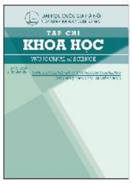
VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce



Original Article

VLSP 2021 - ASR Challenge: Vietnamese Automatic Speech Recognition

Do Van Hai

Thuyloi University, 175 Tay Son, Dong Da, Hanoi, Vietnam

Received 27 December 2021
Revised 18 April 2022; Accepted 5 May 2022

Abstract: Recently, Vietnamese speech recognition has been attracted by various research groups in both academics and industry. This paper presents a Vietnamese automatic speech recognition challenge for the eighth annual workshop on Vietnamese Language and Speech Processing (VLSP 2021). There are two sub-tasks in the challenge. The first task is ASR-Task1 focusing on a full pipeline development of the ASR model from scratch with both labeled and unlabeled training data provided by the organizer. The second task is ASR-Task2 focusing on spontaneous speech in different real scenarios e.g., meeting conversation, lecture speech. In the ASR-Task2, participants can use all available data sources to develop their models without any limitations. The quality of the models is evaluated by the Syllable Error Rate (SyER) metric.

Keywords: VLSP 2021, automatic speech recognition challenge, unlabeled data, semi-supervised training.

1. Introduction

Vietnamese is the official language of Vietnam with more than 76 million native speakers. It is the first language of the majority of the Vietnamese population. Several attempts have been conducted to build Vietnamese automatic speech recognition (ASR) system [1, 2, 3]. In 2013, the National Institute of Standards and Technology, USA released the Open Keyword Search Challenge (Open KWS) for Vietnamese speech. Many approaches have been proposed to improve performance for both keyword search and speech recognition [4-6].

Recently, The International Workshop on Vietnamese Language and Speech Processing (VLSP) has annually organized ASR challenge for Vietnamese. The VLSP Consortium1 regroups all academic and industrial research teams involved in Vietnamese language and speech processing. The first kick-off meeting of this community was in 2005 at the Institute of Information Technology, Vietnam Academy of Science and Technology. The first ASR challenge was organized in VLSP 2018. In this challenge, no training dataset was released by the organizer. Participants used public or their

[*] Corresponding author.
E-mail address: haidv@tlu.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.356

own datasets to develop the models. Only 3 submissions were received. In VLSP 2019-ASR, a 500-hour-dataset was released by the organizer. However, participants could use any additional data to develop the models. In VLSP 2020-ASR, a 250-hour-dataset was released to participating teams to train the models. It was the first time the challenge was divided into two tasks. In Task1, participants had to only use training data provided by the organizer. In Task2, participants could use any resources to train their models. Finally, there were 10 submissions for Task1 and 4 submissions for Task2.

Note that, in all the previous ASR challenges, the data provided by the organizer were with manual transcription and not domain-specific. In the VLSP ASR 2021 challenge, we conducted more challenging and realistic tasks by focusing a specific domain i.e., online lectures. In addition, both labeled and unlabeled data were provided to participating teams. Specifically, the ASR challenge composed of two sub-tasks:

ASR-Task1 focuses on a full pipeline development of the ASR model from scratch. The organizer provided two training datasets. The first dataset is around 241.1 hours of transcribed data. Each participant had to label a part of the dataset before receiving the whole datasets. The second dataset is around 360.7 hours of untranscribed in-domain data. All participants were required to use only this provided data to develop models including acoustic and language models. Any use of another resource for model development was not acceptable.

ASR-Task2 focuses on spontaneous speech in different real scenarios e.g., meeting conversation, lecture speech. For this task, the organization did not provide training data, participants could use all available data sources to develop their models without any limitation. The ASR challenge attracted 47 registrations and 18 final result submissions. Many interesting approaches with remarkable results have been proposed by the participants. This paper presents the challenge description from data preparation to final result submission of participating teams. Moreover, different approaches for Vietnamese ASR will be described in details.

