DOI: http://dx.doi.org/10.26483/ijarcs.v13i6.6935

Volume 13, No. 6, November-December 2022

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

COMPUTATIONAL LINGUISTIC MATERIAL FOR VIETNAMESE SPEECH PROCESSING: APPLYING IN VIETNAMESE TEXT-TO-SPEECH

Pham Van Dong
Faculty of Information Technology
Hanoi University of Mining and Geology
Hanoi, Vietnam

Nguyen Tien Thanh
Viettel CyberSpace Center
Viettel
Hanoi, Vietnam

Mac Dang Khoa
VinBigdata
VinGroup
Hanoi, Vietnam

Tran Do Dat
MOST
Ministry of Science and Technology of Vietnam
Hanoi, Vietnam

Do Thi Ngoc Diep
MICA
Hanoi University of Science and Technology
Hanoi, Vietnam

Vu Thi Hai Ha
Phonetics Department
Vietnam Institute of Linguistics
Hanoi, Vietnam

Dang Thanh Mai
Foreign language department
Hanoi University of Mining and Geology
Hanoi, Vietnam

Abstract: The motivation of this paper is to propose a set of best-quality linguistic materials for Vietnamese speech processing, which can be used for Vietnamese TTS and ASR problems. This proposed material includes: (1) a pronunciation dictionary, which adapts from X-SAMPA, (2) a rule-based grapheme to phoneme for Vietnamese. In order to test and evaluate, we have built a Vietnamese TTS system based on the Merlin engine, using the above materials, and evaluating the quality of speech and the accuracy of pronunciation. The results show that the applicability of these materials is favorable for further research and development on Vietnamese speech processing.

Keywords: Text-to-speech, Dictionary, Grapheme-to-Phoneme, X-SAMPA, computer coding, speech processing, Vietnamese

I. INTRODUCTION

Build a speech processing system consisting of many components, of which one of the primary components is linguistic processing, requiring many linguistic materials such as pronunciation dictionaries, etc. This critical part determines pronunciation (in TTS) and correct speech recognition (ASR)**[1]**.

Vietnamese speech processing has been studied since the early 2000s, and so far, it has achieved many results in two main research areas: speech recognition and synthesis. MICA has researched and created some typical application systems such as Hoa Sung (Speech synthesis in the Vietnamese language), and VOVA (High-quality Vietnamese TTS engine) for Android). Leading Vietnamese companies such as Vais (Vietnam AI System), Viettel[1], Zalo[2], Fpt[3], and the world's leading company Google have created many applications such as Google Text to Speech[4].

Although there are many pieces of research and products, there is no unified set of public linguistic materials for Vietnamese. Usually, product research and development parties build their independent linguistic materials, so sharing, and inheriting are difficult. Therefore, the motivation of this paper is to propose a set of materials with the best quality for Vietnamese processing, which can be used for Vietnamese TTS and ASR problems.

This paper is organized as follows. After presenting the introduction and the linguistic materials in section 2, section 3 will show the experiment. Section 4 presents the evaluation process. Furthermore, the final section presents the conclusions and future directions of the study.Proposed linguistic materials

This section will present the main contributions, including the Vietnamese syllable list, the Vietnamese phone list, and the Vietnamese computer encoding.

[1] https://viettelgroup.ai/service/tts
[2] https://zalo.ai/products/text-to-audio-converter
[3] https://fpt.ai/vi/tts
[4] https://cloud.google.com/text-to-speech

A. Unified XSAMPA phone list for Vietnamese

1) The overview of Vietnamese phonetic

According to A. G. Haudricourt [2], in 1954, the Vietnamese-Muong language group was a language or dialect in the early period of Christianity. Then, through the interaction process with the Chinese language and especially the Tai-Kadai language, the language with a highly developed tone system, the tone system in Vietnamese appeared and today, according to the law of forming tones. The appearance of the tones began around the 6th century (northern part of Vietnamese history) with three tones and steady development around the 12th century (Ly) with six tones. Then some early consonants changed to this day. During the change process, the final consonants change the syllable ends, and the consonant ends change from confusion to tangible.

