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.2017.00095

RAPID DEVELOPMENT OF TEXT TO SPEECH SYSTEM FOR UNSUPPORTED LANGUAGES USING FAKING INPUT APPROACH: EXPERIMENT WITH MUONG

Pham Van Dong[1], Mac Dang Khoa[2], Vu Thi Hai Ha[3], Tran Do Dat[4]

[1] Khoa CNTT, Trường đại học Mỏ Địa chất

[2] Viện nghiên cứu quốc tế MICA, Trường đại học Bách Khoa Hà Nội

[3] Phòng Ngữ âm học, Viện ngôn ngữ học

[4] Văn phòng các chương trình trọng điểm cấp nhà nước, Bộ Khoa học và Công nghệ

phamvandong@humg.edu.vn, dang-khoa.mac@mica.edu.vn, haihavu28@gmail.com, tddat@most.gov.vn

ABSTRACT: *Development of a text-to-speech (TTS) system for a language normally requires a big workload, it need much research on spoken language processing and also linguistic research. For minority languages or under-resourced languages (so called unsupported languages), building a TTS system is very hard work, in some cases it is impossible, due to the lack of available resources and linguistic knowledge. This paper presents the experiment of building a TTS system for Muong language, a minority language of Vietnam, following the "faking input" approach. Based on the phonetic correlation between Muong language and Vietnamese language - a closely related language, a set of phonetic transformation rules was proposed to transform the text input of Muong language to the suitable input of Vietnamese Text-To-Speech system. That allows synthesizing the Muong speech by available Vietnamese TTS systems. The Muong synthesized speech was thus evaluated by Muong peoples in perception tests. The result shows that the synthesized speech is well understood, this suggests the possibility of this approach for rapidly and simply building TTS systems for unsupported languages.*

KEYWORD: *Text-to-speech, faking input approach, unsupported language, Muong language, Vietnamese.*

I. INTRODUTION

Today's speech-processing technology plays an important role in many aspects of human - machine interaction. Many voice interaction systems have been introduced recently allowing users to communicate with devices on a variety of platforms such as smartphones (Apple Siri[1], Google Now[2], Samsung S-Voice[3]) smart car (BMW[4], Ford[5], etc), smart home (Amzone Echo[6], Google Home[7]). In these systems, one of the main components is speech synthesis, or Text-To-Speech (TTS), which can convert input text into speech. Developing a TTS system for a language is not only the deployment of speech processing technique, but also requires linguistic research such as phonetics, phonology, syntax and grammar. Therefore, currently TTS systems are available for 20 or so of the world's major languages, for thousands of other, so called "unsupported" languages, no such technology is available [1]. That is because of lack of critical requirements to build a TTS system such as specific speech data and linguistic knowledge.

Vietnam is a multi-ethnic country with 54 ethnic groups[8]. Of which the Kinh, who speak Vietnamese, is the majority ethnic group accounting for 85.6% of the total population. The remaining 53 ethnic groups are ethnic minorities accounting for 14.4% [2]. It is the abundance of ethnic groups that leads to cultural diversity and especially to the language. Our concern is to provide support to those who speak 53 ethnic languages. Among them, Muong is one of the five largest population ethnic groups in Vietnamese, but it does not yet have the official writing system. That make Muong language is still an under-resourced language in Vietnam, although having more than 1 million speakers. With these reason, Muong language was chosen as the objective for this research. The goal is to build a Text-to-Speech system for Muong rapidly and cheaply.

Our approach to build a TTS for Muong following the idea proposed by Evans, Polyzoaki and Blenkhorn [3]. The idea is to use an existing TTS system for a primary language (basic language - BL) to "emulate" TTS for an unsupported language (target language - TL). In that study, this method was applied to develop fake synthesizers for Greek, Albanian, Czech, Welsh, and several other languages. Using the same technique in 2006 , Harold Somers et al. [1] has developed an experiment with Somali by "faking" on the TTS system of German. In this paper, we will present

[1] https://www.apple.com/ios/siri/

[2] https://en.wikipedia.org/wiki/Google_Now

[3] https://en.wikipedia.org/wiki/S_Voice

[4] https://www.bmw.vn/vi/index.html

[5] http://www.ford.com/

[6] https://en.wikipedia.org/wiki/Amazon_Echo

[7] https://madeby.google.com/home/

[8] According to Vietnamese census results in 2009

the proposal to synthesize the Muong speech by faking some available Vietnamese TTS systems. After the introduction, section II present the proposal of building TTS for Muong. Based on the phonetic correlation between Muong language and Vietnamese language - a closely related language, a set of phonetic transformation rules was proposed to transform the text input of Muong language to the suitable input of Vietnamese Text-To-Speech system. The Muong synthesized speech was thus evaluated by perception test with native listeners, presenting in section III. This paper ends with some discussion and conclusion.

II. FAKING TTS FOR MUONG

The ideal of faking approach for TTS based on the phonetic relation between the BL and the TL. The work of building a faking TTS for an unsupported language includes the following tasks:

- Choosing a BL which is linguistically close to the TL.
- Proposal orthography mapping between BL and TL, based on the phonetic similarly between 2 languages.
- Building the faking TTS for BL by applying the phonetic mapping on the available TTS of BL.

This section will present our work in building a faking TTS for Muong, target language (TL), using Vietnamese as the basic language (BL)

A. The linguistic relation between Vietnamese and Muong

Vietnamese, the official language of Vietnam, is spoken natively by over seventy-five million people in Vietnam and greater Southeast Asia as well as by some two million overseas, predominantly in France, Australia, and the United States. In Vietnam, Muong is one of the five largest population ethnic groups. In terms of language family, Vietnamese and Muong are belongs to the same group Viet Muong belong to the Mon-Khmer branch of the Austroasiatic family [4].

