EasyChair Preprint

№ 7302

A Survey of Vietnamese Automatic Speech Recognition

Cao Hong Nga, Chung-Ting Li, Yung-Hui Li and Jia-Ching Wang

January 5, 2022

A Survey of Vietnamese Automatic Speech Recognition

Cao Hong Nga* Chung-Ting Li† Yung-Hui Li* Jia-Ching Wang*
*Department of Computer Science and Information Engineering, National Central University, Taiwan
*AI Research Center, Hon Hai Research Institute, Taiwan

***Abstract*—In this paper, we survey Vietnamese automatic speech recognition (ASR). The objective of this survey is to provide an overview of the current status and remaining challenges of implementing a Vietnamese ASR. Recently, there are some studies on ASR for Vietnamese language; however, these studies encounter some obstacles and the results obtained compared to other languages such as English or Mandarin are lower. With regards to Vietnamese speech recognition, we will examine the methods of building a system along with speech data and text data collection techniques and available speech resources. We review both the methods applied to acoustic modeling and those used to language modeling. In addition, we convey some directions for future research.**

***Index Terms*—Speech recognition, Vietnamese speech recognition, acoustic model, language model**

I. Introduction

Vietnamese is one of low resource languages with over 90 million native speakers worldwide and the sole official and social language of Vietnam. The Vietnamese writing system was built in the 17^{th} century by Alexandre de Rhodes, a French missionary. This writing system is based on Roman characters with diacritics, which is easier to write and remember than the previous script based on Chinese characters called "chữ Nôm". Today, it is also known as the national language script (chữ Quốc ngữ) [1] [2].

Vietnamese is predominantly monosyllabic language, which means that each word consists of one syllable, e.g., “học” (study). If a word consists of more than one syllable, the syllables are separated by spaces, e.g., “học sinh” (student). Each Vietnamese syllable structure consists of three parts in order: An optional initial consonant - a required vowel with optional diacritic - an optional ending consonant [1] [2] [3] [4]. In particular, the phonemes of Vietnamese language include the following three main components [1] [2] [3] [4]:

- Consonants: Each consonant consists of one, two or three letters like “ngh”, “nh”, “n”. These consonants have their own pronunciation rules except that “ngh” and “ng” share the same sound.
- Vowels: Similar to consonants, vowels are monophthongs ("a"), diphthongs ("ai") or triphthongs ("oai") and have a separate pronunciation for each vowel.
- Tones: Vietnamese has 6 tones: no mark (ngang), acute (sắc), grave (huyền), hook above (hỏi), tilde (ngã), dot below (nặng).

For example, the syllable "học" (study) composes of consonant-vowel-consonant and dot below tone mark, "ăn" (eat) composes of vowel-consonant, "bơi" (swim) composes of consonant-vowel. The writing forms the pronunciation of words.

Recently, there has been various machine learning techniques that can handle Vietnamese ASR; however, the research for this field has some obstacles due to the characteristics of the Vietnamese language. In this paper, we report the recent studies in this subject.

The rest of this paper is organized as follows. Remaining challenges is discussed in Section II. Next, in Section III, we describe approaches to handle Vietnamese speech recognition and processes of collecting Vietnamese speech data and text data. Conclusion and discussion are presented in Section IV.

II. Challenges in building Vietnamese speech recognition

When developing a Vietnamese ASR system, we should concern about the important factors of Vietnamese speech. Some of the key research challenges for building a Vietnamese speech recognition system are as follows:

- Under-resourced language: This means that the Vietnamese language has very low resources that are publicly available for research purposes. Furthermore, most public datasets are generated by reading texts; therefore they are not compatible with real-life scenarios such as conversational or discussion speech recognition.
- Accents: Although there are rules of writing and pronunciation, the accents and words used vary from region to region, which lead to a major challenge when building an ASR system. The accents between regions vary greatly, and can be divided into three main regions: Northern dialect, Central dialect, and Southern dialect. For example, some regions in Northern Vietnam pronounce "l" as "n" and vice versa or most people in the South pronounce the ending consonants in "n" and "ng" the same.
- Speaking style: Like other languages, the intonation and language used in different situations such as communicating on the phone, reporting the news or answering an interview are different. Written text can be easily collected for language models; however, it has a gap with spoken language. In addition, each region uses different words to express the same thing. For example the North people use the word "lợn", the South people use the word "heo" to present the same meaning "pig".