Vietnamese has divided into three main dialects: North, Central, and South. In fact, because of the very long history of development, the North dialect is considered to be a national standard script. It consists of provinces from Northern Vietnam to Thanh Hoa province. The Central dialect refers to that spoken between Nghe An province and Da Nang province. The South dialect is spoken from the South of Da Nang province to the whole Southern province. A mixture of dialects in an area has a boundary between two dialects. For example, the dialect of Thanh Hoa province is nearly similar to the North and the Central ones, sometimes also arranged into the North-central dialect. The differences between dialects are primarily reflected in the tones and consonants. Three main varieties of Vietnamese, North, Central, and South, which are slightly different from each other, are described below.

Table I. Initials in Northen Vietnamese dialect [3]

		Labial	Dental/ Alveolar	Palatal	Velar	Glottal
Nasal		m	n	ɲ	ŋ	
Plosive/Affricate	unaspirated		t	c	k	ʔ
	aspirated		tʰ			
	implosive	ɓ	ɗ			
Fricative	voiceless	f	s		x	h
	voiced	v	z		ɣ	
Approximant			l			

The dialects differ in using tone systems. The Northern dialect uses six tones (level, falling, rising, drop, broken, curve), while the Central and the Southern ones have only five tones, especially some regions of the Central have only four tones. The Central and Southern speakers do not discern clearly between broken and curved. They almost used rising instead of the broken tone. Some North-Central varieties maintain the rising but have a merge of curve and drop. The vowels and consonant pronunciation are also distinct. In the northern dialect, there are six tones.

Northern Vietnamese has no retroflexes /ʈ, ʂ, ʐ/, nor the rhymes /ɑːw, ɑːɛ̆w/. It is the only area that entirely exists the eight final consonants in spelling, including -p, -t, -k, -m, -n, -ng, -u, and - i. d, gi, and r are all pronounced /z/. ch and tr are pronounced /c/, while x and s are pronounced /s/.

Some rural speakers merge /l/ and /n/ into /l/, although this is not considered standard.

Central Vietnamese is known for preserving archaic characteristics in various ways, such as the most significant number of initial consonants, the high amount of ancient words, sub-local tonal variations within the area, and so on (Hoang TC 2003). In the central and southern dialects, retroflexes /ʈ, ʂ, ʐ/, but only five tones exist. Five tones system differs from the Northern Vietnamese system in quantity and quality. There is confusion between the curve and the broken, not distinguishing the broken from the drop. The initial consonant system consists of 24 consonants. In many dialects, there are two aspirated consonants [pʰ, kʰ] instead of two fricatives [f, ɣ]. The front vowel [i, e, ɛ], and back vowel [u, o, ɔ] in the Central dialect tend to shift slightly towards the middle than in the Northern dialect.

Table II. Initials in Central Vietnamese dialect [3]

		Labial	Dental/ Alveolar	Retroflex	Palatal	Velar	Glottal
Nasal		m	n		ɲ	ŋ	
Plosive/Affricate	unaspirated		t	ʈ	c	k	ʔ
	aspirated	pʰ	tʰ			kʰ	
	implosive	ɓ	ɗ				
Fricative	voiceless	f	s	ʂ		x	h
	voiced	v	z	ʐ		ɣ	
Approximant			l				

Southern dialect, where there is less variation in dialect boundaries. In general, there are clearly 5 tones, exist retroflexes /ʈ, ʐ, ʂ/. Confusing the finals -n, -t with -N, -k. The initials system of the Southern dialect has some differences from the initials system in other dialects of Vietnamese. First, the appearance of two the initials /j/ and /w/, in which, /j/ corresponds to three initials /v/, /z/, /ʐ/ in other dialects, while /w/ in the Southern dialect corresponds to the combinations of consonant and glide in other dialects /ʔw/, /hw/, /kw/, /Nw/. Second, confuse /s/ with /ʂ/, /c/ with/ʈ/. There is confusion between the curve and the broken tone.

