VLSP 2018 - Development of a Vietnamese Large Vocabulary Continuous Speech Recognition

Quoc Bao Nguyen, Van Hai Do
Cyberspace Center
Viettel Group
Hanoi, Vietnam
{baonq2, haidv21}@viettel.com.vn

Van Tuan Mai, Quang Trung Le
Cyberspace Center
Viettel Group
Hanoi, Vietnam
{tuanmv2, trunglq12}@viettel.com.vn

Ba Quyen Dam, Manh Dung Do
Cyberspace Center
Viettel Group
Hanoi, Vietnam
{quyendb, dungdm6}@viettel.com.vn

***Abstract*— In this paper, we first present our effort to collect a 500-hour corpus for Vietnamese read speech. After that, various techniques such as data augmentation, RNNLM rescoring, language model adaptation, bottleneck feature, system combination are applied to build the speech recognition system. Our final system achieves a low word error rate at 6.9% on the noisy test set.**

***Keywords*— Vietnamese speech corpus, Vietnamese speech recognition, bottleneck feature, system combination.**

I. Introduction

Vietnamese is the sole official and the national language of Vietnam with around 76 million native speakers[1]. It is the first language of the majority of the Vietnamese population, as well as a first or second language for country's ethnic minority groups.

There were several attempts to build Vietnamese large vocabulary continuous speech recognition (LVCSR) system where most of them developed on read speech corpuses [1-4]. In 2013, the National Institute of Standards and Technology, USA (NIST) released the Open Keyword Search Challenge (Open KWS), and Vietnamese was chosen as the "surprise language". The acoustic data are collected from various real noisy scenes and telephony conditions. Many research groups around the world have proposed different approaches to improve performance for both keyword search and speech recognition [5-7]. In 2017, we presented our effort to collect a Vietnamese corpus and build a LVCSR system for Viettel customer service call center [8] and achieved a promising result on this challenging task.

Recently, the Vietnamese Language and Speech Processing (VLSP) community has organized an evaluation campaign for the Vietnamese speech recognition task. The evaluation data were collected mainly from broadcast news such as VOV, VTV. No training or development data was provided. In this paper, we present our effort to collect 500-hour speech corpus and the process to build a Vietnamese LVCSR speech recognition system. Our final system achieves 6.9% word error rate (WER) on our noisy test set.

The rest of the paper is organized as follows. Section II describes our speech corpus. Section III presents the proposed speech recognition system. Section IV shows the experimental results and Section V concludes the paper.

II. Corpus Description

In this paper, we present our effort to collect a 500-hour read speech corpus which will be used to train our speech recognition system.

Previously, several Vietnamese speech corpora were collected by different research groups [1-4]. However, they are relatively small i.e., less than 100 hours while commercial systems normally use thousands of hours of training data. In Viettel, beside building a speech recognition for telephone conversation such as for call center, we also target on building a commercial system for other applications such as virtual assistance, smart home, etc.

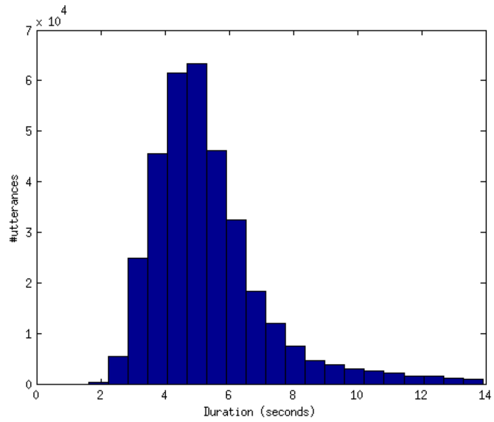

Fig. 1. The distribution of utterance durations.

