Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XI về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 09-10/8/2018
DOI: 10.15625/vap.2018.00013

CROSS-LINGUAL PHONEME RECOGNITION FOR FAMILIAR LANGUAGES: APPLYING TO VIETNAMESE AND MUONG LANGUAGES

Tran Thi Thu Thuy[1,2], Do Thi Ngoc Diep[1,*], Mac Dang Khoa[1], Pham Van Dong[1,2]

[1]International Institute of MICA, Hanoi University of Science and Technology
[2]Faculty of Information Technology, Hanoi University of Mining and Geology

{thu-thuy.tran; ngoc-diep.do; dang-khoa.mac; van-dong.pham}@mica.edu.vn

ABSTRACT: *The automatic phoneme recognition is essential to build speech processing systems for a new language. It requires a lot of knowledge on spoken language processing and linguistic knowledge of this language. For under-resourced languages (e.g. minority languages), a common automatic phoneme recognition system is not available. This paper presents an experiment of phoneme recognition on Muong language, a minority language without writing system in Vietnam. Following the cross-lingual approach, the Muong speech input is transcribed into a sequence of Vietnamese phones, using an acoustic model trained on Vietnamese speech. The result shows that the average of correct recognition rate of Muong phones is more than 50 %. A comparison of correct recognition rates between Muong phone set and Vietnamese phone set can help linguists to confirm the similarities and the distinctions between these two familiar languages.*

Keyword: *phoneme recognition, cross-lingual, under-resourced language, Muong language, Vietnamese language.*

I. INTRODUCTION

The technology of natural language processing including speech processing today has achieved many accomplishments with more extensive applications in the field of human-machine interaction. Developing a speech processing system for a language is not only the deployment of speech processing technique, but also requires specific speech data resources and linguistic knowledge such as phonology, prosody, syntax, grammar, semantics. Therefore, among more than 7000 languages in the world, the speech processing core technologies (i.e. speech recognition, speech synthesis, speech understanding, automatic translation) are available for limited number of well-resourced languages such as English, Chinese, French, Arabic etc. For thousands of other languages, called under-resourced languages, none of such technology is available [1].

Most of these under-resourced languages are the minority languages which are being disappeared due to the low number of native speakers or by being "digested" by the surrounding majority languages. Moreover, half of these languages are unwritten languages (3,188 on the total of 7,097 languages in the world[1]). So the preservation of the language as well as communication enhancement with minority communities are also the issues to be taken in the world and in Vietnam. They typically include some tasks of documentary linguistics such as audio and video recording of native speakers' speech; transcription/annotation/translation of recording speech; preservation and distribution of the resulting materials. Among them, the speech transcription is the very first task to represent the content of speech signal into corresponding text. Especially in the case of unwritten languages, the language speech has to be transcribed in sequence of phonemes, normally using IPA[2] (International Phonetic Alphabet) phone set of this language. This task typically has to be done manually by linguistic experts and it is very time consuming. For example, it can take several hours to transcript few minutes of speech signal [2], [3].

One solution for this time consuming problem of speech transcription is using automatic phoneme recognition technique. This technique is based on the automatic speech recognition (ASR) technology. The input is an utterance but the output is a sequence of phonemes instead of a sequence of words as in ASR. The ASR system uses machine learning method and typically requires hundreds of hours of transcribed speech data to train the models [4]. For under-resourced languages, due to the lack of necessary speech resources, it is impossible to build such an ASR system. A potential solution for this case is using cross-lingual (or cross-language) approach [5], which uses a pre-trained ASR system of a well-resourced language to recognize the speech of another under-resourced language. Based on the similarity between languages, the cross-lingual approach can give promising results with the familiar languages.

