Speech translation for Unwritten language using intermediate representation: Experiment for Viet-Muong language pair

Pham Van Dong[1, 2*], Do Thi Ngoc Diep[2*], Mac Dang Khoa[3,] Vu Thi Hai Ha[4]

[1]Hanoi University of Mining and Geology;
[2]Hanoi University of Science and Technology;
[3]VinBigdata – VinGroup;
[4]Vietnam Institute of Linguistics.
*Corresponding authors: phamvandong@humg.edu.vn; diep.dothingoc@hust.edu.vn
Received 10 Sep 2022; Revised 29 Nov 2022; Accepted 15 Dec 2022; Published 30 Dec 2022.
DOI: https://doi.org/10.54939/1859-1043.j.mst.CSCE6.2022.65-76

ABSTRACT

The paper studies an automatic translation method that translates from the text of a language (L1) to the speech of an unwritten language (L2). Normally the written text is used as the bridge to connect a translation module that translates from the text of L1 to the text of L2 and a synthesis module that generates the speech of L2 from the text. In the case of unwritten language, an intermediate representation has to be used instead of the writing form of L2. This paper proposes the use of phoneme representation because of the intimate relationship between phonemes and speech in one language. The proposed method was applied to the Viet-Muong language pair. The Vietnamese text needs to be translated into Muong language in two dialects, Muong Bi - Hoa Binh and Muong Tan Son - Phu Tho, both unwritten. The paper also proposes a phoneme set for each Muong language and applies them to the problem. The evaluation results showed that the translation quality was relatively high in both dialects (for Muong Bi, the fluency score was 4.63/5.0, and the adequacy score was 4.56/5.0). The synthesized speaking quality in both dialects is acceptable (for Muong Bi, the MOS score was 4.47/5.0, and the comprehension score was 93.55%). The results also show that the applicability of the proposed system to other unwritten languages is promising.

Keywords: Machine translation; Text to speech; Ethnic minority language; Vietnamese; Muong dialects; Unwritten languages; Intermediate representation; Phoneme representation.

1. INTRODUCTION

Recent years deserve to be called the era of information and communication technology. Especially natural language processing (NLP) technology has shown a vital role in supporting human life in many human-machine communication applications. NLP technology has been put into products and services by many significant technology corporations such as Google, Microsoft, Watson, Apple, etc. They primarily focus on the significant languages in the world, such as English, Chinese, Arabic, etc. Among the living languages in the world, there are about 3,074 languages that are not written1. Unwritten languages have not been researched much and suffered many disadvantages leading to their gradual disappearance. So focusing on studying the NLP and machine translation technologies for unwritten languages is still a new and essential task worldwide.

A machine translation system (from text to text or from speech to speech) for unwritten language pairs has been tried in some approaches. The most expensive approach is to build a script for a non-script language [1]. However, this method requires a high cost in terms of time, human resources, and language investment with a large budget and cannot be reused for other languages. Another way, instead of defining a

[1]https://www.ethnologue.com/enterprise-faq/how-many-languages-world-are-unwritten-0

standard script for an unwritten language, an intermediate representation was used rather than the script of the unwritten language in the translation system. The text machine translation module (together with speech recognition and speech synthesis modules in case of speech translation) will operate on these intermediate representations [2, 3]. Among the researches for unwritten languages, phonetic representation of speech had been proposed as one of the representations instead of text [4]. Other intermediate representations have also been proposed, such as using a set of phonological symbols that are recognized from the pure acoustic characteristics of speech signals [5], or meaning representations have also been proposed [6], but the result is still modest. Of course, there is also a proposal for direct translation without using intermediate representation. Nevertheless, this research is recommended for very close language pairs [7].

In this paper, the text of language L1 is translated to the speech of language L2, where L1 is a written language, and L2 is an unwritten language. The translation approach through an intermediate representation at the phonological level was studied because of the intimate relationship between phonemes and speech in one language. In the experiment, Vietnamese and Muong languages are chosen as L1 and L2 languages, respectively. Vietnamese is the official language of Vietnam. Some studies in the field of machine translation for Vietnamese text and other languages have also been focused on since the 2000s, such as English – Vietnamese, French - Vietnamese text machine translation [8, 9], Vietnamese - Japanese text machine translation [10], Google Translate, etc. However, in Vietnam, there are a lot of other minority ethnicities, including unwritten languages. Muong is an unwritten language, and it is a closed language to Vietnamese. The application of modern language processing technologies to the Viet – Muong language pair can bring out socio-economic benefits, and it also opens a new area of research for Vietnamese languages that have the potential to yield many exciting research results. The proposal is also hoped to be applied to other minority ethnicities in Vietnam.