- Recording environments: The recording environment is also a factor affecting the speech recognition results. Each recording situation will have corresponding noise; for example, street noise, meeting room noise, television news noise will be different.

III. VIETNAMESE ASR APPROACHES

There are various studies on development of Vietnamese ASR system. Basically, an ASR system composes of two main parts: (1) the acoustic model predicts the sequence of words for the input audio, (2) the language model is a text-based application which is trained separately from the acoustic model to compute the probability for every possible sequence of words. The structure of an ASR as in Fig. 1. In previous studies, authors often used the Hidden markov Models (HMM) and Gaussian Mixture Model (GMM) for acoustic model. However, with the advent of deep neural network-based architectures, Vietnamese ASR studies have applied these techniques or used hybrid of both statistical model and deep neural network (DNN) model. These novel methods have produced noble results. For language modeling, most studies use n-gram models to compute probability distributions. Besides, there are a few studies using the DNN language model for Vietnamese language. In this section, we will present these studies in chronological order.

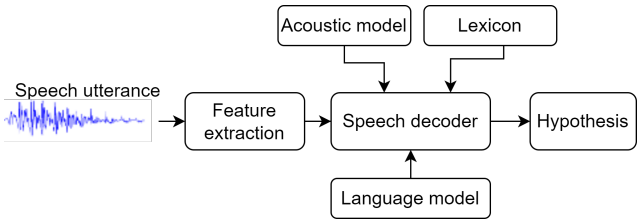

Figure 1. General ASR system diagram

A. Acoustic model

In 2008, Quang et al. used HMM to build acoustic model and recognize Vietnamese tones [5]. The language model applied Good-Turing and Katz back-off smoothing algorithms. During the decoding stage, the score of each word was recalculated based on the acoustic model, language model and tone model with the ratios of alpha, beta, and gamma, respectively. The authors recorded approximately 14.4 hours of speech with 80% paragraph style and 20% conversation style, in which 11.2 hours for training and remaining 3.2 hours for testing. The results in syllable error rate (SER) of the proposed method were 7.6% for paragraph style and 21.9% for conversation style. The research showed that tone score and the use of polysyllabic words improve the model performance.

In 2009, [6] used the Rapid Language Adaptation Tool (RLAT) [7] to train the initial system with HMM and GMM techniques. Then, the model was fine-tuned by applying different methods to handle characteristics of Vietnamese speech such as combining monosyllabic words with multisyllabic words, applying signal adaptation, creating a dialect-dependent recognition. The authors recorded 25 hours of speeches from 120 speakers in Hanoi and Ho Chi Minh cities. The 5-grams language model was used for decoding and achieved the best word error rate (WER) of 11.7%. In the same year, Le et at. proposed different techniques for acoustic model such as applying acoustic–phonetic unit distances, cross-lingual acoustic modeling [8]. The system was applied for Vietnamese and decoded with 3-gram language model with Good–Turing discounting and Katz backoff for smoothing. In addition, the authors proposed to use word/sub-word lattices decomposition and combination for language model. The model achieved the best SER and WER when training with 14-hour speech data of 36.6% and 42.7%, respectively.

In 2014, [9] introduced a method of using DNN-HMMs in combination with pitch features, tone modeling, automatic question generation for low resource language speech recognition. This model was applied to Vietnamese and had an accuracy of 48.32% on the OpenKWS13[1] development set.

Vietnamese is a tonal and monosyllabic language, different tones combined with a word will form different meanings. In 2015, Nguyen et al. proposed a method to convert from grapheme to phoneme that can convert any Vietnamese word into tonal phoneme and then built a Vietnamese speech recognition system [10]. In this study, every syllable was considered to be a combination of three components, namely optional initial consonant, final and tone. The acoustic features used for the model were Mel-frequency cepstral coefficients (MFCC) and pitch. The features vectors were concatenated and applied linear discriminate analysis (LDA) and decorrelated with a maximum likelihood linear transformation (MLLT) to reduce their sizes. The neural network took these features as input data. The language model used for the decoding stage was a trigram with Kneser-Ney smoothing. The training speech dataset consisted of 212 hours with 1267 speakers and the text data for language model included 291k utterances. The WER results on VoiceTraTest (36-minute speech data) and BTECTest (19-minute speech data) were 27.73% and 9.14%, respectively. In this study, the authors proved that the use of phoneme-based pronunciation has a good effect on Vietnamese ASR.

