A Report on the Speech-to-Text Shared Task in VLSP Campaign 2019

Chi Mai Luong
University of Science and Technology of Hanoi
mai.luongchi@gmail.com

Quoc Truong Do
truongdq54@gmail.com

***Abstract*—Automatic Speech Recognition (ASR) is an active research area and studies on it have achieved many impressive results. In English, ASR systems have surpassed human capability on certain situations, and in every year, ASR challenges are organized for companies and universities to participate and comparing system's performance. To catch up with the trend, in 2018, VLSP has organized the very first Vietnamese ASR challenge that has attracted many submissions, and also demonstrated that studies on Vietnamese ASR has also reached to the commercialize level. This year, VLSP continues to organize the 2019 ASR challenges to continue the search for new ideas for Vietnamese ASR development.**

***Index Terms*—conversational speech, language model, combine, asr, speech recognition**

I. Challenge Task Definition

In the ASR task, participants were asked to transcribe automatically Vietnamese audio files into the spoken word sequences. The committee provided the test set only, while the training data for the acoustic and language models was developed by the teams themselves. The test set was delivered on October 5th, 2019, and then each team had 1.5 days to transcribe it. The final result had to submit by 23:59-PM on October 6th, 2019. The Word Error Rate (WER) was measured with references which were human transcripts of the audio files.

II. Participants

There are 30 teams registered and all of them have obtained the data set. However, only 3 teams finally made a submission of their systems, including Zalo, Viettel, and VAIS. Among 3 teams, Viettel and VAIS were also participated in VLSP 2018.

A. Dataset

1) Training data: Collecting speech training data for ASR is time-consuming and costly process. To help participants to start easier, the VLSP organizer has called for data donation and got approximately 500 hours from the Infore company. The data is then distributed to participants under a research purpose agreement.

2) Evaluation data: There are two evaluation data sets were constructed. The first one is the VLSP 2018 data set, where the organizer has refine the data to standardize the reference text better. The second data set is the 2019, which is much larger in size (4 times comparing to the 2018 one). Details of the 2 corpora is listed in the table below.

TABLE I
Evaluation data details

Name	Num. sentences	Hours	Domain
VLSP 2018	756	2	News
VLSP 2019	16K	17	News + Conversation

The VLSP 2018 includes 50%, 40%, and 10% proportion of Northern, Southern, and Central accents; and the dialect distribution of the 2019 are unknown.

B. System description

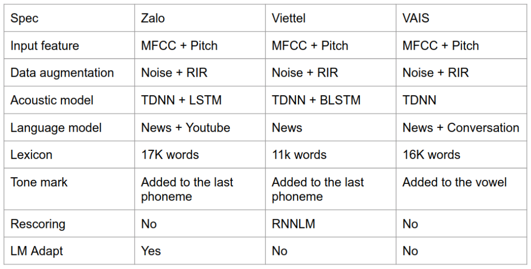
Table 1 shows the system description of all participants. In general, all teams have similar frontend processing which utilizze MFCC and pitch features. This is reasonable since Vietnamese is a tonal language and pitch plays an important role to model tone.

With regards to acoustic modeling, VAIS utilizes TDNN model architecture, while Zalo add another LSTM layer on top to form the TDNN + LSTM model. Viettel takes a more complicate system with BiLSTM to capture both past and future context. The lexicon is similar between Zalo and Viettel teams where they both add tone marks at the final phoneme, and VAIS add it at the vowel. All participants also add foreign words and abbreviations to the lexicon to increase the size and to handle foreign words better.

Regarding language modeling, both Zalo and VAIS utilize news and conversation data while Viettel only use news corpus. All teams train n-gram language models and the Zalo team adopt a language adaptation technique to further improve the system performance.

III. Evaluation Procedure

The organizer provides a scoring server allowing each team to submit their results multiple time. The score, which is word error rate, is returned immediately so that teams can track the performance and adjust the system accordingly.



Spec	Zalo	Viettel	VAIS
Input feature	MFCC + Pitch	MFCC + Pitch	MFCC + Pitch
Data augmentation	Noise + RIR	Noise + RIR	Noise + RIR
Acoustic model	TDNN + LSTM	TDNN + BLSTM	TDNN
Language model	News + Youtube	News	News + Conversation
Lexicon	17K words	11k words	16K words
Tone mark	Added to the last phoneme	Added to the last phoneme	Added to the vowel
Rescoring	No	RNNLM	No
LM Adapt	Yes	No	No

Fig. 1. System specification of all participants

The performance of an ASR system is measured by the amount of mistakes that the system makes, including:

- Deletion: the system cannot recognize words that are spoken in the input audio.
- Substitution: the system miss-recognizes the word to another one.
- Insertion: the system recognizes words that are not spoken in the input audio.

IV. Evaluation results

The results on the VLSP 2018 and 2019 are shown on the Fig.2 and Fig.3, respectively. As we can see, on the 2018 data set, performance of all teams are all below 5% WER. Zalo team has the best performance with the WER of 4.09. Following with Viettel team with WER of 4.4 and finally, VAIS team with 4.85%. Comparing with last year result, Viettel has made a significant improvement by reducing WER by 3% absolute.

On the 2019 data set, we observed significant changes in WER. This is due to the 2019 data set is meant to test the ability of systems in conversation domain, and it is much more difficult comparing to the 2018 data set. VAIS team has achieved the best performance with 15.09% WER, follow up with the Zalo team with 15.99% and Viettel team of 30.71.

There is a big gap between Viettel team and the other twos in the 2019 test set. This result is likely due to the language model that the team Viettel use is highly optimize for the news domain and not being able to handle conversation very well.

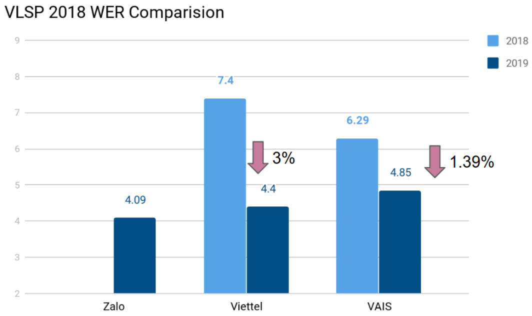


Fig. 2. VLSP 2018 data set performance

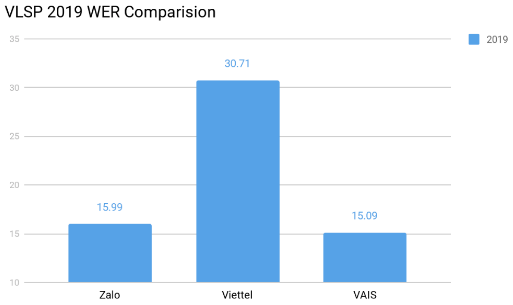


Fig. 3. VLSP 2019 data set performance

To calculate the final score, we sum up the insertion, deletion, and substitution in both test set and derive the WER from that. The result is showed in Table. II.

TABLE II
Average performance

Name	WER
Zalo	14.36
Viettel	27.11
VAIS	**13.7**

V. Conclusion

In this paper, we presented the VLSP 2019 ASR challenges. There are 3 teams submitted their results and demonstrated a significant improvement comparing to the last year challenge. Zalo achived the best performance on the VLSP 2018 while VAIS outperform on the 2019 set. In average, VAIS has archived the best score.