Chuyển đổi hình vị - âm vị trong phân tích văn bản cho hệ thống tổng hợp tiếng nói tiếng Việt

1st Phan Thanh Sơn
Khoa Công nghệ thông tin
Trường Đại học Thông tin liên lạc
Nha Trang, Khánh Hòa, Việt Nam
ptson@tcu.edu.vn

2nd Cao Mạnh Hùng
Khoa Công nghệ thông tin
Trường Đại học Thông tin liên lạc
Nha Trang, Khánh Hòa, Việt Nam
caomanhhung.sqtt@gmail.com

***Tóm tắt*—Đầu vào của một hệ thống tổng hợp tiếng Việt sử dụng tham số thống kê dựa trên mô hình Markov ẩn (Hidden Markov Model, HMM) bao gồm ngữ liệu tiếng nói thu âm trước và dữ liệu văn bản ứng với các câu thu âm, phục vụ cho quá trình huấn luyện các HMM âm vị phụ thuộc ngữ cảnh. Văn bản đầu vào có của hệ thống tổng hợp tiếng nói thể chứa các ký hiệu, các chữ số, các chữ viết tắt và các từ phi tiêu chuẩn. Để có thể sử dụng chúng trong quá trình nhận dạng âm vị phụ thuộc ngữ cảnh trong quá trình huấn luyện và lựa chọn các HMM tương ứng trong quá trình tổng hợp thì trước tiên phải được chuyển đổi sang dạng thức âm vị tương ứng. Trong quá trình xây dựng hệ thống tổng hợp tiếng nói tiếng Việt dựa trên HMM, chúng tôi cũng phát triển một kịch bản (script) nhằm chuẩn hóa văn bản trước khi tiến hành chuyển đổi chuỗi hình vị sang âm vị tương ứng của văn bản, các dấu thanh và nguyên âm tiếng Việt được chuyển đổi sang biểu diễn dưới dạng telex.**

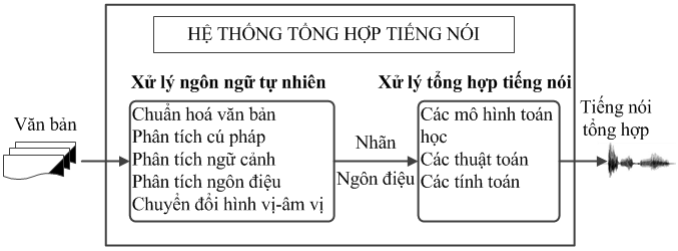
***Từ khóa*—chuyển đổi hình vị-âm vị, mô hình Markov ẩn, tổng hợp tiếng nói, chuẩn hóa, từ vựng**

I. GIỚI THIỆU

Chuyển đổi hình vị - âm vị (Grapheme-to-Phoneme, G2P) là bài toán quan trọng liên quan đến vấn đề xử lý ngôn ngữ, nhận dạng tiếng nói và tổng hợp tiếng nói. Mục đích của chuyển đổi G2P là dự đoán chính xác cách phát âm của một từ mới trong văn bản đầu vào chỉ dựa trên phân tích chính tả.

Tổng hợp tiếng nói là quá trình tạo ra tiếng nói của con người một cách nhân tạo. Tổng hợp tiếng nói từ văn bản (Text-To-Speech, viết tắt là TTS) là quá trình chuyển đổi tự động một văn bản có nội dung bất kỳ thành lời nói. Một hệ thống tổng hợp tiếng nói về cơ bản bao gồm hai khối chức năng: (1) khối phân tích xử lý ngôn ngữ tự nhiên (Natural Language Processing, viết tắt là NLP) hay còn gọi là khối tổng hợp mức cao; và (2) khối xử lý tổng hợp tiếng nói (Speech Synthesis Processing, viết tắt là SSP) có nhiệm vụ tổng hợp tiếng nói hay còn gọi là khối tổng hợp mức thấp (xem Hình 1).