To achieve this target, in the first phase, we collect 500-hour read speech mainly in the northern dialect. Speakers are recruited from our call center. We first collect text from online newspapers and Wikipedia. After cleaning, sentence segmentation is applied and text is then sent to speakers sentence

[1] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

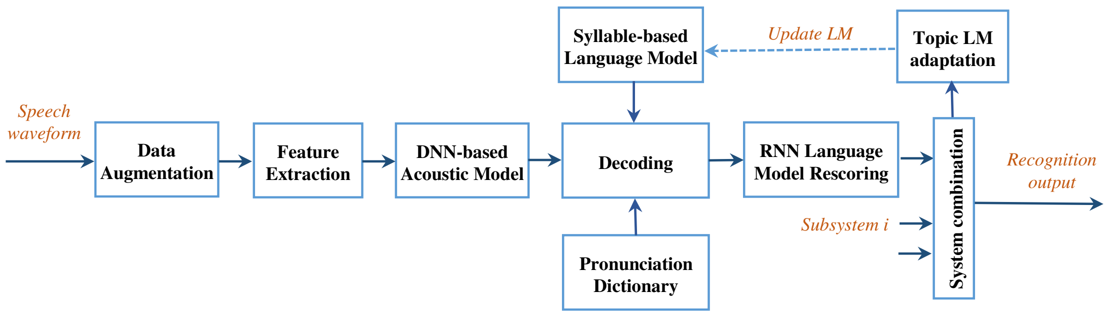

Figure 2. The proposed speech recognition system.

by sentence for speaking and recording. We create a friendly user interface website to help speakers and reviewers to be able to record and supervise easily.

The corpus is recorded with a sampling rate of 16kHz and a resolution of 16 bits/sample. In the corpus, there are 1,424 speakers with totally 343,115 utterances. To improve the corpus quality, each utterance is reviewed by a least one reviewer to warranty speech with good quality and the transcript and speech content are matched.

Figure 1 shows the distribution of utterance durations. The range of duration is from 2 to 14 seconds with the average duration of each utterance is 5.3 seconds.

III. The Proposed System

Our target is to build a speech recognition system which is robust to different recording environments. To achieve to this goal, training data are first augmented by adding various types of noise. Feature extraction is then applied to use for the acoustic model. For decoding, acoustic model is used together with syllable-based language model and pronunciation dictionary. After decoding, recognition output is rescored using RNN language model. The output generated by individual subsystems are combined to achieve further improvement. The recognition output is then used to select relevant text from the text corpus to adapt the language model. The decoding process is then repeated for the second time. In the next subsections, the detailed description of each module is presented.

A. Data Augmentation

To build a reasonable acoustic model, thousands hours of audio recorded in different environments are needed. However, to achieve transcribed audio data is very costly. To overcome this, many techniques have been proposed such as semi-supervised training [9], phone mapping [10], exemplar-based model [11], mismatched crowdsourcing [12]. In this paper, we use a simple approach to simulate data in different noisy environments. Specifically, we collect some popular noise types such as office noise, street noise, car noise, etc. After that noise is added to the clean speech of the original speech corpus with different level to simulate noisy speech. With this approach, we can easily increase the data quantity to avoid over-fitting and improve the robustness of the model against different test conditions.

B. Feature Extraction

We use Mel-frequency cepstral coefficients (MFCCs) [13], without cepstral truncation are used as input feature i.e., 40 MFCCs are computed at each time step which is similar setup in [14]. Since Vietnamese is a tonal language, pitch feature is used to augment MFCC.

Beside MFCC feature, bottleneck feature (BNF) [15] is also considered to build our second subsystem. BNF is generated using a neural network with several hidden layers where the size of the middle hidden layer (bottleneck layer) is very small. With this structure, we can choose an arbitrary feature size without using dimensionality reduction step, independently on the neural network training targets.

C. Acoustic Model

We use time delay neural network (TDNN) and bi-directional long-short term memory (BLSTM) with lattice-free maximum mutual information (LF-MMI) criterion [16] as the acoustic model.

D. Pronunciation Dictionary

Vietnamese is a monosyllabic tonal language. Each Vietnamese syllable can be considered as a combination of initial, final and tone components. Therefore, the pronunciation dictionary (lexicon) needs to be modelled with tones. As in [17], we use 47 basic phonemes. Tonal marks are integrated into the last phoneme of syllable to build the pronunciation dictionary for 6k popular Vietnamese syllables.