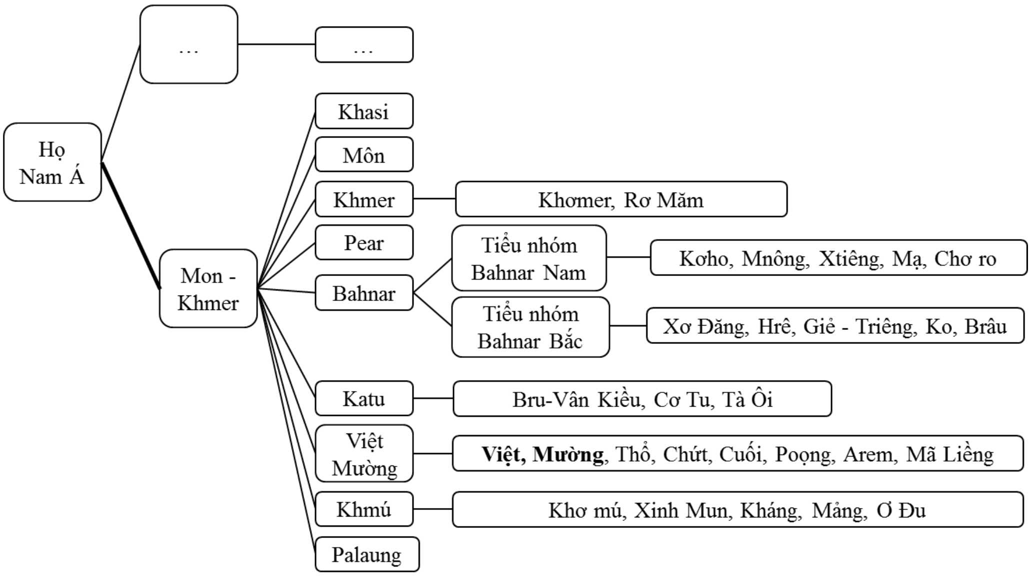
…
…
Họ Nam Á
Khasi
Môn
Khmer
Khơmer, Rơ Măm
Pear
Mon - Khmer
Bahnar
Tiểu nhóm Bahnar Nam
Kơho, Mnông, Xtiêng, Mạ, Chơ ro
Tiểu nhóm Bahnar Bắc
Xơ Đăng, Hrê, Giẻ - Triêng, Ko, Brâu
Katu
Bru-Vân Kiều, Cơ Tu, Tà Ôi
Việt Mường
Việt, Mường, Thổ, Chứt, Cuối, Poọng, Arem, Mã Liềng
Khmú
Khơ mú, Xinh Mun, Kháng, Mảng, Ơ Đu
Palaung

Figure 1. Mon-Khmer branch of the Austroasiatic family [5], 175-176

As one of 5 largest spoken languages in Vietnamese, Muong language have been study since the middle of the last century [6]–[10]. These researches focus on issues such as describing the phonetic characteristics of Muong dialects in localities different [10], [11]; the history of Muong origins [6] or some proposal for transcription and writing [8], [12], [13]. However, until now, Muong does not yet have the official writing system. Muong has an indisputable similarity with Vietnamese, in term of phonology, tone, syntax and vocabulary [14]. Like in Vietnamese, Muong have many dialects. Among them, the Muong dialect of Hoa Binh province is the most widely spoken, and the only Muong dialect have an official proposal for writing system [15]. The Hanoi Vietnamese and Muong dialect of Hoa Binh were chosen as the BL and TL for this experiment of "faking" Text-to-Speech system of Muong language. The following section discusses the phonetic characteristics of the two languages and suggests the phonetic transformation rules for faking method.

B. Phonetic transformation rules for Muong faking TTS

The phonetic transformation rule set is the most important factor of TTS system using faking method. The objective of these rules is to transform the orthography (writing text) of target language to the suitable orthography of BL, in order to using in TTS of BL. The transformation rules are proposed base on the phonetic similar between BL and TL. In this work, based on the phonetic characteristic of Muong and Vietnamese, which have been studied in many linguistic research, the transformation rule between these two languages was proposed. When comparing Vietnamese and Muong, the phonetic element of two languages can be divide into 3 group:

- Equivalent elements: Muong phonemes coincide with phonemes in Vietnamese, so we can use equivalent simulations.
- Closed elements: Muong phonemes are similar to phonemes in Vietnamese, so we can use simulators to replace phonemes approximate.
- Distinct element: Muong phonemes are not found in phonemes in Vietnamese, in this paper we are not deal with this problem, it will be resolved in the future work.

The following will present the proposal for the transformation rules between Muong and Vietnamese in term of consonant, vowel and tone.

