XÁC ĐỊNH CÁC DẤU CÂU CỦA MỘT BÀI NHẬN DẠNG

Trần Anh Tuấn*, Nguyễn Hữu Mộng+, Nguyễn Trọng Khánh++
*Trường Cao đẳng nghề Công nghiệp Thanh Hóa
+Học Viện Kỹ Thuật Quân sự
++Học Viện Công Nghệ Bưu Chính Viễn Thông

Tóm tắt: Trong một bài nhận dạng tiếng Việt được viết lại từ một bài nói thì chưa có một dấu câu nào ngoài dấu cách. Do đó để có một bài viết hoàn chỉnh ta phải xác định và đặt các dấu câu cần thiết cho bài nhận dạng. Bài báo này sẽ khảo sát tình hình sử dụng dấu câu trong tiếng Việt, khảo sát sự phụ thuộc của các dấu câu vào độ dài các khoảng lặng tương ứng và đưa ra được khoảng độ dài của từng khoảng lặng tương ứng với từng dấu câu. Từ đó bài báo đề xuất các thuật toán xác định dấu câu cho các trường hợp bài nói đều và bài nói không đều.

Từ khóa: Tiếng nói tiếng Việt, Dấu câu, Khoảng lặng, Tốc độ nói, Nói đều; Nói không đều, Thuật toán; Nhận dạng, Nhận dạng dấu câu.

I. GIỚI THIỆU

1.1. Tổng quan

Trong lĩnh vực nhận dạng tiếng nói tiếng Việt, trên cơ sở từ một tệp âm thanh là một bài nói được viết lại thành một bài viết. Bài viết này ta gọi là một bài nhận dạng. Đối với một bài nhận dạng tiếng nói tiếng Việt thì chưa có bất kỳ một dấu câu nào ngoài dấu cách, bởi vì khi người ta nói thì không ghi lại được các dấu câu. Do đó, để có được một bài viết hoàn chỉnh thì bắt buộc ta phải xác định và đặt các dấu câu cần thiết cho bài nhận dạng.

Trong một bài nói các dấu câu ẩn tại các khoảng lặng. Từ các cách nói của người Việt ta thấy các dấu câu có liên hệ mật thiết với độ dài các khoảng lặng, tức là các dấu câu khác nhau ứng với các khoảng lặng sẽ có độ dài khác nhau. Để xác định khoảng lặng nào là dấu câu nào thì ta phải xác định được độ dài các khoảng lặng, nếu tính được độ dài tương đối của khoảng lặng thì có thể nhận dạng được đó là dấu câu nào. Để thuận tiện, ta coi độ dài khoảng lặng tương ứng dấu câu là độ dài dấu câu. Trong một đoạn văn bản với tốc độ nói đồng đều thì độ dài các dấu câu được sắp xếp theo một trật tự nhất định, còn trong một đoạn văn bản có tốc độ nói không đều thì thứ tự độ dài các dấu câu có thể thay đổi. Ví dụ, độ dài dấu phẩy ở các đoạn khác nhau với tốc độ nói khác nhau là khác nhau. Chính xác hơn, có hai đoạn văn bản với hai tốc độ nói khác nhau thì độ dài dấu phẩy ở hai đoạn khác nhau và đều nhỏ hơn độ dài dấu chấm tương ứng, nhưng độ dài dấu phẩy ở đoạn này có thể lớn hơn độ dài dấu chấm ở đoạn kia.

Trong các nghiên cứu trước đây [1,2] chúng tôi đã đề xuất các phương pháp tách tiếng, khoảng lặng và áp dụng phương pháp này tính tốc độ nói, độ dài các khoảng lặng. Trong bài báo này chúng tôi xem xét các đặc trưng của các dấu câu trong tiếng Việt, khảo sát sự phụ thuộc của các dấu câu vào độ dài các khoảng lặng tương ứng trong từng đoạn nói đều. Từ các khảo sát thực tế chúng tôi rút ra được khoảng độ dài của từng khoảng lặng tương ứng với từng dấu câu, thống kê tình trạng sử dụng các dấu câu trong các bài nói, bài viết tiếng Việt. Chúng tôi đề xuất thuật toán xác định các dấu câu cho bài nói đều và bài nói không đều.

1.2. Đóng góp chính của bài báo

Trong bài báo này chúng tôi tiến hành khảo sát tình trạng sử dụng các dấu câu trong tiếng nói tiếng Việt; khảo sát dấu câu và độ dài các khoảng lặng tương ứng, đưa ra được các khoảng dấu câu ứng với tốc độ nói khác nhau, từ đó chúng tôi đã đề xuất các thuật toán xác định dấu câu của bài nói đều và bài nói không đều.

