BOOSTING ACTIVE LEARNING FOR SPEECH RECOGNITION WITH NOISY PSEUDO-LABELED SAMPLES

Jihwan Bang[2*], *Heesu Kim*[1,3*], *YoungJoon Yoo*[1], *Jung-Woo Ha*[1]

[1]Clova AI Research, NAVER Corp.
[2]Search Solution Inc.
[3]Dept. of Electrical and Computer Engineering, Seoul National University.
{jihwan.bang, heesu.kim89, youngjoon.yoo, jungwoo.ha}@navercorp.com

ABSTRACT

The cost of annotating transcriptions for large speech corpora becomes a bottleneck to maximally enjoy the potential capacity of deep neural network-based automatic speech recognition models. In this paper, we present a new training pipeline boosting the conventional active learning approach targeting label-efficient learning to resolve the mentioned problem. Existing active learning methods only focus on selecting a set of informative samples under a labeling budget. One step further, we suggest that the training efficiency can be further improved by utilizing the unlabeled samples, exceeding the labeling budget, by introducing sophisticatedly configured unsupervised loss complementing supervised loss effectively. We propose new unsupervised loss based on consistency regularization, and we configure appropriate augmentation techniques for utterances to adopt consistency regularization in the automatic speech recognition task. From the qualitative and quantitative experiments on the real-world dataset and under real-usage scenarios, we show that the proposed training pipeline can boost the efficacy of active learning approaches, thus successfully reducing a sustainable amount of human labeling cost.

Index Terms: speech recognition, active learning, semi-supervised learning, consistency regularization.

1. INTRODUCTION

End-to-End Automatic Speech Recognition (E2E-ASR) models [1, 2, 3, 4] have achieved impressive improvements in Large Vocabulary Automatic Speech Recognition (LVASR). However, although they achieve state-of-the-art performance [5, 6], the methods require more number of samples, decreasing the economical efficiency considering the high labeling cost of the speech data. The cost to annotate labels might be more troublesome in ASR because the cost to transcribe utterances is more expensive due to its error-prone property compared to simple classification problems such as object class for image classification. The reason E2E-ASR models require enormous data stems from the fact that they are trained in end-to-end without strong inductive bias such as explicit acoustic and language models while having lots of model parameters [5]. Therefore, maximizing the training efficiency in labeling cost is necessary for the state-of-the-art E2E-ASR model.

Active Learning (AL) approach, has been studied to reduce the labeling cost by selecting samples most effective for a model training from many unlabeled candidates.The selected samples are annotated by human experts, so *Human-Labeled Samples (HLS)* become the important anchors in training the model. However, the number of HLS is restricted due to the labeling budget, so we usually cannot get sufficient amount of the labeled data for model capacity. Furthermore, even the definitions of effectiveness are different among AL studies, the consensus is that the effective samples for training are in most case unfamiliar and uncertain ones to the current model. Therefore, even HLS complements the model to handle unfamiliar samples, it cannot fully exploit the potential of E2E-ASR models because of constrained labeling budget and bais existing in the selected HLS.

To mitigate such problems in AL without additional labeling cost, we propose to utilize the unlabeled samples which are not selected for HLS. Inspired by *Semi-Supervised Learning (SSL)*, we use the unlabeled samples, relatively familiar and confident in view of the current model contrary to HLS, by generating their pseudo-labels (*Pseudo-Labeled Samples (PLS)*) and appending the samples to the training dataset.

Unfortunately, simply introducing PLS would not lead to the improvement of model performance mainly because of the two reasons. One is that if PLS are selected conservatively, PLS are too familiar to model, so they do not incur any effective variation in model parameters after training. the other is that if PLS are selected speculatively, they are likely to have noisy pseudo-labels, consequently hurting model performance.

Therefore, in this paper, we propose a training pipeline to

*Authors contributed equally to this research. The authors are sorted by alphabetical order.

This work is done while Heesu Kim did internship at Clova AI Research, NAVER Corp.

