Adversarial Augmentation and Adaptation for Speech Recognition

Jen-Tzung Chien and Wei-Yu Sun
Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

***Abstract*—It is crucial to conduct parameter efficient learning to adapt a large-scaled pre-trained backbone model to a downstream task where the desirable performance could be achieved for low-resource automatic speech recognition (ASR). However, the overfitting problem is prone to happen when the model adaptation is performed through fine-tuning individual parameters. The previous studies have explored different data augmentation methods to increase the size of training samples to enrich the data coverage and accordingly alleviate the overfitting issue in model training. In particular, this paper presents a new adversarial training for ASR based on a frozen pre-trained backbone model where the adversarial data augmentation is implemented so that a small amount of controllable parameters in adapters can be sufficiently estimated. In this study, the adversarial speech data are generated by adding the adversarial perturbations when training the adapters in an ASR model. The gradients of the intermediate adversarial examples are accumulated to calculate the augmented speech samples. The experiments on ASR using Common Voice and LibriSpeech datasets show the merit of the proposed adversarial augmentation and adaptation in terms of error rate and model size.**

I. Introduction

Automatic speech recognition (ASR) is built as an optimized computation model which is developed to automatically convert a human speech utterance into the corresponding word sequence. Different from speaker recognition, which attempts to identify the speaker who uttered the speech, ASR aims to recognize the lexical content from a speech signal. Traditionally, ASR was built with the acoustic model based on the hidden Markov model combined with Gaussian mixture model as well as the language model based on n-gram. Nowadays, the deep neural network model combining the acoustic model and language model in an end-to-end manner has been successfully developed to achieve a powerful performance in a sequence-to-sequence domain mapping from speech signal to text string where the massive trainable parameters are configured and the huge amount of training data is required. Importantly, the transformer [1] was exploited to build an end-to-end encoder-decoder framework which has shown dominant performance in a variety of tasks ranging from natural language processing (NLP) to computer vision (CV). Transformer was built according to the scheme of self attention and cross attention over individual and mutual sequences in source domain and target domain. More recently, different types of transformer have been designed to improve ASR by handling various issues in model size, memory capacity and representation efficiency [2], [3], [4]. However, a high-performance transformer rely on a large amount of neural network parameters as well as a huge collection of training utterances. A meaningful approach is to develop an ASR for a downstream task with a specific speaking style or under a different surrounding environment. Such an approach has been applied in many applications where a fine-tuning work was carried out by following a pre-trained backbone model using a limited amount of adaptation data [5].

In general, a large-scaled pre-trained transformer was estimated from a large amount of training data by using supervised, unsupervised and self-supervised learning methods. The data-driven transformer model could be sufficiently trained to cover various rules of syntax and grammar. Continual training is feasible to obtain desirable performance by using a pre-trained model. To customize the transformer to leverage knowledge transfer to fit a specific task using low-resource data in a target domain, it is straightforward to conduct fine-tuning for the encoder and decoder parameters in a pre-trained model. For the pre-trained model in ASR applications, the model is learned from audio signal and applied to calculate the audio features for a downstream task. The main-stream pre-trained ASR models such as wav2vec2.0 [6] and HuBERT [7] are seen as the variants of transformer which are similar in model architecture. In addition, the training procedure of ASR models is similar to that of pre-trained language models. The training objective aims to correctly predict or recover the masked input segments. Nevertheless, purely fine-tuning is easily overfitting in training process and likely dropping in ASR performance. It is meaningful to conduct data augmentation to mitigate the overfitting issue. A simple and general way to data augmentation for ASR could be implemented by applying the SpecAugment [8]. This method treated the Mel spectrum as an image sample. A kind of translational data enhancement was performed. Several consecutive rows and columns on the spectrum matrix were randomly crossed out. A resulting model was trained in a self-supervised way to enhance the capability of learning representation in time and frequency dimensions. There were no additional parameters and calculations. In addition, some other straightforward augmentation methods such as the pitch shift, time shift or noise addition to original audio speech data also worked well. This study deals with the overfitting issue in model construction for speech recognition based on a frozen large-scaled pre-trained backbone model where the small-scaled controllable adapter is learned from low-resource adaptation data [9]. A new adversarial augmentation and adaptation method is proposed to implement a parameter efficient learning approach to improve model robustness in presence of adversarial perturbations. The

training objective is developed to introduce the augmented data to ensure an extended and smoothed decision boundary to meet the local Lipschitz condition [10]. The experiments on various languages show the merit of the proposed method.

