kém nhất. Kết quả nhận dạng phản ánh đúng đặc trưng đường nét thanh điệu của tiếng Việt, trong đó thanh huyền và thanh nặng đều có F0 thấp và đi xuống nên độ phân biệt hai thanh này sẽ không rõ ràng, tương tự thanh sắc có F0 cao và đi lên, nhưng do biến dạng khi phát âm nên nếu không đi lên sẽ lẫn với thanh ngang. Thanh ngã-hỏi có hướng đi xuống rồi đi lên, khác biệt với các thanh còn lại nên có độ chính xác cao, nhưng nếu bị biến dạng không có hướng đi xuống mà chỉ còn hướng đi lên sẽ lẫn với thanh sắc. Độ chính xác của thanh ngang cao nhất 99,2% do đường nét bằng phẳng, dễ phân biệt với hướng đi lên của thanh ngã-hỏi và sắc, và hướng đi xuống của thanh huyền và nặng.

Mặc dù dữ liệu học và thử của chúng tôi khác với [1] và [4], tuy nhiên nó khách quan hơn so với dữ liệu của [1]. Tỷ lệ lỗi xấu nhất của [1] là thanh huyền 47% còn của chúng tôi là thanh nặng 14,3%. So với [4] thì độ chính xác tương đương.

Kết quả thử nghiệm có độ chính xác cao 92,6% cho thấy hướng giải quyết vấn đề gồm làm mịn đường nét F0 của thanh điệu bằng đa thức bậc hai và chuẩn hóa, kết hợp 3 mạng nơron, đồng thời loại bỏ các mẫu xấu là thích hợp cho nhiệm vụ nhận dạng thanh điệu tiếng Việt.

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* **19** (2003) 131–138.

[2] Đoàn Thiện Thuật, *Ngữ âm tiếng Việt*, NXB Đại học và Trung học chuyên nghiệp, 1980.

[3] F. J. Owens, *Signal Processing of Speech*, Macmillan, London, 1993.

[4] Quoc-Cuong Nguyen, E. Castelli, Ngoc-Yen Pham, Tone Recognition for Vietnamese, *http://herakles.imag.fr/castelli/masters-ts/Eurospeech.pdf.*

[5] S. Potisuk, M. P. Harper, J. Gandour, Classification of Thai tone sequences in syllable-segmented speech using the Analysis-by-Synthesis method, *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 1 (1999).

[6] S.-H. Chen, Y.-R. Wang, Tone recognition of continuous mandarin speech based on neural networks, *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 2 (1995).

[7] T. Lee, P. C. Ching, Cantonese syllable recognition using neural networks, *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 4 (1999).

[8] W. Hess, *Pitch Determination of Speech Signals*, Springer-Verlag, 1983.

Nhận bài ngày 14 - 1 - 2003
Nhận lại sau sửa ngày 17 - 11 - 2003