Firstly, in terms of the initial consonant, Muong Hoa Binh has 24 initial consonants /b, k, c, d, g, h, hr, k^h, kl, l, m, n, ŋ, ɲ, p, p^h, r, t, t^h, tl, v, w, s, z/ [15], the Vietnamese has 20 initial consonants which are /ɓ, t, t^h, ɗ, tɕ, k, ʔ, m, n, ŋ, ɲ, f, v, s, z, x, ɣ, h, w, l/ [4]. Closed consonants between Muong and Vietnamese are /b - ɓ, c - tɕ, g - ɣ, k^h – x, p^h – f/, they are part of the reason for the decline in the quality of synthetic speech. The five consonant phonemes in Muong are not in Vietnamese: /p, r, hr, tl, kl/, in writing respectively p, r, hr, tl, kl. The Muong Hoa Binh final consonants consists of 9 consonants /p, t, c, k, m, n, ɲ, ŋ, l/ and 2 approximants 2 /w, j/ [15]. Hanoi Vietnamese licenses eight segments in coda position: three unreleased voiceless obstruent /p, t, k/, three nasals /m, n, ŋ/, and two approximants /j, w/. With consonant tone Ha Noi Vietnamese distinguishes the palate of the mouth /k͡p/ followed by o, u, ô (/oŋ͡m˦/ - ông) and /ŋ̟/ followed by i, ê, a (/sik̟˥/ - xích). With the nasal consonant distinguish the next /ŋ͡m/ followed by o, u, ô (/oŋ͡m˦/ - ông), and /ŋ̟/ followed by i, ê, a (/kiŋ̟˦/ - kinh) [4]. So that, we see only seven consonants /p, t, k, m, n, ɲ, ŋ/ and two semi-vowel /w, j/ in Muong is equivalent to 6 consonants of vietnamese. Two consonant /c, l / we must find alternative equivalent in Vietnamese. This issue will be addressed in the following article. Muong has one medial /w/ is written in w. For example, kwêl khwắn (smoking), khwải (snack); kwa (we), kwải (throw), kwang (clean). Vietnamese has a medial /w/ written by two letters o and u. Example hoa quả [15].

In the vowels system, Muong Hoa Binh has 14 vowel sounds /a, ă, ɤ̆, ɛ, e, i, ɔ, o, ɤ, u, ɯ, iə, ɯɤ, uə/ [15], the Vietnamese have 16 vowels and two vowels: /a, ă, ɤ̆, ɛ, e, i, ɔ, o, ɤ, u, ɯ, iə, ɯə, uə, ɔ̆, ɛ̆/ [16, p. 58]. So that, we can see that the vowel system of Muong and Vietnamese is equivalent in 11 vowels and 2 diphthongs. The difference is that, in the Muong language, the diphthong /ɯɤ/ is a transitive of dipthong ươ, in Vietnamese it is transcribed into /ɯə/. There are several differences in orthography which required us to adapt Muong spelling rules to Hanoi Vietnamese, for example changing êê to ê, oo to o, ôô to ô, uu to u, ưư to ư [15]. As for the vowel system, Vietnamese has 16 vowels [17]. Muong monophthongs, written with the vowel letters *a, e, i, o, u* correspond reasonably equivalent to Hanoi Vietnamese vowels written similarly. Muong has 14 vowels, and the diphthong /ɯɤ/ - ươ is replaced by another transcription /ɯə/. Hanoi Vietnamese has quite complex letter-to-phoneme mappings, especially to do with voicing. The Muong-to- “Hanoi Vietnamese” transliteration process is largely but not entirely automatable, so some manual revision of the texts is necessary.

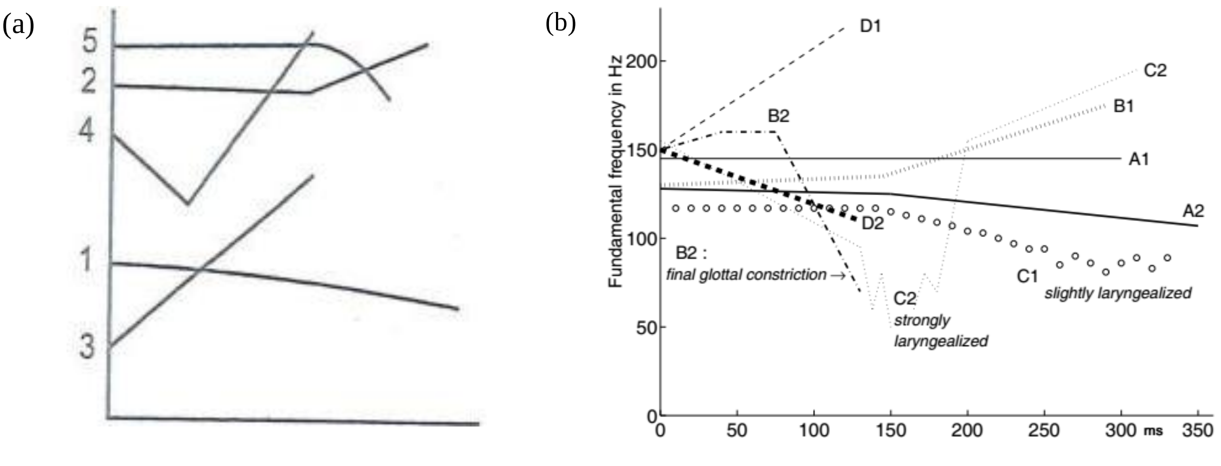

Figure 2. Tonal system of Muong (a) [11, p. 84] and Vietnamese (b) [18]

In terms of tone, most authors when studying the tone of the Muong dialect all say that Muong has 5 tone ([7], [8], [9], [12], [11]). Other authors, including Hoang Thi Chau, who surveyed the Muong language in the epic poem "Đẻ đất đẻ nước” show that, there are only four tones [18]. The local style has made the tone system of the Muong dialect based on the difference in tone of the tone 2, Nguyen Van Tai (generalized) classified the tone system of the Muong dialects into two groups: group 1 - characterized by the break of tone 2, group 2 - characterized by the break of the tone 2 in the high or low tone. In this paper, we use the Muong tonal system described in [15], which contain five tone: 33- Level, 42- Falling, 323 - Falling Rising, 34 - High Rising, 342?- Low Falling. Meanwhile, Hanoi Vietnamese have 8 tones, a six-tone paradigm in open or sonorant final syllables and a two-tone paradigm in syllables ending in an unreleased oral stop (**Figure *2***) [4]. We propose using 5 Vietnamese tones: level (A1), falling (A2), rising (B1), drop (B2), curve (C1) to faking the 5 Muong tones.