The rest of this paper is organized as follows. Section 2 provides information about participants and the processes in the challenge. Section 3 discusses the process of data preparation. Evaluation is described in Section 4. Finally, Section 5 concludes the paper.

2. Participants

For the ASR challenge in VLSP 2021, each team needed to transcribe the audio data. Specifically, after registration, each team received several hours of audio clips recorded from online lectures. They needed to segment a long audio clip into small segments after that manually transcribed those segments. Finally, each team submitted the labeled data in order to receive the whole dataset and participate in the competition. In the competition phase, each team received a training and a development datasets. Participants had 45 days to develop and fine-tune their models. After that the private test sets was released and participating teams had 7 hours to submit their final submissions for the two private tests ASR-Task1 and ASR-Task2.

In summary, 47 teams registered for the ASR challenge. However, there were only 15 teams finish data labeling task. Those 15 teams received the training and development datasets. Finally, there were 9 submissions for Task1 and 9 submissions for Task2 where 8 teams submitted both Task1 and Task2, one team submitted only Task1 and one team submitted only Task2.

After that, final scoreboards and top teams for two tasks were announced; the Top-3 of each task needed to write and submit technical reports. The scoring metric, final scoreboards, and solutions are discussed in Section 4.

3. Data Preparation

In this section, we discuss the process of building the Vietnamese speech recognition datasets for VLSP 2021.

3.1. Dataset description

The ASR-Task1 focuses on recognizing online lectures for different subjects. With the help from student volunteers in Thuyloi University, we collected raw audio data from Youtube. For each Youtube channel, we just collected few clips with the total of duration less than 2 hours to make dataset balance between different speakers. The raw audio data was then divided into 2 parts, one for transcribing, and second one for using as unlabeled training data. The first part was send to participating teams for labeling. The labeling procedure is as follows:

- Any software can be used to label text for speech segments (e.g., Audacity).
- The transcriber selects the speech segments (about a few seconds) and assigns a transcription to the speech segment.
- Only assign labels when transcriber can clearly and surely hear all the words in the segment.
- Write Vietnamese transcription in lowercase, correct spelling with no special characters or punctuation and exactly with the content of the audio.
- For special characters such as @, #, %, etc, need to be written according to the spoken language.
- Need to write the number in words for example: “68” to enter to “six eight” or “sixty eight” depending on the content of audio.
- Foreign words, abbreviations are written according to the original name.

After the labeling data process, 25.5 hours were transcribed by 15 teams. This data was then divided into two subset for training, one for development. Finally, four dataset were released for the challenge.

- 215.6 hours of general domain labeled training set which was inherited from previous challenge.
- 23 hours of in-domain labeled training set which were contributed by participating teams. “In-domain” means that this data is online lectures.
- 2.5 hours of in-domain labeled development set contributed by participating teams.
- 360.7 hours of in-domain unlabeled training set.

4. Evaluation

4.1. Evaluation Metric

In this challenge, we use Syllable Error Rate (SyER) instead of Word Error Rate (WER) to evaluate performance of speech recognition systems. The reason is that in the Vietnamese writing system, spaces are used to separate syllables instead of words. A word can consist of from one to four syllables, and the task to find the boundary between words is not trivial [7]. Syllable Error Rate (SyER) metric is computed as follows.

$$\text{SyER} = \frac{S+D+I}{N} \qquad (1)$$

where

- S is the number of substitutions.
- D is the number of deletions.
- I is the number of insertions.
- C is the number of syllables.
- N is the number of syllables in the reference.

4.2. Results

4.2.1. ASR-Task1

Table 1 shows the SyER of all teams for ASR-Task1. The Lightning, LAB-914-ASR, SMARTCALL, and VC-Tus teams achieved the first four places with the SyER of 8.28%, 11.08%, 12.00%, and 12.41%, respectively. The remaining teams got SyER varying from 16.68% to 35.91% which are significantly worse than the Top-4 teams.

