How to generate Muong speech directly from Vietnamese text: Cross-lingual speech synthesis for close language pair

Pham Van Dong[1, 2,*], Do Thi Ngoc Diep[2], Mac Dang Khoa[3], Nguyen Viet Son[2], Nguyen Van Thinh[3], Nguyen Tien Thanh[4], Tran Do Dat[5]

[1]Hanoi University of Mining and Geology
[2]Hanoi University of Science and Technology
[3]VinBigdata – VinGroup
[4]Viettel CyberSpace Center
[5]Ministry of Science and Technology of Vietnam
*Corresponding author: phamvandong@humg.edu.vn
Received; Revised; Accepted.
DOI:

ABSTRACT

The paper introduces a method for automatic translation of Vietnamese text into Muong speech in two dialects, Muong Bi - Hoa Binh and Muong Tan Son - Phu Tho, which are all unwritten dialects of the Muong language. Due to the very close relationship between the Vietnamese and Muong languages, the translation system was built to look like a cross-lingual speech synthesis system, in which the input is the text of one language (i.e., the Vietnamese) and the output is the speech of another language (i.e., the two Muong dialects). The system used the modern sequence-to-sequence TTS neural models Tacotron2 and WaveGlow. The evaluation results showed a high quality of translation in both dialects (for Muong Bi, the fluency score was 4.61/5.0 and the adequacy score was 4.79/5.0) and also synthesized speech quality in both dialects (for Muong Bi, the MOS score was 4.68/5.0 and the intelligibility score was 94.60%). The results show that the applicability of the proposed system to other minority languages is promising, especially in the case of unwritten languages.

Keywords: Machine Translation, Text to Speech, Ethnic Minority Language, Vietnamese, Muong dialects, unwritten languages, Cross-lingual speech synthesis

1. INTRODUCTION

Natural language processing (NLP) technology plays a broad role in human life, such as in human-machine communication, virtual switchboard, smart home, automatic translation applications, etc. This paper relates to two leading technologies in NLP: Speech synthesis and Machine translation. Speech synthesis, or text-to-speech (TTS), synthesizes intelligible and natural speech from input text [1]. According to [2], developing a TTS system for a language requires not only the deployment of speech processing techniques but also linguistic research such as phonetics, phonology, syntax, and grammar of this language [3]. Currently, TTS is mainly based on a deep neural network, the so-called Neural Speech Synthesis [4], [5]. Machine Translation (MT) is the task of translating a text from a source language into a target language. Currently, Neural Machine Translation (NMT) can perform a translation with an end-to-end neural network [6], [7]. The common feature of these methods is the need for a large amount of data, up to tens of voice recording hours in the TTS task and millions of bilingual sentence pairs for the NMT task. Therefore, there are many difficulties applying current processing technologies to minority or low-resourced languages due to insufficient data and a lack of related research.

Currently, the world has more than 7000 languages, and most of them are minority languages; nearly half of them are unwritten languages[1]. Current Vietnamese TTS and MT technologies

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

mainly support the Vietnamese language, the official language of Vietnam [8], [9]. A minority ethnic community speaks the Muong language, with more than 1.5 million people (top five population in Vietnam) living in Hoa Binh, Phu Tho, and some mountainous areas in Vietnam. The Muong language belongs to the same language family as Vietnamese [10]. The Muong language has little data that has been digitized, and it has no official writing script. Besides, the need for information exchange between Vietnamese and Muong is essential. Therefore, the object of the paper is the Muong language, and the specific goal is to analyze and develop a system that allows translating Vietnamese text to Muong speech (the only means of language representation for the Muong language).