Pitch features and tonal information can help to increase the accuracy of ASR systems of tonal languages including Vietnamese [5], [10], [11]. In [11], MFCC features, pitch features, and tonal information were used for training the system. MFCC features were applied LDA and MLLT, and then trained with triphone GMM acoustic model. The next model was Maximum Mutual Information (MMI). Then speaker dependent was trained by Maximum Likelihood Linear Regression (fMLLR). A hybrid HMM-DNN was used to train as the last model. The 3-gram syllable-based language model was trained with 500MB of text collected from online news and forum. VIVOS dataset, a free Vietnamese ASR corpus, was prepared by the paper's authors and then used to train the acoustic model. This

[1]https://www.nist.gov/system/files/documents/itl/iad/mig/OpenKWS13-EvalPlan.pdf

dataset has a total of 15 hours, of which 14:45 hours are for the training set, the remaining is 0:55 hour for the testing set. The numbers of speakers for the training set and the testing set are 46 and 19, respectively. Among the 56 speakers of the training set, there are 22 male and 24 female speakers. These numbers correspond to the testing set of 12 and 17. The numbers of utterances for the training set and the testing set are 11,660 and 760, respectively. The dataset is available at https://ailab.hcmus.edu.vn/vivos/. The WER of the system was 9.48% for this test set.

Viettel, a Vietnamese multinational telecommunications company, collected Vietnamese telephone conversation speech corpus consisting of 85.8 hours from Viettel's call center [12]. The dataset was divided into two subsets: training set and testing set, with a duration of 70 hours and 15.8 hours, respectively. Speech recognition system started with a voice activity detector (VAD) that detected non-speech phonemes and segmented the audio based on non-speech phonemes with a duration threshold of 1 second. These segments were applied speed perturbation augmentation and feature extraction. The acoustic model were applied GMM with speaker adaptive training (GMM-SAT) and Time Delay Deep Neural Network (TDNN) [13]. A 4-gram language model was used for decoding and the model achieved the WER of 17.44% for the test set.

In 2018, Nguyen et al. collected a 500-hour Vietnamese dataset and used this dataset to build a speech recognition system [14]. The dataset was built by extracting sentences from online newspapers and Wikipedia, then these sentences were read and recorded. In order to adapt to different environments, these audio files were added noise before extracting the features. The features used for the acoustic model were MFCCs and pitch. The acoustic model was built by using TDNN and bi-directional long-short term memory (BLSTM). A 4-gram language model with Kneser-Ney smoothing was used for the decoding process. To improve the final result, the model continued to decode with recurrent neural network language model (RNNLM). The WER for the 3-hour test set extracted from this dataset was 6.9%. In the same year, Do et al. gathered a large amount of Vietnamese speech data with different accents from The North, Central and South of Vietnam and used this data to build an ASR system [15]. The corpus included 1200 hours in a variety of accents and speaking environments. The authors have used two types of input features for the acoustic model, namely MFCC and FBANK features. These features combined with pitch and i-vector to improve system performance. The model was trained with DNN cross-entropy. At the decoding stage, they combined the acoustic model with the 4-gram language model. The system ranked first in performance in the 2018 VLSP challenge with a WER of 6.29% and achieved a WER of 11% in their test set.

In 2019, Huy combined two DNN methods, namely Long Short-Term Memory (LSTM) and TDNN with Connectionist Temporal Classification (CTC) loss function for training Vietnamese ASR [16]. The inputs of the model were MFCCs and coefficients of pitch. These features were applied LDA to reduce the dimensions before fed into DNN model. At the inference stage, the model combined with 3-gram language model to re-score and produce the final result. The speech corpus used for training was the corpus developed by FPT Technology Research Institute, a Corporation in Vietnam (FTRI). The duration of the data set was 2036 hours, including the accents of the North, the Central, and the South regions. To evaluate the performance of the model, the authors used VLSP2018 developed by the Association for Vietnamese Language and Speech Processing and FPT-test set. WERs for VLSP2018 and FPT-Test were 9.71% and 14.41%, respectively.