Following the phonetic characteristics analysis between Vietnamese and Muong above, the phonetic mapping for consonants, phonemes, vowels and tones between Muong and Vietnamese are proposed in the tables **Table *1***.

Table 1. Muong and Vietnamese phonetic comparison (orthography in bold, IPA in italic; Vi: Vietnamese; Mu: Muong)

Group	Equivalent				Closed		Distinct	
	Muong	Viet	Muong	Viet	Muong	Viet	Muong	Viet
Initial consonants	**k, c** */k/*	**k, c, q** */k/*	**t** */t/*	**t** */t/*	**b** */b/*	**b** */ɓ/*	**hr** */hr/*	-
	h */h/*	**h** */h/*	**th** /t^h/	**th** /t^h/	**ch** */c/*	**ch, tr** */tɕ/*	**kl** */kl/*	-
	l */l/*	**l** */l/*	**v** */v/*	**v** */v/*	**đ** */d/*	**đ** */ɗ/*	**p** */p/*	-
	m */m/*	**m** */m/*	**w** */w/*	**u, o** */w/*	**g** */g/*	**g** */ɣ/*	**r** */r/*	-
	n */n/*	**n** */n/*	**x** */s/*	**x** */s/*	**kh** /k^h/	**kh** */x/*	**tl** */tl/*	-
	ng */ŋ/*	**ng, ngh** */ŋ/*	**z** */z/*	**d, gi** */z/*	**ph** /p^h/	**ph** */f/*		
	nh */ɲ/*	**nh** */ɲ/*						
Final consonants	**p** */p/*	**p** */p/*	**nh** */ɲ/*	**nh** */ɲ/*			**ch** */c/*	-
	t */t/*	**t** */t/*	**ng** */ŋ/*	**ng** */ŋ/*			**l** */l/*	-
	c */k/*	**c** */k/*	**w** */w/*	**o, u** */w/*				
	m */m/*	**m** */m/*	**i, y** */j/*	**i, y** */j/*				
	n */n/*	**n** */n/*						
Vowel	**aa, a** */a/*	**a** */a/*	**ôô, ô** */o/*	**ô** */o/*				
	ă */ă/*	**ă** */ă/*	**ơ** */ɤ/*	**ơ** */ɤ/*				
	â */ɤ̆/*	**â** */ɤ̆/*	**uu, u** */u/*	**u** */u/*				
	e */ɛ/*	**e** */ɛ/*	**ưư, ư** */ɯ/*	**ư** */ɯ/*				
	êê, ê*/e/*	**ê** */e/*	**iê** */iə/*	**iê** */iə/*				
	i */i/*	**i** */i/*	**uô** */uə/*	**uô** */uə/*				
	oo, o */ɔ/*	**o** */ɔ/*	**ươ** */ɯɤ/*	**ươ** */ɯə/*				
Glide	**w** */w/*	**u, o** */w/*						
Tones	**Muong tones**				**Vietnamese tones**			
	33 - Level				A1 – Level <Ngang>			
	42 - Falling				A2 – Mid falling <Huyền>			
	324 - Falling Rising				C1 – Low falling < Hỏi >			
	34 - High Rising				B1 - Rising <Sắc>			
	342? - Low Falling				B2 – Low glottalized <Nặng.>			

Based on the above phonetic comparisons, the transformation rules are proposed for mapping from Muong orthography to Vietnamese orthography, which can be read in Vietnamese TTS. For equivalent and close cases, the transformation rules are the simple replacement of Muong item by Vietnamese items in Table 1. For the distinct case, it is impossible to transform these items in Muong to any item in Vietnamese. Therefore, that cases are not considered in this study and will be deal in the future work. Table 2 shows some examples of applying transformation rules to convert the Muong text into input text for Vietnamese TTS.

Table 2 *Examples of applying transformation rules to convert the Muong text into input text for Vietnamese TTS*

Muong Text	Faking text for Vietnamese TTS	English
Ho tang học bài	*Ho tang học bài*	'I'm studying'
Ho phải za ty dộng bầy?	*Ho phải da ty dộng bầy?*	'I'm with you go out?'
Nhà za chiếm từ cúi chăng?	*Nhà da chiếm từ cúi chăng?*	'Your house has many pigs?'

C. Building faking TTS for Muong

The purpose of this work is to find out a simple and cheap way to generate Muong synthesized speech. Therefore, our approach is to develop some independent module, which can convert the Muong transcript to the suitable input of available Vietnamese system. This approach allows Muong TTS to be developed independently, and can works with different Vietnamese TTS systems. **Figure *3*** shows the structure of Muong faking TTS system, which includes 3 main modules.