In this work, the Muong language is considered unwritten. Due to the very close relationship between the Vietnamese and Muong languages, the Muong dialects are "considered" Vietnamese dialects. A "cross-lingual" text-to-speech system was proposed, allowing direct conversion from Vietnamese text to Muong speech. It looks like the speech synthesis system of one language, in which the input is the text of one language (i.e., Vietnamese), and the output is the speech of its dialects (i.e., the two different dialects, Muong Tan Son and Muong Bi Hoa Binh). The proposed system was built based on sequence neural network models successfully applied in TTS and MT. To the best of our knowledge, this work is the first attempt to apply this model to translate from text of one language to the speech of another language. Using the "cross-lingual" proposal, the system can take advantage of available resources in Vietnamese, save a lot of time and cost, and still retrieve a good quality TTS system.

This paper is organized as follows. After presenting the background and related works in section 2, section 3 will show the proposed cross-lingual text-to-speech system. Here the technology of using end-to-end direct-learning neural networks is applied. The section also presents the system implementation, including database building and method implementation. Section 4 presents the evaluation process. Furthermore, the final section presents the conclusions and future directions of the study.

2. BACKGROUNDS AND RELATED WORKS

To build a Vietnamese cross-lingual speech synthesis system, a number of related works have been studied, including research on (2.1) Sequence to Sequence Model application for MT and TTS, (2.2) speech translation for unwritten minority languages, (2.3) linguistic characteristics of Muong language and the relationship with Vietnamese. This section will also detail the direction of cross-lingual speech synthesis technology, with the idea of applying speech synthesis technology to the problem of Vietnamese-Muong text translation (which will be detailed in Section 3).

2.1. Sequence to Sequence modeling in MT and TTS

Sequence to sequence modeling includes all models that map one sequence to another [6]. This model was first introduced and applied to MT in the research of Ilya Sutskever [11]. MT can be described as the task of converting a sequence of words (sentences) in the source language to a vector representation and then applying an invert conversion of this representation into a sequence of words in the target language [6].

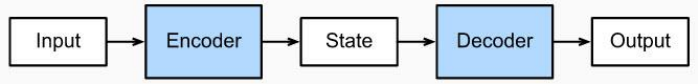

Figure 1. The encoder-decoder architecture

The encoder-decoder architecture has two major components, as shown in *Figure 1*. The first component is an encoder: it takes a variable-length sequence as input and transforms it into a state

with a fixed shape. The second component is a decoder: it maps fixed forms of encoded state to a variable-length sequence.

The attention model is considered very important in a sequence-to-sequence model. Bahdanau [12] added an attention mechanism to the models. It helps to enhance the effect of some essential parts of the input data, even when this part is far from the current fed word while diminishing the effect of the other trivial parts. It sets flexible "soft weights" for parts of the input instead of fixing a typical weight for all input parts at a time step.

Based on the success of the above seq2seq model, many models are proposed and applied for speech processing, especially in the text to speech, as shown in Figure 2.