Table III. Initials in Southern Vietnamese dialect [3]

		Labial	Dental/ Alveolar	Retroflex	Palatal	Velar	Glottal
Nasal		m	n		ɲ	ŋ	
Plosive/Affricate	unaspirated	p	t	ʈ	c	k	ʔ
	aspirated		tʰ				
	implosive	ɓ	ɗ				
Fricative	voiceless	f	s	ʂ		x	h
	voiced		z	ʐ			
Approximant			l	r	j	w	
Rhotic							

The tone of the text is written with the sign: "à" (falling tone), "ã"(broken tone), "ả" (curve tone), "á" (rising tone), "ạ" (drop tone), described in Fig.1.

Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Tone 6
ngang	huyền‘\’	ngã ‘~’	hỏi ‘?’	sắc ‘/’	nặng‘.’

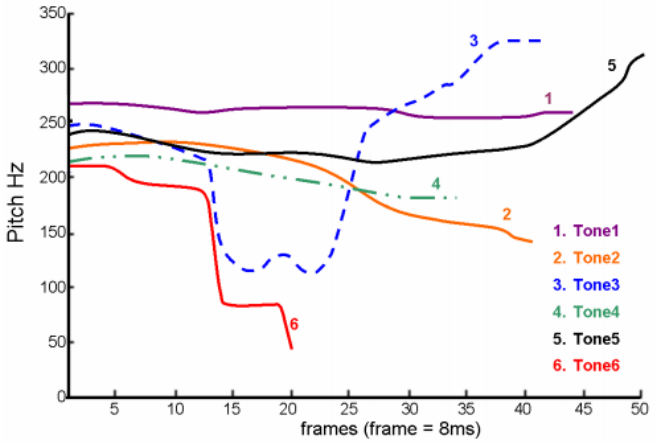

Figure 1 Example of contours of six tones (female subject Pham Ngoc Yen), as described in [4]

Thus, we recommend a listed Vietnamese phone consisting of 19 initial consonants, 12 final consonants, and 13 vowels. Six Vietnamese sounds in our work are recorded in 6 numbers: 0. level tone, 1. falling tone, 2. broken tone, 3. curve tone, 4. rising tone, 5. drop tone.

2) Computer encoding

The International Phonetic Alphabet (IPA[5]) is an alphabetic system of phonetic notation based primarily on the Latin script. The International Phonetic Association devised it in the late 19th century as a standardized representation of speech sounds in written form. IPA has the disadvantage that it is challenging to represent and process on computer code. The Speech Assessment Methods Phonetic Alphabet (SAMPA) is a computer-readable phonetic script using 7-bit printable ASCII characters based on the IPA. The Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA) is a variant of SAMPA developed in 1995 by John C. Wells [5]. It is designed to unify the individual language SAMPA alphabets and extend SAMPA to cover the entire range of characters in the 1993 version of IPA. X-SAMPA is still useful as the basis for an input method for true IPA.

Based on the study of Vietnamese phonetics above, we propose the Vi-XSAMPA phonetic set as an extension of the mapping between IPA and X-SAMPA, serving the coding for the Vietnamese TTS synthesized system due to the following reasons:

- IPA characters have many special characters that make programming difficult, such as ɲ̊, ɲ, ɕ, ʑ …
- Some X-SAMPA characters will cause errors in the list of questions and Merlin programming, such as the @ character, 7 ... because it coincides with the programming keyword

The proposal for the Vietnamese phone set was proposed by [6] and aims to apply to speech synthesis. 25 characters for initial/final consonants and 16 for vowels/diphthongs was used in [6]. In this phone set, many characters make programming difficult. They are ts\, k_+, N_+, 7, i@, u@, and M@.

