Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00060

MỘT GIẢI PHÁP VIỆT HÓA CÁCH PHÁT ÂM CÁC TỪ VỰNG TIẾNG ANH TRONG VĂN BẢN TIẾNG VIỆT

Triệu Thị Ly Ly, Bùi Thanh Sơn, Lê Thị Hà Bình, Ninh Khánh Duy

Khoa Công nghệ thông tin, Trường Đại học Bách khoa, Đại học Đà Nẵng

lylytrieu95@gmail.com, buison559@gmail.com, lethihabinh@gmail.com, nkduy@dut.udn.vn

TÓM TẮT: *Các ứng dụng liên quan đến xử lý ngôn ngữ tự nhiên và tiếng nói cho tiếng Việt thường gặp khó khăn khi văn bản đầu vào chứa các từ ngữ của tiếng nước ngoài, trong đó phổ biến nhất là tiếng Anh. Để có thể xử lý văn bản một cách đầy đủ, các từ ngữ nước ngoài này cần được phiên âm sang một dạng có thể đánh vần được trong tiếng Việt. Bài báo này đưa ra một giải pháp Việt hóa cách phát âm các từ vựng tiếng Anh dựa trên từ điển CMU và các quy tắc ngôn ngữ học. Bằng việc nghiên cứu sự tương đồng về ngữ âm giữa tiếng Anh và tiếng Việt cùng với các quy tắc ghép âm vần, thanh điệu trong tiếng Việt, chúng tôi đã tạo bảng ánh xạ từ một âm vị tiếng Anh sang một âm vị tiếng Việt và triển khai thuật toán tách một chuỗi âm vị tiếng Anh thành các âm tiết tương ứng trong tiếng Việt. Từ đó chúng tôi xây dựng được một công cụ tự động phiên âm một từ vựng tiếng Anh bất kỳ thành chuỗi âm tiết tiếng Việt. Công cụ này đã được tích hợp vào một phần mềm chuyển văn bản thành tiếng nói và cho thấy tính hữu dụng của nó.*

Từ khóa: *Xử lý ngôn ngữ tự nhiên, ngữ âm học, âm vị học, từ điển CMU, bảng phiên âm quốc tế.*

I. ĐẶT VẤN ĐỀ

Trong cuộc sống hằng ngày cũng như trên các phương tiện thông tin đại chúng, phương tiện truyền thông, chúng ta vẫn thường bắt gặp những từ tiếng Anh xuất hiện với tần suất ngày càng nhiều. Trong thời đại hội nhập quốc tế như hiện nay thì xu thế đó là không tránh khỏi, tuy nhiên việc phiên âm các từ này thường mang tính chủ quan, phần đa vẫn dựa vào theo thói quen, không tuân theo bất cứ cơ sở khoa học nào. Mặt khác, việc xuất hiện những từ tiếng Anh này cũng gây khó khăn cho các công nghệ nghiên cứu xử lý tiếng nói, xử lý ngôn ngữ tự nhiên, xử lý văn bản ví dụ như hệ thống chuyển đổi văn bản thành âm thanh, hệ thống nhận diện giọng nói,… Trong một hệ chuyển văn bản tiếng Việt thành tiếng nói, các từ viết bằng tiếng nước ngoài cần được Việt hóa cách phát âm để máy tính có thể chuyển thành tiếng nói của người Việt. Trong một hệ nhận dạng tiếng nói không giới hạn từ vựng dành cho người Việt, các từ không nằm trong tập từ vựng của hệ (out-of-vocabulary words) cần được Việt hóa cách phát âm để máy tính có thể giải mã đoạn tín hiệu âm thanh của từ đó.

Gần đây, đã có nhiều nghiên cứu thực hiện chuyển đổi các từ trong một ngôn ngữ nguồn (tiếng Anh) thành các từ tương đương ngữ âm bằng một ngôn ngữ đích (tiếng Việt) [5]. Bài báo đã đề xuất mô hình và hệ thống để giải quyết vấn đề dựa trên âm vị học tiếng Việt. Tuy nhiên, nghiên cứu chưa có sự tập trung vào việc tách chuỗi âm vị tiếng Anh thành âm tiết phát âm được bằng tiếng Việt.

