Development of Zalo Vietnamese Conversational Speech Recognition for VLSP 2019

Quoc Bao Nguyen[1,2], Ba Quyen Dam[1] and Van Hieu Nguyen[1]
[1]*Zalo Group - VNG Corporation*
[2]*Information and Communication Technology University - Thai Nguyen University*
{baonq6,quyendb,hiennv11}@vng.com.vn

***Abstract*—This report describes the system developed by the Zalo team for automatic speech recognition (ASR) of the VLSP 2019, focusing on conversational speech under noisy conditions. We first present our effort to transcript a 4000-hour corpus for Vietnamese conversational speech. After that the training data was augmented with both background noises and simulated reverberation. We then trained time delay neural network (TDNN) and long-short term memory (LSTM) acoustic models and submitted two systems: (1) a system, which did not apply language model adaptation, achieved a word error rate (WER) of 16.86% on the test2019 set, and (2) a system using language model adaptation achieved a WER of 15.99%.**

***Index Terms*—automatic speech recognition, TDNN-LSTM**

I. Introduction

The International Workshop on Vietnamese Language and Speech Processing (VLSP) is an annual workshop which has the goal of perpetuating the research on Vietnamese language and speech processing, and bringing together researchers and professionals working in this domain. The VLSP workshops are composed of two parts: one part focuses on VLSP software demonstration with technical reports, the other dedicated to evaluations on different tasks of text and speech processing. The speech processing evaluation offers specific tracks for two core technologies namely automatic speech recognition (ASR), and text-to-speech (TTS).

We participated in the ASR track, which was organized to evaluate Vietnamese ASR systems. This year evaluation provided test2018 set and the test2019 set is the transcription of audio coming from segmented youtube audio. The quality of the ASR systems were measured in word error rate (WER).

The organization of the paper is as follows. Section II describes our effort to transcribe a 4000-hour training data. This is followed by Section III, which provides a description of our ASR system. Experiments and results are presented in Section IV.

II. Corpus Description

In this report, we present our effort to collect a 4000-hour conversational speech and 1000 hours of broad cast news speech corpus, which were used to train our speech recognition system. At Zalo, our goal is to build systems for Zalo ecosystem applications such as virtual assistance namely Kiki, Zalo voice message transcription, ...

To achieve the goal, we collected 4000-hour conversational speech with manual transcript and 1000 hours of broad cast news speech with automatic transcription using our best ASR system. We first selected conversational youtube video such as game show, talk show, and then downloaded, and automatically segmented the audio. After that, the segmented audio files were automatically transcribed using our best ASR system. Finally, both of the audio files and their corresponding transcriptions were manually reviewed.

III. System Description

Figure 1 shows our training and decoding scheme of the ASR system. Training data are first augmented with both background noises and simulated reverberation. Feature extraction is then applied to prepare necessary ingredients for the acoustic model training. For decoding, acoustic model is used together with language model and lexicon. After decoding, 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. The detailed description of each module is presented in next subsections.

A. Data Augmentation

Data augmentation is a common strategy adopted to increase the quantity of training data. It is a key point of the state-of-the art techniques for image recognition and speech recognition [1]. In this work, we enriched our training data with a number of noise data sets:

- Google noise: a large-scale collection of human-labeled 10-second sound clips drawn from YouTube videos [2].
- Music: the music files from zing mp3 that were spitted to 10 second segments.
- Zalo noise: the noise files that we collect some popular noise types such as office noise, street noise, car noise, etc.
- Reverb: simulated RIRs as described in [3]

We randomly applied additive noises from the google noise, music and zalo noise datasets separately on training data after doing 3-way speed perturbing. Then reverberation was applied on top of them using the simulated RIRs with room sizes uniformly sampled from 1 meter to 30 meters. The above procedure led to 6 times more augmented training data. The alignments and lattices for these augmented data were obtained from their clean parts.