We have proposed the VI X-SAMPA table in Table IV. Our phonetic set includes 53 characters, developed from X-SAMPA, and covers the entire phonemic set. For tone, the numbers 1 - 5 will be used to represent the tone, unlike the Standard X-SAMPA, which uses some symbols which can be challenging to process. We replaced hard-to-program characters with more good characters. They are: b_< -> b, d_< -> d, t_h -> th, a_X -> aX, O_X -> OX, E_X -> EX, 7 -> ow, 7_X -> aa, i@ ->ie, M@ -> wa, u@ ->uo, k_+ -> kz, k_p -> kp, N_+ -> nz, N_m -> Nm, p -> pc, t -> tc, _M -> 0, _L -> 1, _H -> 4.

Some special symbols in XSAMPA are omitted: _<, for example, b_< => b (similar to b in English), because these two sounds are close, and also want to be able to later share with the English phone board.

Table IV. VIETNAMESE VI-XSAMPA

NO	IPA	Letter	X-SAMPA	VI-XSAMPA
1	ɓ	b, p	b_<	b
2	c	ch, tr	c	c
3	ɗ	đ	d_<	d
4	f	ph	f	f
5	ɣ	g, gh	G	G
6	h	h	h	h
7	k	c, k, q	k	k
8	l	l	l	l
9	m	m	m	m
10	n	n	n	n
11	ɲ	nh	J	J
12	ŋ	ng, ngh	N	N
13	s	x, s	s	s
14	t	t	t	t
15	tʰ	th	t_h	th
16	v	v	v	v
17	z	d, r, gi	z	z
18	χ	kh	X	X
19	w	o, u	w	w
20	a	a	a	a
21	ă	a(u, y), ă	a_X	aX
22	ɔ	o, oo	O	O
23	ɔ̆	o (c, ng)	O_X	OX
24	e	ê	e	e
25	ɛ	e	E	E
26	ɛ̆	a (nh, ch)	E_X	EX
27	ɤ	ơ	7	ow
28	ɤ̆	â	7_X	aa
29	i	i, y	i	i
30	ɯ	ư	M	M
31	o	ô	o	o
32	u	u	u	u
33	iə	ie	i@	ie
34	ɯə	ưa, ươ	M@	wa

[5] https://en.wikipedia.org/wiki/International_Phonetic_Alphabet

35	uə	ua, uô	u@	uo
36	k	c	k	kc
37	ḳ	ch (i, ê, a)	k_+	kz
38	k͡p	c (o,ô,u)	k_p	kp
39	m	m	m	mc
40	n	n	n	nc
41	ŋ	ng	N	Nc
42	ŋ̣	nh (i,ê,a)	N_+	Nz
43	ŋ͡m	ng (o,ô,u)	N_m	Nm
44	p	p	p	pc
45	t	t	t	tc
46	j	i, y	j	j
47	w	u, o	w	wc
48	˧	ngang	_M	0
49	˨	huyền	_L	1
50	˧˩	hỏi		3
51	˧˥	ngã		2
52	˧˥	sắc	_H	4
53	˨˩	nặng		5
54	˧˥	sắc		4
55	˨˩	nặng		5

B. *List of all pronounceable syllables in Vietnamese*

A Vietnamese syllable consists of three obligatory elements: an onset, a tone and a vowel. The syllable may optionally contain an obstruent, nasal, or approximant coda [7].

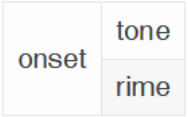

Figure 2. Structure of a syllable

Based on Luong Hieu Thi's word list[6], we have created a combination of 20,190 Vietnamese syllables. In the Figure 1: One-syllable structure of Vietnamese:

• Onset of 27 letters (one letter, two letters, pairs)

• Tone includes six tones (equal, sharp, profound, question, fall, heavy)

• Rime: to generate rimes, the author gives the following table

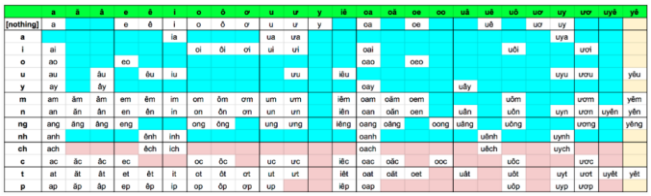

Figure 3. Vietnamese rimes construction table

In the Figure 3: Rimes architecture table with equal tone (blank), Blue areas are rimes that can be combined with six tones, red can combine with sharp and heavy accents (grave, dot), and yellow areas can combine with six tones (but no standing onset).