In [17], the textual data used for language modeling obtained from the Internet and divided into domains to train the language model separately for each domain, then combined these language models with different weights for each domain using interpolation techniques to increase performance of the model. In this study, acoustic model was trained with Gaussian Mixture Model - Hidden Markov Model (GMM-HMM) and TDNN. The best WER for VLSP2018 test set was 4.85% and VLSP2019 test set was 15.09%.

In 2021, [18] compared different platforms including Vais, Vtcc, Fpt, and Google for Vietnamese speech recognition in news, interviews and music. The paper showed that in the news and interview domains, Vais and Viettel yielded good results while Google recognized better in the music domain.

B. Language model

Language model estimates the probabilities of possible word sequences. These probabilities will be combined with the acoustic model likelihoods to yield the best overall hypothesis. There are various types of language model such as n-gram, positional [19], neural network [20]. N-gram and neural network language models are two popular methods used for Vietnamese speech recognition.

1) N-gram: N-gram is a probabilistic language model used to compute the probability of a sequence of words. We can use available toolkits to build a language model such as kenlm[2] and srilm[3]. In [5], [6], [8], [10]–[12], [14], [15], [17], n-gram language model is used as a part of ASR system.

2) Deep neural network: For the language model, we can easily collect large amount of training data for DNN models, the previous results showed that the DNN-based methods gave superior performance compared to the n-gram method [20]. The system in [14] used both n-gram and DNN language models for decoding. We can train the DNN language model from scratch or use a pre-trained model. PhoBert [21], an unsupervised pre-trained language model for Vietnamese, is based on RoBERTa [22] to improve the performance of various NLP tasks. It was trained with 20GB of text and has been applied to various downstream tasks like part-of-speech tagging, named-entity recognition, etc. The results of these downstream tasks outperformed other methods. This pretrained

[2]https://github.com/kpu/kenlm

[3]http://www.speech.sri.com/projects/srilm/

Table I
RECENT STUDIES ON VIETNAMESE SPEECH RECOGNITION

Year	Research	Acoustic model	Language model	Testing set	WER
2019	Vais asr: Building a conversational speech recognition system using language model combination	GMM-HMM + TDNN	N-gram	VLSP2018 VLSP2019	4.85% 15.09%
2019	An End-to-End Model for Vietnamese Speech Recognition	LSTM + TDNN	3-gram	FPT-test VLSP2018	9.71% 14.41%
2018	Development of high-performance and large-scale vietnamese automatic speech recognition system	DNN	4-gram	VLSP2018	6.29%
2018	Development of a Vietnamese large vocabulary continuous speech recognition system under noisy conditions	TDNN+BLSTM	4-gram RNNLM	3-hour-test	6.9%
2017	Development of a Vietnamese speech recognition system for Viettel call center	GMM-SAT + TDNN	4-gram	Viettel-test	17.44%
2016	A non-expert Kaldi recipe for Vietnamese speech recognition system	GMM, MMI, fMLLR + HMM-DNN	3-gram	VIVOS	9.48%

model can also be applied to Vietnamese ASR. PhoBert includes two versions, PhoBERT-base with 135M parameters and PhoBERT-large with 370M parameters. We can accesse PhoBert at https://github.com/VinAIResearch/PhoBERT

IV. DISCUSSION AND CONCLUSION

In this paper, we present researches on Vietnamese ASR and language models which apply for decoding stage of Vietnamese ASR systems. Our study reveals the recent trend to focus on DNN approaches for acoustic modeling and yielded remarkable results. We can observe the results of recent studies from Table I. Most of the studies use the n-gram language model for the decoding phase; however, there are various studies that have used the DNN language model for NLP and achieved remarkable results. We can apply these methods to ASR in the future. We can use pretrained language model for decoding. Besides, to overcome the drawbacks of the system, we also need a large amount of data to cover all the features of Vietnamese speech.