This paper presents the first experiment of phoneme recognition on Muong language, a minority language without writing system in Vietnam, toward to build a more complete speech processing system for Muong language. On one hand, this work is in order to survey the ability of using cross-lingual phoneme recognition in Vietnamese - Muong language pairs. On the other hand, the analysis on cross-lingual phoneme sets can help linguists to confirm the

[1] Ethnologue (www.ethnologue.com), in June 2018
[2] www.internationalphoneticassociation.org
[*] Corresponding author

similarities and the distinctions between these two familiar languages. The paper is organized as follows. Section II gives some comparison between Vietnamese and Muong phonologies. After presenting the general idea of cross-lingual approach in section III, section IV will describe in detail the experiment of Muong phoneme recognition. The results of this experiment are shown and analyzed in section V. This paper ends with some conclusions and the future works.

II. VIETNAMESE AND MUONG LANGUAGE

Vietnam is a multi-ethnic country with 54 ethnic groups. The *Kinh* who speaks Vietnamese is the majority ethnic group accounting for 85.6 % of the total population. The remaining 53 groups are ethnic minorities[3]. Among them, Muong is one of the five largest population ethnic groups in Vietnamese with more than 1 million speakers. Muong have many dialects locating in different provinces in the north of Vietnam such as *Hoa Binh, Phu Tho, Thanh Hoa, Son La* [6] . In terms of language family, Vietnamese and Muong languages belong to the same *Viet-Muong* group*, Mon-Khmer* branch of the Austroasiatic family**.** Therefore, Muong has many similarities with Vietnamese, in term of phonology, tone, syntax and vocabulary [7]. Table 1 shows a comparison of phoneme systems between Vietnamese (Hanoi standard dialect) and Muong Bi dialect (in *Hoa Binh* province, main dialect of Muong). According to this comparison, the phone set of Muong can be divided into the 3 following groups [8]:

- Equivalent phonemes: phonemes coincide with Vietnamese phonemes;
- Closed phonemes: phonemes are acoustically similar to phonemes in Vietnamese;
- Distinct phonemes: phonemes are not found in Vietnamese.

Table 1. Muong and Vietnamese phonetic comparison (in IPA), according to [8]

Group	Equivalent				Closed		Distinct	
	Muong	**Viet**	**Muong**	**Viet**	**Muong**	**Viet**	**Muong**	**Viet**
Initial consonants	/k/	/k/	/t/	/t/	/b/	/ɓ/	/hr/	-
	/h/	/h/	/t^h/	/t^h/	/c/	/tɕ/	/kl/	-
	/l/	/l/	/v/	/v/	/d/	/ɗ/	/p/	-
	/m/	/m/	/w/	/w/	/g/	/ɣ/	/r/	-
	/n/	/n/	/s/	/s/	/k^h/	/x/	/tl/	-
	/ŋ/	/ŋ/	/z/	/z/	/p^h/	/f/		
	/ɲ/	/ɲ/						
Final consonants	/p/	/p/	/ɲ/	/ɲ/			/c/	-
	/t/	/t/	/ŋ/	/ŋ/			/l/	-
	/k/	/k/	/w/	/w/				
	/m/	/m/	/j/	/j/				
	/n/	/n/						
Vowel	/a/	/a/	/o/	/o/				
	/ă/	/ă/	/ɤ/	/ɤ/				
	/ɤ̆/	/ɤ̆/	/u/	/u/				
	/ɛ/	/ɛ/	/ɯ/	/ɯ/				
	/e/	/e/	/iə/	/iə/				
	/i/	/i/	/uə/	/uə/				
	/ɔ/	/ɔ/	/ɯɤ/	/ɯə/				
Glide	/w/	/w/						

This comparison is mostly based on some linguistic researches of Muong language [6], [9]-[12]. In this work, we would like to examine these similarities, as well as the distinctions between Vietnamese and Muong, but in aspect of speech processing. That will be done by using an automatic phoneme recognition of Vietnamese to recognize phoneme sequences in Muong speech following the "cross-lingual" approach, which are presented in the next section.

