Development of a Vietnamese Large Vocabulary Continuous Speech Recognition System under Noisy Conditions

Quoc Bao Nguyen, Van Tuan Mai
Cyberspace Center
Viettel Group Vietnam
{baonq2, tuanmv2}@viettel.com.vn

Quang Trung Le, Ba Quyen Dam
Cyberspace Center
Viettel Group, Vietnam
{trunglq12, quyendb}@viettel.com.vn

Van Hai Do[1,2]
[1]Thuyloi University, Vietnam
[2]Viettel Group, Vietnam
haidv@tlu.edu.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, recurrent neural network language model 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.

CCS CONCEPTS

• Computing methodologies → Speech recognition

KEYWORDS

Vietnamese speech recognition, speech corpus, noisy condition, model adaptation, system combination.

ACM Reference format:

Quoc Bao Nguyen, Van Tuan Mai, Quang Trung Le, Ba Quyen Dam, and Van Hai Do. 2018. Development of a Vietnamese Large Vocabulary Continuous Speech Recognition System under Noisy Conditions. In *SoICT '18: Ninth International Symposium on Information and Communication Technology, December 6–7, 2018, Da Nang City, Viet Nam.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3287921.3287938

1 INTRODUCTION

Vietnamese is the sole official and the national language of Vietnam with more than 76 million native speakers. It is the first language of the majority of the Vietnamese population, as well as the 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 in clean conditions [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]. Recently, 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.

In this paper, we present our another effort to build 500-hour read speech corpus and the process to build a Vietnamese LVCSR speech recognition system. Different from the telephone corpus in [8], in this time, the corpus is recorded with better quality using closed microphones with 16 kHz sampling rate and 16 bit resolution. The aim is to build a speech recognition system for other commercial applications such as virtual assistant, smart home, etc. In these applications, users normally speak with distance microphones and hence noise can significantly reduce the speech quality. To make speech recognition more robust to noise, we introduce data augmentation by adding various types of noise into training data. In addition, other techniques such as language model rescoring, language model adaptation, system combination are also applied. 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 2 describes our speech corpus. Section 3 presents the proposed speech recognition system. Section 4 shows the experimental results and Section 5 concludes the paper.

2 CORPUS DESCRIPTION

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

Previously, several Vietnamese speech corpuses 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 the Viettel group, beside building speech recognition systems for telephone conversation such as for call center [8], we also target on building a

SoICT '18, December 6–7, 2018, Da Nang City, Viet Nam
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6539-0/18/12...$15.00
https://doi.org/10.1145/3287921.3287938

commercial system for other applications such as virtual assistance, smart home, etc.

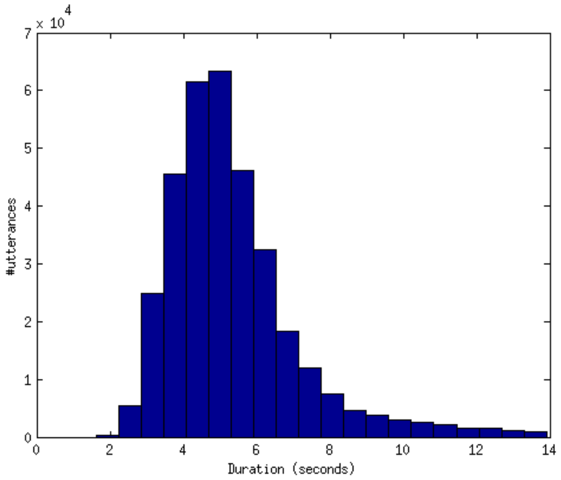

Figure 1: The distribution of utterance durations

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 and normalization, sentence segmentation is applied and text is then sent to speakers sentence by sentence for speaking and recording. All utterances are recorded in clean condition using high quality microphone attached in a computer. We create a website with friendly user interface 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.

Fig. 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.

3 THE PROPOSED SYSTEM

Our target is to build a speech recognition system which is robust to different testing environments. However the training data are recorded in clean conditions. To achieve to the goal, as shown in Fig. 2, 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 recurrent neural network (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.