1.3. Cấu trúc bài báo

Trong bài báo này, chúng tôi tổ chức nội dung như sau. Các nghiên cứu liên quan được trình bày trong mục 2. Trong mục 3 trình bày các dấu câu trong tiếng nói tiếng Việt và tình trạng sử dụng. Mục 4 trình bày dấu câu và khoảng lặng. Các thuật toán xác định dấu câu cho bài nói và kết quả thực nghiệm trình bày trong mục 5. Kết luận, dự kiến xu hướng nghiên cứu được đưa ra trong mục 6.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Dấu câu là một vấn đề quan trọng trong xử lý ngôn ngữ và lời nói. Do đó, trong những năm qua, đã có nhiều công trình nghiên cứu tập trung vào chủ đề khôi phục hoặc dự đoán các dấu câu cho các ngôn ngữ phổ biến như tiếng Anh, tiếng Pháp, tiếng Đức, Trung, Bồ Đào Nha,... Hiện nay có ba cách tiếp cận để dự đoán dấu câu về mặt công nghệ và mô hình.

Đầu tiên, dấu câu được coi là các sự kiện liên từ ẩn [3]. Các mô hình ngôn ngữ n-gram [4] hoặc các mô hình Markov ẩn [5] được sử dụng để khôi phục các dấu câu: A.Gravano và cộng sự [4] đã sử dụng các mô hình ngôn ngữ n-gram để khôi phục dấu câu và viết hoa trên văn bản tiếng Anh; H. Christensen và cộng sự [5] trình bày một mô hình trạng thái hữu hạn thống kê kết hợp giữa các đặc trưng âm thanh, ngôn ngữ và dấu câu.

Hai là, dự đoán dấu câu được xem như một công việc ghi nhãn trình tự [6,7], trong đó dấu câu được gán cho mỗi từ: Kol, J., Lamel, L. [8] phát triển các mô hình dấu

Tác giả liên hệ: Trần Anh Tuấn,
Email: tuankhhtqt@gmail.com
Đến tòa soạn: 26/2/2021, chỉnh sửa: 9/5/2021, chấp nhận đăng: 19/5/2021

câu tự động cho tin tức phát sóng và các cuộc hội thoại phát sóng bằng tiếng Pháp và tiếng Anh. Cả hai tính năng âm thanh và văn bản được kết hợp và thử nghiệm; Batista et al. [9,10] sử dụng phương pháp mô hình hóa entropy tối đa, kết hợp các loại thông tin khác nhau cả từ vựng và âm thanh để phục hồi dấu chấm câu trên tin tức phát sóng Bồ Đào Nha; Lu và Ng [11] đã sử dụng các mô hình CRF (Conditional random fields) nhằm đạt được hiệu suất tốt hơn cho nhiệm vụ xác định dấu câu trên cả tập dữ liệu tiếng Anh và tiếng Trung. Tương tự, Zhao và cộng sự [12] đã điều tra dự đoán dấu câu của Trung Quốc bằng cách ghi nhãn nhiều lần và áp dụng mô hình CRF.

Gần đây, Che et al. [13] đề xuất sử dụng mạng nơ ron học sâu và mạng nơ ron tích chập để dự đoán dấu câu. Kết quả cho thấy phương pháp dựa trên mạng nơ ron thần kinh vượt trội hơn phương pháp dựa trên CRF. Tilk et al. [14] sử dụng bộ nhớ ngắn hạn và mạng thần kinh tái phát hai chiều với cơ chế tập trung để cải thiện hiệu suất dự đoán dấu câu.

Ba là, khôi phục dấu câu được coi là một vấn đề dịch máy đơn ngữ trong đó đầu vào là văn bản không có dấu câu và kết quả đầu ra là văn bản có dấu câu: Cho et al. [15] đã nghiên cứu dự đoán dấu câu cho tiếng Đức-Anh với một hệ thống dịch thuật đơn ngữ và chứng minh kết quả của họ trong các thí nghiệm. Klejch và cộng sự [16,17] đề xuất kiến trúc bộ giải mã-mã hóa RNN (Recurrent Neural Networks) để khôi phục dấu câu. Kiến trúc này tương tự như mô hình được sử dụng cho nhiệm vụ dịch máy. Mặc dù kết quả cho thấy các phương pháp được đề cập ở trên là hiệu quả, Tuy nhiên vẫn còn nhiều điều cần cải thiện.

Ở Việt Nam, việc xây dựng hệ thống tự động dự đoán dấu câu trong văn bản tiếng Việt mới được quan tâm nghiên cứu gần đây, nhóm tác giả Quang H. Pham, Binh T. Nguyen. [18] đã xây dựng hệ thống dự đoán dấu câu đầu tiên của Việt Nam dựa trên các trường ngẫu nhiên có điều kiện, chuỗi tuyến tính (CRFs).

Mới đây nhất, nhóm tác giả Pham T., Nguyen N., Pham Q., Cao H., Nguyen B.[19] đã sử dụng mô hình CRF và phương pháp học sâu để xây dựng hệ thống dự đoán dấu câu cho tiếng Việt dựa trên hai bộ dữ liệu quy mô lớn.

III. CÁC DẤU CÂU TRONG TIẾNG NÓI TIẾNG VIỆT VÀ TÌNH TRẠNG SỬ DỤNG

Chúng tôi khảo sát sự sử dụng các dấu câu sau đây trong văn bản tiếng Việt: phẩy, chấm, chấm phẩy, hai chấm, gạch nối, chấm than, hỏi, cặp ngoặc đơn, ba chấm. Để khảo sát sự sử dụng các dấu câu trên, chúng tôi thiết kế một phần mềm thống kê. Để thể hiện mức độ sử dụng các dấu câu chúng tôi tính tần suất của từng dấu câu trong bài viết (nói).