III. PHONEME RECOGNITION FOR UNDER-RESOURCED LANGUAGE

As mention in section I, the automatic phoneme recognition system is a computer software which can convert the speech signal into a sequence of phonemes. With the available of open automatic speech recognition (ASR) toolkits, the automatic phoneme recognition can be implemented on an ASR system using the phoneme set of one particular language. The ASR system is typical trained on hundreds of hours of transcribed speech data (for acoustic model) and thousands of text sentences (for language model). For under-resourced language, especially unwritten

[3] General statistics office of Vietnam, 2009

language, due to the lack of speech and language resources, building a such of ASR system is impossible. The cross-lingual phoneme recognition technique was proposed to deal with this problem where there is no training data on the target language. The acoustic model pre-trained from a big training corpus of a familiar language was used as initial acoustic model of the target language. And further process of adaptation or improvement of the initial acoustic model will be made after. For the first time, the similarity between two languages' phone lists is determined by a phone mapping technique. Then the phoneme recognition is applied [5].

3.1. Phone mapping

The research in cross-lingual acoustic modeling is based on the assumption that the articulatory representations of phonemes are so similar across languages. So phonemes can be considered as units which are independent from the underlying language [13], [14]. In fact, the concept of "language independent phonemes" occurring in more than one languages (called poly-phonemes to differentiate with mono-phonemes) was firstly introduced by the International Phonetic Association [15] and then in [16].

Firstly, a source-target phoneme mapping table is obtained manually by knowledge-based methods [13], [17], or automatically by data-driven methods [5]. The automatic methods are based on a distance between two phoneme models (compute the distances between Gaussian distributions obtained for each phoneme model). These methods use a variety of distance measure including: entropy-based or log-likelihood based distance [18], Kullback-Leibler distance, Bhattacharyya distance, Euclidean distance [19].

In other approaches, the automatic phone mapping table is generated using confusion matrix [20], [21]. By using small amounts of acoustic data in the target language, the phone mapping table can be automatically created with data-driven methods. A phoneme recognizer in the source language is applied on the development data set of target language which is already transcribed in target language phonemes. Then, the output source phoneme hypotheses are aligned with their target phoneme references frame by frame to count the co-occurrences between a phoneme in source language and target language. By computing the number of times a reference phoneme in the target language that has been confused with a phoneme in source language, the confusion matrix is created. To obtain the final confusion matrix, each entry is normalized by dividing it through the total of occurrences of all corresponding phonemes in source language. Finally, by selecting each phoneme in target language with the correspondence phoneme in source language which has the highest confusion value, the phone mappings are made.

3.2. Cross-lingual phoneme recognition

There were several researches attempt to build cross-lingual acoustic model for under-resourced target language. In [18], the author firstly introduced a statistical distance measure to determine the similarities of sounds of several languages. One of his experiment was using English phoneme models in a German recognizer, instead of the German phoneme models. The cross-lingual model makes correct recognition rate improved for some phonemes but not for the others. However, the cross-lingual model can help in phoneme inventory for a bigger speech recognition system.

The idea was applied again in the work of [22] that used cross language transfer from five languages in the task of German speech recognition. The Turkish language was found fitting better with German phonology among other languages: Croatian, Japanese, Korean, Spanish. The Turkish model gave the word error rate score of 28.4 % while the baseline score in real German model is 15.8 %. Another experiment showed that adding more languages into the multilingual model can improve the quality of recognition system.

This work was extended in [17] to improve the recognition process with language-independent and language-adaptive acoustic models. Especially in this study, the group author introduced three different methods for *multilingual acoustic model combination* which are the language separate method (*ML-sep*); the language mixed method (*ML-mix*) and the language tagged method (*ML-tag*). The combination is realized on mixture weights and Gaussian components per state of the acoustic model. In ML-sep combination method, each language specific phoneme is trained only with data from its own language. In ML-mix combination method, data is shared across different language to train acoustic model of poly-phonemes. No information about language is attached to each poly-phoneme. ML-tag method give another way to share phoneme model across languages. In this method, each phoneme receives a language tag attached in order to preserve to information about the language that the phoneme belongs to. The model combination has some main goals including the reduction of overall amount of acoustic model parameters and the improvement of the model robustness for language adaptation purposes.