II. Adversarial Augmentation and Adaptation

Adversarial learning is known as a powerful machine learning paradigm which is feasible to build a robust model in presence of adversarial perturbation. This study presents an adversarial training algorithm to build a parameter efficient ASR model using adapter [11], [12].

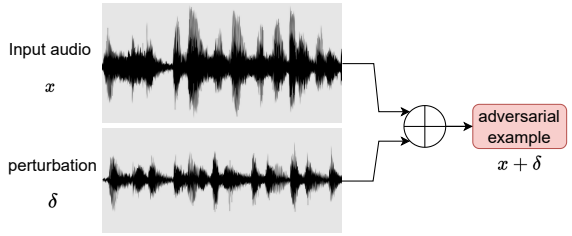

Fig. 1. Illustration of an adversarial example which is seen as a noisy speech signal with an additive perturbation noise.

A. Adversarial robustness

In general, deep learning is prone to build a model θ which is vulnerable to the adversarial example [13]. Such an example $x+\delta$ is basically seen as the addition of an imperceptible perturbation δ on the original example x which likely leads to a wrong prediction of output y. Figure 1 illustrates how an adversarial example of speech signal is observed. The adversarial example is considered as an augmented sample which is merged in training for the adapter parameter θ where a backbone ASR model ϕ is provided and frozen. The perturbation noise δ and the model parameter θ are jointly estimated from a dataset $\mathcal{D}=\{x,y\}$ by following

$$\min_{\theta}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left\{\frac{1}{N}\sum_{i=1}^{N}\max_{\delta_i}\left[\mathcal{L}\left(f_{\theta,\phi}\left(x+\delta_{i-1}\right),y\right)\right]\right. \\ \left.+\,\mathrm{KL}\left(f_{\theta,\phi}(x)\|f_{\theta,\phi}\left(x+\delta_N\right)\right)\right\}. \tag{1}$$

where δ_i denotes the perturbation at iteration i and totally N iterations are run. Correspondingly, the worst-case perturbation with the highest loss function $\mathcal{L}$ is estimated and then incorporated to train the optimal controlled model θ with the lowest loss in $\mathcal{L}$ where the classification outputs $f_{\theta,\phi}$ using original speech x and perturbed speech $x+\delta_N$ in the last iteration N are optimally near by with the smallest Kullback-Leibler (KL) divergence. Basically, this learning process is seen as a two-player game consisting of the players of finding the perturbation δ and then estimating the model θ. An adversarial robustness can be achieved accordingly [14], [15] by merging the adversarial examples $x+\delta$ in training for the controllable adapter θ. Basically, the worst-case perturbation may severely change the output of prediction. The proposed model is trained to improve the robustness of ASR by utilizing the perturbed speech data. Finding the perturbed speech corresponds to implement the generation step also called the adversarial attack. Alternatively, finding the estimated ASR adapter by minimizing the classification loss corresponds to carry out the discrimination step which is also called the adversarial defense. Adversarial training is seen as a process of fulfilling attack and defense for training a robust model.

B. Attack and defense

The key idea of adversarial training is to generate the adversarial examples that maximize the loss and simultaneously use these adversarial examples to train the model by minimizing the loss. Minimax optimization is comparable to fulfill an attack and defense process towards training a robust ASR model via a generation step and a discrimination step. The generation step for finding the perturbation is considered an attack while the discrimination step for training an ASR controllable model is acted as a defense against the attack.

In [16], the so-called fast gradient sign method (FGSM) was proposed as an approach to estimate the perturbation. Considering a supervised learning for a model f_θ from input sample x and its corresponding label y, FGSM is employed to solve Eq. (1) and estimate the perturbation by an iterative updating formula

$$\delta' \leftarrow \delta + \epsilon\cdot\mathrm{sign}(g^\delta) \tag{2}$$

where ϵ denotes the learning rate, δ' denotes the updated perturbation, and g^δ denotes the gradient

$$g^\delta = \nabla_\delta\mathcal{L}(f_{\theta,\phi}(x+\delta),y). \tag{3}$$

FGSM generates a perturbation δ for an addition with the input x. The estimation of perturbation is along the direction of the gradient towards the maximum value of loss function. In addition, an alternative approach to worst-case perturbation where the perturbation is scaled according to a normalized gradient without using the sign function, also known as a fast gradient method (FGM) [17] in a form of