This paper is organized as follows. After presenting the background and related works and the proposed method 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.

2. BACKGROUND

Some related works have been studied to build a translation system through intermediate representation, including (2.1) Related works. In section (2.2), a method to translate the text of one language to the speech of another unwritten language through a suitable intermediate representation has been presented.

2.1. Related works

2.1.1 Machine translation through intermediate representation

Current machine translation technology is aimed at translating from speech to speech. Regarding the general mechanism, this system is a combination of three components or basic technologies: speech recognition, text machine translation, and speech synthesis. Figure 1 depicts how to combine the three technologies through text as a connector. In the paper, the translation starts from the text of L1 to the speech of L2 so the speech recognition module is omitted.

Machine translation automatically transforms a piece of text from a source language into another target language. Methods based on statistical machine translation are being

carried out widely because of their applicability regardless of language or field of translation. From a relatively large database of parallel texts in source and target languages, machine learning algorithms will extract information and statistical "rules" to match text fragments between the two languages based on the calculation of statistical probabilities and selecting the most optimal target language text sentence for an input source sentence [12, 13]. In the past several years, neural network-based machine translation techniques have been presented in the studies of [14-16], etc. The current limitation of this technique is that the amount of data required to train and the computational power required to implement it is very large. There are also some limitations in vocabulary size, sentence length, language complexity, etc. The study [17] also proposed some solutions to use NMT for low training resource conditions, and the results show that it can be equivalent to and slightly better than SMT. However, with the unwritten language, there still needs to be research using NMT applied.

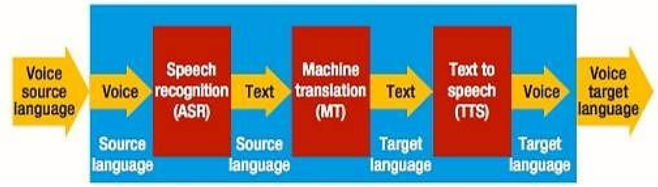

Figure 1. *An ordinary voice translation system [11].*

A speech synthesis system can be considered as a system that can generate speech from the input text (Text to Speech). The system usually consists of two parts. A high-level synthesis module (Natural Language Processing) is responsible for parsing and converting the input text into the appropriate parameter string through text analysis, phonetic analysis, and intonation analysis. The low-level synthesis module (digital signal processing) will receive the appropriate parameters to be analyzed from the high-level module and then feed into the digital signal processing component to generate a corresponding signal waveform. Speech synthesis has gone through many different technologies and approaches, such as concatenative, statistical parametric, and deep learning. Speech synthesis using deep learning techniques has been researched and developed for the past 3-4 years. This method can be hybridized with the synthetic approach using statistical parameters but using neural networks in learning speech parameters. The application scope of neural networks can be only part of the training phase (in combination with HMM) or the whole, using Deep Neural Network (DNN) in the whole process, parameter training for the synthesis system [18, 19].

In translation systems for an unwritten language, the text of the unwritten language needs to be replaced by an intermediate representation. Instead of operating on text, the text machine translation and speech synthesis modules will operate on these intermediate representations. Proposals to use phonological-level intermediate representations in speech processing of non-written languages have been proposed in a number of studies. One of the first experimental studies in automatic speech translation for unwritten languages was performed by [20]. They focus on transcribing the speech database of unwritten languages into a sequence of phones and developing speech processing tools for Basaa, Myene, and Embosi languages [21]. The transcription is done using the automatic phoneme recognition module, then the "word units" in the non-script language are automatically detected from the phoneme sequences by an automatic word separator.

Experimental results have shown that this method is effective and can be applied to many unwritten languages. The work of [22] and [23] investigated the possibility of speech translation based on phonemic representations. Next, a series of other studies revolved around finding an intermediate representation for the speech signal of a non-written language, to replace the text representation in the machine translation problem. Most of these representations are based on automatic phonemic transcription [22-24].

For speech synthesis, to synthesize the speech of a non-script language, [25] uses the phoneme set of a language close to the target language. The authors continue to use advanced integrated techniques such as bootstrapping rotation and further separating word-like structures from phonemic sequences to improve the quality of speech synthesis for non-written languages [26, 27]. The authors use English phonemes in these studies to synthesize German speech (assuming a language without a script). Using English, German, and Marathi phonemic level data, the team, then extended experiments to synthesize Dari, Iraqi, Pashto, Thai, Ojibwe, Inupiaq, and Konkani. The synthetic speech is considered intelligible even though the input training data is non-script. The above result shows that phoneme is one of the best choices for intermediate representations.

