VNU Journal of Science:
Computer Science and Communication Engineering

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

Original Article

# ASR - VLSP 2021: Conformer with Gradient Mask and Stochastic Weight Averaging for Vietnamese Automatic Speech Recognition

Dang Dinh Son*, Le Dang Linh, Dang Xuan Vuong,
Duong Quang Tien, Ta Bao Thang

*Viettel Cyberspace Center, Ton That Thuyet, Cau Giay, Hanoi, Vietnam*

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

**Abstract:** Recent years have witnessed the strong growth of Automatic Speech Recognition (ASR) studies due to its wide range of applications. However, there are few efforts put into the Vietnamese language. This paper introduces an end-to-end approach using Conformer, a combination of Transfomer and Convolution Neural Network, and pseudo labeling for Vietnamese ASR systems. Besides, our approach is equipped with Gradient Mask and Stochastic Weight Averaging method to improve the training performance. The experiment results portrayed that our method achieved the best performance (8.28% Syllable Error Rate) and outperformed all other competitors in Task 1 of the 2021 VLSP Competition on Vietnamese Automatic Speech Recognition.

*Keywords:* Automatic Speech Recognition, Transformer, Conformer, Gradient Mask.

## 1. Introduction[1]

Recently, there has been a considerable shift in the speech community from deep neural network-based hybrid modeling [1] to end-to-end (E2E) modeling [1-4] for automatic speech recognition (ASR). While hybrid models necessitate the disjoint optimization of distinct constituent models such as the acoustic and language models, E2E ASR systems use a single network to directly transform a speech signal sequence into an output token (subwords, or words) sequence.

Modern E2E techniques for ASR systems include:(a) Connectionist Temporal Classification (CTC) [5], (b) Attention-based Encoder-Decoder (AED) [6], and (c) Recurrent Neural Network Transducer (RNN-T) [7, 8]. CTC was the first of these three techniques, and it can map the input speech signal to target labels without requiring any external alignments. However, it suffers from the conditional frame independence assumption

* Corresponding author.
*E-mail address:* sondd9@viettel.com.vn

https://doi.org/10.25073/2588-1086/vnucsce.322

[7]. AED is a generic family of models that were first developed for machine translation but has demonstrated performance in various areas (including ASR [6]). These models, however, are not naturally streaming. Meanwhile, RNN-T extends CTC to eliminate the frame-independence assumption. RNN-T has garnered much attention for industrial applications due to its streaming nature, and it has also succeeded in replacing state-of-the-art hybrid models in many recent circumstances [7-9].

Furthermore, RNN-T models using Conformer [10], a combination of Transformer and Convolution Neural Network, and their variants have achieved remarkable success and been considered state-of-the-art methods in training E2E ASR models. However, Conformers need a large labeled dataset to obtain a good performance. Meanwhile, semi-supervised methods can utilize unlabeled data and only a small amount of labeled to train models effectively. One of the representative cases for this class is pseudo-labeling learning (PLL) [11, 12], which has been successfully demonstrated in many recent studies [13, 14, 15]. The approach trains a seed model with a small set of labeled data and then applies it to infer labels for a larger unlabeled dataset. Next, the combination process of labeled and pseudo-label data is conducted to enrich the training dataset.

However, most ASR models are designed for the English language, and there are few efforts for the Vietnamese language. This paper proposes an RNN-T model which combines Conformer and Pseudo-labeling learning for the Vietnamese language. In particular, we train the Conformer-based RNN-T model on labeled data. Then, this model is used to generate pseudo labels for the unlabeled data. Next, labeled and pseudo-label data are combined into a new dataset. The previously trained model is fine-tuned on this new dataset by the Stochastic Weight Averaging method. However, unlabeled data and labeled data may have a high mismatch domain. As a result, computed pseudo labels on unlabeled data could be wrong. Therefore, to reduce the effect of noises and errors when learning on pseudo-label inputs, Gradient Mask is applied to train the model more effectively. Gradient Mask only allows gradients of pseudo label inputs to backpropagate through the encoder instead of both the encoder and predictor. This helps prevent the effect of wrong labels on the predictor block. Our proposal had the best performance with a Syllable Error Rate (SyER) of 8.28% and ranked first in the 8th International Workshop on Vietnamese Language and Speech Processing (VLSP 2021)[16].