$$\delta' \leftarrow \delta + \epsilon\cdot g^\delta/\|g^\delta\|_2. \tag{4}$$

FGM uses a one-step gradient ascent method to find an adversarial perturbation which may not be optimal. In [18], the projected gradient descent (PGD) was proposed to find an optimal perturbation through an iterative gradient ascent algorithm. In particular, if the perturbation is outside a volume as a norm ball with radius ϵ, i.e. $\|\delta\|_2\le\epsilon$, the perturbation is first projected back to the norm ball to ensure the perturbation is not too large. The solution to PGD is obtained and shown by

$$\delta' \leftarrow \mathrm{Proj}_{\|\delta\|_2\le\epsilon}(\delta+\alpha\cdot g^\delta/\|g^\delta\|_2) \tag{5}$$

where α is a step size. Using the perturbation δ_N in the last iteration N, the adversarial example is obtained as an attack by

$$x' = x + \nabla_\delta\mathcal{L}(f_{\theta,\phi}(x+\delta),y)\big|_{\delta\leftarrow\delta_N}. \tag{6}$$

On the other hand, the defense model is implemented by fulfilling a discrimination step towards estimating the controllable adapter θ for speech classifier in an ASR system. In the implementation, the gradient with respect to adapter θ consists of those for the first term g^θ and the second term in Eq. (1). The first gradient is calculated by initializing the gradient as $g^\theta \leftarrow 0$ and updating it iteratively in the inner loop as

$$(g^\theta)' \leftarrow g^\theta + \frac{1}{N}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\nabla_\theta \mathcal{L}(f_{\theta,\phi}(x+\delta), y)] \quad (7)$$

by N times to calculate the accumulated and normalized g_N^θ in the last iteration N. Then, the model parameter is iteratively updated in outer loop by

$$\theta' \leftarrow \theta - \beta \cdot (g_N^\theta + \nabla_\theta \mathrm{KL}(f_{\theta,\phi}(x) \| f_{\theta,\phi}(x+\delta_N))) \quad (8)$$

where β denotes the learning rate. The optimal adapter θ is trained by minimizing the cross-entropy loss for classification problem in an ASR system where KL divergence is also minimized to regularize the prediction consistency between adversarial sample and original sample.

III. Implementation Algorithm

This paper presents a new adversarial training algorithm to build a defense model for speech recognition where the generation step and discrimination step are implemented to increase the amount of the informative training samples and accordingly enhance the generalization and robustness of a learned model. This model is capable of improving the prediction performance for an ASR system consisting of a frozen pre-trained backbone model and a trainable adapter for a downstream task in a low-resource setting for domain adaptation [19]. The implementation algorithm is formulated and addressed.

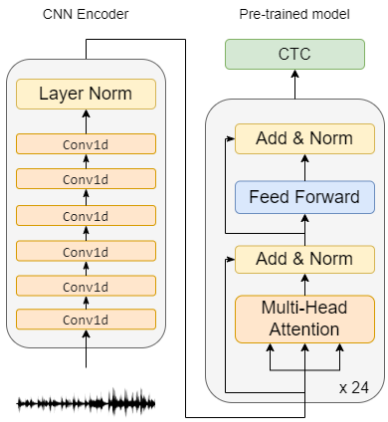

Fig. 2. Illustration of ASR pre-trained backbone model with parameter ϕ where speech signal x is transformed to find word string y. CNN encoder, transformer blocks and CTC classifier are stacked.

A. Model architecture

In this study, the pre-trained wav2vec2.0 [6] and HuBERT [7] were employed as the ASR backbone models. These two models had similar model structures which were trained in different strategies. Basically, wav2vec2.0 quantized the continuous audio data, and was learned to map them to discrete features and predict the masked features for training. The quantization module was used to discretize the output of the feature encoder. This model contained two sets of codebooks, and each codebook contained 320 variables. For each continuous output variable from the feature encoder, wav2vec2.0 identified a code in each set of codebooks, concatenated two codes together, and linearly mapped them to find the final quantization. On the other hand, HuBERT was trained by running two turns of the following procedure including using the k-means to find the clusters of Mel-frequency cepstral coefficient (MFCC) features and then predicting the masked features. After the first training turn was complete, the clustering model was reused to learn the detailed feature representation where the number of clusters in previous clustering model was increased from 100 to 500. The architecture of the pre-trained backbone model is shown in Figure 2. One-dimensional convolutional neural network (CNN) encoder with 7 layers dealt with the down-sampling of the input speech waveform x. The outputs of audio-encoded features were then fed forward to the pre-trained transformer encoder totally stacking 16 transformer blocks of multi-head attention and feedforward network. Then, the connectionist temporal classification (CTC) loss was calculated by using latent information from transformer encoder and then minimized to make prediction of word sequence y.