2.1.2. Viet - Muong language pair

In this paper, the research language pair is the Vietnamese - Muong language pair. Muong is an unwritten language, and it is closely related to Vietnamese. The Muong ethnic group is one of the five ethnic groups with the largest population compared to other ethnic minorities in Vietnam [28]. Some of the linguistic research had been presented for the Muong language [29-33]. However, until now, an agreement on phoneme set for Muong is not set. So the paper also proposes a phoneme set for each Muong language and applies it to the translation problem. After the field research for each dialect in Hoa Binh and Phu Tho provinces, the linguistic information was proposed.

The Vietnamese and Muong syllabic structure has the same five components: onset, glide, nucleus, coda, and tone. The nucleus and tone play an essential role that cannot be absent in syllables. Regarding the phonemic system, Vietnamese, Muong Bi, and Muong Tan Son have many equivalent and different phonemes. For the onset, there are 18 initial consonants in the two Muong dialects, similar to the Vietnamese initial consonants/b, m, t, d, t^h, n, s, z, l, c, ɲ, k, ŋ, ʔ, h, f, χ, ɣ/. There are two consonants /z, ʂ/ are present in Vietnamese but not in Muong. There are four consonants present in the Muong language but not in Vietnamese /p, w, tl (kl), r/. There are two consonants similar to Vietnamese but only in the Muong Tan Son dialect but not in Muong Bi /v, ʈ/. The Muong glide has the same function and position as the Vietnamese glide. Vietnamese has 16 vowels for the nucleus, while Muong has only 14 vowels. Muong language does not have two short vowels /ĕ/ and /ŏ/, like in Vietnamese. Vietnamese have eight codas, including six consonants /p, t, k, m, n, ng, nh/ and two semi-vowels /u, i/. Muong language has 11 codas with the distinction of 2 coda pairs /k/ và /c/; /ŋ/ and /ɲ/ and coda /l/. As for tones, Vietnamese has six tones, and Muong has five tones. There is no high-rising broken tone like in Vietnamese.

The table comparing the phonemic system of the two Muong dialects with Vietnamese has been detailed at link: https://tinyurl.com/dongpv1.

2.2. Proposal Method

Figure 2. *The proposed method of translating Vietnamese text into unwritten ethnic minority languages in Vietnam using intermediate representations.*

The proposed system of translating Vietnamese text into non-written ethnic minority speech using an intermediate representation of phoneme level consists of two components. The first component is the module that automatically translates the Vietnamese text into the phonological representation of the ethnic minority language. The second component is a speech synthesis system based on a sequence of phonological representations of ethnic minority languages. These two components are described in figure 3.

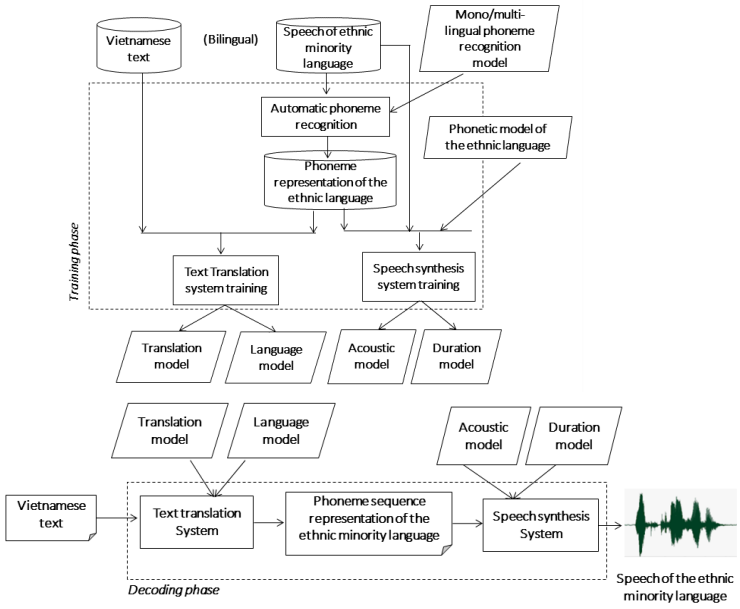

Figure 3. *Training and Decoding of translating Vietnamese text into unwritten ethnic minority speech using an intermediate representation of the phoneme level.*

Derived from a bilingual database of Vietnamese texts - ethnic language speech, the speech data part is transcribed into a phonemic sequence using an automatic phoneme recognizer. After transcribing the speech of ethnic languages, a bilingual database consisting of Vietnamese texts – phonemic representations of ethnic languages are used to train models (translation models, language models) for text translation systems. The database of phonemic representation and corresponding speech is also used to train models of the speech synthesis system. The text-speech translation system is finally combined from these two components using phonemic sequence representation of the ethnic minority language.