Vì vậy, nhóm chúng tôi đã tiến hành nghiên cứu và xây dựng thành công công cụ tự động phiên âm một từ vựng tiếng Anh bất kỳ thành chuỗi âm tiết tiếng Việt, hay nói cách khác là Việt hóa cách phát âm các từ vựng tiếng Anh. Trong quá trình nghiên cứu, chúng tôi đã sử dụng sự hỗ trợ từ một số công cụ như: bộ từ điển CMU, công cụ t2p (text-to-phoneme)… và căn cứ theo Bảng ký hiệu ngữ âm quốc tế - IPA (*International Phonetic Alphabet*). Với cách tiếp cận bằng việc nghiên cứu sự tương đồng về phát âm và ngữ âm giữa tiếng Anh và tiếng Việt cùng với các quy tắc ghép âm, thanh điệu trong tiếng Việt, nhóm đã nghiên cứu và triển khai được thuật toán tách chuỗi âm vị tiếng Anh thành âm tiết phát âm được bằng tiếng Việt và ánh xạ một âm vị tiếng Anh trong CMU sang một âm vị tiếng Việt trong IPA. Từ đó áp dụng các kỹ năng và kỹ thuật lập trình để xây dựng thành công công cụ Việt hóa cách phát âm các từ vựng tiếng Anh.

Bài nghiên cứu này, chúng tôi dựa trên cơ sở sự tương đồng về âm vị học giữa Tiếng Anh và Tiếng Việt. Sự tương đồng này được đánh giá bằng các tiêu chí về phát âm, ngữ âm, kết hợp âm để từ đó đưa ra nền tảng cho việc ánh xạ hai âm vị tương đương giữa hai ngôn ngữ [4].

Bài báo này nhằm mục đích trình bày kết quả nghiên cứu của nhóm chúng tôi với bố cục như sau. Phần II sẽ giới thiệu về bảng Arpabet mà nhóm đã lập được trên cơ sở dựa vào sự tương đồng về phát âm và ngữ âm giữa tiếng Anh và tiếng Việt. Phần III sẽ trình bày phương pháp thực hiện trong đó sẽ nêu ra và phân tích quá trình thực hiện của công cụ. Phần IV đưa ra kết quả thực nghiệm và phần V là kết luận và hướng phát triển cho nghiên cứu trong tương lai.

II. BẢNG ÁNH XẠ TỪ ÂM VỊ ARPABET SANG ÂM VỊ TIẾNG VIỆT

Trong phần này chúng tôi sẽ giới thiệu bảng ánh xạ âm vị từ ARPAbet sang âm vị Tiếng Việt để phục vụ việc phiên âm Tiếng Anh sang Tiếng Việt.

A. Thuật ngữ

1. Âm vị (Phonemes)

Âm vị (Phonemes) là đơn vị nhỏ nhất truyền tải ý nghĩa trong hệ thống âm thanh của một ngôn ngữ [1].

2. TXTTeam

TXTTeam là hệ thống mã sử dụng để ghi các âm vị tiếng Việt trên máy tính.

B. ARPAbet

ARPAbet là hệ thống các mã sao chép âm vị phát triển bởi cơ quan Advanced Research Projects Agency (ARPA) như là một phần của Dự án Thông hiểu Tiếng nói (Speech Understanding Project) (1971-1976) [2].

C. International Phonetic Alphabet

International Phonetic Alphabet (IPA) là một sản phẩm của International Phonetic Association (Hiệp hội Ngữ âm Quốc tế). Mục đích của IPA là ghi lại và sắp xếp âm trong các ngôn ngữ trên thế giới dựa vào những quy tắc của ngữ âm khớp nối (articulatory phonetics principles) [1].

D. Bảng ánh xạ ARPAbet

Ta cần chuyển âm vị của ARPAbet sang IPA và tương tự với âm vị Tiếng Việt, từ đó so sánh, đối chiếu và sắp xếp theo các cặp tương đương nhau.

1. Chuyển đổi âm vị ARPAbet sang âm vị IPA

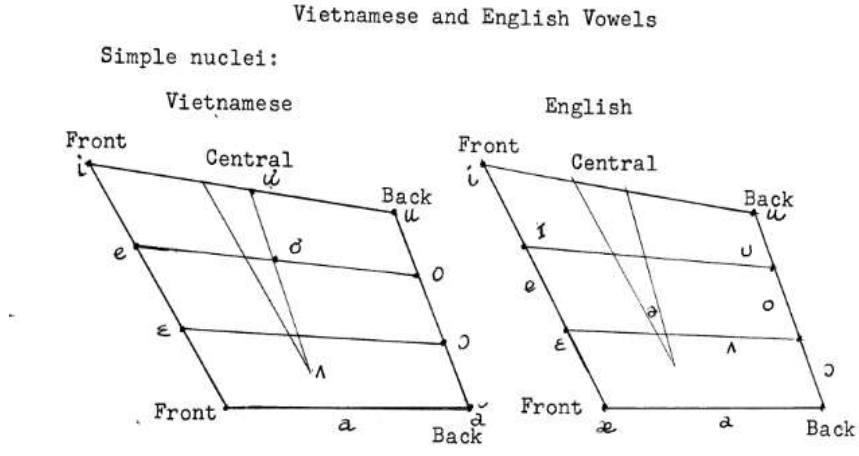
Ta sẽ sử dụng bảng chuyển đổi có sẵn ở nghiên cứu [3].