REFERENCES

[1] André-Georges Haudricourt, "The origin of the peculiarities of the Vietnamese alphabet," *Mon-Khmer Studies*, vol. 39, pp. 89–104, 2010, Translated by Alexis Michaud. Original publication: Haudricourt, André-Georges. 1949. L'origine des particularités de l'alphabet vietnamien. Dân Viêt-Nam 3. 61-68.

[2] Mark Alves, "Linguistic research on the origins of the vietnamese language: An overview," *Journal of Vietnamese Studies*, vol. 1, no. 1-2, pp. 104–130, 2006.

[3] Deborah Hwa-Froelich, Barbara W Hodson, and Harold T Edwards, "Characteristics of vietnamese phonology," 2002.

[4] Laurence C Thompson, "A vietnamese reference grammar (revised edition)," 1991.

[5] Hong Quang Nguyen, Pascal Nocera, Eric Castelli, and Van Loan Trinh, "A novel approach in continuous speech recognition for Vietnamese, an isolating tonal language," in *Proc. Interspeech 2008*, 2008, pp. 1149–1152.

[6] Ngoc Thang Vu and Tanja Schultz, "Vietnamese large vocabulary continuous speech recognition," in *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, 2009, pp. 333–338.

[7] Tanja Schultz and Alan Black, "Rapid language adaptation tools and technologies for multilingual speech processing," *Proc. ICASSP Las Vegas, USA*, 2008.

[8] Viet-Bac Le and Laurent Besacier, "Automatic speech recognition for under-resourced languages: application to vietnamese language," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1471–1482, 2009.

[9] Shifu Xiong, Wu Guo, and Diyuan Liu, "The vietnamese speech recognition based on rectified linear units deep neural network and spoken term detection system combination," in *The 9th International Symposium on Chinese Spoken Language Processing*, 2014, pp. 183–186.

[10] Van Huy Nguyen, Chi Mai Luong, and Tat Thang Vu, "Tonal phoneme based model for vietnamese lvcsr," in *2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2015, pp. 118–122.

[11] Hieu-Thi Luong and Hai-Quan Vu, "A non-expert kaldi recipe for vietnamese speech recognition system," in *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, 2016, pp. 51–55.

[12] Quoc Bao Nguyen, Van Hai Do, Ba Quyen Dam, and Minh Hung Le, "Development of a vietnamese speech recognition system for viettel call center," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.

[13] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth annual conference of the international speech communication association*, 2015.

[14] Quoc Bao Nguyen, Van Tuan Mai, Quang Trung Le, Ba Quyen Dam, and Van Hai Do, "Development of a vietnamese large vocabulary continuous speech recognition system under noisy conditions," in *Proceedings of the Ninth International Symposium on Information and Communication Technology*, 2018, pp. 222–226.

[15] Quoc Truong Do, Pham Ngoc Phuong, Hoang Tung Tran, and Chi Mai Luong, "Development of high-performance and large-scale vietnamese automatic speech recognition systems," *Journal of Computer Science and Cybernetics*, vol. 34, no. 4, pp. 335–348, 2018.

[16] Van Huy Nguyen, "An end-to-end model for vietnamese speech recognition," in *RIVF*, 2019, pp. 1–6.

[17] Quang Minh Nguyen, Thai Binh Nguyen, Ngoc Phuong Pham, and The Loc Nguyen, "Vais asr: Building a conversational speech recognition system using language model combination," *arXiv preprint arXiv:1910.05603*, 2019.

[18] Hai Thanh Diep, Thanh Thi My Nguyen, Bich Ngoc Le, and Quy Xuan Dao, "Evaluation of vietnamese speech recognition platforms," in *2021 The 5th International Conference on Machine Learning and Soft Computing*, 2021, pp. 141–146.

[19] Yuanhua Lv and ChengXiang Zhai, "Positional language models for information retrieval," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 299–306.

[20] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin, "A neural probabilistic language model," *The journal of machine learning research*, vol. 3, pp. 1137–1155, 2003.

[21] Dat Quoc Nguyen and Anh Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1037–1042.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.