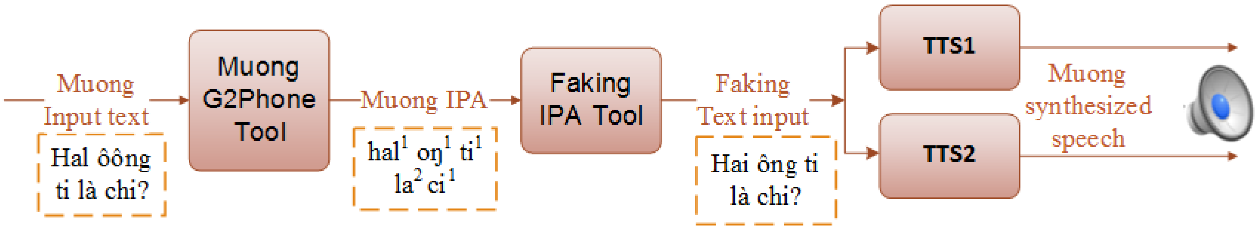

Figure 3. Faking TTS for Muong

Firstly, The Muong G2Phone Tool is to convert the Muong text into IPA of Muong. The input is a Muong writing text, using the Muong Hoa Binh writing system proposal in 2016 [15]. The output of this module is the Muong phonetic transcription in IPA. Muong G2Phone Tool is developed in Java, using G2P rules set is based on phonetic research of Muong [11]. Secondly, the Muong Faking IPA module is to convert the Muong IPA into IPA transcribed in Vietnamese and then convert it into transcription text so that Vietnamese TTS can be read. The faking rules set is based on the proposal transformation rules mention in section II.

For the Vietnamese TTS system, with the independent module structure, the system can be applying with different Vietnamese TTS systems (as in Figure 3). There are two main current approaches in speech synthesis: unit selection approach and statistical parametric approach (HMM/DNN). The available Vietnamese TTS system are also follow these two approaches, such as VOS TTS (Voice Of Southern Vietnam) [9], MICA TTS [10], vnSpeak[11] (unit selection technique); VAIS TTS[12], OpenFPT TTS[13] (statistical parametric technique). According to [19], the unit selection approach has advantages of producing high quality at the waveform level, because it concatenates speech waveforms directly. This technique is also easy to implement [20], and has been researched in a longer history. By contrast, statistical parametric approaches, which generates the average of some set of similarly sounding speech segments [19], cannot be compared with unit selection in producing natural voice [21]. For testing Muong faking TTS, we chose the Vietnamese TTS system of both technique, in order to examine whether the TTS technique affect to the sound quality of Muong synthesize Speech. We also chose the Vietnamese TTS as web service for convenience in using and pairing the module. Finally, two TTS web service for Vietnamese, which both support generate Hanoi Vietnamese speech, were chosen for our experiment. They are MICA TTS service (unit selection technique - TTS1) [22], [28], [29] and OpenFPT TTS (statistical parametric technique - TTS2). The faking results of the built-in system with 2 Vietnamese TTS will be tested in the section later.

III. EVALUATION

The objective is to examine whether the native Muong listener understands the faking Muong speech and how do they judge the quality of Muong faking speech.

A. Testing materials

The testing material were design to examine the transformation rules proposed in section II. The test data so is divide into 3 groups:

- Group 1 – Faking tones testing: For examine whether the proposal Vietnamese tone can “fake” the Muong tones. Five tones in Muong were set in 5 syllables, with simple structure (Consonant-Vowel) and put in a same containing sentence, as in **Table 3**.

Table 3. Testing material for faking tone

Muong Tone	Vietnamese tone for faking	Containing sentence (in Muong)	IPA	Faking text	Vietnamese meaning
33 - Level	A1 – Level <Ngang>	**ka**	kaA1	ca	gà
42 - Falling	A2 – Mid falling <Huyền>	**mè**	mɛA2	mè	mè
34 - High Rising	C1 – Low falling < Hỏi >	**ná**	naB1	ná	nỏ
324 - Falling Rising	B1 - Rising <Sắc>	**tẻ**	tɛC1	tẻ	đẻ
342? - Low Falling	B2 – Low glottalized <Nặng.>	**mệ**	meB2	mệ	mẹ

- Group 2: Phone transformation testing, 5 equivalent phonemes will be tested in five sentences requiring the listener to clearly state the words that he or she has just heard, and evaluate the sentence quality.

[9] http://www.ailab.hcmus.edu.vn/
[10] http://mica.edu.vn/vova/
[11] http://www.vnspeak.com
[12] https://vais.vn/en/text-to-speech-service/
[13] http://ngtts.stis.vn/#/demo

Table 4. Testing material for faking phone (the concerning phonemes in **bold**)

Muong word	Transcription IPA	Transformation (IPA)	Faking text for TTS	Vietnamese meaning
bang	**b**aŋ1	**ɓ**aŋA1	**b**ang	con hoẵng
cha	**c**a1	**tɕ**aA1	**c**ha	vườn
gế	**g**e4	**ɣ**eB1	**gh**ế	ghế
kha	**k**^{h}a1	**x**aA1	**k**ha	vợt bắt cá
phui	**p**^{h}ui1	**f**uiA1	**ph**ui	vui

- Group 3: General testing, with the remaining phonemes left, we took out a set of 5 sentences for the Muong people to listen and record their sentences just heard, and evaluate the quality of the sentence.

Table 5. Testing material for remaining phonemes

Muong sentence	Muong IPA	IPA transcription	Faking text	Vietnamese sentence
Chú mua của oi?	cu4 muə1 cuə3 ɔi1	cuB1 muəA1 cuəC1 ɔiA1	Chú mua của oi?	Anh mua của ai?
Ho tang cúm lọ	hɔ1 taŋ1 cum4 lɔ5	hɔA1 taŋA1 cumB1 lɔB2	Ho tang cúm lọ	Tôi đang sẩy lúa
Cải chi ni?	kai3 ci1 ni1	kaiC1 ciA1 niA1	Cải chi ni?	Cái gì đây
Da bí thía nó à?	da1 bi4 t^{h}ia4 nɔ4 a2	daA1 biB1 t^{h}iəB1 nɔB1 aA2	Da bí thía nó à?	Mày bị làm sao thế?
Ở lái ăn cơm hái	ɤ3 lai4 ăn1 kɤm1 hai4	ɤC1 laiB1 ănA1 kɤmA1 haiB2	Ở lái ăn cơm hái	Ở lại ăn cơm nhé

Fifteen sentences of 3 group above were set as input of faking TTS system for Muong with 2 Vietnamese TTS (as mention in section II). Totally, the output are 30 synthesized utterances (15 sentences x 2 TTS techniques). These utterances are stored as audio files to using in the perceptual test.