3.1 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 training speech of the original speech corpus with different level to simulate noisy speech as in Eq (1).

$$\boldsymbol{x'_t = x_t + \alpha * n_t} \tag{1}$$

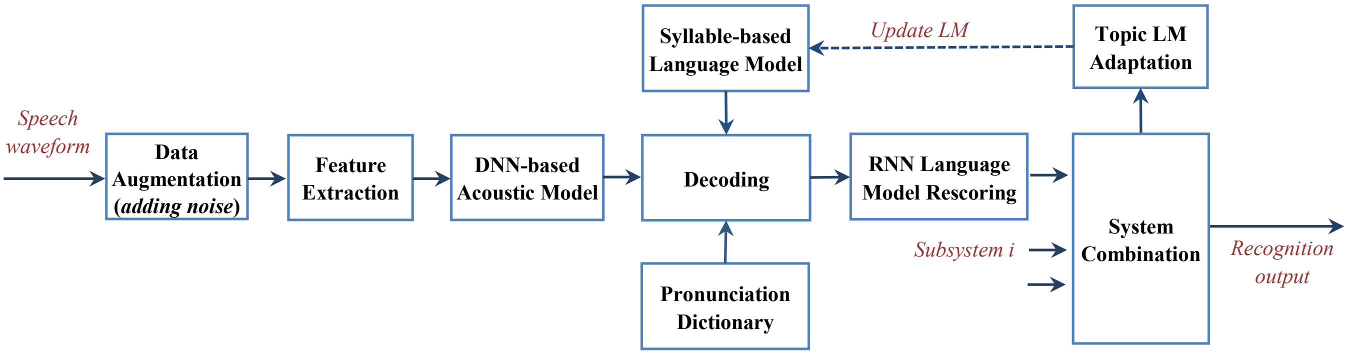

Figure 2: The proposed speech recognition system

where $\boldsymbol{x_t}$ is clean speech, $\boldsymbol{n_t}$ is noise, $\boldsymbol{\alpha}$ is scaling factor to adjust the noise level, $\boldsymbol{x'_t}$ is 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.

3.2 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 10 ms time step which is similar setup in [14]. Since Vietnamese is a tonal language, augment MFCC with pitch feature can significantly improve performance [8].

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.

3.3 Acoustic Model

Acoustic model is used to model the feature distribution among different phonemes. 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.

3.4 Pronunciation Dictionary (Lexicon)

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 words and abbreviations, we select 5k popular foreign words and abbreviations from web newspapers. These words are then manually pronounced in the Vietnamese pronunciation before converting them into Vietnamese phone sequences. As a result, the total number of words in our lexicon is about 11k words. This lexicon is used for training as well as for decoding. Note that a foreign word or an abbreviation can be pronounced in different ways by Vietnamese speakers. To make the speech recognition system can operated with different users, we try to collect most of popular pronunciations for each foreign word or abbreviation. For example: "*YouTube*" can be pronounced as "*diu túp*", "*du túp*", "*diu tu bi*", etc.

3.5 Language Model

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

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

3.6 System Combination

As described in Subsection 3.2, we have two subsystems i.e., the first subsystem uses MFCC feature while the second subsystem uses bottleneck feature. The combination of information from different 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].

3.7 Language Model Adaptation

The recognition output of our system has a relatively low word error rate (WER). Hence, from decoded text, we are able to 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 decoded by the system. 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. One disadvantage of this approach is that it is hard to apply to real-time decoding.

4 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 types with signal to noise ratio (SNR) from 15-40 dB. Note that the noise is different from the noise added in the training set. System performance is evaluated using word error rate (WER).

4.1 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 1, 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 with different signal to noise ratio (SNR). Obviously, this makes model more robust with noisy conditions and hence we achieve a low WER at 10.3%.

Table 1: Effect of data augmentation to system performance

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

4.2 RNNLM Rescoring

As shown in Table 2, by applying RNNLM rescoring technique, we can achieve 1.4% absolute improvement. It shows the effectiveness of using RNN language modelling.

Table 2: Effect of RNNLM rescoring to system performance

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

4.3 System Combination

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

As shown in Table 3, using BNF does not provide a good performance as MFCC. However, it provides complementary information and hence we can gain by combining them.

Table 3: Bottleneck feature and system combination

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

4.4 Language Model Adaptation

As shown in Table 4, 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 4: Effect of language model adaptation to system performance

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

5 CONCLUSION

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. 2005. Vietnamese large vocabulary continuous speech recognition. In *Proc. Annual Conference of the International Speech Communication Association* (INTERSPEECH), 492–495.

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

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

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

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

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

[7] I-Fan Chen, Nancy F. Chen, and Chin-Hui Lee. 2014. 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).

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

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

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

[11] Van Hai Do, Xiong Xiao, Eng Siong Chng, and Haizhou Li. 2014. 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), 6–10.

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

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

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

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

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

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

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

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

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

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