Figure 1 illustrates the SyER of submitting teams for different audio clips in the ASR-Task1 test set. It can be seen that the Lightning team achieved the lowest SyER over almost all clips in the test set. The last two clips were recorded in highly noisy and reverberant environments, specifically in a big lecture theater with a long distance from speakers to the microphone. This resulted in high SyERs for all participation teams at those two clips.

Table 2 summarizes the approaches of the Top-3 teams i.e., Lightning, LAB-914-ASR, and SMARTCALL. All three teams used data augmentation techniques to make their models more robust. SpecAugment [8], a simple and effective data augmentation technique was used for the Lighting and LAB-914-ASR teams. In addition, the LAB-914-ASR team also used speed perturbation [9] to change the speed of the audio signal. The SMARTCALL team used a traditional method [10] by adding noise and reverbration into the training data.

For speech feature for ASR systems, both the Lightning and LAB-941-ASR teams used high resolution fbank feature with 80 dimensions. The SMARTCALL team used a lower resolution fbank with 40 dimensions however it was augmented with pitch feature. Using pitch has been demonstrated to achieve better ASR performance for tonal languages like Vietnamese [11].

Table 1. SyER given by different teams for Task1:

Rank	Team	Organization	SyER
1	Lightning	Viettel Cyberspace Center	8.28%
2	LAB-914-ASR	Hanoi University of Science and Technology	11.08%
3	SMARTCALL	SMARTCALL.,JSC	12.00%
4	VC-Tus	VCCorp Corporation	12.41%
5	VB_ASR	VinBrain.,JSC	16.68%
6	D2_Speech	Hanoi University of Science and Technology	21.01%
7	DAL	VNG Corporation	21.29%
8	CHC-79	Hanoi University of Science and Technology	22.08%
9	eve	VietAI, ProtonX	35.91%

It is different from the previous VLSP ASR challenges, this year participating teams were supplied a substantial amount of unlabeled data i.e., 360.7 hours. However, only the Lightning and LAB-914-ASR teams leveraged this unlabeled data to improve their models. The Lightning team used Gradient Mask [12] to update model with pseudo-label. LAB-914-ASR used unlabeled training data to make a pre-train model before fine-tuning it with labeled data [13]. About the acoustic model - a crucial of every speech recognition system, the Top-3 teams used different approaches. The Lightning team used Conformer architecture [14] which is an effective combination of convolutional neural networks and Transformer architecture. LAB-914-ASR used Transformer [15] as their acoustic model while SMARTCALL used the traditional hybrid HMM/TDNN +LSTM model [16].

Language model is also an important module in ASR. The Lightning team used aninternal language model which is implicitly incorporated into the Conformer architecture. LAB-914-ASR and SMARTCALL used simple 6-gram and 4-gram language models, respectively. However, after the 1st-pass decoding, SMARTCALL used RNN language model to re-score to improve their result. Note that all three teams only used training transcription provided by the organizer to train their language models.

Lexicon or pronunciation dictionary is a bridge between acoustic and language model in the hybrid ASR architecture. A comprehensive lexicon with 19k entries was used by SMARTCALL. However, the both Lightning and LAB-914-ASR followed the end-to-end fashion [17], hence characters were modeled directly without using a lexicon.

The ASR-Task1 dataset focuses on the online lectures for different subjects. Hence, there are many abbreviation and loan words in the transcription. Before the private test set was released, all participating teams were provided a list of abbreviation and loan words in the test set together with their possible pronunciations. The Lightning team model handled this issue naturally by their effective tokenizer. Both the LAB-914-ASR and SMARTCALL teams used text normalization technique to converte abbreviation and loan words from spoken representation into writing form.

In the Top-3 teams, only LAB-914-ASR used ensemble technique to improve their performance by combining three different models. While both the Ligtning and SMARTCALL teams just used a single system.