Recently, in order to help the French linguists process language documentation for Yongning Na language, an unwritten Sino-Tibetan language with less than 50,000 speakers in Southwest China, a simple phoneme recognition system were built in [23]. A cross-lingual model based on ML-sep combination method from [17] was built from 5-hour speech data of five other languages (English, French, Chinese, Vietnamese and Khmer) to determine to what extent Na sounds similar to sounds found in these five languages could be accurately recognized. Although the correct error rate at first pass was high, there were some clues that the method was reasonable.

The cross-lingual phoneme recognition therefore can be seen as the first step in approaching a new target language where there is no training data. In this work, the under-resourced and unwritten Muong language is considered as the target language to apply the "cross-lingual" phonemes recognition method above. With the available of phone mappings between Muong and Vietnamese as in the Table 1, we try to experiment on the phoneme recognition for Muong language using Vietnamese phoneme recognizer.

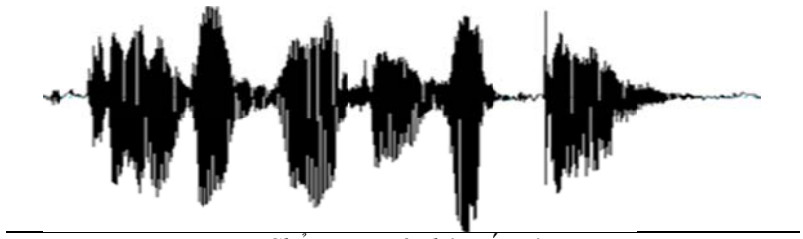
IV. EXPERIMENT

This section presents the experiment of Muong phoneme recognition using "cross-lingual" approach. The experiment includes two main tasks:

- Building an automatic phoneme recognition system for Vietnamese: the baseline system.
- Appling the baseline system to recognition phoneme sequences of Muong speech.

4.1. Baseline system: Vietnamese phoneme recognition

The baseline system was built as a simple Vietnamese phoneme recognition system. This system was developed using an open source ASR framework (CMU SphinX[4]) and training with VNSpeechCorpus [20] consisting of 20 hours of recorded speech with more than 30 speakers from North of Vietnam. Unlike training for a typical ASR system, the training process of acoustic model for phoneme recognition system uses the input of speech signal transcribed in phonemes, instead of normal words. Therefore, all the transcriptions in training corpus were converted into phone sequences, as shown in Figure 1.



Chẳng việc gì phải gấp cả

că4ŋ viə8k zi2 fa4j ɣə̌7p ka4

c aX NG se v ie k se z i se f a j se G 7X p se k a

Figure 1. Example of training speech and transcription in Vietnamese word *(top)*, in IPA (*middle*) and in XSAMPA (*bottom*)

The XSAMPA transcriptions were used as the input for training the model. Note that we used "*se*" to represent the boundary between syllable, and it was considered like a single phoneme. In this pilot work, we did not take into account the tones of Muong. So the transcription for training (in XSAMPA) had no information of tones (see Figure 1). The system was trained using the default configuration of CMU SphinX toolkit, with 3-state HMMs left-to-right, 64 Gaussian mixtures, pronunciation dictionary in phone-phone template. We deal with context-dependent acoustic model for only the phones.

After training process, the baseline system was firstly evaluated in Vietnamese phoneme recognition. The test set for this evaluation consisted of 774 Vietnamese speech sentences (corresponding to more than one hour of speech) with phone transcriptions. These sentences were put into the baseline system to get the outputs of 774 corresponding phoneme sequences. The correct recognition rate was calculated by comparing output phoneme sequences and the correct phone transcription of input speech. Overall, the average of correct recognition rate of Vietnamese phone is about 74.5 %.

4.2. Cross-lingual phoneme recognition for Muong

After building the Vietnamese phoneme recognition system (the baseline), we used this system to recognize the Muong phoneme sequences of Muong speech. This task was done simply, thanks to the phonology comparison between Viet-Muong language presented in the section II. The input of baseline system was Muong speech. A simple conversion of the Vietnamese phone to Muong phone was applied on the output of system.