Khối tổng hợp mức cao có nhiệm vụ chuyển đổi chuỗi các ký tự văn bản đầu vào thành một dạng chuỗi các nhãn ngữ âm đã được thiết kế trước của hệ thống TTS. Khối tổng hợp mức thấp sẽ chọn ra các tham số thích hợp từ tập các giá trị tần số cơ bản, phổ tín hiệu, trường độ âm thanh (bao gồm âm vị, âm tiết). Sau đó, tiếng nói ở dạng sóng tín hiệu sẽ được tạo ra bằng một kỹ thuật tổng hợp [14].



Hình 1. Sơ đồ chức năng tổng quát của một hệ thống TTS.

Trong phần tổng hợp, văn bản đầu vào bất kỳ được tổng hợp bằng cách lựa chọn các mô hình ngữ âm đã xây dựng trong quá trình huấn luyện các HMM phụ thuộc ngữ cảnh. Quá trình này đòi hỏi văn bản đầu vào phải được gán nhãn phụ thuộc ngữ cảnh, các nhãn này hàm chứa các thông tin ngôn điệu và ngữ âm của văn bản, giúp cho hệ thống có thể tạo ra tiếng nói tổng hợp đạt chất lượng giống với tiếng nói tự nhiên. Trong hệ thống TTS, mô-đun G2P có nhiệm vụ xác định phiên âm của văn bản đầu vào. Ngoài ra, G2P còn đảm nhiệm cả việc chuẩn hoá văn bản đầu vào và ánh xạ giữa cách thực biểu diễn hình vị sang cách thực biểu diễn âm vị tương ứng của một âm tiết [13].

Trong nhiều trường hợp, cách viết và cách phát âm của một âm vị không giống nhau, nên cần phải có phép ánh xạ để chuyển đổi sang cách phát âm thống nhất. Trong phần 2 của bài báo này, chúng tôi tập trung mô tả các bước chuẩn hoá văn bản đầu vào trong khối xử lý ngôn ngữ tự nhiên (NLP) của hệ thống tổng hợp tiếng nói (TTS). Phần 3 mô tả quá trình chuyển đổi chuỗi văn bản nguyên thuỷ thành chuỗi phát âm của các âm vị tương ứng, đây chính là chức năng chính của mô đun G2P. Các kết quả thử nghiệm và đánh giá được trình bày trong phần 4 của bài báo này.

II. CHUẨN HOÁ CHUỖI VĂN BẢN ĐẦU VÀO

Bước đầu tiên của quá trình thiết kế nhãn phụ thuộc ngữ cảnh là chuyển đổi các âm vị từ dạng ban đầu (biểu diễn dạng ký tự) sang dạng ký hiệu để xử lý, giai đoạn này được gọi là biến đổi từ hình vị sang âm vị (G2P). Trong bài báo này và các nghiên cứu của mình, chúng tôi thống nhất sử dụng font

Unicode và biểu diễn các nguyên âm tiếng Việt với các thanh điệu theo kiểu gõ Telex.

Văn bản đầu vào của hệ thống TTS tiếng Việt có thể là chuỗi văn bản bất bất kỳ. Chuỗi này có thể chứa một số từ (âm tiết) không thuần văn bản (không chuẩn) như: các từ viết tắt, tên riêng có nguồn gốc từ ngôn ngữ khác, số điện thoại, các giá trị số thập phân, phân số, chuỗi các con số liên quan đến thời gian, đơn vị tiền tệ, ký hiệu toán học, hoá học hoặc các ký hiệu, biểu tượng ngôn ngữ. Các mô-đun xử lý ngôn ngữ tự nhiên (NLP) được thiết kế để có thể xử lý các từ không chuẩn đó và kết quả là ta thu được chuỗi văn bản chuẩn hoá, mà cách phát âm có thể được lấy từ tập các quy tắc G2P. Mô-đun G2P có nhiệm vụ ánh xạ chuỗi các ký tự (hình vị) của các âm tiết (các từ) trong chuỗi văn bản thành chuỗi các phát âm của các âm vị.