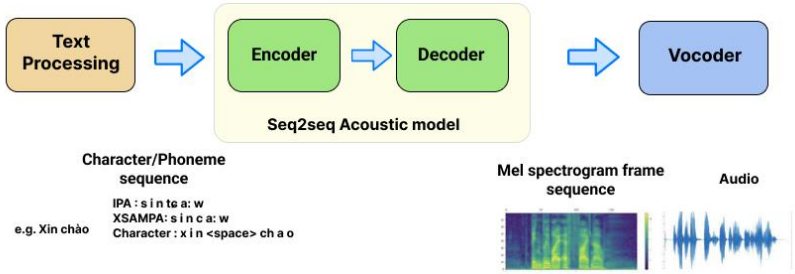

Figure 2. The adaptation of the encoder-decoder architecture to TTS

The text processing will process input text, including tasks like normalizing text, POS Tagger (it does the Part Of Speech tagging of the input text), tokenizing a sentence into words, pronunciation (it breaks the input text into phonemes based on pronunciation). The encoder acts to encode the input text sequence into a compact hidden representation, and the decoder consumes it at every decoding step. The decoder is an autoregressive recurrent neural network that predicts the Mel spectrogram from the encoded input sentence one frame at a time. The neural vocoder is cascaded to convert the sequence of the acoustic features to the audio waveform [13], [14]. The Char2Wav and Tacotron models are two famous neural TTS systems. The Char2Wav speech synthesis system was developed by the group [15]; the goal is to build a system that is trained with a string as an input. Char2Wav consists of two components: an encoding-decoding model with an attention mechanism and a neuronal vocoder. Tacotron is an end-to-end speech synthesis system from input text [16]. By 2017, Tacotron had been developed to version 2 [17], overcoming the disadvantage of converting acoustic feature spectrogram to sound waveform. The seq2seq mechanism in the Tacotron model will be described in detail in the following section.

2.2. *Speech translation for under-resourced and unwritten languages*

Under-resourced/unwritten language processing is a new research field that has been paid attention to in the past few years and has received modest results [18]. They have essential meanings because, besides bringing speech communication technologies to minority ethnic communities, products applying this technology also contribute to the conservation of endangered languages.

The first and most expensive approach is constructing a script for an unwritten language. In 2006, the TRANSTAC project of the US Department of Defense created a system that allows translating short conversations in combat between English-speaking soldiers and Iraqiui Arabic-speaking soldiers [19]. This project requires the definition of a new writing system for Iraqiui Arabic based on the Standard Arabic script. However, this method requires a high cost in terms of time, expert human resources, and budget investment and cannot be reused for other languages. The following approach is to construct an intermediate representation instead of the text. The recognition, speech synthesis, and text machine translation modules in the system will operate on these intermediate representations instead of text (Figure 3). Most of these representations are

proposed based on phonemic transcription [20], adapted phoneme transcription from one or more close languages [21], or based on a set of symbols that recognize the phonologically acoustic characteristics of speech signals [22].

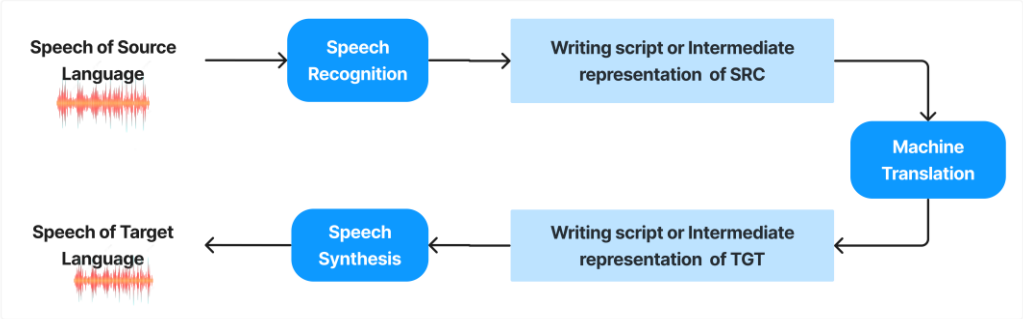

Figure 3. Model of the speech-to-speech machine translation system using intermediate representation for unwritten language

2.3. Relationship between Vietnamese and Muong language

The Muong language belongs to the Viet-Muong language group, Mon Khmer branch, and Austroasiatic family [10]. The Muong and Vietnamese languages have the exact origin and have a very close relationship. Thus, the Muong language has many similarities with the Vietnamese language. For example, the grammatical relationship of words in a Muong sentence is also indicated by word order or by word order plus vanity [23].