- For the equivalent phoneme group and closed phoneme group (as in Table 1): These phonemes in Muong are mapped to the corresponding phonemes in Vietnamese. The phoneme /*p*/: not exist in original Vietnamese phone set, however this phone appear in many Vietnamese loan word, for example "cục pin " - /*kukp̂ pi1n*/. So it can be considered as an equivalent phoneme.

[4] https://cmusphinx.github.io/

- For the distinctive phonemes in Muong: The phoneme /r/ does not exist in Hanoi standard Vietnamese, and always be pronounced as /z/ (i.e "cái rổ" - /*kaj zo*/). The phonemes /*hr*/, /*kl*/, /*tl*/ are the specific phonemes of Muong and difficult to transform to any phoneme in Vietnamese. Thus, these cases are not considered in this study and will be dealt in the future work.

For evaluation, we used a test set of Muong speech consisting of 100 utterances spoken by one female speaker from Muong television program, *Hoa Binh* Radio and Television. These speech signals were manually transcribed into Muong phoneme sequences by two linguists who have experiences with Muong Bi language to get the references. These speech signals of Muong speech were put into the baseline phone recognition system to get the Vietnamese phoneme output. The mapping rules above were applied to convert all Vietnamese phoneme outputs of the test set to the Muong phonemes. These outputs in Muong phonemes were compared with the reference transcriptions of input speech to evaluate the performance of the system.

V. RESULTS

The objectives of result analysis are to (1) evaluate the performance of phoneme recognition system built on Vietnamese phoneme set on Muong speech; and (2) examine the similarities, the distinctions and the confusions of phoneme systems between these two familiar languages, using speech processing application.

5.1. Phoneme recognition

The correct recognition rates of Muong and in Vietnamese phonemes are presented in Figure 2 in two groups: equivalent phonemes and closed phonemes. Globally, most of phonemes are recognized in both of languages. In Vietnamese, all of phonemes are recognized with the correct recognition rate from 30 % to 90 %. The correct recognition rate of phonemes in Muong is varied from 10 % to 70 %. In most of cases, the correct recognition rate of Muong is 15 % to 25 % lower than in that of Vietnamese. This is a fairy good result and shows that the phoneme model of Vietnamese can be used to recognize most of phonemes in Muong language.