Tất cả các từ không chuẩn phải được chuyển đổi sang các hình vị (grapheme) tiếng Việt tương ứng, trước khi các mô-đun G2P ánh xạ chúng thành âm vị (phoneme). Dựa trên bộ từ điển phiên âm và ngữ cảnh văn bản tiếng Việt, mô-đun này sẽ đưa ra quyết định hình thức phát âm của một từ không chuẩn trong văn bản đầu vào cần tổng hợp. Ví dụ, một dãy các chữ số sẽ được đọc như số điện thoại hay là một giá trị tiền tệ.

III. CÁC PHƯƠNG PHÁP CHUYỂN ĐỔI G2P

Hình vị thường có hình thức cấu tạo một âm tiết, tức là mỗi hình vị trùng với âm tiết, trên chữ viết mỗi hình vị được viết thành một chữ. Hình vị trong tiếng Việt có thể một mình đóng vai trò như một từ cũng có thể làm thành tố cấu tạo từ, nhưng nó chỉ được phân xuất ra nhờ phân tích bản thân các từ [6].

Có nhiều nghiên cứu chuyển đổi G2P từ cổ điển đến hiện đại, tuỳ thuộc vào bài toán cụ thể mà chọn lựa phương pháp chuyển đổi thích hợp nhất. Thống kê trung bình các mô hình là kỹ thuật mới nhất, kỹ thuật này được sử dụng để kiểm tra, đánh giá các phương pháp khác, một đánh giá toàn diện về kỹ thuật chuyển đổi từ chữ viết sang phát âm (letter to sound) được mô tả chi tiết bởi [2].

Chuyển đổi hình vị-âm vị là quá trình xác định chuỗi âm vị chính xác hoặc tốt nhất khi biết hình vị của một âm tiết. Theo Taylor [8], có ba cách tiếp cận cơ bản trong bài toán chuyển đổi G2P: quy tắc âm vị học (luật chữ viết), phương pháp tiếp cận hướng dữ liệu (điều khiển dữ liệu) và các phương pháp thống kê.

A. Luật chữ viết

Đây là cách tiếp cận đầu tiên và cũ nhất, được các nhà phát triển hệ thống soạn ra dựa theo kinh nghiệm và tri thức về lĩnh vực ngôn ngữ học của mình. Thông thường, các luật này có dạng cảm ngữ cảnh và được viết lại dưới dạng A/X/B -> y, nghĩa là ký tự X được ánh xạ thành âm vị y khi X nằm trong ngữ cảnh sau ký tự A và trước ký tự B. Một ví dụ chuyển đổi G2P của dạng này là sp/g/i -> d, ký tự g đứng đầu âm tiết và trước ký tự i chuyển thành âm vị d (ký tự gi trong chữ gìn, giếng trong tiếng Việt), hay a/c/sp -> kc, ký tự c chuyển thành âm vị kc khi nó đứng cuối âm tiết (âm cuối).

Cách tiếp cận này sử dụng các luật để tạo ra cách phát âm cho một âm tiết. Các luật này thường được tạo ra một cách tự động nhờ phân tích thống kê từ điển phát âm [8].

B. Điều khiển dữ liệu

Phương pháp tiếp cận thứ hai được gọi là tiếp cận "điều khiển dữ liệu" (data-driven), đây là phương pháp máy học với thuật toán được học các luật từ dữ liệu một cách tự động. Có ba kỹ thuật hiệu quả nhất của phương pháp này là: sử dụng cây quyết định, phát âm theo các âm tương đồng và mạng nơ ron. Trong kỹ thuật cây quyết định, các truy vấn của cây thường được hình thành giống như các luật cảm ngữ cảnh, và thuật toán lựa chọn các nút trên cây quyết định cũng giống như thuật toán chọn một luật. Sự khác nhau cơ bản nằm trong phương thức hình thành nên các luật; kỹ thuật này là hình thành từ chữ viết, kỹ thuật kia là học từ chính dữ liệu.