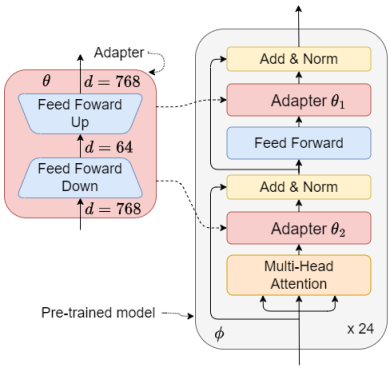

Fig. 3. Illustration of a stacked encoder where each frozen pre-trained transformer block ϕ is augmented with two controllable adapters $\{\theta_1, \theta_2\}$.

In model configuration, there were two adapters with parameter $\theta = \{\theta_1, \theta_2\}$ which were inserted in individual transformer blocks of a pre-trained backbone model ϕ. There were 24 blocks in an encoder. One adapter was connected with the multi-head attention and the other one was added to the feedforward layer as shown in Figure 3 where each adapter was shaped as a bottleneck structure with two feedforward layers with dimension from 768 to 64 and then back to 768. The parameters of the adapted transformer blocks were frozen except those parameters of adapters that were trainable. For each downstream task, only those adapter parameters θ were learned from relatively few training utterances. This adaptation strategy dramatically reduced the size of controllable parame-

ters and considerably increased the parameter efficiency for a downstream ASR task. The resulting model was compact for transfer learning where the numbers of training utterances and adjusted parameters were balanced.

B. Augmentation and adaptation

This paper presents an adversarial training method which is developed to generate the augmented examples in an adversarial attack step [20], [21]. The proposed adversarial algorithm eliminates the overhead cost of generating adversarial examples. A loopy optimization procedure with an inner loop and an outer loop is implemented as attack and defense, respectively. A robust ASR model with a pre-trained backbone model ϕ and a controllable adapter model θ is constructed. Finding the adversarial examples is implemented as the generation step for data augmentation in an inner loop via the gradient ascent algorithm while estimating the controllable adapter layers θ is performed as the discrimination step for classifier estimation in an outer loop via the gradient descent algorithm. In particular, the projection in PGD method is run by N times in inner loop to estimate the adversarial perturbation δ. The augmented data using the perturbation δ_N in the last iteration N is used to update the classifier or correspondingly the adapter θ in the outer loop as given in Eq. (6). A detailed implementation for this loopy procedure for adversarial augmentation and adaptation (AAA) is shown in Algorithm 1.

Algorithm 1 Adversarial augmentation & adaptation process

require: training samples $\mathcal{D} = \{x, y\}$, frozen pre-trained weights ϕ, adapter weights θ, perturbation δ, perturbation radius ϵ, adversarial step size α, learning rate β

  for $t = 1, \ldots, M$ **do** ▷ outer loop minimization

    initialize $g_0^\theta \leftarrow 0$, $g^\delta \leftarrow 0$, randomize δ_0

    for $i = 1, \ldots, N$ **do** ▷ inner loop maximization

      $g_i^\theta \leftarrow g_{i-1}^\theta + \frac{1}{N}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\nabla_\theta \mathcal{L}(f_{\theta,\phi}(x+\delta_{i-1}), y)]$

      $g^\delta \leftarrow \nabla_\delta \mathcal{L}(f_{\theta,\phi}(x+\delta_{i-1}), y)$