2. Chuyển đổi âm âm vị Tiếng Việt sang âm vị IPA

Bảng chuyển đổi âm vị Tiếng Việt sang âm vị IPA đã được thực hiện dựa trên các nghiên cứu [4]. Từ đó nhóm chúng tôi đã thiết lập bảng ghi âm vị Tiếng Việt bằng âm vị IPA đồng thời sử dụng các mã ASCII để ghi các âm vị Tiếng Việt đó, tạo sự thuận lợi cho việc lập trình sau này.

3. Tổng hợp bảng ánh xạ từ âm vị Tiếng Việt sang âm vị Tiếng Việt

Nhận thấy rằng một số âm vị trong Tiếng Việt và Tiếng Anh có cùng âm vị IPA, chúng tôi coi chúng là một cặp tương đương. Tuy nhiên luôn tồn tại sự khác biệt giữa ngữ âm trong hai ngôn ngữ, Tiếng Việt tồn tại những âm vị mà không xuất hiện trong Tiếng Anh và điều này cũng diễn ra theo hướng ngược lại. Để giải quyết khó khăn này, chúng tôi buộc lòng dựa trên sự tương đồng về phát âm và ngữ âm được trình bày như Hình 1.



Hình 1. Phân tích các âm vị Tiếng Anh và Tiếng Việt [4].

4. Bảng ánh xạ

Bảng 1 trình bày cách chuyển đổi âm vị Tiếng Việt sang âm vị IPA đã được thực hiện dựa trên các nghiên cứu [4]. Từ đó nhóm chúng tôi đã thiết lập bảng ghi âm vị Tiếng Việt bằng âm vị IPA đồng thời sử dụng các mã ASCII để ghi các âm vị Tiếng Việt đó, tạo sự thuận lợi cho việc lập trình sau này.

Bảng 1. Ánh xạ âm vị ARPAbet sang âm vị Tiếng Việt

Nguyên âm		
Arpabet	IPA	TXTTeam
AO	ɔ	O
AA	ɑ	A
IY	I	I
UW	U	u
EH	ɛ	E
...	...	...
Phụ âm		
Arpabet	IPA	TXTTeam
P	p	P
B	B	B
T	T	t
D	D	d
K	K	k
...	...	...

III. PHƯƠNG PHÁP THỰC HIỆN

Việc phiên âm một từ vựng tiếng Anh bất kỳ thành chuỗi âm tiết tiếng Việt thực sự phức tạp và mơ hồ vì người Việt vốn phiên âm các từ tiếng Anh đó thường theo thói quen, không tuân theo bất kỳ quy tắc nào. Hơn nữa còn dựa trên quan điểm chủ quan cá nhân cũng như ảnh hưởng của vùng miền, giọng địa phương. Điều đó dẫn đến việc một từ tiếng Anh có thể có nhiều cách phiên âm khác nhau, ví dụ "Arsenal" có thể được phát âm là "a – xê - nan" hoặc cũng có thể là "a – xê – nồ" hay "ác – xê – nan"…

Trong nghiên cứu này, chúng tôi đề xuất một giải pháp để Việt hóa cách phát âm các từ vựng tiếng Anh tuân theo các nguyên tắc dựa trên các đặc trưng của tiếng Việt. Hình 2 trình bày mô hình tổng quát trình tự thực hiện của công cụ chúng tôi đã xây dựng.

Hình 2. Mô hình tổng quát

Trong phần tiếp theo chúng tôi trình bày chi tiết phương pháp thực hiện từng bước trong mô hình trên.

A. Công cụ t2p

1. Định nghĩa

t2p là một domain package trong Perl dùng để xây dựng những quy tắc biến tự vị thành âm vị dựa trên từ điển phát âm. Nói cách khác, nó xây dựng quy tắc biến chữ cái thành âm đê phát âm một từ cho trước dựa trên ví dụ là những từ đã được phát âm trước đó.