C. Thống kê

Phương pháp thứ ba là tiếp cận thống kê. Đây cũng có thể nói là điều khiển dữ liệu, nhưng kỹ thuật này không những chỉ học từ dữ liệu mà còn sử dụng đặc tính thống kê của dữ liệu. Một kỹ thuật thuộc phương pháp này là sử dụng chung n-gram và chuỗi âm vị trong bài toán dịch máy thống kê giữa hai ngôn ngữ [1]. Sử dụng HMM trong phương pháp tiếp cận thống kê là một kỹ thuật được đề xuất bởi Paul Taylo [8]. Trong kỹ thuật này, các âm vị là các trạng thái ẩn, các bước chuyển trạng thái giữa các âm vị được mô tả như là xác suất mà một âm vị đi theo ngay sau một âm vị khác và các hình vị quan sát được. Cũng giống như trong nhận dạng tiếng nói, hệ thống này sẽ tạo ra chuỗi âm vị có xác suất cao nhất ứng với các quan sát hình vị.

IV. THIẾT KẾ MÔ-ĐUN G2P

Trong hệ thống tổng hợp tiếng nói, mô-đun G2P chuyển đổi chuỗi văn bản chuẩn tắc về mặt chính tả thành chuỗi biểu diễn cách phát âm các âm vị đại diện. Vì vậy, có thể nói G2P là mô-đun cơ bản trong hệ thống TTS, mô-đun chuẩn hóa văn bản sẽ nhập một chuỗi từ vào mô-đun G2P. Quá trình chuyển đổi hình vị - âm vị của chuỗi từ có thể được thực hiện bằng cách sử dụng luật Letter to Sound (L2S). Các luật L2S được dựa trên một từ điển phát âm, chứa ánh xạ cách phát âm của một từ thành một chuỗi các âm.

Tiếng Việt thuộc ngữ hệ phương Nam, dòng Nam Á, ngành Môn-Khơ me [3], là loại ngôn ngữ được xếp vào loại hình đơn lập (isolate), không biến hình, đơn âm tiết, ranh giới âm tiết trùng với ranh giới hình vị, và có thanh điệu, sử dụng các ký tự La tinh để ghi chữ viết và các ký hiệu phụ để ghi dấu thanh. Thanh điệu đóng một vai trò rất quan trọng lên toàn bộ âm tiết tiếng Việt. Thanh điệu trong âm tiết là âm vị siêu đoạn tính (thể hiển trên toàn bộ âm tiết). Do đó, đặc trưng về thanh điệu thể hiện trong tín hiệu tiếng nói không rõ nét như các thành phần khác của âm tiết. Những âm tiết với chuỗi âm vị tương tự có thể mang ý nghĩa khác nhau nếu chúng kết hợp với thanh điệu khác nhau [12]. Do đó, thanh điệu là yếu tố cần phải được xem xét một cách cẩn thận trong hệ thống tổng hợp giọng nói đối với ngôn ngữ có thanh điệu.

Nhìn về mặt ghi âm: âm tiết tiếng Việt có cấu tạo chung là: phụ âm-vần. Vần trong tiếng Việt lại được cấu tạo từ các âm vị nhỏ hơn, trong đó có một âm vị chính là nguyên âm. Theo tác giả Đoàn Thiện Thuật [3], xét về mặt ngữ âm-âm vị học thì một âm tiết tiếng Việt có cấu tạo như bảng *1*.

Bảng 1
CẤU TRÚC CỦA ÂM TIẾT TIẾNG VIỆT

Thanh điệu			
[Âm đầu]	Vần		
	[Âm đệm]	Âm chính	[Âm cuối]

Mỗi ngôn ngữ có tập hợp các âm vị riêng của nó, tuỳ vào từng hệ ngôn ngữ cụ thể mà có số lượng khoảng từ 20 đến 60. Ví dụ tiếng Anh có thể biểu diễn bằng khoảng 42 âm vị, tiếng Việt khoảng 46 âm vị (12 nguyên âm đơn: a, ă, â, o, u, ...; 3 nguyên âm đôi: ie, uo, wa; 1 âm đệm: w; 22 phụ âm đầu: k, l, m, ph, ... và 8 phụ âm cuối: p, t, k, m, n, ng, uz, iz) và 6 thanh điệu (ngang, huyền, ngã, hỏi, sắc, nặng) biểu diễn bằng giá trị số từ 0 đến 5 [11]. Tiếng Việt có khoảng 2376 âm tiết cơ bản và 6492 âm tiết có thanh điệu [9], do đó nếu thanh điệu được xác định một cách chính xác thì không những bảo đảm tính chất nghe rõ của tiếng nói tổng hợp mà tính tự nhiên và ngữ điệu cũng được nâng cao. Trong bài báo này, thanh điệu được trích rút thông qua các tham số phổ tín hiệu ($MFCC$) và tần số cơ bản (F_0), trong quá trình gán nhãn âm vị phụ thuộc ngữ cảnh, bên cạnh việc gán nhãn thanh điệu cho âm tiết hiện tại, chúng tôi còn xem xét đến ngữ cảnh của các thanh điệu của hai âm tiết trước và sau, âm tiết hiện tại, cùng với vị trí của nó trong toàn câu.