The automatic phoneme recognizer is built using phoneme recognition models of this language, or a language close to the non-script language or a multilingual phoneme-recognition model from many languages close to non-scripted languages.

In the case of the Viet-Muong language pair, due to the inexistent of a phoneme recognition model for the Muong language, a new one has been trained from a small

number of manually annotated speech. Given the technologies and data at the moment, using an automatic phoneme recognizer to transcribe audio files of a non-script language is a machine learning method. However, its accuracy absolutely cannot be achieved. Therefore, the output of phoneme sequence still needs to be corrected by linguists so that the transliteration database has the highest accuracy. Using automatic phoneme recognizers can be considered a pre-processing step for linguists in transcribing, reducing their time and effort.

3. EXPERIMENT

For experimentation, the main following tasks have been performed:

- Building bilingual data on Vietnamese Text and Muong's speech in two dialects;
- Building the SMT of Vietnamese text into a phonological representation of Muong;
- Building Muong TTS using the phone sequence of Muong.

3.1. Database building

To build a bilingual database including Vietnamese text and Muong speech, the process follows three steps below.

a) A Vietnamese text database of 20,000 sentences was collected from online newspapers to maximize vocabulary and balance word distribution. There are around 160,000 words, with an average of 8 comments in each sentence. Of those, 7000 words are unique. Due to the shortage of human resources for labeling, text data collection is still relatively limited.

b) Muong's speech corresponding to this Vietnamese text was recorded in sound-proof rooms. Four Muong native speakers, two males, and two females, from 2 dialects (Muong Bi – Hoa Binh and Muong Tan Son – Phu Tho) were chosen to record the database. All speakers are Muong radio broadcasters with good, clear, and coherent voices. The speakers read each Vietnamese sentence in the collection of 20.000 sentences and then speak them in Muong speech. The male voices of two dialects were used to train the system (the female voices are reserved for other phonetic studies). Detail of the text and speech corpora building can be referred to in [7].

c) Automatic transcription: Firstly, the phoneme recognition model for each Muong dialect was built. The 5000 sentence pairs of Vietnamese text and Muong speech were randomly selected for each dialect, and the Muong speech part was transcripted manually by the linguist according to the proposal phoneme set. For each speech, there are four levels of data labelling. Level 1 is a Vietnamese sentence, level 2 is Vietnamese's words, level 3 is Muong's tone, and level 4 is Muong's phone corresponding to the Muong speech, as shown in figure 4. The phoneme recognition model was built for these 5000 Muong speech and phoneme representation pairs using the Kaldi toolkit[2]. The phoneme recognition model was applied to the rest of the 15,000 Muong speech. Finally, a post-editing was done by Muong Linguists to correct the wrong phonemes according to the heard speech and the proposed phoneme set. After this step, bilingual corpora of 20,000 Vietnamese texts and the corresponding phoneme representation sequences in each Muong dialect were built and ready for the training step. Table 1 presents some examples of the training database for the machine translation module.

[2] https://kaldi-asr.org

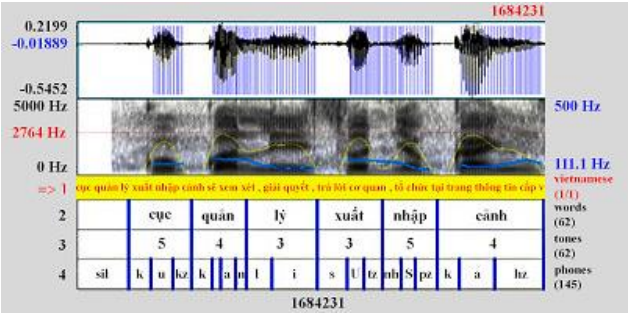

Figure 4. *The result after manual annotation.*

Table 1. *Examples of labelling Vietnamese text into an intermediate representation of Muong Bi and Muong Tan Son phonemes.*