      $\delta_i \leftarrow \text{Proj}_{\|\delta\|_2 \le \epsilon}(\delta_{i-1} + \alpha \cdot g^\delta / \|g^\delta\|_2)$

    end for

    $\theta_{t+1} \leftarrow \theta_t - \beta \cdot (g_N^\theta + \nabla_\theta \text{KL}(f_{\theta,\phi}(x) \| f_{\theta,\phi}(x+\delta_N)))$

  end for

output: adapter weights θ_M

In this learning procedure, the gradient g^δ for perturbation δ is updated N iterations in the inner loop for maximization of cross-entropy loss function to cope with classification problem in ASR. The gradient g^θ for adapter θ is also updated N times to find g_N^θ and then combined with that for KL divergence between prediction outputs from original utterance x and perturbed utterance $x+\delta_N$. In the implementation, the gradient information g^δ is recycled to update the perturbation δ_i in every inner loop i, which means that the intermediate updatings of gradients for perturbation g^δ are reused. Therefore, we generate several augmented samples for data augmentation in one adversarial training in outer loop t. The loopy algorithm is performed to find adapter model θ_M which is obtained by running M iterations in outer loop t. The proposed method is inspired by the variants of the FREE algorithm in [22], [23] which were developed in computer vision and natural language processing. Considering this trick, this paper proposed a new AAA algorithm for speech recognition. The original sample x and the perturbed sample x' in the last iteration N as shown in Eq. (6) are both used when updating the adapter θ through

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f_{\theta,\phi}(x), y) + \mathcal{L}(f_{\theta,\phi}(x'), y)]. \tag{9}$$

The gradient with respect to θ is computed by using original sample x and adversarial sample x' with equal weighting.

C. Tradeoff between accuracy and robustness

In addition to FGM [17] and PGD [18], another adversarial training algorithm called the tradeoff-inspired adversarial defense via surrogate-loss minimization (TRADES) [24] was also implemented for comparison with the proposed AAA method. Using TRADES, the same setting as PGD in Eq. (6) was taken into account to generate the adversarial examples in accordance with maximizing the KL divergence between the logits of the prediction models by using original samples and adversarial samples in

$$\begin{aligned}\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[&\mathcal{L}(f_{\theta,\phi}(x), y) \\ &+ \lambda \cdot \max_{\|\delta\|_2 \le \epsilon} \text{KL}(f_{\theta,\phi}(x) \| f_{\theta,\phi}(x+\delta))\big].\end{aligned} \tag{10}$$

where $\lambda = 1$ is specified.

The attack and defense in adversarial training correspond to perform the generation and discrimination steps, which are adjustable to tradeoff between adversarial robustness and classification accuracy, respectively. The proposed AAA method carries out the adapter-based solution to adversarial robustness via Eq. (1) which comparably implements the tradeoff between two terms in Eq. (10) based on TRADES with a adjustable hyperparameter λ. The first term as a cross-entropy loss controls the accuracy while the second term as a KL divergence handles the robustness. Maximizing the KL divergence aims to generate the worst-case adversarial examples preserving the strongest protection to attack while minimizing the cross-entropy loss focuses on finding the best adapter to achieve the lowest word error rate (WER) for speech recognition.

Similar issue of balancing the tradeoff between accuracy and robustness was also mentioned in CV and NLP domains [10], [23], [25]. The second term of AAA in Eq. (1) is seen as a regularization condition for model generalization. This term encourages the prediction of adversarial example $x + \delta_N$ with the worst-case perturbation δ_N by using the learned adapter θ to get close to that of original example x. The resulting class boundary is likely fitted to the Lipschitz condition so that a smoothed boundary is determined to assure a better generalization [10]. This study presents a new approach to adversarial robustness in low-resource setting based on adversarial augmentation and adapter estimation. The KL term is optimized in an adversarial manner to get closer to the Lipschitz condition for a better model robustness or generalization.

IV. Experiments

This paper presents an adversarial augmentation and adaptation for speech recognition in two low-resource downstream tasks under two different pre-trained backbone models.

A. Experimental settings

As shown in Figure 2, a CTC classifier was stacked after an ASR pre-trained model based on 16 transformer blocks for word prediction where the audio embedding was calculated from a speech signal using CNN encoder. Word error rates (WERs) were measured to evaluate the ASR performance. Table I shows the settings of pre-trained backbone model based on wav2vec2.0 [6] and HuBERT [7]. The number of parameters in a backbone models is 315M while that of adapter is only 3.1M. AAA only adjusts the parameters of adapter instead of fine-tuning all backbone parameters as done in the previous methods. A parameter efficient learning is assured in using AAA method. Table II reports the settings of two speech datasets including Common Voice [26] and LibriSpeech (100 hours) [27] for Turkish and English, respectively. The number of samples or utterances is shown. In addition to the baseline result, this study also implemented the fine-tuning of the pre-trained model by using the other adversarial data augmentation methods where PGD [18], FGM [17] and TRADES [24] were included. Baseline method was implemented via a purely fine-tuning scheme. In the implementation, the perturbation δ was estimated and added to the audio signal as the augmented noisy speech signal. This strategy carried out an approach to noise-based data augmentation in a way of increasing the difficulty of correctly classifying the input audio signal into its corresponding features and word tokens based on the pre-trained model augmented with the controllable adapter.