Our local field research shows that the Muong language's syllable structure is the same as that of the Vietnamese language. The structure of the Vietnamese and Muong syllables has five components: onset, glide, nucleus, coda, and tone. The nucleus and tone play an essential role that cannot be absent in the syllable. In the phonemic system, Vietnamese, Muong Bi, and Muong Tan Son have many similar and different phonemes. For the first consonant, there are 18 first consonants in two Muong dialects that are similar to Vietnamese first consonants, including /b, m, t, d, th, n, s, z, l, c, ɲ, k, ŋ, ʔ, h, f, χ, ɣ/. The two consonants /z, ʂ/ are present in Vietnamese but not in Muong. There are four consonants in Muong but not in Vietnamese /p, w, tl (kl), r/. /v, ʈ/ are two consonants like in Vietnamese but exist only in the Muong Tan Son dialect but not in Muong Bi. The Muong accompaniment has the same function and position as the Vietnamese one. Vietnamese has 16 vowels, while Muong has only 14 vowels. The Muong language does not have two short vowels /ɛ̆/ and /ɔ̆/, like in Vietnamese. Vietnamese has eight final sounds for the last sound, including six consonants /p, t, k, m, n, ng, nh/ and two semi-vowels /u, i/. The Muong language has 11 final sounds, with two pairs of final sounds (/k/ and /c/). For tones, Vietnamese has six tones, and Muong has 5. There is no high-rising broken tone like in Vietnamese.

Based on the similarities between the two languages, Muong can be considered a Vietnamese dialect in our processing system. The reason is that in one language, there are many dialects. These dialects use the same script and differ only in the vocabulary and pronunciation of some words. Thus, the problem of synthesizing Vietnamese speech into different dialects can be considered as using the same Vietnamese text input, and the output is the different voices of different dialects (different pronunciations of words). In the case of the unwritten Muong language, the hypothesis that the Muong language is a "dialect" of Vietnamese can help to convert the problem of translating Vietnamese text to Muong speech into a speech synthesis problem from Vietnamese text to speech of another dialect of Vietnamese, i.e., Muong speech. Part 3 will detail the proposed method for this paper.

3. PROPOSED METHOD AND EXPERIMENT

Starting from the previous hypothesis, the translation system from Vietnamese text to the Muong speech turns into a speech synthesis system from the Vietnamese text into the Muong speech. The SOTA end-to-end speech synthesis technology was applied to the pair of languages. The proposal to use direct conversion also comes from the initial success of the inverse problem, which maps directly from speech to text presented in [24], [25]. An attempt to directly translate a source speech signal into target-language text was proposed by [24]. The idea of building an end-to-end speech-to-text translation system without using source language text in the learning or decoding process, thereby reducing the need for source language transcription, is the main idea in [25].

In the proposed system, text in the source language (Vietnamese with script) will be mapped to the corresponding output speech of the target language (Muong speech without the script).

3.1. Model architechture

The system model will follow the model architectural as in the Figure 4.