Input Vietnamese text sentence	English meaning	Muong Bi phoneme intermediate	Muong Tan Son phoneme intermediate
Để khắc phục tình trạng thiếu nước ngọt sinh hoạt, người dân miền Tây có nhiều cách như tích trữ nước mưa.	People in the West have many ways to overcome the shortage of fresh water for daily life, such as storing rainwater.	ti khAkz fukz , tihz tlagz thieuz dakz ngocz , sihz hwatz , ngU@iz z$nz mienz t$iz , ko tU kacz , nh@ tikz trU dakz , mU@	tE khAkz fukz tihz tragz thieuz rakz ngwacz , sihz hwatz , ngU@iz z$nz mienz t$iz , ko nhieuz kacz , nhU ticz trU rakz , mU@
Chung kết cuộc thi Đại sứ du lịch Quảng Trị đã diễn ra tại thành phố Đông Hà, Quảng Trị.	The final round of the Quang Tri Tourism Ambassador contest took place in Dong Ha city, Quang Tri.	cugz kEtz kuokz thi , daiz sU , zu licz , kwagz tri , ta zienz tha @ thahz fO dOgz ha kwagz tli	cugz kEtz kuokz thi , daiz sU , zu licz , kwagz tri , taiz zienz ha taiz thahz fO dOgz ha , kwagz tri

3.2. System development

3.2.1. Text to phone translation

The MOSES[3] toolkit (with GIZA ++) was used to build a translation system with the default configuration parameters. The Text To Phone module is built with limited training data (20,000 parallel samples). Although it is possible to use NMT with the improvements suggested in the paper [17], the authors decided to use SMT with the MOSES framework due to its simplicity in implementation and computational resources. At the same time, according to the paper [17], the results of NMT with SMT in this low resource condition have no significant difference. It is unnecessary to use a complex model like Transformer because it is easy to overfit the model, and the Moses model is a simple probabilistic model that is suitable for small amounts of data and only requires 20,000 sentences of machine translation data. After some text pre-processing, the training data includes 16,785 pairs of sentences for Muong Bi and 12,899 pairs of sentences for Muong Tan Son (the testing data includes 200 pairs of sentences). In order to determine whether the generated sound is significantly influenced by the input noisy

[3] http://www2.statmt.org/moses/

text or not, we aim to create a straightforward intermediate representation machine translation model with an acceptable level of accuracy. The quality of the system that translates Vietnamese text into the phonological representation of the Muong language was evaluated with the BLEU score. The BLEU score for the Vietnamese text - Muong Bi phoneme representation translation system was 42.93%. For Muong Tan Son, the BLEU score was 63.29% because Muong Tan Son sounds more like Vietnamese, so the linguistics can have consistency in the annotation. In general, the quality of the translation system was pretty good.

3.2.2. Phone to Sound Conversion

This module was implemented similarly to a text-to-speech system, with a phone sequence as input. The system model will follow the model architectural as in the figure 5.

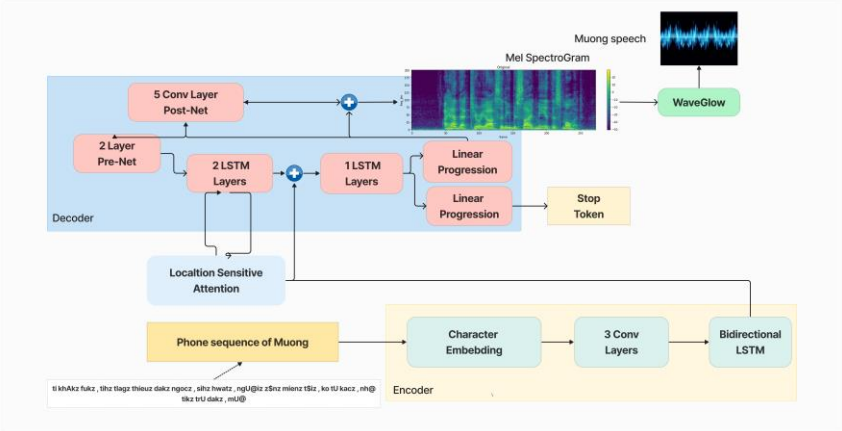

Figure 5. *System Architecture.*

Our model has used Tacotron 2 model [19]. The network consists of an encoder and a decoder with attention. The encoder converts the phone sequence of Muong into a hidden feature representation, and the decoder is to predict a spectrogram. Input is represented using a learned 512-dimensional character embedding. The output of the final convolutional layer is to generate the encoded features. We use a content-based tanh attention decoder, where a stateful recurrent layer generates an attention query at each decoder time step. That query is combined with the context vector and fed into the decoder RNN, consisting of GRU cells with residual connections; these connections help speed up the convergence of the model. The output of the decoder is an 80-band Mel-scale spectrogram.