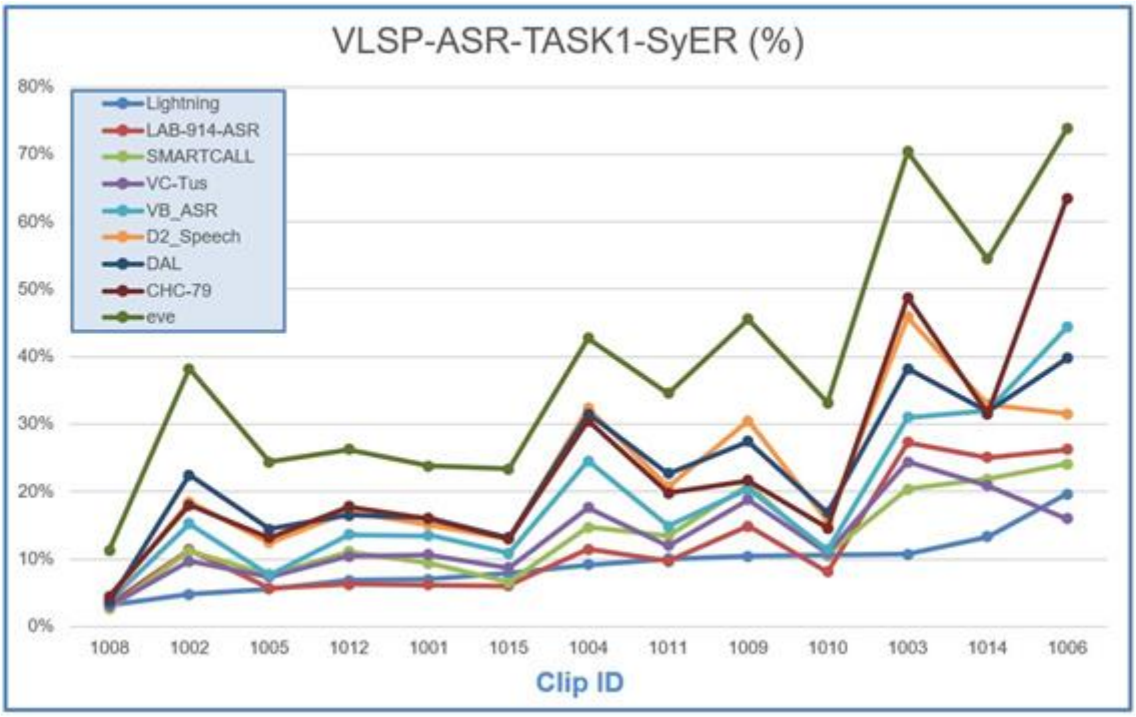


Figure 1. SyER given by submitting teams for different audio clips in the ASR-Task1 test set.

4.2.2. ASR-Task2

In Task2, participating teams could use any resources to build their ASR systems without limitation. The SyER of all teams participating the ASR-Task2 is represented

in Table 3. The Lightning, Rikkei-ASR, and VC-Tus teams achieved the first three places with the SyER of 4.17%, 6.72%, and 8.83%, respectively. While the remaining teams got significantly higher SyER, varying from 9.88% to 28.60%. Figure 2 illustrates the SyER of participating teams for different audio clips in the ASR-Task2 test set. It can be seen that the Lightning team achieved the lowest SyER over all clips in the test set.

For the first two clips, most of the teams achieved very low error rate, even some results closed to 0%. In contrast, the last two clips were recorded in very low signal-to-noise ratio environments and speaking rate was relatively fast with many technical words. It is also observed that at last two clips, the gap between the Top-3 and the remaining teams is significantly bigger. It demonstrates that the Top-3 models are very robust in obstacle conditions.