Chúng tôi xem xét hai loại tần suất: tần suất theo số tiếng (f_{tieng}), và tần suất theo số dấu câu (f_{dau}). Tần suất theo số tiếng bằng số lần sử dụng dấu câu trên tổng số tiếng, tần suất theo dấu câu bằng số lần sử dụng trên tổng số dấu câu. Thực hiện cho nhiều bài viết khác nhau chúng tôi thu được bảng thống kê sau đây.

Bảng I. Thống kê tình trạng sử dụng các dấu câu trong tiếng Việt

TT	Văn bản	Tổng số tiếng	Số lượng dấu câu									
			,	.	;	:	-	!	?	" "	( )	…
1	Văn bản mô tả	85	12	3	0	0	0	0	0	0	0	0
		f_{tieng}	14,1%	3,5%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
		f_{dau}	80,0%	20,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
2	Văn bản tường thuật	203	19	8	0	0	0	0	0	1	0	0
		f_{tieng}	9,4%	3,9%	0,0%	0,0%	0,0%	0,0%	0,0%	0,5%	0,0%	0,0%
		f_{dau}	67,9%	28,6%	0,0%	0,0%	0,0%	0,0%	0,0%	3,6%	0,0%	0,0%
3	Văn bản giáo dục	169	13	7	0	0	0	0	0	0	1	0
		f_{tieng}	7,7%	4,1%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,6%	0,0%
		f_{dau}	61,9%	33,3%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	4,8%	0,0%
4	Bài báo pháp luật	618	37	11	6	1	11	0	6	1	0	1
		f_{tieng}	6,0%	1,8%	1,0%	0,2%	1,8%	0,0%	1,0%	0,2%	0,0%	0,2%
		f_{dau}	50,0%	14,9%	8,1%	1,4%	14,9%	0,0%	8,1%	1,4%	0,0%	1,4%
5	Bài báo giáo dục	477	22	19	0	0	5	0	0	3	1	0
		f_{tieng}	4,6%	4,0%	0,0%	0,0%	1,0%	0,0%	0,0%	0,6%	0,2%	0,0%
		f_{dau}	44,0%	38,0%	0,0%	0,0%	10,0%	0,0%	0,0%	6,0%	2,0%	0,0%
6	Bài báo đời sống	533	40	24	0	0	0	0	0	5	2	0
		f_{tieng}	7,5%	4,5%	0,0%	0,0%	0,0%	0,0%	0,0%	0,9%	0,4%	0,0%
		f_{dau}	56,3%	33,8%	0,0%	0,0%	0,0%	0,0%	0,0%	7,0%	2,8%	0,0%
7	Bài phát biểu	3061	297	65	21	3	3	0	0	8	1	1

	hội nghị	f_{tieng}	9,7%	2,1%	0,7%	0,1%	0,1%	0,0%	0,0%	0,3%	0,0%	0,0%
		f_{dau}	74,4%	16,3%	5,3%	0,8%	0,8%	0,0%	0,0%	2,0%	0,3%	0,3%
8	Bài phát biểu giáo dục	424	34	6	0	0	4	5	0	0	0	0
		f_{tieng}	8,0%	1,4%	0,0%	0,0%	0,9%	1,2%	0,0%	0,0%	0,0%	0,0%
		f_{dau}	69,4%	12,2%	0,0%	0,0%	8,2%	10,2%	0,0%	0,0%	0,0%	0,0%
9	Bài phát biểu của Thủ tướng	2002	89	98	6	7	0	5	5	6	0	2
		f_{tieng}	4,4%	4,9%	0,3%	0,3%	0,0%	0,2%	0,2%	0,3%	0,0%	0,1%
		$f_{dấu}$	40,8%	45,0%	2,8%	3,2%	0,0%	2,3%	2,3%	2,8%	0,0%	0,9%
10	Giáo trình kỹ thuật	11207	436	470	2	174	80	0	1	0	114	3
		f_{tieng}	3,9%	4,2%	0,0%	1,6%	0,7%	0,0%	0,0%	0,0%	1,0%	0,0%
		f_{dau}	34,1%	36,7%	0,2%	13,6%	6,3%	0,0%	0,1%	0,0%	8,9%	0,2%
11	Giáo trình xã hội	12757	604	449	51	98	51	0	12	16	29	18
		f_{tieng}	4,7%	3,5%	0,4%	0,8%	0,4%	0,0%	0,1%	0,1%	0,2%	0,1%
		f_{dau}	45,5%	33,8%	3,8%	7,4%	3,8%	0,0%	0,9%	1,2%	2,2%	1,4%
12	Giáo trình giáo dục	39462	1650	1782	69	217	439	2	16	6	294	114
		f_{tieng}	4,2%	4,5%	0,2%	0,5%	1,1%	0,0%	0,0%	0,0%	0,7%	0,3%
		f_{dau}	36,0%	38,8%	1,5%	4,7%	9,6%	0,0%	0,3%	0,1%	6,4%	2,5%

Từ bảng thống kê trên đây ta thấy rõ mức độ sử dụng các dấu câu. Dấu phẩy sử dụng nhiều nhất rồi đến dấu chấm. Các dấu khác sử dụng rất ít, ít nhất là các dấu ba chấm, ngoặc đơn. Trong một bài nói phổ biến chỉ có các dấu câu sau đây xuất hiện: phẩy, chấm, hỏi, than. Do đó ta chỉ xét bài nói với các dấu câu này.