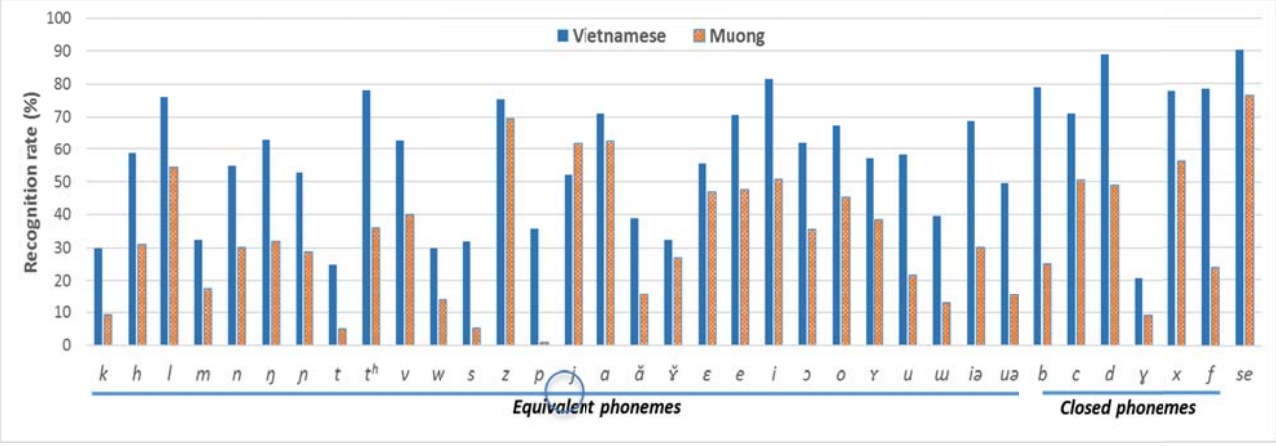


Figure 2. Correct recognition rates of Vietnamese and Muong phonemes

In more detail, there is no significant different in phonemes recognition of Muong between the equivalent group and closed group. Both contains well-recognized phonemes (correct recognition rate >50 %) and hardly-recognized phonemes (correct recognition rate <10 %). In the equivalent group, phonemes are well recognized are */l/, /z/, /j/, /a/, /i/,* while in the closed group they are phonemes /*c*/, /*x*/. The phonemes */p/, /t/, /k/, /s/* are difficult to recognized in the equivalent group and the same for /ɣ/ phoneme in closed group. The phonemes /*t*/, /*s*/ have very low correct recognition rate (5 %). The phonemes /*p*/ is nearly unrecognizable. And even for other languages these phonemes are hard to be recognized automatically. The case of /*j*/ phoneme is interesting. The correct recognition rate of /*j*/ in Muong language is 62 % which is even better than the correct recognition rate of /*j*/ in Vietnamese (52 %). Perhaps, in Vietnamese (Hanoi dialect), this phoneme is typically pronounced as /*z*/ phoneme (e.g. the word "*giáo dục*" is pronounced like "*dáo dục*"), but this case is not appeared in Muong language. The syllable boundary ("*se*") has a very high correct recognition rate for both of languages (80 %- 90 %). That confirms the similarity of syllable structure between two languages (monosyllabic languages). We suppose that the acoustic model for these phonemes with a simple cross-lingual training processed from Vietnamese speech can be used to apply to Muong language.

5.2. Phonemes confusion

The confusion matrices of Muong phonemes recognition were presented in Figure 3, by computing the confusion recognition rate (in %) from one phonemes to the other phonemes. According to this confusion matrix, some phonemes of Muong are well recognized and have no confusion with other phonemes such as /*l*/, /*z*/, /*j*/, /*a*/ and syllable

boundary (*se*). Three phonemes /*n*/, /*ŋ*/, /*ɲ*/ have the reciprocal confusion. Some phonemes have a good correct recognition rate, however sometime they are confused to another phoneme such as /*m*/ to /*b*/, /*u*/ to /*o*/, /*b*/ to /*d*/, /*f*/ to /*d*/. Phoneme /*t*/ is nearly unrecognizable and strongly be confused to phoneme /*d*/. Two phonemes /*p*/ and /ɣ/ are also unrecognizable, however it is difficult to find a major confusion with other phonemes. Actually, in the output of phoneme confusion matrix, the phoneme /*p*/ have a very high non-recognition rate of 41.4 %. That means in half of cases, the phonemes /*p*/ in Muong speech cannot be assigned to any other phonemes in the acoustic model of Vietnamese speech. The confusion recognition rate will be more accuracy if the test set is balanced in phoneme. So in the next step we will continue to analyze this problem.

VI. CONCLUSIONS

This paper presents the first experiment of building a phonemes recognition for Muong, an under-resourced and unwritten language in Vietnam. Following the cross-lingual approach, an acoustic model trained on Vietnamese speech were applied to convert Muong speech input into a sequence of Muong phonemes. The fairy good result (an average of more than 50 % correct recognition rate) shows that this is a potential approach, which can be quickly applied to create an automatic phoneme recognition for a minority language without available training data. The result analysis also shows some similarities and the distinctions between Vietnamese and Muong languages. Some interesting cases were found in the phonemes recognition and confusion analyses which need more study in the future.

As a pilot study, the result in this experiment will be the basic for our work in Muong language processing. The future work will also deal with some remain problems such as: (1) processing the distinct phonemes in Muong, (2) studying the effect of language model in cross lingual phonemes recognition, (3) taking into account the dialect and tonal information of Muong language, and also (4) adaptation of Vietnamese acoustic model to Muong acoustic model.