Table 2. Techniques used for different modules in the Top-3 teams (Task1):

Module	Lightning[18]	LAB-914-ASR[19]	SMARTCALL[20]
Data augmentation	SpecAugment	SpecAugment +	Adding noise +
		speed perturbation	reverberation
Feature	80fbank	80fbank	40fbank+pitch
Unlabeled data usage	Gradient Mask	Pretraining+self-training	-
Acoustic Model	Conformer	Transformer	HMM/TDNN+LSTM
		(wav2vec 2.0)	
Language Model	Internal	6-gram	4-gram + RNN
Lexicon	-	-	19k words
Abbreviation & Loan			
words processing	Direct modeling	Text normalization	Text normalization
Ensemble	No	Yes	No

Table 3. SyER given different teams for Task2:

Rank	Team	Organization	SyER
1	Lightning	Viettel Cyberspace Center	4.17%
2	Rikkei-ASR	RIKKEI.AI	6.72%
3	VC-Tus	VCCorp Corporation	8.83%
4	LAB-914-ASR	Hanoi University of Science and Technology	9.88%
5	VB_ASR	VinBrain.,JSC	13.19%
6	D2_Speech	Hanoi University of Science and Technology	14.09%
7	CHC-79	Hanoi University of Science and Technology	18.05%
8	DAL	VNG Corporation	18.99%
9	eve	VietAI, ProtonX	28.60%

Now we discuss about the techniques which the Top-3 teams used for Task2. Table 4 summarizes the approaches of the Top-3 teams i.e., Lightning, Rikkei-ASR, and VC-Tus. It is different from Task1, in Task2, the participating teams could use their own training data to build their systems.

The Lightning, Rikkei-ASR, and VC-Tus teams used 2000, 3000, and 400 hours of audio training data to train their acoustic models, respectively. Both the Lightning and Rikkei-ASR teams just used training transcription to train the language model without using additional text resource. Data augmentation is a technique to help model more robust under different environments. In this task, all Top-3 teams used SpecAugment [8] to augment the training data. In addition, VC-Tus augmented data by changing speed, volume, pitch of the input audio. For speech feature, all three teams used high resolution fbank feature with 80 dimensions. About the acoustic model, the Lightning team used Conformer architecture [14] while Transformer architecture [15] was used by Rikkei-ASR and VC-Tus. For language modeling, the Lightning team used an internal language model which is implicitly incorporated into the Conformer architecture while the rest two teams used simple 4-gram and 5-gram language models. All three teams built their ASR models based on end-to-end architecture [17] and hence no lexicon was needed. Similar to Task1, in Task2, there were many abbreviation and loan words in the private test set. This issue was handled directly in the Lightning team by their effective tokenizer. VC-Tus used text normalization technique to converted abbreviation and loan words from spoken representation to writing form. Also note that all Top-3 teams just used their single system to produce the result.