IV. DẤU CÂU VÀ KHOẢNG LẶNG

Xác định các dấu câu trong một bài nói, tức là xác định khoảng lặng nào là dấu câu nào. Khi xét các dấu câu ta cũng phải xét cả dấu cách vì về mặt âm thanh dấu cách cũng là một khoảng lặng như các dấu câu. Theo thống kê thực tế như trong mục 3 ta chỉ xét các dấu câu sau: dấu cách, dấu phẩy, dấu chấm, dấu hỏi và dấu than. Do đó, một bài viết bất kỳ hoàn toàn có thể không cần dùng các dấu câu còn lại.

Khảo sát dấu câu và độ dài các khoảng lặng tương ứng ta thấy độ dài các dấu câu là khác nhau nếu người nói chuẩn phổ thông, tức là nói bình thường, không kéo dài, không bỏ âm. Tuy nhiên sự khác biệt này là tương đối. Trong hai đoạn có tốc độ nói khác nhau thì độ dài của dấu phẩy của đoạn này có thể lớn hơn độ dài của dấu chấm ở đoạn kia, trong khi đó, ở cùng một đoạn thì độ dài dấu chấm luôn lớn hơn độ dài dấu phẩy. Chính vì vậy để nhận dạng các dấu câu ta phải xác định trước các nhóm tốc độ với các ngưỡng độ dài của các dấu câu.

Thực hiện các thống kê nhiều lần cho 120 bài nói nói phổ thông (giọng nam miền Bắc: 60 bài; giọng nữ miền Bắc: 60 bài) với từng tốc độ nói khác nhau ta thu được các khoảng độ dài của các dấu câu như các bảng sau đây.

Bảng II. Thống kê khoảng độ dài của các dấu câu với tốc độ nói v=2,7 âm tiết/giây

STT	Dấu câu	Nhỏ nhất	Lớn nhất
1.	Dấu cách	0,150162	0,155888
2.	Dấu phẩy	0,251334	0,258715
3.	Dấu chấm	0,306226	0,309851
4.	Dấu hỏi	0,310510	0,314882
5.	Dấu than	0,315840	0,317012

Bảng III. Thống kê khoảng độ dài của các dấu câu với tốc độ nói v=3,8 âm tiết/giây

STT	Dấu câu	Nhỏ nhất	Lớn nhất
1.	Dấu cách	0,0898624	0,090011
2.	Dấu phẩy	0,1316931	0,133849
3.	Dấu chấm	0,1890556	0,189282
4.	Dấu hỏi	0,1896322	0,189865
5.	Dấu than	0,1899563	0,189974

Bảng IV. Thống kê khoảng độ dài của các dấu câu với tốc độ nói v=5,2 âm tiết/giây

STT	Dấu câu	Nhỏ nhất	Lớn nhất
1.	Dấu cách	0,011051	0,011632
2.	Dấu phẩy	0,061963	0,066014
3.	Dấu chấm	0,094658	0,095337
4.	Dấu hỏi	0,097008	0,097225
5.	Dấu than	0,098977	0,099663

V. THUẬT TOÁN XÁC ĐỊNH CÁC DẤU CÂU

5.1. Nói đều và nói không đều

Con người khi nói thì có lúc nhanh, lúc chậm và nói chung với tốc độ khác nhau. Tốc độ nói đã được phân tích và tính toán trong [2]. Để áp dụng các thuật toán đã đề xuất cho việc xác định các dấu câu trong một bài nói ta cần phải phân biệt hai dạng bài nói: nói đều và nói không đều. Ta thống nhất rằng, một bài nói đều là bài nói mà tốc độ nói của các đoạn là cùng nhóm trong số ba nhóm tốc độ nói mà ta đã đề cập ở bài viết trước đây [2]. Tốc độ của cả bài nói là tốc độ trung bình của của các đoạn. Bài nói không phải là bài nói đều là một bài nói không đều. Một bài nói không đều có các đoạn với tốc độ nói khác nhau. Tuy nhiên ta hiểu rằng, trong một bài nói có thể có nhiều đoạn có cùng tốc độ nói trung bình cùng nhóm, và tất nhiên, chúng là những đoạn không liền nhau. Nếu có hai đoạn liền nhau mà có chung nhóm tốc độ nói thì gộp chúng thành một.

5.2. Thuật toán xác định dấu câu trường hợp nói đều

Giả sử ta có một bài nói gồm nhiều đoạn ngắn và các đoạn đều có cùng tốc độ nói hay các tốc độ nói gần như nhau. Ta có thể hiểu trường hợp này như sau. Giả sử đoạn đầu tiên kéo dài trong 1 phút với tốc độ nói là *v*. Khi đó, nếu các đoạn 1 phút sau đó cũng có tốc độ gần *v* thì cả bài nói có thể coi là có tốc độ đều và tốc độ đều đó chính là tốc độ trung bình của tất cả các đoạn và gần bằng *v*.