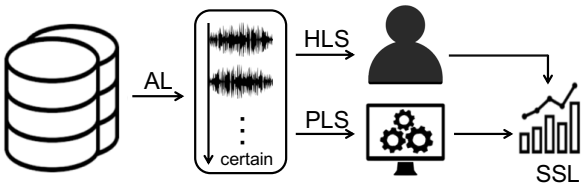


Fig. 1: An overview of the training pipeline proposed in this paper.

overcome the limitations of both AL and SSL.To this end, we introduce *Consistency Regularization (CR)* [7, 8, 9] technique which regularizes the side-effect of noisy pseudo-labels by forcing a model to predict consistently on both of genuine and distorted versions of a sample. The experimental results suggest that our new training pipeline can fully utilize PLS in training. The overview of the training pipeline proposed in this paper is depicted in Figure 1.

CR is mostly applied in computer vision tasks [7, 8, 9] and CR has not been actively considered in ASR task because of its incompatibility to distortions (i.e., augmentations), which has been reported to be effective for vision tasks. However, we show that appropriate augmentations, which do not corrupt essential linguistic information, can enable CR on utterances for ASR. By introducing loss for CR to train objective with the carefully configured augmentations, our training pipeline restores the degradation of performance caused by restricted labeling budget in AL without any additional labeling cost. Consequently, it achieves a significant reduction of the labeling cost, as well as minimizing the performance degradation.

To validate the proposed training pipeline, we conduct extensive experiments on the real-world samples acquired from services deployed to end-users, which provides voice search and voice control for IoTs. The amount of collected samples is about 500 hours of utterances recorded from various devices and users. Comparing with conventional AL on such a dataset, our proposed training pipeline boosts the performance of the model by 12.76% and 4.02% in Character-level Error Rate (CER) when the labeling budget is 1/3 and 1/10 of unlabeled samples, respectively.

In summary, our contributions to achieving such an objective can be summarized as threefold: 1) this work adopts consistency regularization on samples with noisy pseudo-labels in E2E-ASR model training to boost the effect of active learning for label-efficiency. 2) we configure the feasible augmentations for utterances to adapt consistency regularization for ASR, and 3) we verify and analyze the efficacy of the proposed training pipeline including consistency regularization with extensive real-world data and realistic environments.

2. RELATED WORKS

Active Learning for ASR: Studies on AL can be categorized into three major approaches in how they select the samples to be annotated by human experts: uncertainty-based approaches [10, 11, 12, 13, 14, 15], diversity-based approaches [16, 17], and expected-model-change approaches [18, 19]. However, for ASR, predicting uncertainty or diversity for utterances is more difficult than those of images, because transcription is configured as a sequence of labels. It is required to compute uncertainty or diversity for a sample by jointly considering all labels consisting of the sample. The studies [20, 21] demonstrate that the length-normalized path-probability from the decoder in E2E-ASR model can successfully represent the uncertainty of a sequence of labels, and the works [18, 19] propose the approximate metrics representing the expected-model-change for ASR task.

Semi-Supervised Learning with Pseudo Labeling: SSL [22, 23, 24] provide practical ways to data-hungry deep neural network models by extending training dataset from enormous unlabeled samples without supervision. One of their main approaches is generating pseudo-labels [23, 24] for unlabeled samples by models. Moreover, the works [25, 26, 27] have studied ways to adopt such a pseudo-labeling approach to ASR, and present competitive or even superior results [26] with well-designed training algorithms.

There have been some works [11, 12] trying to achieve synergies from combined AL and SSL. However, the task considered in [11] was a type of call-type classification assigning one or more independent call-type(s) to each utterance, not adequate for ASR aligning a sequence of labels to each utterance. Furthermore, the work [12] considered an only acoustic model by maximizing the lattice entropy reduction, while we target E2E-ASR model consisting of both acoustic model and language model.

Consistency Regularization: Recently, Consistency Regularization (CR) techniques [28, 29, 7, 8, 30] have been actively studied for SSL. They achieve the state-of-the-art results in situation of extremely small ratio of dataset are HLS and the other samples in dataset are PLS. Because CR forces model to keep their prediction even distortions are applied to input samples, it additionally impose an unsupervised objective to the supervised objective using labels. Still, the studies of consistency regularization have not been popular in the ASR task because of the inherent fragility of the utterances on distortions against the robustness of images under distortions. To resolve such problems, we introduce the appropriate augmentations which distorting acoustic features of utterances while minimizing the distortions on their linguistic information so that ASR models can enjoy the same benefits from CR as it did in those of computer vision tasks.