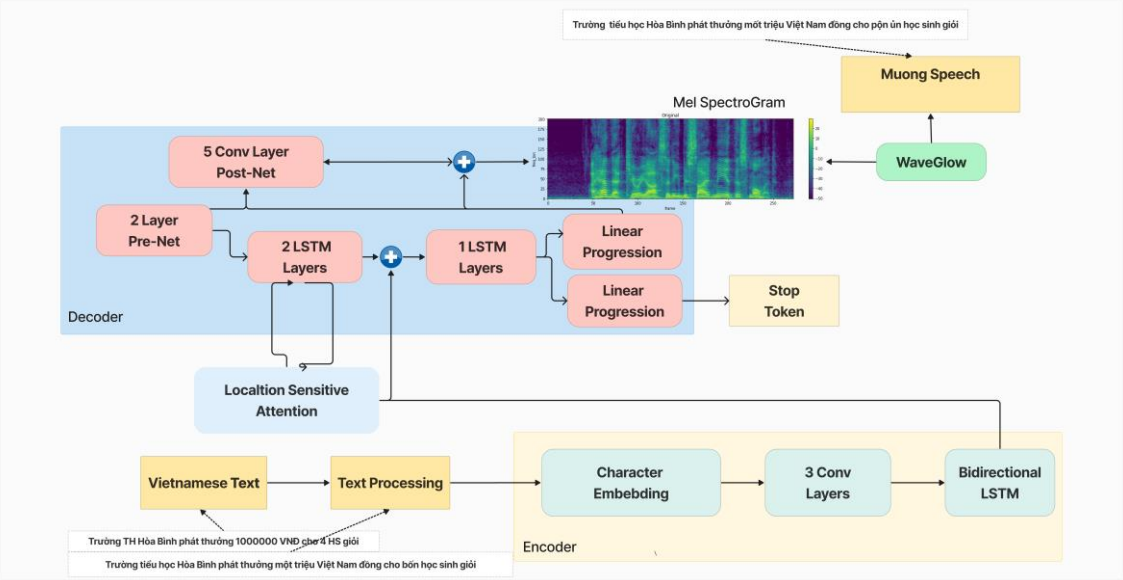

Figure 4. System Architecture

The Text Processing module is to convert non-orthographic items into speakable words. This is known as normalizing the text from various symbols, numbers, dates, abbreviations, and other non-orthographic entities of the text into a standard orthographic phonetic transcription suitable for later phonetic conversions. It also needs to parse spaces, punctuation, and other delimiters to determine the document structure.

Our model keeps using Tacotron 2's original model. The network consists of an encoder and a decoder with attention. The encoder converts a character sequence into a hidden feature representation, and the decoder is to predict a spectrogram. Input characters are 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.

In the vocoder, WaveGlow [26], an alternative to Tacotron's original WaveNet model, was used. We use 12 coupling layers and 12 invertible 1x1 convolutions. The coupling layer networks (W N) each have eight layers of dilated convolutions with 512 channels used as residual connections and 256 channels in the skip connections. The WaveGlow network was trained on a 4 GPU NVIDIA GTX 2080Ti using randomly chosen clips of 16,000 samples.

3.2. Database building

The proposed models were trained on pairs of bilingual Vietnamese text and Muong audio speech in two dialects, respectively.

3.2.1. Text data buiding

The Vietnamese text database was built before recording the Muong speech database. The selected domain is the news field. To ensure the quality of the translation, the Vietnamese text has to balance phonemic and lexical distribution as in reality. Nearly 4 million Vietnamese text sentences were collected from the general Vietnamese news field (Vietnamnet, Dantri, etc.) and around 900,000 Vietnamese text sentences from Muong local news publishers (Hoa Binh newspaper[2], Phu Tho newspaper[3]). The local news data helps to collect words commonly used in the Muong regions. The raw data is then preprocessed (separate sentences, remove short sentences (under five tokens) and long sentences (over 120 tokens), and remove non-Vietnamese sentences (which contain more than 50% of foreign language words). The original text collection contained around 4.9 million sentences and was considered the word-reality distribution in the domain. A random extraction algorithm that ensures the balance of phonemic and syllable distribution as in the original text collection was applied to extract a set of 20,000 sentences from the original text collection. Twenty thousand sentences are divided into two sets: one consists of 5,000 sentences extracted from 900,000 sentences in Muong newspapers, and the other includes 15,000 sentences extracted from 4 million sentences in general newspapers.

3.2.2. Recording

The speech corpus 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 a good, clear, and coherent voice. 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).

3.2.3. Data processing

Along with the speech recording, the speech data was processed to normalize energy, remove noise, long pauses, pronunciation errors, or unexpected errors encountered during recording. The faulty speech was required to be recorded again. Each speaker was required to record all sentences of good quality. Audio files are recorded at a speech sampling rate of 44.1 kHz. The audio files are then converted to the 22.05 kHz speech sampling rate to match the system's training input. Each collection for the male voice corresponds to more than 1800 minutes of the audio signal after post-processing. The Vietnamese text data was preprocessed (normalize non-standard words: punctuations, numbers, acronyms, words, upper/lowercase characters, etc.) to get the appropriate representation of a sentence as a string of Vietnamese words.