Theo thống kê thì khi người việt nói chuẩn phổ thông, không kéo dài giọng, nói gọn rõ ràng rành mạch thì độ dài các dấu câu tăng dần như sau:

1- dấu cách. 2 - dấu phẩy. 3- dấu chấm. 4- dấu hỏi. 5- dấu than. Ở đây ta xét thêm dấu cách vì về mặt âm thanh dấu cách cũng là một khoảng lặng như các dấu câu.

Bài toán: Giả sử bài nói đều có tốc độ *v* cùng nhóm thống kê biết các khoảng dấu câu (a_1,b_1), (a_2,b_2), (a_3,b_3), (a_4,b_4), (a_5,b_5) tương ứng với các dấu câu phổ biến trên đây. Để xác định các dấu câu của bài nói với tốc độ đều, ta có thuật toán sau:

Thuật toán 1: Xác định các dấu câu của bài nói với tốc độ đều

Đầu vào: Bài nói đều (tệp âm thanh *.wav) và bài nhận dạng tương ứng.

Đầu ra: Bài nhận dạng với các dấu câu (bài nhận dạng hoàn chỉnh).

Thuật toán gồm hai giai đoạn: xác định dấu câu, đặt các dấu câu.

+ *Xác* định *các dấu câu của bài nói*:

1. Tách các tiếng của bài nói thu được hai dãy số:

$$tt_1, tt_2, ..., tt_{m+1};$$

$$tl_1, tl_2, ..., tl_m,$$

trong đó, *m* là số khoảng lặng, *m*+1 là số âm tiết của bài nói.

2. Tính tổng thời gian nói (thời gian của bài nói):

$$T = \sum_{k=1}^{m} (tt_k + tl_k) + tt_{m+1}. \quad (1)$$

3. Tính tốc độ trung bình của bài nói:

$$v = T / (m+1). \quad (2)$$

4. Xác định dải tốc độ (a,b) của bài nói và các cận $(c_1,d_1),(c_2,d_2),(c_3,d_3)$ của các dấu câu:

- Nếu $a_1 \le v \le b_1$ thì $a = a_1, b = b_1$ và

$$c_1 = a_{11}, d_1 = b_{11};$$
$$c_2 = a_{12}, d_2 = b_{12};$$
$$c_3 = a_{13}, d_3 = b_{13};$$
$$c_4 = a_{14}, d_4 = b_{14}.$$

- Nếu $a_2 \le v \le b_2$ thì $a = a_2, b = b_2$ và

$$c_1 = a_{21}, d_1 = b_{21};$$
$$c_2 = a_{22}, d_2 = b_{22};$$
$$c_3 = a_{23}, d_3 = b_{23};$$
$$c_4 = a_{24}, d_4 = b_{24}.$$

- Nếu $a_3 \le v \le b_3$ thì $a = a_3, b = b_3$ và

$$c_1 = a_{31}, d_1 = b_{31};$$
$$c_2 = a_{32}, d_2 = b_{32};$$
$$c_3 = a_{33}, d_3 = b_{33};$$
$$c_4 = a_{34}, d_4 = b_{34}.$$

5. Xác định các dấu câu $dc_1, dc_2, ..., dc_m$ của bài nói:

Với $k = 1,2,...,m$ kiểm tra:

$$\text{nếu } \begin{cases} c_1 \le tl_k \le d_1, & dc_k = ""; \\ c_2 \le tl_k \le d_2, & dc_k = ","; \\ c_3 \le tl_k \le d_3, & dc_k = "."; \\ c_4 \le tl_k \le d_4, & dc_k = "?". \end{cases}$$

+ Đặt các dấu câu cho bài nhận dạng là một văn bản tiếng Việt (*chưa có dấu câu*):

Với $k = 1,2,...,m$ thay khoảng trống thứ *k* bằng dấu câu dc_k.

{Ta có thể cài đặt vòng lặp này cùng vòng lặp trong bước 5 của giai đoạn xác định dấu câu, tức là, xác định được ở đâu thì đặt ở đó.}

6. Kết thúc.

5.3. Thuật toán xác định dấu câu trường hợp không đều

Ta thống nhất chỉ xét các bài nói với các đoạn khác nhau mà tốc độ nói của từng đoạn chỉ là một trong ba nhóm sau đây: chậm, phổ thông, nhanh. Theo thống kê ở [2] thì ba nhóm tốc độ đó là:

- Nhóm thứ nhất: Từ 2,5 âm tiết đến 2,7 âm tiết.

- Nhóm thứ hai: Từ 3,5 âm tiết đến 3,7 âm tiết.

- Nhóm thứ ba: Từ 4,5 âm tiết đến 4,7 âm tiết.

Các nhóm tốc độ này không kề nhau là do cách nói của người Việt: trong một đoạn nói đều thì tốc độ nói của từng khúc không khác nhau nhiều; còn với các đoạn khác nhau: chậm, phổ thông hay nhanh lại khác biệt rõ ràng.

Để tiện cho việc trình bầy ta kí hiệu ba nhóm tốc độ trên lần lượt là: V_1, V_2, V_3. Các cận của các nhóm ta kí hiệu là: v_{11}, v_{12}, v_{21}, v_{22}, v_{31}, v_{32}.