B. Experiment protocol

The testing protocol was designed following the testing method for synthesized speech, proposed by ITU [30]. This test was design for 2 purposes:

- Examine the intelligibility of Muong faking speech: whether the listener can understand exactly the content of the testing sentence.
- Evaluate the quality of the synthesize sound: How the listener judges the natural of Muong faking speech

The test is set up in a quiet room, using a normal headset at a basic hearing level. Seven Muong natives (2 male, 5 females, with a mean age of 25) participated in the test. Five of the participants speak the dialect of Muong Hoa Binh, two participants speak the dialect of Muong Tan Son, Phu Tho province, all of them know Muong Hoa Binh writing system. The Muong participant all know Vietnamese also. For testing, the listener will listen to each sentence one to three times. After listening, the listeners were asked to:

- (1) write down the sentence they heard in Muong text and in Vietnamese meaning. That will be use for the intelligibility factor.
- (2) and giving them the quality assessment scores. The score is calculated on the following scales: 5 - Very good (like natural voice), 4 - Pretty (quite natural), 3- Moderate (Acceptable), 2 - Poor (Hard to hear), 1 - Bad (Inaudible).

C. Results

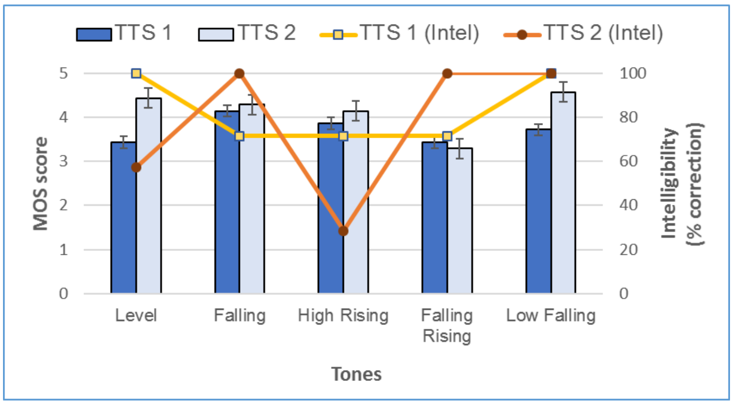

Figure 4. Intelligibility and MOS Results for Muong faking tones

Figure 4 shows result of intelligibility test and MOS test [30] of the Muong synthesized speech on 1st group (Tone Testing). The result shows that basically the Muong tone was successfully identified, only the Falling tone, the High Rising tone and the Falling – Rising tone were recognized at 71%, 71% and 57% respectively. In general, for the intelligibility tone test, both TTS systems are well recognized. TTS1 has higher recognition rates on Falling Rising tone and Low Falling tone. For the MOS tone test, the average rating of TTS1 is 3.71, TTS2 is 4.14. This result shows that both TTS systems are highly regarded for their quality. In the figure, the level, falling, falling rising, low falling tone can be replaced by equivalent, high rising tone has low recognition rate to find the appropriate alternative.

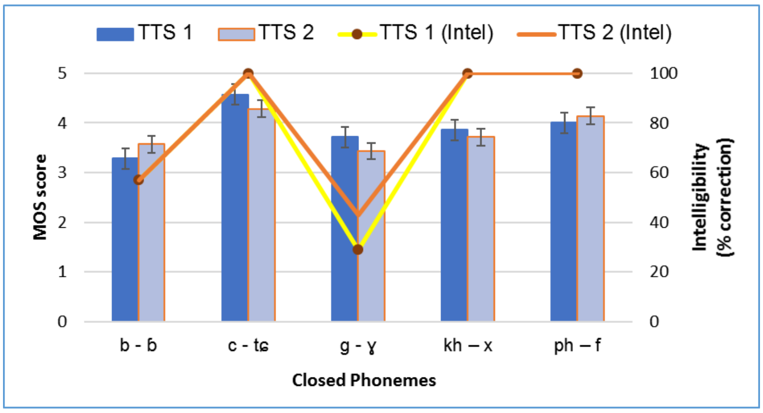

Figure 5. Intelligibility and MOS Test Result for faking close phonemes

The results of test group 2, the closed phoneme test is presented in the Figure 5. Three pairs of phonemes are nearly equivalent /c - tɕ, kh – x, ph – f/ with a 100% identity, pair of phonemes /b - ɓ/ have a mean of 57%, the rate of identification of the phoneme pairs /g - ɣ/ is rather low, 29% with TTS1 and 43% with TTS2. The average quality rating of TTS1 was 3.89, of TTS2 was 3.83. This is a fairly high score, and both systems produce roughly equivalent results. Thus, closed pairs of / c - tɕ, kh - x, ph - f / can be used as substitutes, the / b - ɓ / phonetic pair may be studied further. In combination, the / g - ɣ / closed phoneme pairs have a low understanding rate, so we must use a different approach instead of the one used.