TABLE I
SETTINGS OF THE PRE-TRAINED BACKBONE MODELS.

backbone	attention head	layer	hidden state	feature layer	dropout
wav2vec2.0	16	24	1024	7	0.1
HuBERT	16	24	1024	7	0.1

TABLE II
SETTINGS OF THE DATASETS FOR SPEECH RECOGNITION.

dataset	training set	testing set	language
Common Voice	3478	1647	Turkish
LibriSpeech	31242	2620	English

TABLE III
SETTINGS FOR THE ADVERSARIAL DATA AUGMENTATION.

method	ϵ	α	inner loop	norm type
baseline	N/A	N/A	N/A	N/A
PGD	1	0.3	3	ℓ_2
FGM	1	0.3	1	ℓ_2
TRADES	1	0.1	3	ℓ_2
AAA	1	0.1	3	ℓ_2

Furthermore, Table III lists the hyperparameter settings of using different methods for adversarial data augmentation including perturbation radius ϵ, adversarial step size α, number of inner-loop iterations N and the norm type which was fixed as ℓ_2 norm. In case of $N = 3$, the amount of augmented data were increased three times. In the experiments, the computation times in different settings were measured for two ASR tasks by using a single GPU, NVIDIA GeForce RTX 3090, with 24G RAM where the CPU was configured with an Intel®Core™i9-10900K CPU @ 3.70GHz. The batch sizes of speech utterances were 32 and 4 for Common Voice and LibriSpeech, respectively.

TABLE IV
WERS (%) AND TRAINING TIMES USING WAV2VEC2.0.

	CommonVoice		LibriSpeech	
	WER	training time	WER	training time
baseline	35.5	55min	4.25	14hr
PGD	35.2	2hr 6min	3.99	17hr 50min
FGM	34.5	1hr 22min	4.05	14hr 22min
TRADES	34.2	2hr 13min	4.18	18hr 20min
AAA	31.7	2hr 8min	3.94	18hr

TABLE V
WERS (%) AND TRAINING TIMES USING HUBERT.

	CommonVoice		LibriSpeech	
	WER	training time	WER	training time
baseline	51.3	59min	8.86	13hr 51min
PGD	48.4	2hr 8min	4.83	18hr 4min
FGM	51.8	1hr 19min	5.59	14hr 34min
TRADES	48.8	2hr 12min	5.02	18hr 25min
AAA	47.8	2hr 6min	4.77	17hr 57min

TABLE VI
AN EXAMPLE OF PREDICTED WORDS WHERE THE WORDS IN BLUE REPRESENT THE ERRORS. WAV2VEC2.0 WAS USED.

Label: he has grave doubts whether sir frederick leighton's work is really greek after all and can discover in it but little of rocky ithaca	
baseline	he has graved doubts whether sir frederick layton's work is ready greek after all and can discover in it but little of rocky ithica
AAA	he has grave doubts whether sir frederick laytons work is really greek after all and can discover in it but little of rocky ithaca

B. Experimental results

Tables IV and V compare the results of WER and training time by using wav2vec2.0 and HuBERT as the pre-trained backbone models, respectively. The number of outer-loop iterations M was varied and sufficient to assure convergence in using different methods. From the results, it is found that various adversarial data augmentation methods are beneficial in two ASR downstream tasks on the basis of pre-trained backbone models by using wav2vec2.0 and HuBERT. Wave2vec2.0 performs better than HuBERT in ASR using two different speech datasets. The adversarial augmentation and adaptation by using AAA consistently achieves lower

WERs than the adversarial data augmentation and fine-tuning by using PGD, FGM and TRADES by using different pre-trained backbone models under different speech datasets. In terms of computation time, the proposed AAA is more efficient than PGD and TRADES, but less efficient than FGM. FGM only ran one inner-loop iteration which is smaller than three inner-loop iterations in implementation of AAA. In addition, Table VI shows an example of comparing the predicted word sequences by using baseline and the proposed AAA where wav2vec2.0 was considered as the backbone model. The true label sequence is provided. It is obvious that AAA improves the ASR performance in this example.