The rest of the paper is organized as follows. Section 2 presents the corpus description while our method is described in section 3. Section 4 shows our experiments and evaluation and gives out discussions. Finally, conclusions are drawn in section 5.

## 2. Corpus Description

Three datasets are provided by the organizer of the 8th International Workshop on Vietnamese Language and Speech Processing (VLSP 2021). The first dataset contains about 200 hours of open domain transcribed data and 21 hours of labeled in-domain data. The second contains about 300 hours of untranscribed in-domain data. The corpus duration distribution is explicitly described in Figure 1.

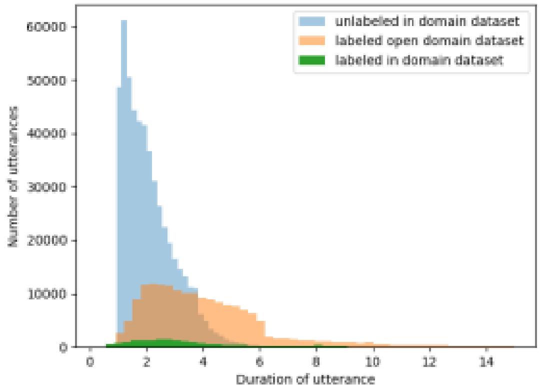


Figure 1. The distribution of utterance duration.

## 3. Methodology

Let A = {Xi; Yi} be a labeled dataset and B = {Xj} be a large unlabeled dataset. We first train a seed transducer model M on the labeled dataset A. Subsequently, we use this seed acoustic model M to generate pseudo labels on dataset B to obtain B'= {Xj; Y'j}. We then combine B' with A to form a new dataset C = B' ∪ A. However, labels (text) generated on B may be wrong due to the mismatch domain, affecting the predictor's performance. Therefore, when updating the model parameters from the pseudo-label data in B', Gradient Mask as described in 3.2 is applied to block gradient flows from pseudo label data into the predictor network. The optimization process using Stochastic Weight Averaging Algorithm (SWA) is repeatedly executed until the Syllable Error Rate on the validation dataset is converged.

The transducer model, Gradient Mask, and Stochastic Weight Averaging details are described below.

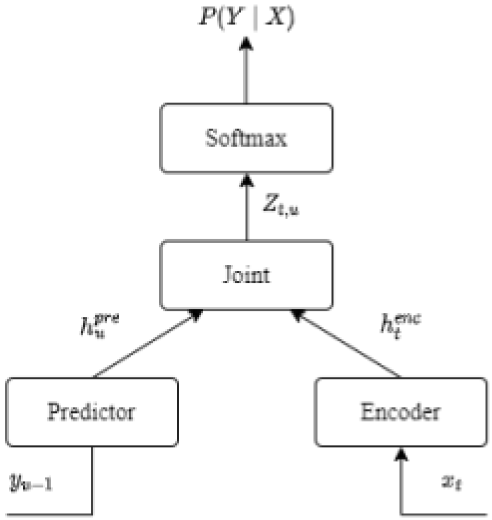


Figure 2. Transducer model

### 3.1. Transducer Model

The transducer model described in Figure 2 contains a combination of a prediction network, an encoder, and a joint network. The encoder adopts Conformer blocks which proved its effectiveness for ASR systems in many recent studies [10, 17, 18]. Conformer integrates Transformer and Convolution Neural Network together to overcome shortcomings of constitutive models. Transformers are unable to extract fine-grained local feature patterns. On the other hand, convolution requires more layers or parameters to capture global information. By combining convolutions and transformers, Conformer can learn both location-based local features and content-based global interactions in input speech signal sequences. The predictor uses a 1-layer LSTM. Input of the encoder and predictor are acoustic features $x_1, x_2, \ldots, x_T$ ], and text labels [$y_1, y_2, \ldots, y_U$ ], respectively.

Next, output of the encoder henc and the predictor hpre are fed to the joint network.

$$Z_{t,u}^{joint} = f^{joint}(h_t^{enc}, h_u^{pred}) \qquad (1)$$

The probability distribution over vocabulary is calculated by a softmax layer.