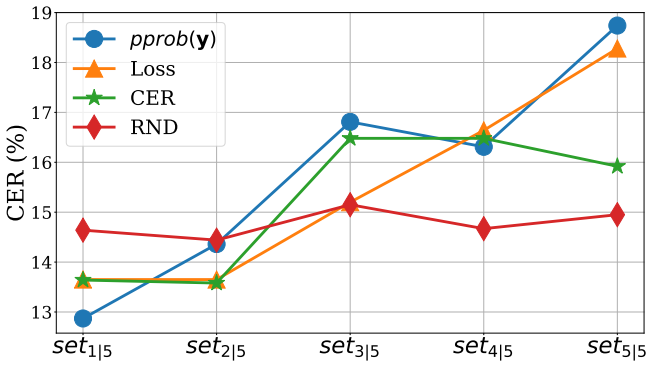


Fig. 2: Character-Level Error Rate (CER) of model trained by each subset, which is split from all unlabeled samples in equal sizes (386.5/5 = 77.3 hours) after sorted by each uncertainty metric. $set_{1|5}$ consists of the most uncertain samples according to each uncertainty metric.

3. UNCERTAINTY-BASED ACTIVE LEARNING

3.1. Length-Normalized Path-Probability

Here, we use an uncertainty-based approach in AL since it is relatively simple, but leads to a substantial reduction in labeling cost with marginal performance degradation [13, 15]. Therefore, we select the most uncertain (i.e., the most effective for training) samples from unlabeled samples. However, unlike single-label classification tasks where the uncertainty can be calculated simply with the top-1 class posterior probability such as image-classification, ASR requires to consider the joint probability of a sequence of labels (i.e., transcription).

The path-probability for a decoded sequence of labels calculated at the decoder of E2E-ASR models is the most straight-forward metric since it represents the joint probability of the decoded labels. Moreover, we can easily improve the quality of path-probability through beam-search decoding considering multiple paths during decoding and several normalization techniques including length-normalization [20, 21]. Therefore, we utilize the beam-search decoding with width=5 and design a length-penalty ($lp(\mathbf{y})$) for the length-normalization following [31]. The length-normalized path-probability ($pprob(\mathbf{y})$) for a given sample ($\mathbf{x}$) is calculated by E2E-ASR model as following equation 1.

$$pprob(\mathbf{y}) = \max_{\mathbf{y}} \log P(\mathbf{y}|\mathbf{x}) \;/\; lp(\mathbf{y}) \tag{1}$$

$$lp(\mathbf{y}) = \frac{(5+|\mathbf{y}|)^{1.2}}{(5+1)^{1.2}} \tag{2}$$

Here, $\mathbf{x} = (x_1, x_2, ..., x_T)$ is acoustic features of an input utterance and $\mathbf{y} = (y_1, y_2, ..., y_L)$ is the decoded sequence of labels by E2E-ASR model. The log-joint-probability ($\log P(\mathbf{y}|\mathbf{x})$) is divided by the length-penalty ($lp(\mathbf{y})$) for the length-normalization.

3.2. Selecting Samples to Be Annotated (HLS)

Before selecting samples to be annotated from unlabeled samples, we first calculate $pprob(\mathbf{y})$ over all unlabeled samples using the current model. After calculating the probability, we select the samples having the lowest $pprob(\mathbf{y})$ as many as a labeling budget allows, then annotate them via human experts. After the annotation, they become HLS in the training dataset.

3.3. Comparison of Uncertainty Metrics

To verify the superiority of our uncertainty metric ($pprob(\mathbf{y})$), we compare $pprob(\mathbf{y})$ with the oracle uncertainty metrics such as supervised loss (*Loss*) and performance metric (*CER*) as upper bound, which require ground-truth labels, and random sampling *RND* as lower bound.

Figure 2 illustrates CERs on the validation set over models trained from training subsets (i.e., five subsets as depicted along the x-axis) where each subset is configured by equally dividing the samples sorted in descending order according to each uncertainty metric. That is, $set_{1|5}$ contains top 1/5 uncertain samples and $set_{5|5}$ contains bottom 1/5 uncertain samples for each uncertainty metric. Hence, CERs from models trained with each $set_{1|5}$ represent the lowest CER for each metric with 1/5 labeled samples. As seen in Figure 2, $pprob(\mathbf{y})$ shows the lowest CER at $set_{1|5}$ and it also shows the expected CER changes across five subsets where CER monotonically increases as less uncertain subsets are used for training. In contrast to $pprob(\mathbf{y})$, *CER* and *RND* show the unexpected changes across the subsets since they might not measure a joint probability of decoded labels, instead just measuring the discrepancy of predictions w.r.t ground-truth without considering the dependency between labels in a sequence.