In order to build the dictionary for foreign, we select 5k popular foreign words from web newspapers. These words are then manually pronounced in the Vietnamese pronunciation. As a result, the total number of words in our lexicon is about 11k words. This lexicon is used for training as well as decoding.

E. Language Model

A syllable-based language model is built from 900MB web text collected from online newspapers. 4-gram language model with Kneser-Ney smoothing is used after exploring different configuration.

To get further improvement, after decoding, recurrent neural network language model (RNNLM) is used to rescore decoding lattices with a 4-gram approximation as described in [18].

F. System Combination

As described above, we have two subsystems i.e., the first subsystem uses MFCC feature while the second system uses bottleneck feature. The combination of information from different ASR subsystems generally improves speech recognition accuracy. The reason for this advantage is explained by the fact that different subsystems often provide different errors. In this paper, we examine the combination of our two subsystems using the minimum Bayes risk (MBR) decoding method described in [19], which we view as a systematic way to perform confusion network combination (CNC) [20].

G. Language Model Adaptation

The recognition output of our system has a relatively low word error rate (WER). Hence, from decoded text, we can know about the topic of the input utterances. This is especially important when we have no domain information.

Our algorithm is implemented as follows. The in-domain language model is constructed by using the recognition output. After that sentences from the general text corpus (900MB in this paper) are selected based on a cross-entropy difference metric. Detailed description about this selection algorithm can be referred in [21]. Finally, about 200MB text which have the most relevant to the recognition output are selected to build the adapted language model. The decoding process is then repeated with the new language model.

IV. Experiments

To evaluate our system performance, a test set is selected from our 500 hour corpus which is separated from the training set. The test set contains 2000 utterances with around 3 hours of audio. To simulate the real condition, the test set is added different noise with signal to noise ratio (SNR) from 15-40 dB.

A. Data Augmentation

We first examine the effect of data augmentation to the system performance. In this case MFCC feature is used. As shown in Table I, by applying data augmentation brings a big improvement. When the original training data are used only i.e., without data augmentation, the system is only trained with clean speech while test set is noisy. Hence, the model cannot recognize efficiently. By applying data augmentation, the original training data is multiplied by 11 times by adding various types of noise. Obviously, this makes model more robust with noise conditions and hence we achieve a low WER at 10.3%.

TABLE I. Effect of data augmentation to system performance.

Data augmentation	Word Error Rate (%)
No	28.2
Yes	10.3

B. RNNLM Rescoring

As shown in Table II, by applying RNNLM rescoring technique, we can achieve 1.4% absolute improvement.

TABLE II. Effect of RNNLM rescoring to system performance.

RNNLM Rescoring	Word Error Rate (%)
No	10.3
Yes	8.9

C. System Combination

The systems in the previous subsections are trained using MFCC feature. In this subsection, we investigate the effect of using bottleneck feature and its usefulness in system combination.

As shown in Table III, using BNF does not provide a good performance as MFCC. However, it provides complementary information and hence we can gain by combining them.

TABLE III. Bottleneck feature and system combination.

Subsystem	Word Error Rate (%)
Subsystem 1 (MFCC)	8.9
Subsystem 2 (BNF)	9.5
Combined system	8.1

D. Language Model Adaptation

As shown in Table IV, by applying language model adaptation, a significant WER reduction is achieved. It can be explained that the algorithm only chooses relevant (in-domain) sentences, while mismatched (out-domain) sentences which can be harmful to language model are discarded.

TABLE IV. Effect of language model adaptation to system performance.

Language model adaptation	Word Error Rate (%)
No	8.1
Yes	6.9

V. Conclusions

In this paper, we have described our 500-hour speech corpus. Various techniques such as data augmentation, RNNLM rescoring, language model adaptation, bottleneck feature, system combination were then applied. Our final system achieves a low word error rate at 6.9% on the noisy test set.

In the future, we will enlarge the speech corpus to cover most of the popular dialects in Vietnamese with different aging ranges as well as enlarge the text corpus to make our system more robust and achieve even better performance.