We have trained two network models of Tacotron2 and WaveGlow to build the phone sequence of Muong to Muong speech. The training steps of the Tacotron2 and WaveGlow networks used the default parameter settings of the original networks. The training dataset contains 20,000 bilingual phone sequences of Muong-Muong speech pairs of a sentence. One thousand files were randomly used for validation, another 1000 files for testing and the remaining files were fed into the train. All models were trained on a GPU, NVIDIA GTX 2080Ti, with batch sizes of 16. The acoustic model converged after 100k steps, while the vocoder converged after 100k steps.

4. EVALUATION

The purpose of the evaluation is to evaluate the quality of the translation system in two categories: the quality of machine translation and the quality of the output

synthesized Muong speech.

Typically, the quality of the automatic translation of the text can be evaluated automatically by comparing the output text of the translation system with the manually translated text by the human using some standard metrics such as BLEU, NIST, WER, etc. However, in our case, the output of the translation system is not a text but a speech in an unwritten language. So the objective/automatic evaluation scores for translation can not be calculated. Human annotators evaluated the quality of the translation with two traditional criteria: adequacy and fluency [34]. The adequacy criterion is to rate the amount of meaning expressed in a Vietnamese text that is also expressed in the Muong speech after translation. The fluency criterion asks annotators to rate the well-formedness of the Muong speech in the Muong language. This criterion indicates whether the Muong speech after translation follows the Muong grammar or not. The adequacy rate includes five levels (none-1, little-2, much-3, most-4, all-5). The fluency rate includes five levels (incomprehensible-1, disfluent-2, non native-3, good-4, flawless-5).

The output Muong speech quality of the translation 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 four criteria were conducted through perceptual experiments with listeners. The system was tested in a low-noise environment with two sets of participants: people from the Muong ethnic group in Tan Son district, Phu Tho province, and people from the Muong ethnic group in Tan Lac district, Hoa Binh province. Each group of participants consisted of 10 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 100 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; 10 different people will hear each sentence.

Table 2. *Vietnamese text to Muong speech translation system evaluation result.*

Evaluation Criteria		**Muong Bi**	**Muong Tan Son**
Translation quality	*Fluency (0-5)*	4.63	4.88
	Adequacy (0-5)	4.56	4.65
Output speech quality	*MOS (0-5)*	4.47	4.32
	Intelligibility (%)	93.55%	92.17%

Participants can listen to the translation results once or again if needed. Then participants will rate the four 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 table 2.

The fluency scores of 4.63 for Muong Bi and 4.88 for Muong Tan Son show that the output sentences produced have a high degree of fluency, almost equivalent to the fluency of the Muong language. The adequacy scores of 4.56 for Muong Bi and 4.65 for Muong Tan Son also show that the translation sentences contain most of the original Vietnamese sentence content, and rare information was lost. Both results prove that the quality of the automatic translation system from Vietnamese text to Muong speech is highly appreciated.

For synthetic Muong speech quality, the MOS scores for Muong Bi and Muong Tan Son were set to 4.47 and 4.32, respectively. The high scores indicate that the output speech was almost as natural as human speech. The intelligibility scores of 93.55% for Muong Bi and 92.17% for Muong Tan Son also show that the output speech was easy to understand and listen to. Both criteria show that Muong's speech's output is of good quality. It helps to evaluate also the proposed phoneme set can be good for these two Muong dialects.

One exciting remark here is that all of Muong Tan Son's rating scores are higher than those of Muong Bi. This can be explained by the fact that Muong Tan Son is closer to Vietnamese than Muong Bi (in the vocabulary, for example). The evaluation results show that the Vietnamese-Muong translation system can achieve high results in both translation quality and synthesized speech quality.

5. CONCLUSIONS

This paper presents our machine translation work for Unwritten language using intermediate representation. The text of a language (L1) can be translated into the speech of an unwritten language (L2) using the phoneme sequence of L2 as the intermediate representation instead of its text. An experiment on translating Vietnamese text into Muong speech in two dialects has been conducted. A phoneme set for each Muong language was proposed and applied to the problem. The subjective assessment results of people in the two regions show that the automatic translation system from Vietnamese text to Muong speech has good translation quality, and the speech output quality is highly appreciated.

The results of this paper are encouraging, especially for non-close-related language pairs, because using an SMT module can help in learning the translation even between far language pairs. Future work can be to apply the automatic translation method from Vietnamese text to other unwritten languages in Vietnamese. Furthermore, a fascinating further investigation will be planned for the extension method that can be applied for pairs of languages belonging to a different language family. We can further improve the results by testing more NMT models for the Text To Phone stage or using end-to-end translation from Vietnamese text to Speech Muong.