Theo các khảo sát đã thực hiện thì một đoạn nói đều cùng tốc độ có nghĩa là tốc độ các câu trong đoạn gần như nhau mà không phải lúc nào cũng bằng nhau, tuy nhiên khác nhau không đáng kể. Vì bài nói không có các dấu câu nên ta không thể xác định sơ bộ số lượng âm tiết của từng câu. Do vậy ta tạm coi một câu tượng trưng có 10 âm tiết để tiện cho việc phân loại tốc độ. Về nguyên tắc với 2 âm tiết thì cũng có thể tính được tốc độ nói (tất nhiên là tốc độ nói của bài nói 2 âm tiết).

Ta giả thiết thêm bài nói có ít nhất 10 âm tiết và nếu có nhiều hơn thì tốc độ trung bình của các đoạn đều thuộc vào các khoảng nêu trên.

Với các giả thiết trên đây, ta có thuật toán xác định dấu câu cho bài nói không đều sau:

Thuật toán 2: Xác định dấu câu của bài nói không đều

Đầu vào: Bài nói (tệp âm thanh *.wav) và bài nhận dạng (bài text tiếng việt không dấu câu)

Đầu ra: Bài viết (bài nhận dạng với các dấu câu)

1. Tách bài nói không đều thành các đoạn khác nhau mà mỗi đoạn có tốc độ trung bình thuộc một trong ba nhóm tốc độ như đưa ra trên đây. Hai đoạn kề nhau phải có tốc độ trung bình khác nhau.

2. Áp dụng thuật toán gán dấu câu bài nói đều cho từng đoạn đã tách.

3. Kết thúc.

Để tách bài nói thành nhiều đoạn mà mỗi đoạn là một bài nói với tốc độ trung bình thuộc một trong ba nhóm tốc độ nêu trên, trước tiên ta tách bài nói bằng thuật toán tách âm tiết và khoảng lặng [2], ta thu được dãy độ dài các âm tiết $tt_1, tt_2, \ldots, tt_{m+1}$ và độ dài các khoảng lặng là $tl_1, tl_2, \ldots, tl_m$. Như vậy ta có m+1 âm tiết.

Đặt $M = m+1$, T là một mảng n+1 phần tử, n là số đoạn của bài nói, V là mảng ghi lại nhóm tốc độ của các đoạn trong bài nói. Ta hiểu đoạn thứ nhất là các âm tiết từ T(1) đến T(2), đoạn thứ hai từ T(2)+1 đến T(3), …, đoạn cuối từ T(n-1)+1 đến T(n). Như vậy: T(1) = 1, T(2) = số âm tiết của đoạn thứ nhất, T(3) là tổng số âm tiết của đoạn thứ nhất và đoạn thứ hai, tương tự T(n) = m + 1. Ta kí hiệu V(n) là số hiệu nhóm tốc độ của các đoạn âm thanh bài nói. Nếu đoạn thứ i có tốc độ trung bình v_k thuộc nhóm thứ nhất thì ta gán V(i) = 1, nếu v_k thuộc nhóm thứ 2 thì V(i) = 2, nếu v_k thuộc nhóm thứ 3 thì V(i)= 3. Kết thúc thuật toán là đưa ra được các dãy T(n) và V(n). Vậy ta có thuật toán sau đây:

Thuật toán 3: Tách bài nói không đều thành các đoạn khác nhau

Đầu vào: Bài nói (tệp âm thanh *.wav)

Đầu ra: Các đoạn của bài viết với các nhóm tốc độ khác nhau (T(n), V(n))

1. $M = m$+1; n = 1; T(1) = 5; tách 10 âm tiết tính tốc độ trung bình v của 10 âm tiết và ghi lại khoảng tốc độ của 5 âm tiết này, tức là tính V(n); $M = M$ - 10.

2. Nếu M <= 5 thì tính tốc độ v của đoạn các âm tiết còn lại và tính nhóm tốc độ trung bình của nó và xác định nhóm tốc độ v_k của v.

+ Nếu v_k = V(n) thì T(n) = T(n) + M; chuyển sang bước 5.

+ Nếu $v_k \neq$ V(n) thì n= n+1; T(n) = T(n-1) + M; chuyển sang bước 5.

3. Nếu M > 10 thì tách 10 âm tiết; $M = M$ - 10. Tính tốc độ trung bình của 10 âm tiết vừa tách và tính nhóm tốc độ của 10 âm tiết là v_k.

+ Nếu v_k = V(n) thì T(n) = T(n) + 5.

+ Nếu $v_k \neq$ V(n) thì $n = n + 1$ và T(n) = T(n-1) + 5.

4. Nếu M <= 5 thì chuyển lên bước 2, còn nếu M > 5 chuyển lên bước 3.

5. Kết thúc.

5.4. Thực nghiệm

5.4.1. Phương pháp xây dựng bộ dữ liệu thực nghiệm

Trong nghiên cứu này, các bài nói được thu trong phòng thu âm, lồng tiếng chuyên nghiệp với hệ thống cách âm, lọc nhiễu tốt. Mỗi bài được lưu thành một file wav, tín hiệu thu được lấy mẫu ở tần số 16000Hz và 16 bit cho một mẫu.