[2] http://www.baohoabinh.com.vn/en/
[3] https://baophutho.vn/

3.3 Model trainning

We have trained two network models of Tacotron2 and WaveGlow to build a Vietnamese text to Muong speech translation system. 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 Vietnamese text-Muong speech pairs of a sentence. Three hundred files were used for validation, 300 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 was converged after 100k steps, while the vocoder was converged after 100k steps.

The model, after training, was deployed in a web-based service system allowing users to access the service of translation from Vietnamese text to Muong speech in two dialects[4]. Users can choose the dialect and input type as a text or an uploaded document. The translation result was displayed in the format of the Muong audio file for each sentence. Users can hear or download the result files.

Figure 5. Web service system of the Vietnamese text – Muong speech translation system

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 BLUE, 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 [27]. 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).

[4] http://mica.edu.vn/MuongTTS

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 28 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 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 1.

Table 1. Vietnamese text to Muong speech translation system evaluation result

Evaluation Criteria		**Muong Bi**	**Muong Tan Son**
Translation quality	*Fluency (0-5)*	4,61	4,99
	Adequacy (0-5)	4,79	4,99
Output speech quality	*MOS (0-5)*	4,68	4,98
	Intelligibility (%)	94,60%	99,70%

The fluency scores of 4.61 for Muong Bi and 4.99 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 4.79 for Muong Bi and 4.99 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 4.68 and 4.98, respectively. The high scores indicate that the output speech was almost as natural as human speech. The intelligibility scores of 94.6% for Muong Bi and 99.7% 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.

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). However, the MOS rating of 4.98 for Muong Tan Son is too high. This might be because listeners in Tan Son were too eager to work with an exciting system for the Muong people. The evaluation results show that the Vietnamese-Muong translation system can achieve high results in both translation quality and synthesized speech quality.

4. CONCLUSION

This paper presents the first work of machine translation from Vietnamese text into Muong speech, one of Vietnam's unwritten minority ethnic languages. The proposed system was built using end-to-end text-to-speech neural network technology. In the training process, instead of

entering text and voice in the same language, the input data consists of Vietnamese text and the voice of the Muong minority ethnic language. Bilingual corpora of 20 thousand pairs of Vietnamese text and Muong speech in each dialect, Muong Bi-Hoa Binh and Muong Tan Son-Phu Tho were constructed. The results of the subjective assessment 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 auspicious, especially for close-related language pairs. Therefore, future work will continue to apply the automatic translation method from the text of one language to the speech of another close-related language. For example, several close-related languages can be selected for testing, such as Tay-Nung, Vietnam-Tho, Mnong-Stieng, etc. 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.

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

[2] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A Survey on Neural Speech Synthesis," *ArXiv210615561 Cs Eess*, Jul. 2021, Accessed: Apr. 27, 2022. [Online]. Available: http://arxiv.org/abs/2106.15561

[3] F. de Saussure, *Course in General Linguistics*. Columbia University Press, 2011.

[4] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A review of deep learning based speech synthesis," *Appl. Sci.*, vol. 9, no. 19, p. 4050, 2019.

[5] Z. Mu, X. Yang, and Y. Dong, "Review of end-to-end speech synthesis technology based on deep learning," *ArXiv Prepr. ArXiv210409995*, 2021.

[6] G. Neubig, "Neural machine translation and sequence-to-sequence models: A tutorial," *ArXiv Prepr. ArXiv170301619*, 2017.

[7] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *ArXiv Prepr. ArXiv14061078*, 2014.

[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] T. Do Dat, E. Castelli, L. X. Hung, J.-F. Serignat, and T. Van Loan, "Linear F0 contour model for Vietnamese tones and Vietnamese syllable synthesis with TD-PSOLA," 2006. Accessed: Dec. 15, 2016. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.492.5987&rep=rep1&type=pdf

[10] M. Ferlus, "Langues et peuples viet-muong," *Monkhmer Stud.*, pp. 7–28, 1996.