Một ví dụ sau khi áp dụng t2p cho một từ tiếng anh "Steven", ta được kết quả là "S T IY V AH N"

2. Nguyên lý hoạt động

t2p nhận một từ điển phát âm, trong bài nguyên cứu này, chúng tôi sử dụng từ điển CMU và xây dựng Cây quyết định (Decision Tree) để tạo mô hình cho các từ.

Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật (series of rules). Về bản chất thì nó giống như câu lệnh "if then else". Điều này được thể hiện rõ trong Hình 3.

```
if ($att{'L'} eq 'H') {
  if ($att{'L1'} eq 'A') {
    if ($att{'R1'} eq 'A') {
      if ($att{'L3'} eq 'G') {
        if ($att{'R3'} eq '-') {
          return 'HH';
        }
        return '_';
      }
      if ($att{'L3'} eq 'H') {
        return '_'; # unique at depth 4
      }
      if ($att{'L3'} eq 'U') {
        if ($att{'L2'} eq 'J') {
          return 'AE'; # unique at depth 5
        }
        return 'HH';
      }
```

Hình 3. Kết quả cây quyết định được xây dựng bởi t2p

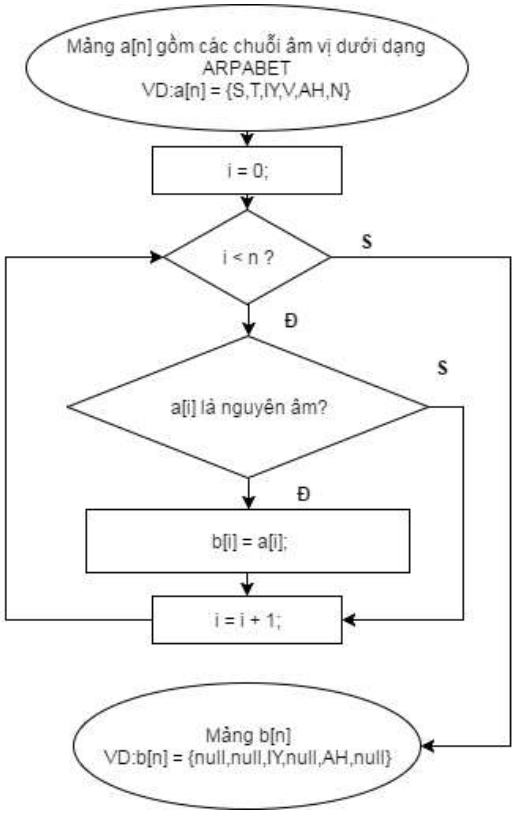
Nhờ vào cây quyết định, chúng ta có thể cho ra kết quả chuỗi Arpabet đối với một từ nằm ngoài từ điển (out-of-vocabulary words).

B. Thuật toán tách âm vị thành âm tiết

Âm tiết là đơn vị cơ bản của Tiếng Việt [5]. Muốn phát âm được từ Tiếng Anh ta cần cắt các âm vị trong từ đó thành từng âm tiết Tiếng Việt có thể đọc được. Công cụ để biến từ Tiếng Anh thành chuỗi âm vị là t2p phát triển bởi CMU [6]. Mặc dù đã có ánh xạ âm vị ARPAbet sang âm vị tiếng Việt tuy nhiên chúng tôi vẫn sử dụng ARPAbet để thuận tiện cho việc lập trình.