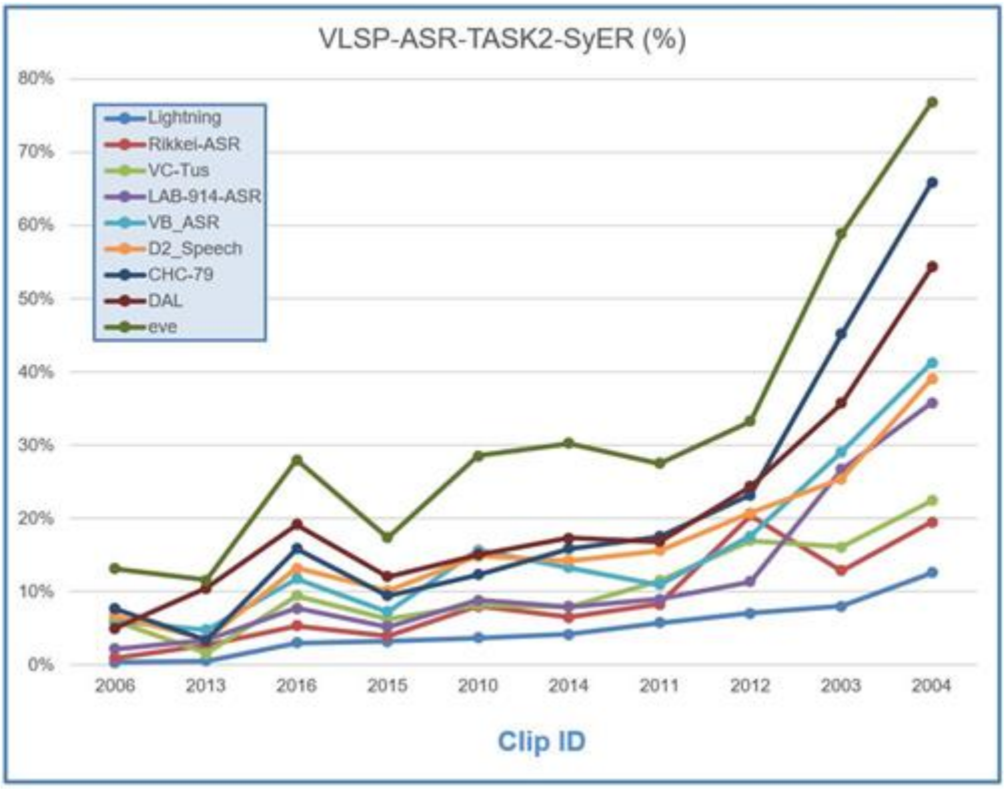


Figure 2. SyER of submitting teams for different audio clips in the ASR-Task2 test set.

5. Conclusion

In this paper, we summarized the organization of the ASR challenge at VLSP 2021. The challenge included two sub-tasks i.e., ASR-Task1 and ASR-Task2 with 18 submissions in total.

Task1 focused on a full pipeline development of the ASR model for online lectures from scratch. In this task, three datasets were released for training i.e., transcribed general-domain, transcribed in-domain, and untranscribed in-domain datasets. There were 9 submissions for this task with the best result of 8.28% SyER. Task2 focused on spontaneous speech in different real scenarios e.g., meeting conversation, lecture speech, etc. Participants could use all available data sources to develop their models without any limitation. Finally, 9 teams submitted their results with the best SyER of 4.17%. In conclusion, the ASR challenge in

Table 4. Techniques used for different modules in the Top-3 teams (Task2):

Module	Lightning[21]	Rikkei-ASR[22]	VC-Tus[23]
Training data (audio)	2000 hours	3000 hours	400 hours
Training data (text)	Training Transcription	Training Transcription	Not mentioned
Data augmentation	SpecAugment	SpecAugment	SpecAugment+changing
			speed, volume, pitch
Feature	80fbank	80fbank	80fbank
Acoustic Model	Conformer	Transformer	Transformer
Language Model	Internal	4-gram	5-gram
Lexicon	-	-	-
Abbreviation & Loan			
words processing	Direct modeling	Not mentioned	Text normalization
Ensemble	No	No	No

VLSP 2021 has attracted a lot of attention from the speech community with 47 registrations and 18 final submissions. Various interesting approaches had been conducted to handle both labeled and unlabled data such as self-supervised learning, gradient mask, etc. Through the challenge, we also can see the domination of end-to-end architectures over the traditional hybrid HMM/DNN approach. The best teams achieved very low SyERs which are comparable to the state-of-the-art ASR performances for other languages. In the next VLSP ASR challenges, we can focus on building models with limited training data which can be applied to various minority languages in Vietnam.

References

[1] T. V. Tat, D. T. Nguyen, M. C. Luong, J.-P. Hosom, Vietnamese large vocabulary continuous speech recognition, in: Proceeding of Eurospeech Conference, 2005, pp. 1172–1175.

[2] Q. Vu, K. Demuynck, D. V. Compernolle, Vietnamese automatic speech recognition: The flavor approach, in: International Symposium on Chinese Spoken Language Processing, Springer, 2006, pp. 464–474.