References

[1] Thang Tat Vu, Dung Tien Nguyen, Mai Chi Luong, and John-Paul Hosom, "Vietnamese large vocabulary continuous speech recognition," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2005, pp. 492–495.

[2] Quan Vu, Kris Demuynck, and Dirk Van Compernolle, "Vietnamese automatic speech recognition: The flavour approach," in Proc. *the 5th International Conference on Chinese Spoken Language Processing (ISCSLP),* 2006, pp. 464–474.

[3] Tuan Nguyen and Quan Vu, "Advances in acoustic modeling for Vietnamese LVCSR," in Proc. *International Conference on Asian Language Processing (IALP)*, 2009, pp. 280–284.

[4] Ngoc Thang Vu and Tanja Schultz, "Vietnamese large vocabulary continuous speech recognition," in Proc. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2009.

[5] Nancy F. Chen, Sunil Sivadas, Boon Pang Lim, Hoang Gia Ngo, Haihua Xu, Bin Ma, and Haizhou Li. "Strategies for Vietnamese keyword search," in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4121-4125.

[6] Tsakalidis, Stavros, Roger Hsiao, Damianos Karakos, Tim Ng, Shivesh Ranjan, Guruprasad Saikumar, Le Zhang, Long Nguyen, Richard Schwartz, and John Makhoul. "The 2013 BBN Vietnamese telephone speech keyword spotting system," in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7829-7833.

[7] I-Fan Chen, Nancy F. Chen, and Chin-Hui Lee, "A keyword-boosted sMBR criterion to enhance keyword search performance in deep neural network based acoustic modeling," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014.

[8] Quoc Bao Nguyen, Van Hai Do, Ba Quyen Dam, Minh Hung Le, "Development of a Vietnamese Speech Recognition System for Viettel Call Center," in Proc. *Oriental COCOSDA*, pp. 104-108, 2017.

[9] Haihua Xu, Hang Su, Eng Siong Chng, and Haizhou Li. "Semi-supervised training for bottle-neck feature based DNN-HMM hybrid systems," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014.

[10] Van Hai Do, Xiong Xiao, Eng Siong Chng, and Haizhou Li, "Context-dependent phone mapping for LVCSR of under-resourced languages," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2013, pp. 500–504.

[11] Van Hai Do, Xiong Xiao, Eng Siong Chng, and Haizhou Li, "Kernel Density-based Acoustic Model with Cross-lingual Bottleneck Features for Re-source Limited LVCSR," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014, pp. 6–10.

[12] Van Hai Do, Nancy F. Chen, Boon Pang Lim and Mark Hasegawa-Johnson, "Multi-task Learning using Mismatched Transcription for Under-resourced Speech Recognition," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 734-738, 2017

[13] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[14] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015, pp. 3214-3218.

[15] F. Grezl, M. Karafiat, S. Kontar, and J. Cernock, Probabilistic and bottleneck features for LVCSR of meetings," in Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007, vol. 4 pp. 757-760.

[16] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI", in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2751–2755, 2016.

[17] Quoc Bao Nguyen, Tat Thang Vu, and Chi Mai Luong, "The Effect of Tone Modeling in Vietnamese LVCSR System," *Procedia Computer Science* 81 (2016): 174-181.

[18] Xunying Liu, Yongqiang Wang, Xie Chen, Mark Gales, and P. C. Woodland, "Efficent lattice rescoring using recurrent neural network language models," in Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[19] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes Risk Decoding and System Combination Based on a Recursion for Edit Distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802 – 828, 2011.

[20] G. Evermann and P. C. Woodland, "Posterior Probability Decoding, Confidence Estimation and System Combination," in Proc. *Speech Transcription Workshop*, 2000.

[21] P. Bell, H. Yamamoto, P. Swietojanski, Y. Z. Wu, F. McInnes, C. Hori, and S. Renals, "A Lecture Transcription System Combining Neural Network Acoustic and Language Model," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2013