### 3.2. Gradient Mask

For each acoustic input X = [$x_1, x_2, \ldots, x_t$ ] fed to the encoder module, a sequence mask mask = [$m_1, m_2, \ldots, m_T$] is generated randomly to represent the mask positions of X, with $m_t$ is 1 if features are masked at time t, and otherwise mt is 0. Then, The output of the encoder f enc is as follows:

$$h^{enc} = f^{enc}((\sim mask) \cdot X) \qquad (2)$$

In back-propagation process, gradients are back-propagated to the encoder as follows:

$$grad_{h^{enc}} = (\sim mask) \cdot grad_{h^{enc}} \qquad (3)$$

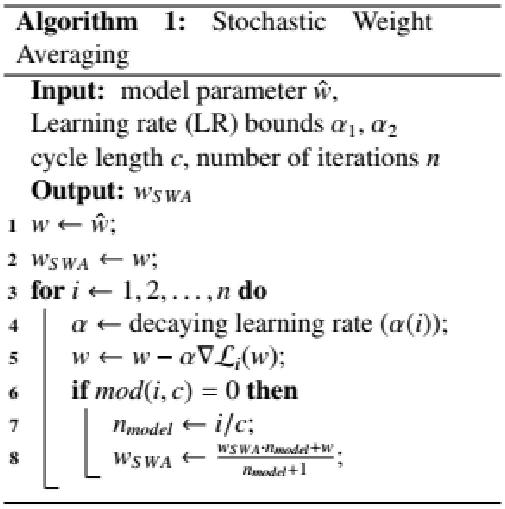
Notably, for the predictor, when back-propagation is executed, the gradient flow from the pseudo-label data into the predictor network is blocked to avoid effect of wrong labels. This process can be expressed in this function:

$$h_{t,u}^{joint} = f^{joint}(h_t^{enc}, sg(h_u^{pred})) \qquad (4)$$

where sg(x) ≡ x, $\frac{d}{dx}$sg(x) = 0 is the stop gradient operator. The objective function is still the same transducer loss where we try to minimize $P(Y|X)$ of all alignment paths. As a result, the trained model will be less affected by noise and wrong labels while learning a strong acoustic representation.

### *3.3. Stochastic Weight Averaging*

We use the Stochastic Weight Averaging (SWA) [19] instead of Stochastic Gradient Descent (SGD) to optimize our model. The reason is that SWA can avoid overfitting issues and reduce computational resources, which improves the accuracy of training the model. A common finding to explain success is that by averaging weights after the SGD process with periodic or high constant learning rates, broader optimizations can be found, which leads to better generalization. The SWA procedure is described in Algorithm 1.



**Algorithm 1:** Stochastic Weight Averaging

```
Input: model parameter $\hat{w}$,
Learning rate (LR) bounds $\alpha_1, \alpha_2$
cycle length $c$, number of iterations $n$
Output: $w_{SWA}$
$w \leftarrow \hat{w}$;
$w_{SWA} \leftarrow w$;
for $i \leftarrow 1, 2, \ldots, n$ do
    $\alpha \leftarrow$ decaying learning rate ($\alpha(i)$);
    $w \leftarrow w - \alpha \nabla \mathcal{L}_i(w)$;
    if $mod(i, c) = 0$ then
        $n_{model} \leftarrow i/c$;
        $w_{SWA} \leftarrow \frac{w_{SWA} \cdot n_{model} + w}{n_{model} + 1}$;
```

## 4. Experiments

### *4.1. Setup*

To evaluate the system performance on the datasets, we computed 80-channel Mel-filterbanks with window size 25ms and stride 10ms. On the labeled dataset, we apply SpecAugment [20] with mask parameter F = 27, and 10 time masks with maximum time-mask ratio pS = 0:05, where the maximum size of the time mask is set to pS the length of the utterance. The probability that a time step i is masked in the Gradient Mask method is 0:065. We set the mini-batch ratio from labeled data to pseudo-label data to 2:3, the same as the data ratio.