Looking at Figure 3, we see that each rime will consist of 2 components, called prefix and suffix (prefixes and suffix).

- Step 1: Create the rimes first, based on the tone field in the file meta_data.json and the color area that the rime belongs to (blue, red, or yellow) to determine the correct tone.

- Step 2: After having rhyme, we will generate a syllable by combining it with the onset (except for the yellow zone because the rimes in this region are not associated with the onset)

Code for each cell in the rimes construction table, from top to bottom and from left to right. As a result, we got 20,190 Vietnamese syllables. All the source code and instructions we have published on GitHub[7].

C. *Vietnamese G2P*

Grapheme to phoneme conversion is a key component of the phonetic analysis block. The task of this component is to "convert lexical orthographic symbols to phonemic representation" (i.e., phonemes - basic units of sound) along with "possible diacritic information (e.g., stress placement)" or lexical tones in tonal languages [8]. There are two ways to implement G2P, G2P based on rules and G2P based on a statistical model, training on a pronunciation dictionary (e.g., phonetisaurus [8]). For Vietnamese, which has strict pronunciation rules and can be codified, developing a G2P rule base for Vietnamese is feasible and effective. vPhon[7] is a software that takes UTF-8 Vietnamese orthography and returns UTF-8 output in the IPA alphabet for three significant dialects of Vietnamese: Northern (Hà Nội), Central (Huế), and Southern (Sài Gòn) speech.

We have modified the file north.py of vPhon to include the VI-X-SAMPA transcription, as suggested in the above section. We also edited the code in the vPhon.py file to make the output correctly. All changes are described in detail on my github[9]. Here is an example: "Chúng tôi, những người đang làm sản phẩm BaamBoo Tra Từ, cảm thấy rất tự hào được làm việc cùng các bạn" (We, who are making BaamBoo Tra Tu products, feel very proud to be working with you). This example is transcribed by our VI-X-SAMPA vphone as: "4 c u Nm 0 t o j, 2 J M Nc 1 N wa j 0 d a Nz 1 l a mc 3 s a nc 3 f aa mc BaamBoo 0 c a 1 t M, 3 k a mc 4 th aa j 4 z aa tc 5 t M 1 h a wc 5 d wa kc 1 l a mc 5 v ie kc 1 k u Nm 4 k a kz 5 b a nc". The VI-X-SAMPA vphon will make it easy to convert G2P to X-SAMPA, good support for programming.

D. *Vi-ProDict*

Based on modified Vphon[10], a list of words by Luong Hieu Thi[11], and the proposal of VI-X-SAMPA, we have created a

6 https://gist.github.com/hieuthi/0f5adb7d3f79e7fb67e0e499004bf558

7 https://github.com/phamvandong/All-Vietnamese-syllables-2022

8 https://github.com/AdolfVonKleist/Phonetisaurus

9 https://github.com/phamvandong/vPhon-master-MICA-master-modify

10 https://github.com/kirbyj/vPhon

complete phonetic dictionary of more than 18,000 Vietnamese words following the process below, the entire source code and detailed documentation shared on my Github[12]:

Here are some results of our dictionary:

- ba 0 b a
- bai 0 b a j
- bam 0 b a mc
- ban 0 b a nc
- bang 0 b a Nz
- banh 0 b EX Nz
- bao 0 b a wc