V. CẢI TIẾN NHÃN PHỤ THUỘC NGỮ CẢNH

Bước đầu tiên của quá trình thiết kế nhãn phụ thuộc ngữ cảnh là chuyển đổi các âm vị từ dạng ban đầu (biểu diễn dạng ký tự) sang dạng ký hiệu để xử lý, giai đoạn này được gọi là biến đổi từ hình vị sang âm vị (grapheme-to-phoneme - G2P). Ở đây, chúng tôi thống nhất sử dụng font Unicode và biểu diễn các nguyên âm tiếng Việt với các thanh điệu theo kiểu gõ Telex. Thể hiện ánh xạ các âm vị tiếng Việt được minh hoạ trong các bảng *2*, *3*, *4* và *5*:

Thông tin ngữ cảnh cần thiết cho quá trình gán nhãn (hình 2) dữ liệu tiếng nói tiếng Việt được cải tiến từ [11], có mang thông tin ngôn điệu:

- Mức âm vị:
  - Hai âm vị trước, âm vị hiện tại, hai âm vị phía sau;
  - Vị trí hiện tại của âm vị trong âm tiết (tính từ đầu và từ cuối âm tiết);
- Mức âm tiết:
  - Thanh điệu của hai âm tiết trước, âm tiết hiện tại, hai âm tiết phía sau;
  - Số lượng âm vị trong âm tiết trước, âm tiết hiện tại, âm tiết sau;
  - Vị trí của âm tiết trong từ hiện tại (tính từ đầu và từ cuối từ);
  - Mức độ trọng âm (thể hiện ngữ điệu);

Bảng 2
ÁNH XẠ HÌNH VỊ SANG ÂM VỊ CỦA ÂM ĐẦU

Ký tự	Âm vị	Ví dụ	Ký tự	Âm vị	Ví dụ
b	b	buồn bã	ph	ph	phất phới
đ	dd	đẫy đà	v	v	vùng vằng
t	t	tan tác	x	x	xa xôi
th	th	thơm tho	d	d	dễ dàng
tr	tr	trục trặc		gi	giếng, giề
ch	ch	châm chước		g trước /i,ie/	gìn, giếng
k	k trước /i,e,ê/	kênh kiệu	l	l	lấp lánh
	c trước /u,o,a,ơ,ư/	cầu cạnh	s	s	sạch sẽ
	q trước âm đệm	quần quật	r	r	rậm rạp
m	m	mĩ miều	kh	kh	khập khiễng
n	n	nườm nượp	g	gh trước /i,e,ê/	
nh	nh	nhí nhố		g không trước i	
ng	ngh trước /i,e,ê/	nghiêm nghị	h	h	hồi hộp
	ng t/h còn lại	ngây ngô	p	p nước ngoài	sa pa, pác bó

Bảng 3
ÁNH XẠ HÌNH VỊ SANG ÂM VỊ CỦA NGUYÊN ÂM ĐƠN

Ký tự	Âm vị	Ví dụ	Ký tự	Âm vị	Ví dụ
y	i (sau âm đệm)	huy, suy	ô	oo	ô tô, lô nhô
i	i (t/h còn lại)	tin, tích	o	o	lon ton
ê	ee	ếch, bệnh	oa	o trước kc, ngz	xóc, vòng
e	e	xem xét	ư	uw	hừng hực
a	ea trước ch,nh	chanh, trách	ơ	ow	lơ tơ mơ
	a	lang thang	â	aa	rầm rập
	aw trước u, y	lau, tay	ă	aw	ăn năn, nằm
u	u	súng, cú, tù			