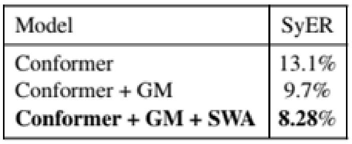
The filterbank features are first passed into two blocks of 2D-Conv layers, in which time reduction layers are added after each block to down-sample the frame rate to 4 before passing into the encoder. The encoder model consists of 16 layers of Conformer block, where we set the model dimension to 640 with eight attention heads, 31 kernel sizes in convolution block, and the same setting as Conformer-L [10]. We use LSTM as our predictor, and the LSTM predictor contains one layer with 640 units and a projection layer with 640 units. The Transducer's collaborative network is designed as a simple feed-forward layer. The total number of parameters is about 166M.

We use the learning rate warmup for the first 10k updates up to 1e-4 peak, then apply learning rate decay for model training. The model is trained on 8 NVIDIA A100 GPUs, a batch size of 128. Training time with this current setup is 30 hours. The training process is separated into 2 phases. In the first phase, the model is trained with 20 epochs using Adam. The second phase contains ten training epochs using Adam, the last two applied with SWA. The reason is that although Adam or SGD is a good optimization algorithm, it may not converge if the selected learning rate is too large [21]. Additionally, using SWA at the last stages helps reduce variance in solution quality and enhance convergence.

### 4.2. Result and Analysis

We implement three stages on the provided datasets. First, we train a supervised baseline with only 221 hours (21 in-domain hours and 200 open-domain hours) of labeled data. This trained model is then used to generate the pseudo-labels on unlabeled data. Next, we train the previous model on the generated pseudo-labels with the gradient mask method. Finally, we extend the second model with SWA improvement. The result is presented in Table 1.

Table 1. Syllable error rate (SyER):



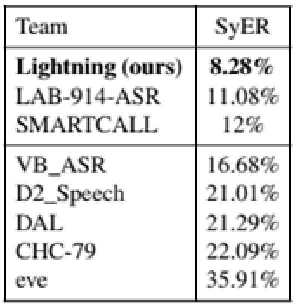
| Model | SyER |
|---|---|
| Conformer | 13.1% |
| Conformer + GM | 9.7% |
| **Conformer + GM + SWA** | **8.28%** |

The first model is trained only on the transcribed dataset, ignoring the second untranscribed. It is considered a competitive baseline for pre-labeled data sets with 13.1% SyER.

Next, our second implementation outperforms the baseline with 9.7% SyER, benefiting from the enormous untranscribed in-domain dataset. This proved the effectiveness of the Gradient Mask and semi-supervised method.

The last integrated second model with an SWA gives the best performance with 8.28% SyER. As the number of labels and data is limited, models often suffer from overfitting problems. SWA preeminently proved a better generalization than traditional training on these provided datasets and handled the overfitting problem.

Table 2. Compare our proposal with all othercompetitors in the 2021 VLSP:



| Team | SyER |
|---|---|
| **Lightning (ours)** | **8.28%** |
| LAB-914-ASR | 11.08% |
| SMARTCALL | 12% |
| VB_ASR | 16.68% |
| D2_Speech | 21.01% |
| DAL | 21.29% |
| CHC-79 | 22.09% |
| eve | 35.91% |

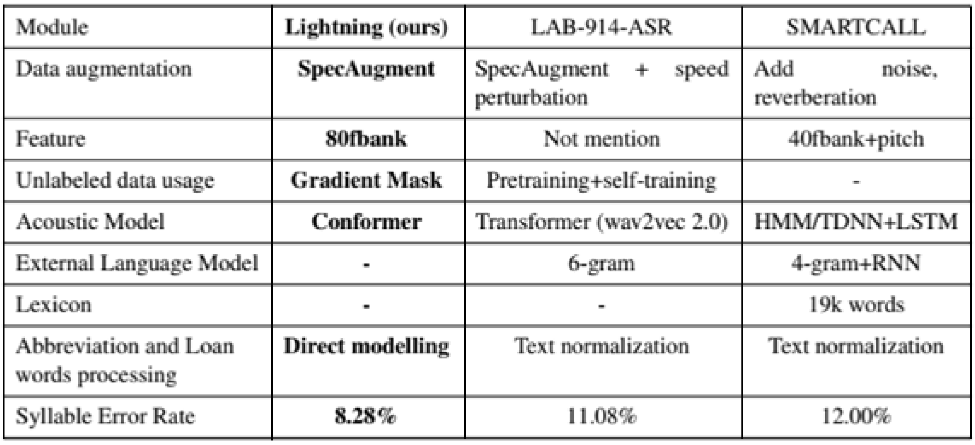
Furthermore, our proposal ranks first in the 2021 VLSP competition on Vietnamese ASR, as shown in Table 2. Our proposal has huge improvement gaps of 25% and 33%, respectively, compared to the second and third places. In methodology comparison, as shown in Table 3, our proposal does not use any external language model, lexicon, or text normalization. This proved the complete superiority of our proposed method compared to all other competitors.