4. SEMI-SUPERVISED LEARNING WITH CONSISTENCY REGULARIZATION

4.1. Pseudo-Labeling

To boost the training efficiency over AL in section 3, we exploit the samples which are not selected for HLS and remained in unlabeled state. Since they do not have labels and there is no additional budget for labeling after annotating HLS, we generate pseudo-label for each sample by model. We use the most probable decoded labels ($\tilde{\mathbf{y}}$) defined in equation 3 as pseudo-label for a sample.

$$\tilde{\mathbf{y}} = \arg\max_{\mathbf{y}} \log P(\mathbf{y}|\mathbf{x}) \;/\; lp(\mathbf{y}) \tag{3}$$

However, the pseudo-labels are likely to be not only less informative but noisy compared to labels of HLS, so PLS would not contribute to model training, or it rather hinders model performance by giving incorrect information to model [32]. Therefore, we decide to introduce consistency regularization on PLS in model training. It means that E2E-ASR models should predict consistent decoded labels ($\mathbf{y}$), regardless of whether *data augmentations* are applied to PLS or not.

4.2. Data Augmentation for Utterances

Basically, applying effective data augmentations on training samples improves the robustness of the trained model on a variety of sample conditions that will face in real usage, since the model was already exposed on extensive distortions during training. However, to achieve such an effect, data augmentations should maintain the essential semantics of samples that determining their labels while maximally distorting non-essential parts as possible.

Contrary to images where reshaping operations such as scaling, flipping, and rotating hardly change the essential semantics for determining labels, the essential semantics for linguistic information contained in speech is much vulnerable to such basic reshaping operations [33]. Furthermore, the faults in an early part of decoding incur subsequent faults in the following decoding processing.

Because of such reasons, effective data augmentations for utterances are a critical part to adopt CR for ASR. Therefore, we examine two acoustic-specific augmentations; changing playing speed (SPEED) [34] and pitch-shifting (PITCH) [33], which effectively improve the robustness of ASR models while not destructing the essential semantics of utterances. In addition, domain-independent data augmentations such as adding white noise to a sample and randomly masking parts of a sample [35] are also considered, so we examine two additional data augmentations; Adding White Gaussian Noise (AWGN) and Specaugment [6] showing significant improvement in E2E-ASR training.

4.3. Consistency Regularization Loss

CR can be realized by adding an unsupervised training objective (i.e., loss, $\mathcal{L}_{CR}$) for training. Such objective plays a role of regularization against existing supervised training objective ($\mathcal{L}_{SUP}$), thus complementing the supervised training objective for better robustness and generalization performance [7]. Therefore, we conjecture that $\mathcal{L}_{CR}$ would alleviate the side-effects incurred by $\mathcal{L}_{SUP}$ on PLS with noisy pseudo-labels.

As mentioned before, our resultant training objective consists of two objectives: the supervised loss ($\mathcal{L}_{SUP}$) on both HLS and PLS and the unsupervised loss ($\mathcal{L}_{CR}$) on PLS. The supervised loss is defined in equation 4 following the standard cross-entropy (H) loss as did in [1].

$$\mathcal{L}_{SUP} = \frac{1}{\sum_{n=1}^{B} L_n} \sum_{n=1}^{B} \sum_{l=1}^{L_n} H(y_{n,l}, P(\hat{y}_{n,l}|\mathbf{x_n}))), \quad (4)$$

where B is the size of mini-batch, and L_n is the length of n-th sample. $y_{n,l}$ represents ground-truth labels in form of hard label and $P(\hat{y}_{n,l}|\mathbf{x_n})$ represents the posterior probability from the model.