Có 120 bài nói (giọng miền Bắc) được thu âm, gồm 60 nữ và 60 nam là các phát thanh viên, giáo viên được lựa chọn theo các tiêu chí: có độ tuổi từ 22 đến 50 tuổi, có phân bố cân bằng giữa giọng nam và giọng nữ, nói đều và nói không đều, có kinh nghiệm và biểu đạt tốt, nói rõ ràng, rành mạch. File dữ liệu thu xong được xử lý trước bằng cách sử dụng công cụ cắt bỏ hết khoảng lặng ở đầu và cuối bài nói.

Kịch bản thu âm được xây dựng gồm 120 bài viết tiếng Việt theo các tiêu chí sau:

- Kịch bản thu được thiết kế với ngữ cảnh để các phát thanh viên, giáo viên biểu đạt một cách rõ ràng, rành mạch nhất.

- Kịch bản thu được thiết kế với các bài nói đều và nói không đều.

5.4.2. Kết quả thực nghiệm

A. Bài nói đều: Thực hiện thuật toán 1 cho một bài nói với tốc độ đều ta được kết quả như bảng V.

Bài nhận dạng như sau: ”mỗi năm chúng ta sử dụng hàng triệu bao ni lông những cái bao ni lông dùng đó sẽ đi đâu trong khi nó không phân hủy được do đó chúng ta phải thiêu huỷ chúng và khi đó chúng ta sẽ tạo ra một lượng lớn khí thải độc hại ảnh hưởng đến môi trường và con người thật nguy hại”

Bảng V. Kết quả xác định dấu câu với bài nói đều

Khoảng lặng	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Dấu câu		,									.					
Khoảng lặng	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Dấu câu					?							,				
Khoảng lặng	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Dấu câu				.												
Khoảng lặng	49	50	51	52	53	54	55	56	57	58	59	60	61	62		
Dấu câu											.			!		

B. Bài nói không đều: Chọn một bài nói không đều gồm ba đoạn: nhanh, phổ thông, chậm và sử dụng thuật toán 2, thuật toán 3 ta được kết quả như bảng VI.

Bài nhận dạng như sau: “hè đến khi ve râm ran hát khúc ca trong từng tán lá hè đến khi hoa phượng đỏ thắm góc sân trường mùa hè đến mang theo bao sung sướng của tuổi học trò chúng tôi có thể thở phào nhẹ nhõm và háo hức chờ những chuyến đi chơi xa cùng gia đình khi những chú ve sầu bắt đầu hát ca những điệu nhạc du dương khi những chùm phượng vĩ nở đỏ rực cả góc sân trường đó là hình ảnh báo hiệu mùa hè mới bắt đầuôi mùa hè đang về đấy hầu hết học sinh ai cũng yêu mùa hè trong đó có tôi"

Bảng VI. Kết quả xác định dấu câu với bài nói không đều

Khoảng lặng	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Dấu câu													,			
Khoảng lặng	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Dấu câu							.									
Khoảng lặng	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Dấu câu			.													
Khoảng lặng	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
Dấu câu							.									
Khoảng lặng	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
Dấu câu					,											
Khoảng lặng	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
Dấu câu	,											.	!			
Khoảng lặng	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	
Dấu câu		!									,				.	

Nhận xét: Chúng tôi đã tiến hành thử nghiệm các thuật toán đề xuất trên đây với nhiều bài nói khác nhau trong các trường hợp nói đều và nói không đều. Các kết quả nhận được đều phù hợp với thực tế chính tả của các bài nói; các vị trí các khoảng lặng trong bài nhận dạng đều trùng khớp với vị trí các dấu câu trong bài viết hoàn chỉnh. Do khuôn khổ bài báo có hạn nên chúng tôi chỉ đưa ra một số minh họa như trên.

VI. KẾT LUẬN

Trong bài báo này, chúng tôi đã khảo sát tình trạng sử dụng các dấu câu trong tiếng nói tiếng Việt, đưa ra được các khoảng dấu câu ứng với tốc độ nói khác nhau, từ đó chúng tôi đã đề xuất các thuật toán xác định dấu câu của bài nói đều và bài nói không đều. Các thuật toán đề xuất không chỉ đạt được độ chính xác cao, mà còn đạt được tốc độ xử lý rất nhanh phù hợp với các ứng dụng thực tế. Tuy nhiên điểm hạn chế của thuật toán là chỉ xem xét trường hợp nói phổ biến dùng các dấu câu: dấu cách, dấu phẩy, dấu chấm, dấu hỏi, dấu than. Trong thời gian tới chúng tôi sẽ tiếp tục nghiên cứu cải thiện thuật toán để mở rộng xem xét hết các trường hợp nói khác nữa.

TÀI LIỆU THAM KHẢO

[1] Tuan Tran Anh, Mong Nguyen Huu, Khanh Nguyen Trong. A Method for Automatic Vietnamese Speech Segmentation. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-11, September 2019.