Finally, we also examine the effect of speech input on the effectiveness of ASR models in the 2021 VLSP. The results are shown in Figure 3. The findings showed that accent, noise, and reverberation in speech inputs have a high impact on the quality of ASR models. The higher noise and reverberation is in the dataset, the bigger gap between ASR models. ASR models recognized southern accent speech inputs harder than other accents. Our proposal (Lightning team) achieved a better performance than other models on southern accent and noise speech inputs.

## 5. Conclusion

This paper sums up our approach to the latest Vietnamese datasets provided in the 2021 VLSP competition. The task is practical and close to the actual implementation of ASR products because the size of the transcribed dataset is small and often outnumbered by the amount of unlabeled, untranscribed data. We presented our three Conformer-based implementations. We use the labeled dataset to train a Conformer-based transducer model and then apply it with Gradient Mask on the pseudo-labels. Next, we train this model on the whole dataset with Stochastic Weight Averaging. The best results reached 8.28% SyER and achieved first place in the ASR competition of VLSP. Our future works include reducing the number of parameters and model size. Our model is huge-sized, and the dataset sizes are not always sufficient to avoid overfitting. Reducing model size while maintaining its accuracy is an effective way to handle these issues.

Table 3. Compare our approach with second and third places (Source: the 2021 VLSP conference report):



| Module | **Lightning (ours)** | LAB-914-ASR | SMARTCALL |
|---|---|---|---|
| Data augmentation | **SpecAugment** | SpecAugment + speed perturbation | Add noise, reverberation |
| Feature | **80fbank** | Not mention | 40fbank+pitch |
| Unlabeled data usage | **Gradient Mask** | Pretraining+self-training | - |
| Acoustic Model | **Conformer** | Transformer (wav2vec 2.0) | HMM/TDNN+LSTM |
| External Language Model | - | 6-gram | 4-gram+RNN |
| Lexicon | - | - | 19k words |
| Abbreviation and Loan words processing | **Direct modelling** | Text normalization | Text normalization |
| Syllable Error Rate | **8.28%** | 11.08% | 12.00% |

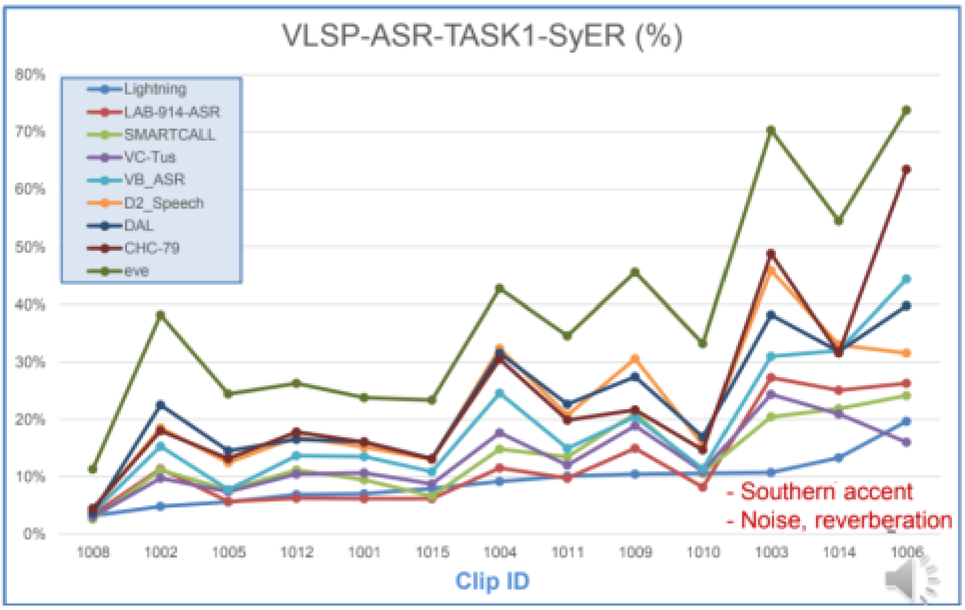


Figure 3. Effect of speech input on the effectiveness of ASR models (Source: the 2021 VLSP conference report)

## References

[1] Y. Miao, M. Gowayyed, F. Metze, Eesen: End-To-End Speech Recognition Using Deep Rnn Models And Wfst-Based Decoding, in: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, 2015, pp. 167–174.