Acknowledgement: *This work was supported by the Vietnamese national science and technology project: "Re-search and development automatic translation system from Vietnamese text to Muong speech, apply to unwritten minority languages in Vietnam" (Project code: ĐTĐLCN.20/17).*

REFERENCES

[1]. J. Riesa, B. Mohit, K. Knight, and D. Marcu, *"Building an English-Iraqi Arabic machine translation system for spoken utterances with limited resources,"* in Ninth International Conference on Spoken Language Processing, (2006).

[2]. L. Besacier, B. Zhou, and Y. Gao, *"Towards speech translation of non written languages,"* in 2006 IEEE Spoken Language Technology Workshop, pp. 222–225, (2006).
[3]. G. Adda *et al.*, *"Breaking the unwritten language barrier: The BULB project,"* Procedia Comput. Sci., vol. 81, pp. 8–14, (2016).
[4]. Y.-F. Cheng, H.-S. Lee, and H.-M. Wang, *"AlloST: Low-resource Speech Translation without Source Transcription."* arXiv. (2021). http://arxiv.org/abs/2105.00171
[5]. P. K. Muthukumar and A. W. Black, *"Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis,"* in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2594–2598, (2014).
[6]. O. Scharenborg *et al.*, *"Speech Technology for Unwritten Languages,"* IEEEACM Trans. Audio Speech Lang. Process., vol. 28, pp. 964–975, (2020).
[7]. V. Đ. Phạm *et al.*, *"How to generate Muong speech directly from Vietnamese text: Cross-lingual speech synthesis for close language pair,"* J. Mil. Sci. Technol., no. 81, (2022).
[8]. N.-H. Doan, *"Generation of Vietnamese for French-Vietnamese and English-Vietnamese Machine Translation,"* in Proceedings of the 8th European Workshop on Natural Language Generation - Volume 8, Stroudsburg, PA, USA, pp. 1–10 (2001).
[9]. D. Thi Ngoc Diep, L. Besacier, and E. Castelli, *"Improved Vietnamese-French Parallel Corpus Mining Using English Language,"* in IWSLT, (2010).
[10]. D. Thi-Ngoc-Diep, M. Utiyama, and E. Sumita, *"Machine translation from Japanese and French to Vietnamese, the difference among language families,"* in 2015 International Conference on Asian Language Processing (IALP), pp. 17–20, (2015).
[11]. T. Duarte, R. Prikladnicki, F. Calefato, and F. Lanubile, *"Speech recognition for voice-based machine translation,"* IEEE Softw., vol. 31, no. 1, pp. 26–31, (2014).
[12]. P. Koehn *et al.*, *"Moses: Open source toolkit for statistical machine translation,"* in Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, pp. 177–180, (2007).
[13]. R. Zens, F. J. Och, and H. Ney, *"Phrase-based statistical machine translation,"* in Annual Conference on Artificial Intelligence, pp. 18–32, (2002).
[14]. K. Cho *et al.*, *"Learning phrase representations using RNN encoder-decoder for statistical machine translation,"* ArXiv Prepr. ArXiv14061078, (2014).
[15]. I. Sutskever, O. Vinyals, and Q. V. Le, *"Sequence to sequence learning with neural networks,"* Adv. Neural Inf. Process. Syst., vol. 27, (2014).
[16]. M.-T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba, *"Addressing the rare word problem in neural machine translation,"* ArXiv Prepr. ArXiv14108206, (2014).
[17]. R. Sennrich and B. Zhang, *"Revisiting low-resource neural machine translation: A case study,"* ArXiv Prepr. ArXiv190511901, (2019).
[18]. J. Shen *et al.*, *"Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,"* in 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4779–4783, (2018).
[19]. Y. Wang *et al.*, *"Tacotron: Towards end-to-end speech synthesis,"* ArXiv Prepr. ArXiv170310135, (2017).
[20]. L. Besacier, B. Zhou, and Y. Gao, *"Towards speech translation of non written languages,"* in 2006 IEEE Spoken Language Technology Workshop, pp. 222–225, (2006).
[21]. G. Adda *et al.*, *"Breaking the Unwritten Language Barrier: The BULB Project,"* Procedia Comput. Sci., vol. 81, pp. 8–14, (2016), doi: 10.1016/j.procs.2016.04.023.
[22]. J. Jiang, Z. Ahmed, J. Carson-Berndsen, P. Cahill, and A. Way, *"Phonetic representation-based speech translation,"* in Proceedings of Machine Translation Summit XIII: Papers, (2011).
[23]. Z. Ahmed, J. Jiang, J. Carson-Berndsen, P. Cahill, and A. Way, *"Hierarchical phrase-based mt for phonetic representation-based speech translation,"* in Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers, (2012).