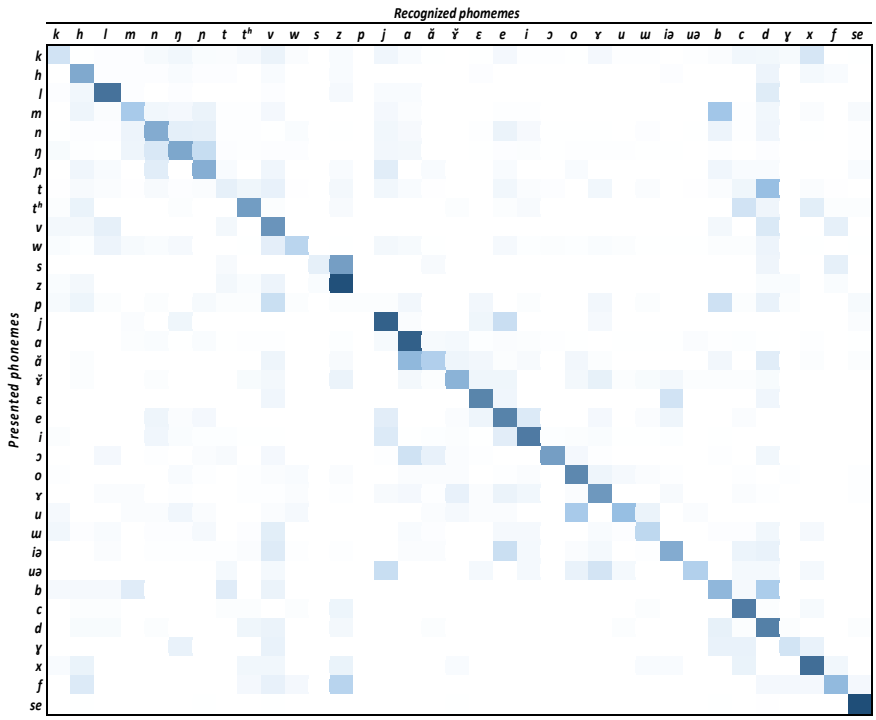


Figure 3. Phonemes confusion matrix of Muong (*The darker cell presents the higher rate of recognition/confusion*)

VII. ACKNOWLEDGEMENT

This research is funded by the Hanoi University of Science and Technology (HUST) under project number T2016-PC-186.

VIII. REFERENCES

[1] L. Besacier, E. Barnard, A. Karpov, and T. Schultz. "Automatic speech recognition for under-resourced languages: A survey". *Speech Communication*, vol. 56, pp. 85-100, 2014.

[2] B. Michailovsky, M. Mazaudon, A. Michaud, S. Guillaume, A. François, and E. Adamou. "Documenting and researching endangered languages: the Pangloss Collection". *Language Documentation and Conservation*, vol. 8, pp. 119-135, 2014.

[3] O. Niebuhr and A. Michaud. "Speech data acquisition: the underestimated challenge". *KALIPHO-Kieler Arbeiten zur Linguistik und Phonetik*, vol. 3, pp. 1-42, 2015.

[4] G. Hinton *et al.*. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2012.

[5] T. Schultz and A. Waibel. "Multilingual and crosslingual speech recognition" in *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, 1998, pp. 259-262.

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

[7] A. G. Haudricourt. "La place du vietnamien dans les langues austroasiatiques". *Bulletin de la Société de Linguistique de Paris*, vol. 49, no. 1, pp. 122-128, 1953.

[8] V. D. Pham, D. K. Mac, T. H. H. Vu, and D. D. Tran. "Rapid development of text to speech system for unsupported languages using faking input approach: experiment with Muong". presented at the FAIR 11 - Fundamental and Applied IT Research, Da Nang, 2017.

[9] Nguyễn Minh Đức. "Một vài nét về các thổ ngữ của tiếng Mường Hòa Bình" in *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" in *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 Khang, Bùi Chỉ và Hoàng Văn Hành. *Từ điển Mường - Việt*. Hà Nội: Nhà xuất bản Văn hóa Dân tộc, 2002.

[12] Trần Trí Dõi. *Một vài vấn đề nghiên cứu so sánh - lịch sử nhóm ngôn ngữ Việt Mường*. Hà Nội: Nhà xuất bản Đại học quốc gia Hà Nội, 2011.

[13] J. Köhler. "Comparing three methods to create multilingual phone models for vocabulary independent speech recognition tasks" in *Multi-Lingual Interoperability in Speech Technology*, 1999.