  - Khoảng cách đến âm tiết có trọng âm trước và đến âm tiết có trọng âm sau;
- Mức từ:
  - Loại từ (Part-of-speech) của từ trước, từ hiện tại và từ phía sau;
  - Số lượng âm tiết trong từ trước, từ hiện tại và từ phía sau;
  - Vị trí của từ trong cụm từ;
  - Số lượng từ trong cụm từ trước, sau tính từ vị trí hiện tại;
  - Khoảng cách đến từ trước và từ sau tính từ vị trí hiện tại;
- Mức cụm từ:

Bảng 4
ÁNH XẠ HÌNH VỊ SANG ÂM VỊ CỦA ÂM ĐỆM VÀ NGUYÊN ÂM ĐÔI

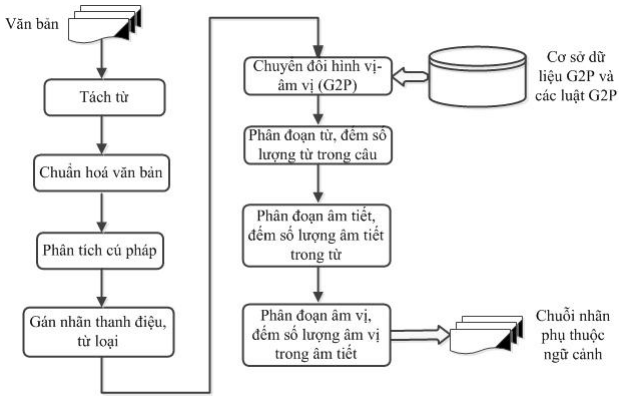
Ký tự	Âm vị	Ví dụ	Ký tự	Âm vị	Ví dụ
o	w trước /a,ă,e/	hoa hoè, hoạ hoằn	yê	ie (khi trước có âm đệm hoặc âm cuối là u)	yêu, uyển chuyển
u	w t/h còn lại	huy, tuần, huệ	uô	uo (khi sau có âm cuối)	muộn, tuồn tuột
ia	ie (trước không có âm đệm và sau không có âm cuối)	kia, thìa, bia	ua	uo (khi sau không có âm cuối)	mua, vua chúa
ya	ie (khi trước có âm đệm)	khuya	ưa	wa (khi không có âm cuối)	mưa, vừa vừa
iê	ie (khi trước không có âm đệm và sau có âm cuối)	tiên tiến, tiếng, tiết	ươ	wa (khi có âm cuối)	ương bướng

Bảng 5
ÁNH XẠ HÌNH VỊ SANG ÂM VỊ CỦA ÂM CUỐI

Ký tự	Âm vị	Ví dụ
p	pc	chập, tấp nập
t	tc	cắt, lắt nhắt
m	mz	đom đóm
n	nz	lan man
ch	kc (đứng sau các âmvị /i,a,ê/i)	thích,sạch
c	kc (t/h còn lại)	được,việc

  - Số lượng âm tiết, từ trong cụm từ trước, cụm từ hiện tại và cụm từ phía sau;
  - Vị trí của cụm từ hiện tại trong câu nói;
  - Thẻ ToBI của của cụm từ hiện tại;
- Mức câu nói:
  - Số lượng âm tiết, từ, cụm từ trong câu nói;

Bảng *6* minh hoạ một số phiên âm trong mô-đun chuyển đổi âm vị-hình vị, sử dụng trong hệ thống tổng hợp tiếng Việt dựa trên HMM của chúng tôi:



Hình 2. Vị trí của mô-đun G2P trong lược đồ quá trình gán nhãn phụ thuộc ngữ cảnh.