[2] Trần Anh Tuấn, Nguyễn Hữu Mộng, Nguyễn Trọng Khánh. “Tốc độ nói và các khoảng lặng trong bài nói tiếng Việt”. Tạp chí khoa học công nghệ Thông tin và Truyền thông, Học viện Công nghệ Bưu chính Viễn thông, ISSN: 2525-2224, Số 04B (CS.01)2020, tr. 92-100.

[3] E. Liu, Y. nd Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, “Enriching speech recognition with automatic detection of sentence boundaries and disfluencies,” IEEE Trans Audio Speech Language Process, vol. 14, no. 5, pp. 1526–1540, 2006.

[4] A. Gravano, M. Jansche, and M. Bacchiani, “Restoring punctuation and capitalization in transcribed speech,” in ICASSP, 2009, pp. 4741–4744.

[5] Heidi Christensen, Yoshihiko Gotoh, and Steve Renals, “Punctuation annotation using statistical prosody models,” Proc Isca Workshop on Prosody in Speech Recognition and Understanding, pp. 35–40, 2001.

[6] N. Ueffing, M. Bisani, and P. Vozila, “Improved models for automatic punctuation prediction for spoken and written text,” in INTERSPEECH, 2013, pp. 3097–3101.

[7] P. elasko, P. Szymaski, J. Mizgajski, A. Szymczak, Y. Carmiel, and N. Dehak, “Punctuation prediction model for

conversational speech," in INTERSPEECH, 2018, pp. 2633–2637.

[8] J. Kol and L. Lamel, "Development and evaluation of automatic punctuation for french and english speech-totext," in INTERSPEECH, 2012, pp. 1376–1379.

[9] F.Batista, D.Caseiro, N.Mamede, and I.Trancoso, "Recovering capitalization and punctuation marks for automatic speech recognition: Case study for portuguese broadcast news," Speech Communication, vol.50, no.10, pp. 847–862, 2008.

[10] J. Driesen, A. Birch, S. Grimsey, S. Safarfashandi, J. Gauthier, M. Simpson, and S. Renals, "Automated production of true-cased punctuated subtitles for weather and news broadcasts," pp. 2146–2147, 2014.

[11] W. Lu and H.T. Ng, "Better punctuation prediction with dynamic conditional random fields.," in EMNLP, 2010, pp. 177–186.

[12] Zhao, Y., Wang, C., Fu, G.: "A CRF sequence labeling approach to Chinese punctuation prediction." In: Pacific Asia Conference on Language, Information and Computation (2012).

[13] X. Che, C. Wang, H. Yang, and C. Meinel, "Punctuation prediction for unsegmented transcript based on word vector," in LREC, 2016, pp. 654–658.

[14] O. Tilk and T. Alumae, "Lstm for punctuation restoration in speech transcripts," in INTERSPEECH, 2015, pp. 683–687.

[15] E. Cho, J. Niehues, and A. Waibel, "Segmentation and punctuation prediction in speech language translation using a monolingual translation system," pp. 252–259, 2012.

[16] O. Klejch, P. Bell, and S. Renals, "Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches," in Spoken Language Technology Workshop, 2016, pp. 433–440.

[17] O. Klejch, P. Bell, and S. Renals, "Sequence-tosequence models for punctuated transcription combining lexical and acoustic features," in ICASSP, 2017, pp. 5700–5704.

[18] Pham, Q.H., Nguyen, B.T. (2014), Cuong, N.V.: Punctuation prediction for Vietnamese texts using conditional random fields. In: ACML Workshop: Machine Learning and Its Applications in Vietnam, pp. 1–9.

[19] Pham T., Nguyen N., Pham Q., Cao H., Nguyen B. (2020) Vietnamese Punctuation Prediction Using Deep Neural Networks. In: Chatzigeorgiou A. et al. (eds) SOFSEM 2020: Theory and Practi ce of Computer Science. SOFSEM 2020. Lecture Notes in Computer Science, vol 12011. Springer, Cham.

PUNCTUATION PREDICTION IN A VIETNAMESE SPEECH

Abstracts: In a Vietnamese identity article rewritten from a speech, there is no punctuation but space. Therefore, in order to have a complete article, we must identify and put the necessary punctuation for identification. This article will examine the use of punctuation in Vietnamese, examine the dependence of punctuation marks on the corresponding silent lengths and give the length of each silence corresponding to each punctuation . Since then, the paper proposes algorithms for determining punctuation for the case of regular speech and irregular speech.

Keywords: Vietnamese Voice, Punctuation, Silence, Speech Speed, Evenly, Irregular, Algorithm,Recognition, Punctuation Recognition.

Trần Anh Tuấn, Nghiên cứu sinh tại Học viện Công nghệ Bưu chính Viễn thông. Hiện đang công tác tại Trường Cao đẳng nghề Công nghiệp Thanh Hóa. Lĩnh vực nghiên cứu: Xử lý ảnh và nhận dạng tiếng nói.

Email:Tuankhhtqt@gmail.com

Nguyễn Hữu Mộng, Giảng viên Học viện kỹ thuật Quân sự. Lĩnh vực nghiên cứu: Toán tin ứng dụng, công nghệ nhận dạng.

Email:nghm06@yahoo.com

Nguyễn Trọng Khánh, Giảng viên Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: Hệ thống thông tin, công nghệ mô phỏng. Email:khanhnt82@gmail.com