1. Tìm nguyên âm

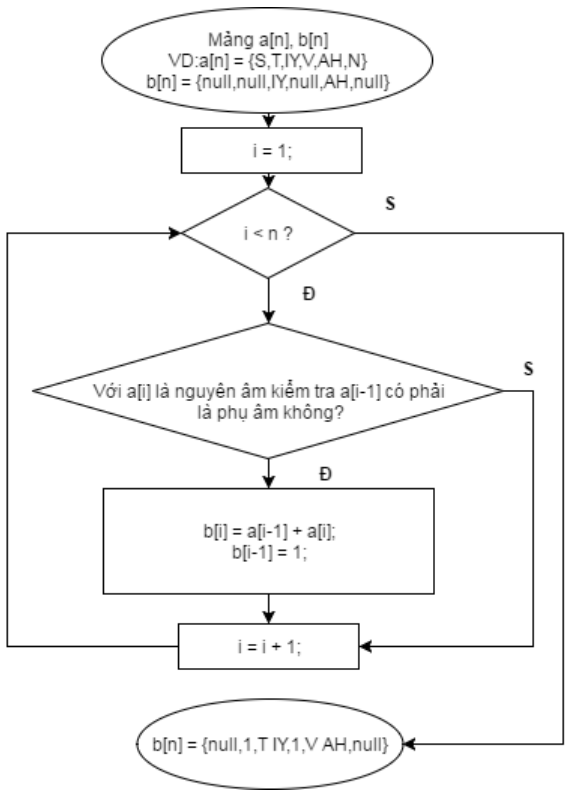
Thành phần chính, bắt buộc phải có của âm tiết là nguyên âm nên chúng tôi sử dụng nguyên âm làm điểm mốc [5]. Xét "Steven" cho một chuỗi âm vị dưới dạng ARPAbet được đưa vào một mảng có dạng {S,T,IY,V,AH,N}. Nhận thấy "IY" và "AH" là nguyên âm, chúng tôi đưa chúng vào một mảng rỗng với vị trí giống mảng ban đầu {null,null,IY,null,AH,null}. Thuật toán được trình bày như Hình 4.



Hình 4. Sơ đồ khối tìm nguyên âm

2. Ghép phụ âm bắt đầu âm tiết vào nguyên âm

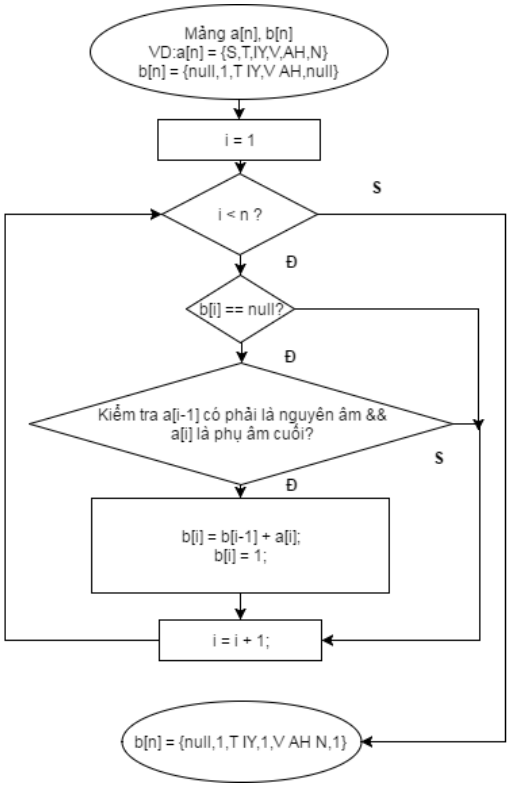
Mọi phụ âm bắt đầu âm tiết được ghép vào nguyên âm vì hầu hết các âm tiết trong Tiếng Việt bắt đầu bằng phụ âm [5]. Vì "T" và "V" là phụ âm và đứng trước các nguyên âm, chúng được ghép vào nguyên âm liền kề sau. Vị trí của hai phụ âm đó được đánh dấu bằng số 1 trong mảng. Thuật toán được trình bày như Hình 5.



Hình 5. Sơ đồ khối ghép phụ âm đầu

3. Ghép phụ âm cuối âm tiết vào nguyên âm:

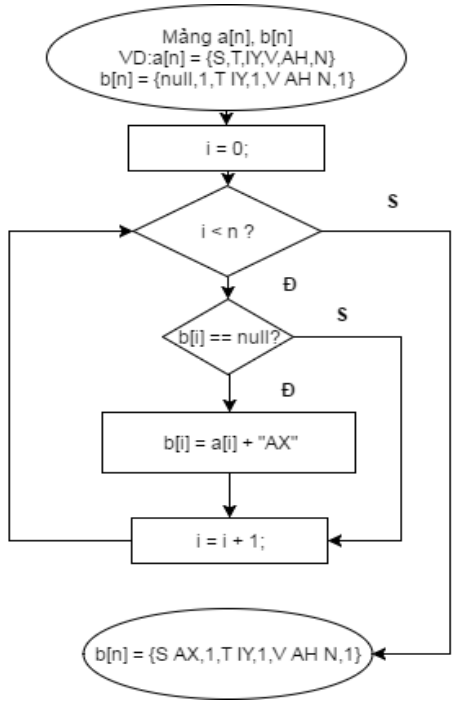
Các phụ âm cuối nếu tồn tại được xác định và ghép vào nguyên âm để hoàn chỉnh âm tiết [5]. Nếu liền kề trước phụ âm cuối đó là nguyên âm, ghép chúng lại với nhau đồng thời đánh dấu vị trí của phụ âm cuối. Ở đây "N" là phụ âm cuối và được ghép vào "AH". Thuật toán được trình bày như Hình 6.