The process of creating a dictionary is described in detail in Figure 4

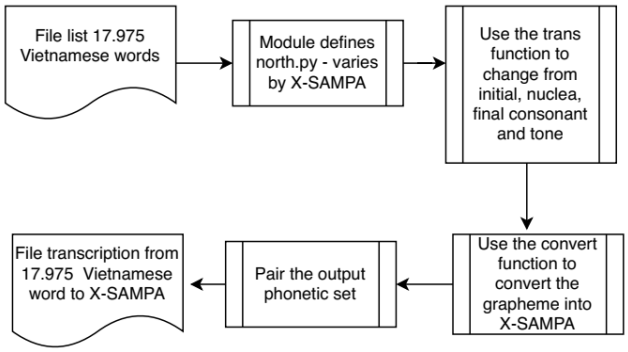

Figure 4. Create dictionary process

II. EXPERIMENT IN VIETNAMESE TTS

This section aims to applying those materials with one of the fundamental problems of Speech Processing: Text to speech. We also check if the materials can help the TTS system correctly pronounce Vietnamese sounds, especially difficult, rare, and often mispronounced: an, ang, eng, éc, eng, eat, etc. We have prepared the data and trained the TTS system using Merlin Tool.

A. Data

The training database is a voice database of more than 3 hours of recording, divided into 2.360 sentences, Le Diem recording voice. The recording file is in a studio with a sampling frequency of 16khz.

B. Merlin Toolkit

The Merlin [9] is a toolkit for neural network-based speech synthesis. Fig.3 is a standard Merlin DNN synthesis architecture. The system takes linguistic features as input and employs neural networks to predict acoustic features, which are then passed to a vocoder to produce the speech waveform. Various neural network architectures are implemented, including a standard feedforward neural network, mixture density neural network, recurrent neural network (RNN), and long short-term memory (LSTM) recurrent neural network, amongst others. The toolkit is Open Source, written in Python, and extensible.

Standard Merlin DNN synthesis architecture is described in detail in.

[11] http://www.hieuthi.com/blog/2017/03/21/all-vietnamese-syllables.html

[12] https://github.com/phamvandong/dictionary_xsampa_for_TTS

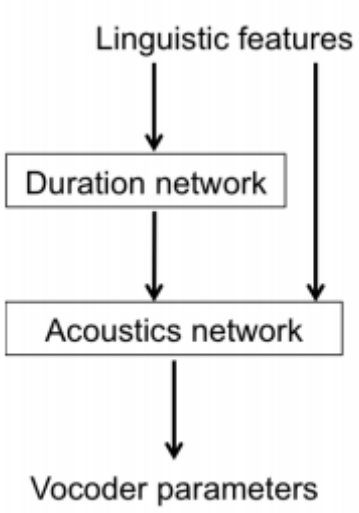

Figure 5. Standard Merlin DNN synthesis architecture [10]

C. Training TTS

For testing this article, we used Merlin, changed the dictionary and Question List files, and tested and evaluated the results. The system uses a 60-dimensional vector of Mel coefficients containing spectral envelope information, a 25-dimension vector of aperiodicities, and a logarithm of F0 [9]. In the training phase, these vectors are used as the acoustic model deep neural network output. In the synthesis phase, these feature vectors (predicted by the DNN acoustic model) are used as input for the vocoder to synthesize speech signals.

III. EVALUATION

The evaluation aims to evaluate the quality of output synthesized Vietnamese speech.

The output Vietnamese speech quality of the TTS system was evaluated according to two synthetic speech quality assessment standards. The naturalness of speech was assessed using the MOS (Mean Opinion Score) criterion and rated with five levels (bad-1, poor-2, fair-3, good-4, excellent-5). The intelligibility criterion refers to the ability to fully convey content through synthetic speech, measured as a percentage of the content intelligible ranging from 0% (worst) to 100% (best).

All these assessments for two criteria were conducted through perceptual experiments with listeners. The system was tested in a low-noise environment with 26 Vietnamese people, balanced between men and women, between the ages of 18 and 70, with no hearing or vision impairments or diseases. All test participants do not participate in the training data-building process. The entire testing process will be guided and supervised by technical staff. During the test, each participant will take turns testing ten pre-designed questionnaires. Each questionnaire comprises five Vietnamese sentences selected randomly from an original set of 200 sentences in 10 different fields: culture, society, international, health, law, sport, agriculture, economy, education, tourism, and politics. These sentences were new and did not exist in the training data. Sentences were distributed among listeners. Each sentence in the original will get the same number of evaluations; 7 different people will hear each sentence.