[11] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[12] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *ArXiv Prepr. ArXiv14091259*, 2014.

[13] W.-C. Huang, T. Hayashi, S. Watanabe, and T. Toda, "The sequence-to-sequence baseline for the voice conversion challenge 2020: Cascading asr and tts," *ArXiv Prepr. ArXiv201002434*, 2020.

[14] O. Watts, G. E. Henter, J. Fong, and C. Valentini-Botinhao, "Where do the improvements come from in sequence-to-sequence neural TTS?," in *2019 ISCA Speech Synthesis Workshop (SSW)*, 2019, vol. 10, pp. 217–222.

[15] J. Sotelo *et al.*, "Char2wav: End-to-end speech synthesis," 2017.

[16] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," *ArXiv Prepr. ArXiv170310135*, 2017.

[17] 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)*, 2018, pp. 4779–4783.

[18] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Special Issue on Processing Under-Resourced Languages-Speech Communication Journal." Elsevier, 2014.
[19] J. Riesa, B. Mohit, K. Knight, and D. Marcu, "Building an English-Iraqi Arabic machine translation system for spoken utterances with limited resources," 2006.
[20] J. Jiang, Z. Ahmed, J. Carson-Berndsen, P. Cahill, and A. Way, "Phonetic representation-based speech translation," *13th Mach. Transl. Summit*, 2011.
[21] T. Kempton, R. K. Moore, and T. Hain, "Cross-Language Phone Recognition when the Target Language Phoneme Inventory is not Known.," in *INTERSPEECH*, 2011, pp. 3165–3168. Accessed: Aug. 18, 2016. [Online]. Available: https://pdfs.semanticscholar.org/fe54/2f3b674368ab32b5d137cb68c76b07bc1b01.pdf
[22] 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)*, May 2014, pp. 2594–2598. doi: 10.1109/ICASSP.2014.6854069.
[23] Nguyễn Văn Tài, *Ngữ âm tiếng Mường qua các phương ngôn*. Hà Nội: NXB Từ điển Bách khoa, 2005.
[24] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 949–959.
[25] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *ArXiv Prepr. ArXiv161201744*, 2016.
[26] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3617–3621.
[27] LDC, "Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5," 2005.

TÓM TẮT

Cách tạo tiếng nói tiếng Mường trực tiếp từ văn bản tiếng Việt: Tổng hợp tiếng nói đa ngôn ngữ cho cặp ngôn ngữ họ gần

Bài báo giới thiệu phương pháp dịch tự động văn bản tiếng Việt sang tiếng nói tiếng Mường ở hai phương ngữ Mường Bi - Hòa Bình và Mường Tân Sơn - Phú Thọ, đều là hai phương ngữ chưa có chữ viết chính thức của tiếng Mường. Do mối quan hệ rất chặt chẽ giữa tiếng Việt và tiếng Mường, hệ thống phiên dịch được xây dựng giống như một hệ thống tổng hợp tiếng nói đa ngôn ngữ, trong đó đầu vào là văn bản của một ngôn ngữ (ví dụ tiếng Việt) và đầu ra là tiếng nói của một ngôn ngữ khác (ví như tiếng nói của hai phương ngữ Mường). Hệ thống sử dụng mô hình mạng nơ-ron sequence-to-sequence TTS hiện đại đó là Tacotron2 và WaveGlow. Đánh giá đạt được cho Mường Bi là : Tính trôi chảy - 4,61/5, Tính đầy đủ - 4,79/5, Tính tự nhiên trên thang điểm MOS - 4,68/5, Độ dễ hiểu - 94,60%. Các kết quả nhận được cho thấy khả năng áp dụng của hệ thống đề xuất cho các ngôn ngữ thiểu số khác là đầy hứa hẹn, đặc biệt là trong trường hợp ngôn ngữ chưa có chữ viết.

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.