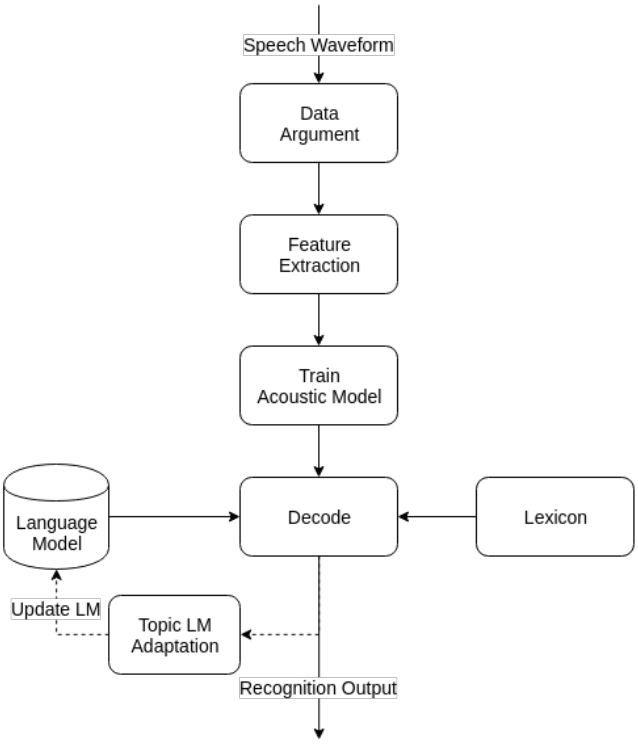

Fig. 1. The Vietnamese conversational ASR systems of Zalo for VLSP2019.

B. Feature Extraction

We used Mel-frequency cepstral coefficients (MFCCs) [4], without truncation, as input feature. A 40^{th}-order MFCCs were computed at each 10 ms time step which is similar setup in [5]. Since Vietnamese is a tonal language, augmented both MFCCs and pitch can significantly improve the recognition performance [6].

C. Acoustic Model

Acoustic model was used to model the feature distribution among different phonemes. We used time delay neural network (TDNN) and long-short term memory (LSTM) as an acoustic model. The model was trained with lattice-free maximum mutual information (LF-MMI) criterion denoted in Equation (1):

$$F_{LF-MMI} = \sum_{n=1}^{N} log \frac{P(O_n|L_n)^k P(L_n)}{\sum_L P(O_n|L)^k P(L)} \quad (1)$$

where L_n is the phone sequence of the n-th utterance and $P(L)$ is the phone language model estimated by HMM-GMM phone alignments.

D. Language Model

The based language model was built from our 189MB Youtube transcription. A 4-gram language model with Kneser-Ney smoothing [7] was used after exploring different configurations. Based on the language model, we did the the language model adaptation by using the recognition output decoded by the system. After that sentences from the our general text corpus (32GB of text form baomoi.com website, 10GB of text from forums,comments and book) were selected based on a cross-entropy difference metric. Detailed description about this selection algorithm can be referred in [8]. Finally, about 350MB text, which have the most relevant to the recognition output, were selected to build the adapted language model. The decoding process was then repeated with the new language model.

E. Lexicon

Vietnamese is a monosyllabic tonal language. Each Vietnamese syllable can be considered as a combination of initial, final and tone components. Therefore, lexicon needs to be modelled with tones [9]. We used 48 basic phonemes. The tonal marks were concatenated to the last phoneme of a syllable. This resulted in a lexicon of 7000 Vietnamese syllables.

To build the lexicon for foreign words and abbreviations, we selected 10k popular foreign words and abbreviations from web newspapers. These words were then manually pronounced in the Vietnamese pronunciation before converting them into Vietnamese phone sequences. As a result, our lexicon has a total number of 17000 words.

IV. Experiments

During development, we evaluated our system on a test set, which was selected from our 5000 hour corpus. The test set contained 17000 utterances with around 20 hours of audio (zalo-dev20) and test2018, test2019 test set that released by the VLSP organizers. Table I lists the performace of our systems in terms of the word error rate (WER).

TABLE I
Experiment results

System	WER(%)		
	zalo-dev20	***test2018***	***test2019***
TDNN-LSTM	5.46	4.18	16.89
TDNN_LSTM+LMadapt	-	4.09	15.99

V. Conclusion and Future works

In this paper, we presented our Vietnamese conversational ASR systems for the 2019 VLSP evaluation. We have described our 5000-hour speech corpus. The key solutions of our systems included data augmentation, language model adaptation, TDNN-LSTM acoustic model. The final result achieved a WER of 15.99% on the 2019 evaluation set and a WER of 4.09% on the 2018 evaluation set.

In the future, we want to improve language model using RNN language models for rescoring as well as applying more state-of-the art techniques to improve the acoustic model of our systems.

REFERENCES

[1] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, Audio augmentation for speech recognition, in Interspeech 2015, 2015, pp.3586-3589.

[2] J. F. Gemmeke et al., "Audio Set: An ontology and human-labeled dataset for audio events," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 776–780. doi: 10.1109/ICASSP.2017.7952261

[3] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 5220-5224.

[4] 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.

[5] 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.

[6] 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.

[7] Frankie James. 2000. Modified Kneser-Ney Smoothing of N-Gram Models. Technical Report. RIACS.

[8] 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).

[9] 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.