The unsupervised loss is also defined in equation 5.

$$\mathcal{L}_{CR} = \frac{1}{\sum_{n=1}^{B} L_n} \sum_{n=1}^{B} \sum_{l=1}^{L_n} H(\tilde{y}_{n,l}, P(\hat{y}_{n,l}|\mathcal{A}(\mathbf{x_n}))) \quad (5)$$

It measures the inconsistency using cross-entropy (H) between pseudo-labels ($\tilde{y}_{n,l}$) from genuine input features ($\mathbf{x_n}$) and their predictions ($P(\hat{y}_{n,l}|\mathcal{A}(\mathbf{x_n}))$) from augmented input features ($\mathcal{A}(\mathbf{x_n})$). Note that we augment input features ($\mathbf{x_n}$) with augmentation function ($\mathcal{A}$) and we update pseudo-labels continuously per predefined period (Δ) in epoch while expecting that the noisiness of pseudo-labels will decrease as the training proceeds.

By integrating the supervised loss and the unsupervised loss, the resultant loss is defined as in equation 6

$$\mathcal{L} = \mathcal{L}_{SUP} + \lambda \mathcal{L}_{CR} \quad (6)$$

where λ is scaling constant for the unsupervised loss.

5. EVALUATION

5.1. Experiment Setup

Sample Pool: We validate the efficacy of our proposed training pipeline boosting AL on the real-world environment where unlabeled samples are redundant and the labeling budget is constrained. To reflect such an environment in our experiments, we prepare a sample pool containing 496 hours of samples collected from being deployed end-user applications. Firstly, we extract the 110 hours samples, which are collected ahead of the other samples, from the sample pool as initial dataset and annotate them to train a *initial model*. The left 386 hours samples are unlabeled and will be used as either of HLS or PLS according to the proposed training pipeline. Additionally, we prepare 56 hours of samples collected after the sample pool for a test. Note that we always include the initial dataset in training sets.

Model: Our model follows a variant version of LAS [1] model proposed in [36]. We stack three layers of bidirectional-LSTM for an encoder and two layers of unidirectional-LSTM for a decoder with location-aware attention module [37]. The hidden size of all LSTMs is set to 512. We generate spectrograms from the samples using the hamming window with $200ms$ window-length, $100ms$ stride-length, then use them as the input acoustic features.

Table 1: Measured CER (%) (Lower is better) on test dataset over various labeling budgets, which are represented by the portion out of total unlabeled samples (Initial: use only the initial dataset, Full Budget: labeling all unlabeled samples). The columns represent the training pipelines 'HLS' denotes the case only using HLS, '+PLS' denotes the case joining PLS in training without CR loss and, '+PLS-τ' means adding the preliminary filtering with a threshold on '+PLS'. '+CR-*X*' denotes adding CR loss with *X* augmentation for PLS.

Labeling Budget	Initial	HLS	+PLS	+PLS-τ	+CR-S	+CR-P	+CR-A	+CR-SA	Full Budget
38.6h (1/10)	15.60	12.07	18.97	17.82	10.95	11.03	10.77	**10.53**	8.74
57.0h (1/7)		11.41	18.05	17.65	10.61	10.61	10.46	**10.40**	
77.0h (1/5)		10.70	17.46	15.49	10.21	10.35	10.09	**9.86**	
137.0h (1/3)		9.96	14.95	11.38	9.80	9.79	9.80	**9.56**	

Training: For model training, we utilized ADAM optimizer with a learning rate 0.003 for the initial model training and 0.001 for later training pipeline with $512(B)$-sized mini-batches. The learning rate was divided by 1.1 for every epoch over 50 epochs for initial model training and 30 epochs for the other parts of the training pipeline. The norm of gradients was clipped to 400 for training stability. Furthermore, we applied SpecAugment [6] to the initial dataset during training initial model, but stopped using it after then since it adds unpleasant instability during training with noisy pseudo-labels. To prevent an unrecovered degradation caused by abnormal samples, we cut the unlabeled samples whose uncertainty ($pprob(\mathbf{y})$) exceeding the predefined threshold (τ) when it is required and call it *preliminary filtering*.

Augmentations: When applying CR, we used four augmentation techniques configured in section 4.2; SPEED (**S**), PITCH (**P**), AWGN (**A**), and SpecAugment (**SA**). They distort the samples by fast-forwarding $1.5\times$, shifting two half-steps when an octave is divided into twelve half-steps, adding Gaussian noise with SNR=5, and masking spectrograms with (40, 27, 2, 2) hyperparameters which are the width of time masking and frequency masking, the number of time masks, and the number of frequency masks, respectively.