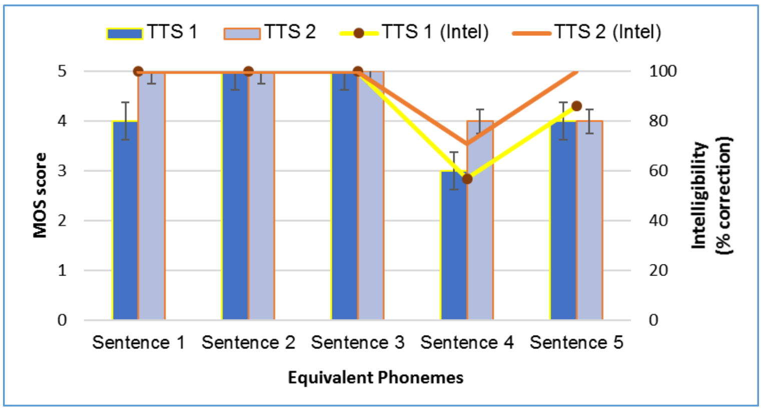

Figure 6. Intelligibility and MOS Test Result for Equivalent phonemes

The results of test group 3, similar phonemes test are shown in the **Figure 6**. One hundred percent absolute rate with sentences 1, 2, 3. Sentence 5 has a high recognition rate of 100% with TTS2, 86% with TTS1. Sentence 4 has an above average identification rate of 57% with TTS1 and 71% with TTS2. The average quality rating of TTS1 is 4.41, of TTS2 is 4.26. This is a fairly high score, and both systems produce roughly equivalent results. The general MOS test results for 3 test group. The average quality rating of TTS1 is 3.90, of TTS2 is 4.08. This is a fairly high score, and both systems produce roughly equivalent results. A set of phonemes of equivalent phonemes has high recognition rates and quality scores indicating that one can use this method to develop a faking TTS.

IV. CONCLUSIONS

The results show that, the Muong faking synthesis system is understandable but not immediate in some cases. Some synthesized speeches are unclear but in general, most synthesized speeches are intelligible for listeners. Participants feel the faking voice is similar to Vietnamese, and they found that the sentence does not have the intonation as some dialects of the Muong language. This study followed closely the new writing system of Muong - Hoa Binh; obtained results showed that cause of similarity of phonemes system, the fake-approach based synthetic voices are understandable and perceptive. The evaluation of voice quality scores (MOS test) is quite high for the Muong people who take the test. The study also uses two high quality TTS1 and TTS2 speech synthesis systems. Experimental results also show that in some cases, the TTS1 system scores higher scores but in other cases the TTS2 system scores higher scores. This is also a hint that the researcher can continue to choose the direction of development to go deeper into solving the more difficult problems in the emulation technique. In general, both TTS systems have produced satisfactory results. From this research, we can see that this approach can be applied experimentally to quickly create a TTS system for the languages of some other Vietnamese ethnic minorities.

The number of testers is few, so the results may not be convergent. It is because finding people test for Muong language is very difficult and constrained, such as having to dialect, know Vietnamese, write letters etc. This study only treats similar elements. The distinct elements will require extensive intervention in the TTS system, such as adding phonemes in the TTS training database, which will be investigated in the future work.

REFERENCES

[1] H. Somers, G. Evans, and Z. Mohamed, "Developing speech synthesis for under-resourced languages by 'faking it': An experiment with Somali," *Pap. Submitt. LREC*, 2006.

[2] Ban chỉ đạo Tổng điều tra dân số và nhà ở Trung ương, *Tổng điều tra dân số và nhà ở Việt Nam năm 2009: Kết quả toàn bộ*. Hà Nội: Nxb Thống kê, 2010.

[3] D. G. Evans, K. Polyzoaki, and P. Blenkhorn, "An approach to producing new languages for talking applications for use by blind people," in *International Conference on Computers for Handicapped Persons*, 2002, pp. 575–582.

[4] J. P. Kirby, "Vietnamese (Hanoi Vietnamese)," *J. Int. Phon. Assoc.*, vol. 41, no. 3, pp. 381–392, Dec. 2011.

[5] Trần Trí Dõi, *Ngôn ngữ các dân tộc thiểu số ở Việt Nam*. Hà Nội: Nhà xuất bản Đại học Quốc gia Hà Nội, 2016.

[6] J. Cusinier, *Les Mường: Géographie humaine et sociologie*, vol. 45. Institut d'Ethnologie, 1948.

[7] Nguyễn Phan Cảnh, "Khảo sát về thanh điệu tiếng Mường (phương ngữ Mường Bi) trong các từ tách rời," *Thông báo Khoa học*, vol. 1, p. 36, 1962.

[8] Nguyễn Kim Thản, "Vài nét về hệ thống âm vị tiếng Mường và phương án phiên âm tiếng Mường," *Ngôn Ngữ*, vol. 1, 1971.

[9] Nguyễn Minh Đức, "Một vài nét về các thổ ngữ của tiếng Mường Hòa Bình," trong *Tìm hiểu ngôn ngữ các dân tộc thiểu số ở Việt Nam*, Hà Nội: Nhà xuất bản Khoa học xã hội, 1972.

[10] Nguyễn Văn Tài, "So sánh hệ thống ngữ âm tiếng Mường một số vùng quanh Hòa Bình," trong *Tìm hiểu ngôn ngữ các dân tộc thiểu số ở Việt Nam*, vol. 1, Hà Nội: Nxb Khoa học xã hội, 1972.

[11] Nguyễn Văn Tài, *Ngữ âm tiếng Mường qua các phương ngôn*. Hà Nội: Nxb Từ điển Bách khoa, 2005.

[12] Nguyễn Như Ý, "Dự thảo phương án chữ Mường." Tọa đàm Viện Ngôn ngữ học, 1994.