V. Conclusions

This paper has presented an adversarial augmentation and adaptation algorithm to enhance the robustness of speech recognition in presence of perturbation noises. The overfitting issue in low-resource setting was handled through a frozen pre-trained backbone model combined with the learnable adapter. A parameter efficient learning algorithm was proposed to carry out an approach to ensure the adversarial robustness in speech recognition. The proposed method was illustrated as an attack and defense strategy through a procedure of data generation and model discrimination where the tradeoff between robustness and accuracy was adjustable for speech recognition, respectively. The relation of the proposed method to the other schemes to adversarial data augmentation was illustrated. A straightforward solution to an end-to-end data augmentation was exploited. In the experiments on a number of investigations over different languages and backbone models showed that the proposed method to adversarial robustness based on adapter performed better than the other conventional methods based on fine-tuning in terms of word error rates in speech recognition. The proposed method was more efficient in terms of parameter size than the other methods in most of settings. Future study on extending this method to the other kinds of downstream applications will be explored.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] J.-T. Chien and S.-E. Li, "Contrastive Disentangled Learning for Memory-Augmented Transformer," in *Proc. of Annual Conference of International Speech Communication Association*, 2023, pp. 2958–2962.

[3] J.-T. Chien and Y.-H. Huang, "Bayesian Transformer Using Disentangled Mask Attention," in *Proc. of Annual Conference of International Speech Communication Association*, 2022, pp. 1761–1765.

[4] C.-H. Leong, Y.-H. Huang, and J.-T. Chien, "Online Compressive Transformer for End-to-End Speech Recognition," in *Proc. of Annual Conference of International Speech Communication Association*, 2021, pp. 2082–2086.

[5] L.-J. Yang, C.-H. H. Yang, and J.-T. Chien, "Parameter-Efficient Learning for Text-to-Speech Accent Adaptation," in *Proc. of Annual Conference of International Speech Communication Association*, 2023, pp. 4354–4358.

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

[7] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[8] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[9] Y. Chen, X. Yang, H. Zhang, W. Zhang, D. Qu, and C. Chen, "Meta adversarial learning improves low-resource speech recognition," *Computer Speech & Language*, vol. 84, pp. 101576, 2024.

[10] Y.-Y. Yang, C. Rashtchian, H. Zhang, R. R. Salakhutdinov, and K. Chaudhuri, "A closer look at accuracy vs. robustness," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8588–8601, 2020.

[11] C.-Y. He and J.-T. Chien, "Learning adapters for code-switching speech recognition," in *Proc. of Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2023, pp. 1–6.

[12] B. Aditya, M. Rohmatillah, L.-H. Tai, and J.-T. Chien, "Attention-guided adaptation for code-switching speech recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 10256–10260.

[13] J.-T. Chien and Y.-A. Chen, "Towards a unified view of adversarial training: A contrastive perspective," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 5365–5369.

[14] C.-T. Chu, M. Rohmatillah, C.-H. Lee, and J.-T. Chien, "Augmentation strategy optimization for language understanding," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7952–7956.

[15] Z. Wang and J. H. L. Hansen, "Towards improving synthetic audio spoofing detection robustness via meta-learning and disentangled training with adversarial examples," *IEEE Access*, 2024.

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[17] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial training methods for semi-supervised text classification," in *Proc. of International Conference on Learning Representations*, 2016.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. of International Conference on Learning Representations*, 2018.

[19] L. Li, M.-W. Mak, and J.-T. Chien, "Contrastive adversarial domain adaptation networks for speaker recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 5, pp. 2236–2245, 2022.

[20] H. Wang, Z. Jin, M. Geng, S. Hu, G. Li, T. Wang, H. Xu, and X. Liu, "Enhancing pre-trained ASR system fine-tuning for dysarthric speech recognition using adversarial data augmentation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 12311–12315.

[21] J.-T. Chien and W.-Y. Sun, "Adversarial augmentation for adapter learning," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2023, pp. 1–7.

[22] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[23] C. Zhu, Y. Cheng, Z. Gan, S. Sun, T. Goldstein, and J. Liu, "FreeLB: Enhanced adversarial training for natural language understanding," in *Proc. of International Conference on Learning Representations*, 2019.

[24] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. of International Conference on Machine Learning*, 2019, pp. 7472–7482.

[25] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao, "SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2177–2190.

[26] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. of Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210.