Participants can listen to the voice results once or again if needed. Then participants will rate the two criteria according to their subjective feelings. The final criteria score for the system was defined as the average value of the evaluation results for all sentences, all hearings, and all participants. The results of the evaluation process are summarized in 54.

Table V. TTS system evaluation result

Output speech quality	**Vietnamese**
MOS (0-5)	4,2
Intelligibility (%)	88%

The MOS scores for Vietnamese were set to 4.2. The high scores indicate that the output speech was almost as natural as human speech. The intelligibility scores of 88% for Vietnamese also show that the output speech was easy to understand and listen to. Both criteria show that Vietnamese's speech's output is of good quality.

The evaluation results show that the Vietnamese TTS system can achieve high results in synthesized speech quality.

IV. CONCLUSION

This paper has summarized and proposed Vietnamese speech processing materials, including a list of more than 18.000 Vietnamese words; VI-X-SAMPA transliteration to encode Vietnamese TTS. We have created a Vietnamese question list with an extensive study of phonetics, creating an 808-dimensional matrix in TTS using the Merlin tool.

We also tested the above materials by using Merlin to build a Vietnamese speech synthesis system. The results are evaluated with a MOS score of 4.2 and an accuracy of 88%, which shows that it is highly satisfactory when applying these materials to problems of Vietnamese speech processing.

V. REFERENCES

[1] P. Taylor, "Text-To-Speech Synthesis," *Camb. Univ. Press*, 2009.

[2] A.-G. Haudricourt, "La place du vietnamien dans les langues austroasiatiques," *Bull. Société Linguist. Paris*, vol. 49, no. 1, pp. 122–128, 1953.

[3] "Phương Ngữ Học Tiếng Việt (NXB Đại Học Quốc Gia 2009) - Hoàng Thị Châu - 287 Trang | PDF," *Scribd*. https://www.scribd.com/document/534836166/Ph%C6%B0%C6%A1ng-Ng%E1%BB%AF-H%E1%BB%8Dc-Ti%E1%BA%BFng-Vi%E1%BB%87t-NXB-%C4%90%E1%BA%A1i-H%E1%BB%8Dc-Qu%E1%BB%91c-Gia-2009-Hoang-Th%E1%BB%8B-Chau-287-Trang (accessed Dec. 14, 2022).

[4] Q. C. Nguyen, "Reconnaissance de la parole en langue Vietnamienne," PhD Thesis, Grenoble INPG, 2002.

[5] J. C. Wells, "Computer-coding the IPA: a proposed extension of SAMPA," *Revis. Draft*, vol. 4, no. 28, p. 1995, 1995.

[6] N. T. T. Trang, C. D'ALESSANDRO, A. RILLIARD, and T. Do Dat, "HMM-based TTS for Hanoi Vietnamese: issues in design and evaluation," in *14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, 2013, pp. 2311–2315.

[7] J. Kirby, "Kirby, James. vPhon: a Vietnamese phonetizer." Nov. 15, 2016. Accessed: Nov. 21, 2019. [Online]. Available: https://github.com/kirbyj/vPhon

[8] T. T. T. Nguyen, "HMM-based Vietnamese Text-To-Speech: Prosodic Phrasing Modeling, Corpus Design System Design, and Evaluation," Paris 11, 2015. Accessed: May 27, 2017. [Online]. Available: http://www.theses.fr/2015PA112201

[9] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System.," in *SSW*, 2016, pp. 202–207.

[10] Z. Malisz, H. Berthelsen, J. Beskow, and J. Gustafson, "Controlling Prominence Realisation in Parametric DNN-Based Speech Synthesis.," in *INTERSPEECH*, 2017, pp. 1079–1083.