Bảng 6
CHUYỂN ĐỔI HÌNH VỊ-ÂM VỊ MỘT SỐ ÂM TIẾT TIẾNG VIỆT

Hình vị	Âm vị	Phiên âm	Hình vị	Âm vị	Phiên âm
ạch	[achj]	ea kc sp	chèo	[cheof]	tr e uz sp
ẩn	[aanr]	aa nz sp	phiến	[phieens]	ph ie nz sp
ập	[aapj]	aa pc sp	phiếu	[phieeus]	ph ie uz sp
bậc	[baacj]	b aa kc sp	khuỷu	[khuyur]	kh w i uz sp
bấm	[baams]	b aa mz sp	khùng	[khungf]	kh u ngz sp
bầu	[baauf]	b aa uz sp	nghiên	[nghieen]	ng ie nz sp
chững	[chuwngx]	tr uw ngz sp	lây	[laay]	l aa iz sp
lãng	[langx]	l a ngz sp	chỉnh	[chinhr]	tr i ngz sp

VI. THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Mô-đun G2P được thiết kế xây dựng và cài đặt trong thử nghiệm và đánh giá kết quả của nghiên cứu [13]. Các thử nghiệm trong bài báo này được thực hiện trên bộ dữ liệu 510 câu thu âm tiếng Việt (một giọng nam, một giọng nữ, phương ngữ Bắc bộ), cân bằng đủ số âm vị và thanh điệu. Các câu thu âm mang nhiều yếu tố ngữ nghĩa và ngữ điệu. Chúng tôi sử dụng 400 câu cho tập huấn luyện và 110 câu còn lại để tổng hợp và đánh giá kết quả. Tất cả dữ liệu tiếng nói thu âm đều được lấy mẫu ở 48 kHz, kênh đơn (mono chanel) và mã hóa ở định dạng PCM 16 bit, sau đó tín hiệu tiếng nói được chuyển đổi về tần số lấy mẫu ở 16 kHz, định khung 40 ms với cửa sổ Hamming và độ dịch khung là 8 ms trước khi đưa vào hệ thống để huấn luyện.

Các $MFCC$ và F_0 được tính toán cho từng câu nói thu âm nhờ sử dụng bộ công cụ SPTK [15]. Các nhãn phụ thuộc ngữ cảnh của hai bộ dữ liệu tiếng nói tiếng Việt được sinh ra tự động từ các văn bản tương ứng nhờ sử dụng bộ phân tích văn bản tiếng Việt [11], sau đó cải tiến đặc tính thanh điệu thông qua gán nhãn phụ thuộc ngữ cảnh lại một cách thủ công. Ngoài ra, chúng tôi sử dụng kỹ thuật phân cụm ngữ cảnh dựa trên cây quyết định để huấn luyện các HMM phụ thuộc ngữ cảnh tương ứng với từng tham số $MFCC$, F_0 và các thành phần tuần hoàn khác.

VII. KẾT LUẬN

Trong bài báo này, chúng tôi giới thiệu một nghiên cứu thiết kế mô-đun chuyển đổi âm vị - hình vị, phục vụ cho quá trình cải tiến thiết kế nhãn phụ thuộc ngữ cảnh cho văn bản đầu vào, có sử dụng đặc trưng thanh điệu, từ loại và ngữ điệu cho hệ thống tổng hợp tiếng nói thống kê dựa trên HMM, phát triển cho tổng hợp tiếng Việt.

Kết quả thử nghiệm trong [13] cho thấy rằng, các thông tin ngữ cảnh liên quan đến thanh điệu, từ loại và ngữ điệu trong văn bản làm tăng đáng kể tính tự nhiên của tiếng nói tổng hợp. Tuy nhiên, trong bài báo này, chúng tôi mới chỉ xem xét đến nhãn phụ thuộc ngữ cảnh ở mức âm vị và mức âm tiết,

trong tương lai, nghiên cứu có thể phát triển cho các nhãn phụ thuộc ngữ cảnh đến mức từ (từ ghép, từ láy, cụm từ) và mức đoạn văn bản. Thử nghiệm mới chỉ áp dụng trên bộ dữ liệu thu âm giọng nam và nữ phương ngữ Bắc bộ, tiếp theo có thể tiến hành trên dữ liệu thu âm phương ngữ khác.