[2] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, N. Jaitly, A Comparison Of Sequence-To-Sequence Models For Speech Recognition., in: Interspeech, 2017, pp. 939–943.

[3] C. C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, Nguyen, Z. Chen, A. Kannan, R. J. Weiss, Rao, E. Gonina, State-of-the-art speech recognition with sequence-to-sequence models, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 4774–4778.

[4] J. Li, R. Zhao, Z. Meng, Y. Liu, W. Wei, Parthasarathy, V. Mazalov, Z. Wang, L. He, Zhao, Y. Gong, Developing RNN-T Models Surpassing High-Performance Hybrid Models with Customization Capability, in: Proc. Interspeech 2020, 2020, pp. 3590–3594. doi:10.21437/Interspeech.2020-3016.

[5] A. Graves, N. Jaitly, Towards End-To-End Speech Recognition With Recurrent Neural Networks, in: International conference on machine learning, PMLR, 2014, pp. 1764–1772.

[6] J. K. Chorowski, D. Bahdanau, D. Serdyuk, Cho, Y. Bengio, Attention-Based Models For Speech Recognition, Advances in neural information processing systems 28.

[7] J. Li, R. Zhao, H. Hu, Y. Gong, Improving Rnn Transducer Modeling For End-To-End Speech Recognition, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2019, pp. 114–121.

[8] M. Jain, K. Schubert, J. Mahadeokar, C.-F. Yeh, K. Kalgaonkar, A. Sriram, C. Fuegen, L. Seltzer, Rnn-t for latency controlled asr with improved beam search, arXiv preprint arXiv:1911.01629.

[9] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473.

[10] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y.Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, Conformer: Convolution-Augmented Transformer For Speech Recognition, in: Interspeech 2020, ISCA, 2020, doi:10.21437/interspeech.2020-3015.

[11] V. Manohar, H. Hadian, D. Povey, S. Khudanpur, Semi-Supervised Training Of Acoustic Models Using Lattice-Free MMI, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018. doi:10.1109/icassp.2018.8462331.

[12] S. Karita, S. Watanabe, T. Iwata, A. Ogawa,M. Delcroix, Semi-Supervised End-To-End Speech Recognition, in: Interspeech 2018, ISCA, 2018. doi:10.21437/interspeech.2018-1746.

[13] S. H. K. Parthasarathi, N. Strom, Lessons From Building Acoustic Models With A Million Hours Of Speech, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019. doi:10.1109/icassp.2019.8683690.

[14] J. Kahn, A. Lee, A. Hannun, Self-Training For End-To-End Speech Recognition, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020. doi:10.1109/icassp40776.2020.9054295.

[15] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, Q. V. Le, Improved Noisy Student Training For Automatic Speech Recognition, in: Interspeech 2020, ISCA, 2020. doi:10.21437/interspeech.2020-1470.

[16] D. V. Hai, ASR Challenge: Vietnamese Automatic Speech Recognition, VLSP, 2021.

[17] S. Li, M. Xu, X.-L. Zhang, Efficient Conformer-Based Speech Recognition with Linear Attention, arXiv preprint arXiv:2104.06865.

[18] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, Recent Developments On Espnet Toolkit Boosted By Conformer, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 5874–5878.

[19] H. Song, M. Kim, D. Park, Y. Shin, J.-G. Lee, Learning From Noisy Labels With Deep Neural Networks: A Survey, 2020, arXiv:arXiv:2007.08199.

[20] 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:1904.08779.

[21] H. Guo, J. Jin, B. Liu, Stochastic Weight Averaging Revisited, arXiv preprint arXiv:2201.00519.