Hình 6. Sơ đồ khối ghép phụ âm cuối

4. Bổ sung nguyên âm:

Như vậy sau khi thiết lập các tổ hợp âm vị phụ âm đầu – nguyên âm – phụ âm cuối, mảng âm vị ban đầu chỉ còn lại những phụ âm rời rạc. Việc thiếu nguyên âm khiến chúng không thể phát âm, chúng tôi xử lí điều này bằng việc bổ sung nguyên âm "AX". Ở đây phụ âm rời rạc là "S". Thuật toán được trình bày như Hình 7.



Hình 7. Sơ đồ khối bổ sung nguyên âm

C. Ánh xạ và tiền xử lý ngoại lệ

Ở bước này, sau khi triển khai thuật toán tách âm vị ở III.B, chuỗi âm tiết được xử lý để tuân thủ quy tắc ghép âm vần của tiếng Việt, sau đó ánh xạ sang chuỗi âm tiết có thể phát âm trong tiếng Việt.

1. Tiền xử lý ngoại lệ

Chuỗi âm tiết thu được còn ngoại lệ cần tiền xử lý trước khi ánh xạ đó là trường hợp các cặp nguyên âm và phụ âm cuối không đi liền được với nhau. Vì số lượng các nguyên âm và phụ âm cuối trong tiếng Việt là hữu hạn và không nhiều [7], nên để xử lý vấn đề này, chúng tôi thống kê tất cả các nguyên âm và phụ âm cuối trong cột Arpabet (Bảng 1) sau đó tổ hợp tất cả các trường hợp có thể xảy ra để đưa ra kết luận và hướng giải quyết.

a) Sắc hóa

Trong tiếng Việt, một từ kết thúc bằng các âm bật hơi (k, p , t) thì luôn được thêm dấu sắc hoặc dấu nặng trong đó số từ thêm dấu sắc chiếm đa số. Chính vì lí do đó, trong trường hợp xuất hiện âm bật hơi đứng cuối âm tiết, việc "sắc hóa" (thêm thanh sắc vào âm tiết) sẽ đưa âm tiết về đúng quy tắc tiếng Việt và có thể phát âm được.

b) Xử lý các cặp nguyên âm và phụ âm cuối không thể đi đôi với nhau

Dựa vào sự tương đồng một cách tương đối về phát âm giữa các âm vị trong tiếng Việt [4], nhóm đưa ra hướng giải quyết cho vấn đề này bằng cách thay thế một trong hai âm vị bằng một âm vị có cách phát âm tương đồng. Quy tắc như sau:

- AH→AX (â → ơ)
- NG, ENG → EN (ng → n)
- AX, IX, ER, AXR, UH R → AH (ơ → â)

2. Ánh xạ

Sau bước tiền xử lý ở 3.3.1, ta thu được chuỗi âm tiết hoàn chỉnh, tuân thủ các quy tắc ghép âm vần trong tiếng Việt. Thao tác cuối cùng đó là ánh xạ từ chuỗi âm tiếng Anh (Arpabet) sang chuỗi âm tiết có thể phát âm trong tiếng Việt dựa trên bảng ánh xạ (Bảng 1).

IV. THỰC NGHIỆM

A. Dữ liệu từ điển

Chúng tôi sử dụng từ điển *cmudict_SPHINX_40.dic* để xây dựng cây quyết định cho t2p. Đây là một từ điển dùng cho mục đích nhận dạng tiếng nói. Từ điển gồm 133031 dòng với mỗi dòng ứng với một từ hoặc kí tự trong Tiếng Anh cùng với cách phát âm của chúng. Bộ âm vị sử dụng trong từ điển gồm 39 âm vị ARPAbet và một âm vị SIL (âm câm).

B. Kết quả thực nghiệm

Chúng tôi đã tiến hành thử nghiệm việc phiên âm từ tiếng anh sang IPA dựa trên phương pháp đề ra ở Phần III. Bên cạnh đó, chúng tôi viết phần mềm chuyển văn bản thành tiếng nói [7] thực hiện việc phát âm để kiểm chứng kết quả đạt được trên các từ Tiếng Anh thông dụng thường gặp trên các trang báo mạng của Việt Nam. Kết quả thực hiện được thể hiện trong Hình 8 và Bảng 2.