Tóm lại, với những cải tiến này, chúng ta có thể tổng hợp được tiếng nói với các đặc điểm thay đổi về ngữ điệu, trọng âm, thanh điệu. Trong tương lai, chúng tôi tập trung nghiên cứu, áp dụng các yếu tố ngữ cảnh và điều kiện phân cụm ngữ cảnh, cải tiến quá trình xử lý văn bản tự động thông qua phân tích sâu về ngữ âm tiếng Việt để đạt được mục tiêu chất lượng tiếng nói tổng hợp tốt hơn và tổng hợp tiếng nói với các đặc tính âm học khác nhau.

TÀI LIỆU THAM KHẢO

[1] Chen Stanley F., "Conditional and joint models for Grapheme-tophoneme conversion," Proc. in Eurospeech, Geneva, Switzerland, tr.105–108, 2003

[2] Damper, R. I., Marchand, Y., Adamson, M. J. and Gustafson, K., *A comparison of letter-to-sound conversion techniques for English text-to-speech synthesis,* Proceedings of the Institute of Acoustics 20(6), 1999.

[3] Đoàn Thiện Thuật, *Ngữ âm tiếng Việt*, NXB Đại học Quốc gia Hà Nội, 2003.

[4] H. Zen, K. Tokuda, and A.W. Black, *Statistical parametric speech synthesis,* Speech Communication, 51, Issue 11, pp. 1039-1064, 2009.

[5] K. Tokuda, H. Zen, J. Yamagashi, T. Masuko, S. Sako, A.W. Black, and T. Nose, *HMM-based speech synthesis system (HTS)*, version 2.3, 2012. http://hts.sp.nitech.ac.jp/ (accessed and downloaded December, 2012).

[6] Lê Đình Tư và Vũ Ngọc Cân, Nhập môn ngôn ngữ học, Nhà xuất bản Giáo dục, Hà Nội, 2009.

[7] Nguyễn Thị Minh Huyền, Vũ Xuân Lương, Lê Hồng Phương, *Sử dụng bộ gán nhãn từ loại xác suất QTAG cho văn bản tiếng Việt*, Báo cáo hội thảo ICT.rda, 2003.

[8] P. Taylor, *Grapheme-to-Phoneme conversion using Hidden Markov Models,* Proc. Interspeech, Lisbon, Portugal, tr. 1973-1976, 2005.

[9] Phu Ngoc Le, Eliathamby Ambikairajah, Eric H.C. Choi, *Improvement of Vietnamese Tone Classification using FM and MFCC Features*, Computing and Communication Technologies RIVF '09, 2009.

[10] Phan Thanh Son, Vu Tat Thang, "HMM-based Vietnamese Speech Synthesis using MFCC and F0" , Chuyên san CNTT, Tạp chí Khoa học và Kỹ thuật, số 150, Học viện Kỹ thuật quân sự, Số 150(1), tr. 147-155, 2012.

[11] Thang Tat Vu, Mai Chi Luong, Satoshi Nakamura, *An HMM-based Vietnamese Speech Synthesis System*, Proc. Oriental COCOSDA, 2009.

[12] T.T Vu, T.K. Nguyen, H.S. Le, C.M. Luong, *Vietnamese tone recognition based on MLP neural network,* Proc. Oriental COCOSDA, 2008.

[13] Thanh-Son PHAN, Tu-Cuong DUONG, Anh-Tuan DINH, Tat-Thang VU, Chi-Mai LUONG, "Improvement of Naturalness for an HMM-based Vietnamese Speech Synthesis using the Prosodic information," The 10th IEEE-RIVF International Conference on Computing and Communication Technologies, Hanoi, Vietnam, 2013.

[14] Cambridge University Engineering Department, HTK 3.5 beta 2, http://htk.eng.cam.ac.uk/, 1989 [cập nhật 23-8-2017]

[15] Department of Computer Science, Nagoya Institute of Technology: Speech Signal Processing Toolkit, SPTK 3.11. Reference manual, http://sourceforge.net/projects/sp-tk/, Japan, 12- 2003. [cập nhật 25-12-2017]