[3] N. T. Vu, T. Schultz, Vietnamese large vocabulary continuous speech recognition, in: 2009 IEEE Workshop on Automatic Speech Recognition & Understanding, IEEE, 2009, pp. 333–338.

[4] I.-F. Chen, N. F. Chen, C.-H. Lee, A keyword-boosted smbr criterion to enhance keyword search performance in deep neural network based acoustic modeling, in: Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[5] S. Tsakalidis, R. Hsiao, D. Karakos, T. Ng, Ranjan, G. Saikumar, L. Zhang, L. Nguyen, Schwartz, J. Makhoul, The 2013 bbn vietnamese telephone speech keyword spotting system, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 7829–7833.

[6] N. F. Chen, S. Sivadas, B. P. Lim, H. G. Ngo, Xu, B. Ma, H. Li, et al., Strategies for vietnamese keyword search, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 4121–4125.

[7] D. Dien, H. Kiem, N. Van Toan, Vietnamese word segmentation., in: NLPRS, Vol. 1, 2001, pp. 749–756.

[8] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, Specaugment: A Simple Data Augmentation Method For Automatic Speech Recognition, arXiv preprint arXiv:1904.08779.

[9] T. Ko, V. Peddinti, D. Povey, S. Khudanpur, Audio Augmentation For Speech Recognition, In: Sixteenth Annual Conference Of The International Speech Communication

Association, 2015.

[10] D. Snyder, G. Chen, D. Povey, Musan: A Music, Speech, And Noise Corpus, arXiv preprint arXiv:1510.08484.

[11] Q. B. Nguyen, V. T. Mai, Q. T. Le, B. Q. Dam, H. 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.

[12] S. Ling, C. Shen, M. Cai, Z. Ma, Improving pseudo-label training for end-to-end speech recognition using gradient mask, arXiv preprint arXiv:2110.04056.

[13] Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in Neural Information Processing Systems,Vol. 33 , 2020, pp. 12449–12460.

[14] Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, Conformer: Convolution-augmented transformer for speech recognition, arXiv preprint arXiv:2005.08100.

[15] L. Dong, S. Xu, B. Xu, Speech-Transformer: A No-Recurrence Sequence-To-Sequence Model for Speech Recognition, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5884–5888.

[16] V. Peddinti, Y. Wang, D. Povey, S. Khudanpur, Low Latency Acoustic Modeling Using Temporal Convolution And Lstms, IEEE Signal Processing Letters, Vol. 25, No. 3, 2017, pp. 373–377.

[17] D. Wang, X. Wang, S. Lv, An Overview Of End-To-End Automatic Speech Recognition, Symmetry, Vol. 11, No. 8, 2019, 1018.

[18] D. S. Dang, D. L. Le, X. V. Dang, Q. T. Duong, T. Ta, ASR: Conformer with Gradient Mask and Stochastic Weight Averaging for Vietnamese Automatic Speech Recognition., VLSP 2021.

[19] V. T. Pham, D. C. Le, T. T. T. Nguyen, ASR: Semi-supervised ensemble model for Vietnamese Speech Recognition at VLSP 2021 Automatic Speech Recognition Shared Task., VLSP 2021 .

[20] V. T. Mai, B. Q. Dam, Q. B. Nguyen, ASR: The Smartcall's ASR Systems for VLSP 2021, VLSP 2021 -

[21] D. S. Dang, D. L. Le, X. V. Dang, Q. T. Duong, T. Ta, ASR: Automatic Speech Recognition with blank label deweighting on open dataset., VLSP 2021.

[22] T. T. Truong, ASR: An Effective Transformer-based Approach to VLSP 2021 ASR Task., VLSP 2021.

[23] A. D. Trinh, V. S. Dang, V. T. Do, V. V. Ngo, ASR: Vietnamese Automatic Speech Recognition with Transformer, VLSP 2021.