Hình 8. Công cụ phiên âm từ vựng tiếng Anh và đọc ra thành âm thanh

Bảng 2. Kết quả thực nghiệm các từ Tiếng Anh thông dụng

Từ tiếng Anh	Phiên âm IPA (theo từ điển Oxford)	Phiên âm của công cụ t2p	Phát âm tiếng Việt tìm được	Phát âm tiếng Việt thường dùng
FACEBOOK	/'feɪsbʊk/	F EY S _ B _ _ K	phây sơ bơ cơ	phây búc
SMARTPHONE	/'smɑːrtfoʊn/	S M AA R T F _ OW N _	xơ mát phâu nơ	xờ mát phôn
ROBOT	/'roʊbɑːt/	R OW B AA T	râu bát	rô bốt
MODEL	/'mɑːdl/	M AA D AH L	ma đơ lơ	mô đồ

Từ tiếng Anh	Phiên âm IPA (theo từ điển Oxford)	Phiên âm của công cụ t2p	Phát âm tiếng Việt tìm được	Phát âm tiếng Việt thường dùng
GOOGLE	/ˈguːgl/	G UW _ G AH L	gu gơ lơ	gu gồ
MICROPHONE	/ˈmaɪkrəfoʊn/	M AY K R AH F _ OW N _	mai cơ rơ phâu nơ	mai cờ rô phôn
LIVESTREAM	/ˈlaɪv striːm/	L IH V _ S T R IY _ M	li vơ sơ tơ rim	lai xờ trim
THAILAND	/ˈtaɪlænd/	_ _ _ AY L AE N D	ai len đơ	thái lan
INTERNET	/ˈɪntərnet/	IH N T _ ER N EH T	in tơ nét	in tơ nét
SERVER	/ˈsɜːrvər/	S _ ER V _ ER	sơ vơ	sơ vờ
CLIENT	/ˈklaɪənt/	K L AY AH N T	cơ lai ân tơ	cờ lai ần
VALENTINE	/ˈvæləntaɪn/	V AE L AH N T IY N _	ve lân tin	va len thai
COMMENT	/ˈkɑːment/	K AH M _ EH N T	cơ men tơ	còm men
OVERNIGHT	/ˌoʊvərˈnaɪt/	OW V _ ER N AY _ _ T	âu vơ nai tơ	âu vờ nai
CONFIRM	/kənˈfɜːrm/	K AH N F _ ER M	cân phơm	còn phơm
PLAYLIST	/ˈpleɪlɪst/	P L EY _ L IH S T	pơ lây li sơ tơ	pờ lay lít
SEARCH	/sɜːrtʃ/	S _ _ ER CH _	sơ chơ	xớt
TENNIS	/ˈtenɪs/	T EH N _ IH S	te ni sơ	ten nít
POSTER	/ˈpoʊstər/	P OW S T _ ER	pâu sơ tơ	pót tơ
FESTIVAL	/ˈfestɪvl/	F EH S T AH V AH L	phe sơ tơ vơ lơ	phéc ti van

Chúng tôi nhận thấy rằng những từ Tiếng Anh được du nhập vào Việt Nam từ lâu như Google, model, robot,… thường có cách phát âm tiếng Việt rất khác so với cách phát âm Tiếng Anh của chính nó, vì vậy phát âm tiếng Việt do tìm được cũng sai lệch so với cách phát âm thường dùng. Lí do là vì chúng đã được Việt hóa để cách phát âm trở nên tự nhiên, gần gũi với Tiếng Việt. Ngoài ra, các từ Tiếng Anh gồm nhiều phụ âm nhưng không có nguyên âm đi kèm như microphone, livestream,... cũng bị sai lệch về hai cách phát âm Tiếng Việt. Điều này xảy ra vì sự khác biệt giữa cấu trúc hai ngôn ngữ Tiếng Anh và Tiếng Việt: Tiếng Việt buộc mỗi phụ âm phải có một nguyên âm kèm theo. Thêm nữa, ở một số trường hợp công cụ t2p thất bại trong việc phiên âm dẫn đến hiện tượng thiếu nguyên âm hoặc phụ âm của từ tiếng Anh cũng ảnh hưởng lớn cách phát âm Tiếng Việt của từ đó, tuy nhiên tỉ lệ này khá thấp: chỉ 2/100 từ là Facebook và Thailand. Ngoài những trường hợp trên, mặc dù cách phát âm tiếng Việt tìm được của các từ còn chưa hoàn toàn tự nhiên, nhưng đã miêu tả khá giống cách phát âm chuẩn Tiếng Anh.