[14] J. Schalkwyk. "Multi-lingual speech recognition with cross-language context modeling". Dec-2006.

[15] P. Ladefoged. "The revised international phonetic alphabet". *Language*, vol. 66, no. 3, pp. 550-552, 1990.

[16] O. Andersen, P. Dalsgaard, and W. Barry. "Data-driven identification of poly-and mono-phonemes for four European languages" in *Third European Conference on Speech Communication and Technology*, 1993.

[17] T. Schultz and A. Waibel. "Language-independent and language-adaptive acoustic modeling for speech recognition". *Speech Communication*, vol. 35, no. 1-2, pp. 31-51, 2001.

[18] J. Kohler. "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds" in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 1996, vol. 4, pp. 2195-2198.

[19] J. J. Sooful and E. C. Botha. "Comparison of acoustic distance measures for automatic cross-language phoneme mapping" in *Seventh International Conference on Spoken Language Processing*, 2002.

[20] V. B. Le, D. D. Tran, L. Besacier, E. Castelli, and J. F. Serignat. "First steps in building a large vocabulary continuous speech recognition system for Vietnamese" in *RIVF 2005*, 2005.

[21] W. Byrne *et al.*. "Towards language independent acoustic modeling" in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, 2000, vol. 2, pp. II1029-II1032.

[22] T. Schultz and A. Waibel. "Adaptation of pronunciation dictionaries for recognition of unseen languages" in *Proc. SPIIRAS International Workshop on Speech and Computer, St. Petersburg*, 1998, pp. 207-210.

[23] T. N. D. Do, A. Michaud, and C. Eric. "Towards the automatic processing of Yongning Na (Sino-Tibetan): developing a 'light' acoustic model of the target language and testing'heavyweight'models from five national languages" in *4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2014)*, 2014, pp. 153-160.

[24] O. Adams, T. Cohn, G. Neubig, and A. Michaud. "Phonemic transcription of low-resource tonal languages" in *Australasian Language Technology Association Workshop 2017*, 2017, pp. 53-60.

NHẬN DẠNG ÂM VỊ CHÉO CHO CÁC NGÔN NGỮ CÙNG HỌ, ÁP DỤNG CHO TIẾNG VIỆT VÀ TIẾNG MƯỜNG

Trần Thị Thu Thúy, Đỗ Thị Ngọc Diệp, Mạc Đăng Khoa, Phạm Văn Đồng

TÓM TẮT: *Nhận dạng âm vị tự động cho một ngôn ngữ là bài toán cần thiết trong xây dựng các hệ thống xử lý tiếng nói đối với một ngôn ngữ mới. Nó yêu cầu các kiến thức về xử lý tiếng nói và kiến thức ngôn ngữ học. Với các ngôn ngữ ít nguồn tài nguyên (ví dụ như các ngôn ngữ dân tộc thiểu số), hệ thống nhận dạng âm vị tự động chung chưa có sẵn. Bài báo này trình bày một thử nghiệm xây dựng hệ thống nhận dạng âm vị tự động cho tiếng Mường, một ngôn ngữ thiểu số chưa có chữ viết ở Việt Nam. Dựa trên cách tiếp cận "cross-lingual", đầu vào tiếng nói Mường được phiên âm tự động thành chuỗi âm vị tiếng Việt dựa trên một mô hình âm học được huấn luyện sẵn trên dữ liệu tiếng nói tiếng Việt. Kết quả đánh giá cho thấy tỷ lệ nhận dạng âm vị đúng trung bình với tiếng Mường là trên 50 %. So sánh về tỷ lệ nhận dạng đúng giữa bộ âm vị của tiếng Mường và tiếng Việt cho những kết quả thú vị, giúp các nhà ngôn ngữ học đánh giá các đặc điểm chung và khác biệt giữa hai ngôn ngữ gần gũi này.*

Từ khóa: *Nhận dạng âm vị, chéo ngôn ngữ, ngôn ngữ nghèo tài nguyên, tiếng Mường, tiếng Việt.*