[13] Nguyễn Văn Khang, "Về cách ghi phiên âm tiếng Mường trong 'Đẻ đất đẻ nước' và những vấn đề đặt ra khi làm chữ Mường," in *Kỷ yếu "Trao đổi khoa học về chữ Mường,"* Hà Nội, 1994.

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

[15] UBND tỉnh Hòa Bình, "Quyết định 2295/QĐ-UBND phê chuẩn Bộ chữ dân tộc Mường Hòa Bình 2016." Ủy ban nhân dân tỉnh Hòa Bình, 08-Sep-2016.

[16] T. T. T. Nguyen, "HMM-based Vietnamese Text-To-Speech: Prosodic Phrasing Modeling, Corpus Design System Design, and Evaluation," Paris 11, 2015.

[17] T. T. Đoàn, *Ngữ âm tiếng Việt*, Nhà xuất bản Đại học Quốc gia Hà Nội. 1997.

[18] Hoàng Thị Châu, *Xây dựng bộ chữ phiên âm cho các dân tộc thiểu số ở Việt Nam*. Hà Nội: Nxb Văn hóa dân tộc, 2001.

[19] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[20] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Springer Science & Business Media, 2007.

[21] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.

[22] D. D. Tran, "Synthèse de la parole à partir du texte en langue vietnamienne," Grenoble, INPG, 2007.

[23] V. H. Quan and C. X. Nam, "Phrase-based concatenation for Vietnamese TTS," *J. Inf. Technol. Anh Commun. Vietnam.*, vol. 1.

[24] A. T. Dinh, T. S. Phan, T. T. Vu, and C. M. Luong, "Vietnamese HMM-based speech synthesis with prosody information," in *Eighth ISCA Workshop on Speech Synthesis*, 2013.

[25] S. T. Phan, T. T. Vu, and M. C. Luong, "Extracting MFCC, F0 feature in Vietnamese HMM-based speech synthesis," *Int. J. Electron. Comput. Sci. Eng.*, vol. 2, no. 1, pp. 46–52, 2013.

[26] T. T. Vu, M. C. Luong, and S. Nakamura, "An HMM-based Vietnamese speech synthesis system," in *Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on*, 2009, pp. 116–121.

[27] N. T. T. Trang, T. Do Dat, R. Albert, A. Christophe, and P. T. N. Yen, "Intonation issues in HMM-based speech synthesis for Vietnamese," in *The 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14),* pp. 98–104.

[28] T. Van Do, D.-D. Tran, and T.-T. T. Nguyen, "Non-uniform unit selection in Vietnamese speech synthesis," in *Proceedings of the Second Symposium on Information and Communication Technology,* 2011, pp. 165–171.

[29] D.-K. Mac and D.-D. Tran, "Modeling Vietnamese Speech Prosody: A Step-by-Step Approach Towards an Expressive Speech Synthesis System," in *Trends and Applications in Knowledge Discovery and Data Mining*, Springer, 2015, pp. 273–287.

[30] I. T. Union, "Methods for subjective determination of transmission quality," *ITU-T Recomm.*, p. 800, 1996.

PHÁT TRIỂN NHANH HỆ THỐNG TỔNG HỢP TIẾNG NÓI CHO CÁC NGÔN NGỮ CHƯA ĐƯỢC HỖ TRỢ: THỬ NGHIỆM VỚI TIẾNG MƯỜNG

Phạm Văn Đồng, Mạc Đăng Khoa, Vũ Thị Hải Hà, Trần Đỗ Đạt

TÓM TẮT: *Phát triển các hệ thống tổng hợp tiếng nói từ văn bản (TTS) cho một ngôn ngữ thông thường đòi hỏi nhiều công sức, như xây dựng bộ dữ liệu tiếng nói chuyên biệt cho hệ thống, đồng thời cũng cần kết hợp nhiều kết quả nghiên cứu về mặt ngôn ngữ (ví dụ: ngữ âm học, âm vị học, ngôn điệu) cho ngôn ngữ cần tổng hợp. Với những ngôn ngữ dân tộc thiểu số hay ít nguồn tài nguyên (gọi là các ngôn ngữ không được hỗ trợ), việc xây dựng một hệ thống TTS lại càng gặp nhiều khó khăn, đôi khi là không thể thực hiện được do sự thiếu hụt các nguồn tài nguyên và các kiến thức ngôn ngữ học. Bài báo này trình bày sự thử nghiệm xây dựng hệ thống tổng hợp tiếng nói cho tiếng Mường, một tiếng dân tộc thiểu số ở Việt Nam, theo hướng tiếp cận "giả đầu vào". Dựa trên những nghiên cứu về sự tương quan ngữ âm giữa tiếng Mường và tiếng Việt (là những ngôn ngữ cùng thuộc nhóm Việt Mường), một tập các luật chuyển đổi ngữ âm được đề xuất để biến đổi văn bản chữ viết tiếng Mường thành đầu vào thích hợp cho hệ thống tổng hợp tiếng nói tiếng Việt. Việc này cho phép tạo ra tiếng nói tổng hợp tiếng Mường từ hệ thống TTS tiếng Việt sẵn có. Tiếng nói tổng hợp tiếng Mường đầu ra của hệ thống sau đó được đánh giá bởi bài đánh giá cảm thụ của người Mường bản địa. Kết quả đánh giá cho thấy tiếng nói Mường tổng hợp hoàn toàn có thể hiểu được. Kết quả cho thấy khả năng của hướng tiếp cận này trong việc xây dựng hệ thống TTS cho các ngôn ngữ chưa được hỗ trợ một cách đơn giản và nhanh chóng.*