[24].F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, *"Pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment,"* in International Conference on Statistical Language and Speech Processing, pp. 260–272, (2013).
[25].S. Palkar, A. W. Black, and A. Parlikar, *"Text-To-Speech for Languages without an Orthography,"* in Coling, (2012).
[26].S. Sitaram, S. Palkar, Y.-N. Chen, A. Parlikar, and A. W. Black, *"Bootstrapping text-to-speech for speech processing in languages without an orthography,"* in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7992–7996.
[27].S. Sitaram, G. K. Anumanchipalli, J. Chiu, A. Parlikar, and A. W. Black, *"Text to speech in new languages without a standardized orthography,"* in Proceedings of 8th Speech Synthesis Workshop, Barcelona, (2013).
[28].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), (in Vietnamese).
[29].Nguyễn Văn Tài, *"Ngữ âm tiếng Mường qua các phương ngôn"*. Hà Nội: Nxb Từ điển Bách khoa, (2005), (in Vietnamese).
[30].Trần Trí Dõi, *"Một vài vấn đề nghiên cứu so sánh - lịch sử nhóm ngôn ngữ Việt - Mường"*. Hà Nội: Nxb Đại học Quốc gia Hà Nội, (2011) , (in Vietnamese).
[31].Nguyễn Kim Thản, *"Vài nét về hệ thống âm vị tiếng Mường và phương án phiên âm tiếng Mường,"* Ngôn Ngữ, vol. 1, (1971), (in Vietnamese).
[32].M. E. Barker, M. A. Barker, and L. Assessment, *"Mường-Vietnamese-English dictionary",* https://www.sil.org/resources/archives/35773
[33].Nguyễn Như Ý, *"Dự thảo phương án chữ Mường."* Tọa đàm Viện Ngôn ngữ học, (1994), (in Vietnamese).
[34].LDC, *"Linguistic data annotation specification: Assessment of fluency and adequacy in translations*. *Revision 1.5,"* (2005).

TÓM TẮT

Dịch tiếng nói cho ngôn ngữ chưa có chữ viết sử dụng biểu diễn trung gian: Thử nghiệm cho cặp ngôn ngữ Việt-Mường

Bài báo nghiên cứu một phương pháp dịch tự động từ văn bản của một ngôn ngữ (L1) sang tiếng nói của một ngôn ngữ chưa có chữ viết (L2). Thông thường, văn bản đã viết được sử dụng làm cầu nối để kết nối một mô-đun dịch chuyển từ văn bản của L1 sang văn bản của L2 và một mô-đun tổng hợp tạo ra tiếng nói của L2 từ văn bản. Trong trường hợp ngôn ngữ không có chữ viết, một biểu diễn trung gian phải được sử dụng thay cho chữ viết của L2. Bài báo này đề xuất việc sử dụng biểu diễn âm vị vì mối quan hệ mật thiết giữa âm vị và lời nói trong một ngôn ngữ. Phương pháp đề xuất được áp dụng cho cặp ngôn ngữ Việt - Mường. Văn bản tiếng Việt cần được dịch sang tiếng Mường ở hai phương ngữ là Mường Bi - Hòa Bình và Mường Tân Sơn - Phú Thọ, đều chưa có chữ viết. Bài báo cũng đề xuất bộ âm vị cho mỗi phương ngữ tiếng Mường nêu trên và áp dụng chúng vào bài toán thử nghiệm. Kết quả đánh giá cho thấy chất lượng dịch khá cao ở cả hai phương ngữ (đối với Mường Bi, điểm lưu loát là 4.63/5.0 và điểm đầy đủ là 4.56/5.0) và chất lượng tiếng nói tổng hợp ở cả hai phương ngữ cũng khá tốt (đối với Mường Bi, điểm MOS là 4.47/5.0 và điểm hiểu rõ là 93.55%). Kết quả cũng cho thấy khả năng ứng dụng của hệ thống đề xuất đối với các ngôn ngữ chưa có chữ viết khác là đầy hứa hẹn.

Từ khoá: Dịch tự động; Tổng hợp tiếng nói; Ngôn ngữ thiểu số; Tiếng Việt; Các phương ngữ tiếng Mường; Ngôn ngữ chưa có chữ viết; Tổng hợp tiếng nói đa ngôn ngữ.