Ở nghiên cứu này, chúng tôi không đưa ra đánh giá về độ chính xác bởi vì việc phiên âm trên các phương tiện thông tin đại chúng ở Việt Nam mang tính chủ quan, phần nhiều dựa trên thói quen và mặt chữ. Do đó thiếu nguồn dữ liệu tham chiếu để thực hiện việc đánh giá định lượng.

C. Hướng giải quyết các hạn chế:

Với các hạn chế còn tồn tại trong việc Việt hóa phát âm từ Tiếng Anh đã nêu trên, chúng tôi đề ra một số giải pháp hậu xử lí. Đối với các từ Tiếng Anh có nhiều phụ âm đứng riêng, chúng tôi cân nhắc bỏ đi các phụ âm đứng ở cuối khi phiên âm bởi Tiếng Việt không phát âm các phụ âm cuối. Về sự thất bại của công cụ t2p trong một số trường hợp, một phương pháp phù hợp để giải quyết là liên tục cập nhật các phiên bản từ điển mới của công cụ t2p. Từ đó, cây quyết định xây được trở nên chính xác và hoàn thiện hơn, giảm tỉ lệ thất bại trong việc phiên âm.

V. KẾT LUẬN

Trong bài báo này chúng tôi đã tóm tắt kết quả nghiên cứu và phát triển công cụ tự động phiên âm một từ vựng tiếng Anh bất kỳ thành chuỗi âm tiết tiếng Việt. Việc nghiên cứu công cụ Việt hóa cách phát âm các từ vựng tiếng Anh này là một tiếp cận có nhiều ứng dụng thiết thực trong xử lý tiếng nói và xử lý ngôn ngữ tự nhiên. Trong tương lai, chúng tôi sẽ tiếp tục nghiên cứu để cải tiến công cụ, hướng tới đưa học máy và bổ sung đánh giá của người sử dụng vào hệ thống để tăng độ chính xác và hoàn thiện cho sản phẩm.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi đề tài mã số T2017-02-93 của Trường Đại học Bách khoa, Đại học Đà Nẵng.

TÀI LIỆU THAM KHẢO

[1] Hoang Gia Ngo, Nancy F. Chen, Sunil Sivadas, Bin Ma, Haizhou Li, “A Minimal-Resource Transliteration Framework for Vietnamese”, Annual Conference of the International Speech Communication Association (INTERSPEECH), 2014.

[2] Hoang Thi Quynh Hoa, “A Phonological Contrastive Study of Vietnamese and English”, Master’s thesis, Texas Technological College, 1965.

[3] John Kominek, “TTS From Zero: Building Synthetic Voices for New Languages”, PhD thesis, Carnegie Mellon University, 2009.

[4] The CMU Pronouncing Dictionary: http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[5] Luis Pedro Hurtarte Caceres, "Translation of Poetry´s Structures to Architecture", Master's thesis, Bauhaus-University in Weimar, 2016.

[6] Kevin Lenzo, "t2p: Text-to-phoneme converter builder":
http://www.cs.cmu.edu/afs/cs.cmu.edu/user/lenzo/html/areas/t2p.

[7] Duy Khanh Ninh, Yoichi Yamashita, "F0 parameterization of glottalized tones in HMM-based speech synthesis for Hanoi Vietnamese", IEICE Transactions on Information and Systems, Vol. E98-D, No. 12, 2015.

A SOLUTION FOR ENGLISH TRANSLITERATION TO VIETNAMESE

Trieu Thi Ly Ly, Bui Thanh Son, Le Thi Ha Binh, Ninh Khanh Duy

***ABSTRACT**: In regard of applications of natural language processing and speech processing, text normalization is a vital problem since the input text often consists of many non-standard words such as abbreviations, numbers and foreign words. This paper introduces a solution for English transliteration to Vietnamese. By researching the pronunciation similarity and phonetic equipvalent, we establish an algorithm that split English phonemes into syllables that is able to be pronounced in Vietnamese and create English phonemes in CMU dictionary to Vietnamese phonemes mapping. Based on that, a tool is built to spell an arbitrary English word into Vietnamese syllables. This tool has been incorporated into a text-to-speech software and therefore demonstrates its usefulness.*