5.2. Comparison of Training Pipeline

The training pipelines we compare here are using only HLS (HLS), appending PLS without CR (+PLS and +PLS-τ), and with CR (CR-{+S, A, P, SA}, **Ours**) across the various labeling budgets represented by the portion of total unlabeled samples. So, the samples up to the labeling budget from the unlabeled dataset become HLS and the others become PLS. We use $\lambda = 1$ and $\Delta = 1$ here, which were set to utilize PLS aggressively in training.

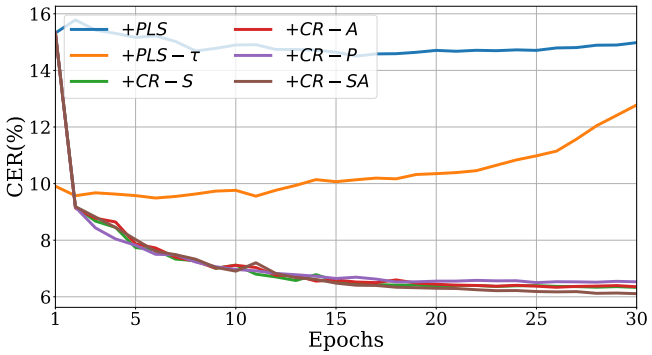
Table 1 summarizes the resultant CERs measured over pipelines. Firstly, we can see that the proposed +CR-*X*s outperform the other pipelines over every labeling budgets, and especially +CR-SA achieves the best CERs among +CR-*X*s. Secondly, we can see that +PLS achieves the worse CERs than those of HLS in all labeling budgets even it sees additional samples (i.e. PLS) during training. It explains our hypothesis mentioned in section 4.1 that noisy pseudo-labeled samples rather hinders the model training. To resolve such a problem, we try adopting the preliminary filtering mentioned at section 4.3 to filter out the samples having too noisy pseudo-labels, so +PLS-τ with $\tau = -0.5$ achieves the better CERs, but it still is worse than those of our proposed +CR-*X*s utilizing the noisy samples effectively other than abandoning them. We discuss this observation in the following subsection.

The gains of +CR-*X*s over the other pipelines are impressive as the labeling budget is smaller. For example, +CR-SA reduces 1.54%p with 1/10 budget, but only reduced 0.4%p with 1/3 budget compared to HLS. That is because the portion of PLS is more dominant under less labeling budget.

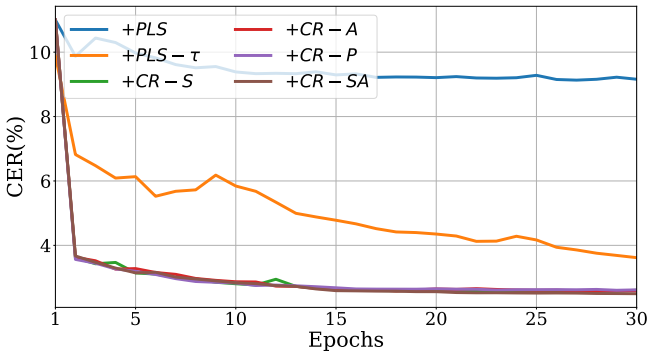
5.3. Efficacy of Consistency Regularization

We conjectured that consistency regularization alleviates side-effects from noisy pseudo-labeled samples. To verify the conjecture, we analyze the error (P-CER) of generated pseudo-labels over the various periods, which is measured by computing CER between pseudo-labels and ground-truth labels.

Figure 3 shows P-CER over training epochs. We can see that +PLS and +PLS-τ have the apparently worse P-CER than those of +CR-Xs and this relationship resembles that of CERs reported in Table 1. Moreover, P-CER of +PLS does not show any improvement throughout training epochs. Such observations have confirmed that utilizing noisy pseudo-labels does not directly improve training efficiency. When we apply the preliminary filtering on +PLS, +PLS-τ shows the better P-CER and promising dynamics in 1/3 labeling budget, but it does not work in the case of 1/10 labeling budget where the more difficult samples have remained in the unlabeled dataset. On the other hand, the goodness of P-CER for +CR-Xs supports the conjecture that CR loss in training objectives regularize the supervised loss, thus alleviating the side-effects caused by noisy pseudo-labels.



(a)



(b)

Fig. 3: The error of pseudo-labels (P-CER) measured in CER (%) with (a) 1/7 and (b) 1/3 of labeling budget.

5.4. Heterogeneous Domains

We also verify the efficacy of our proposed training pipeline on the more realistic and harsh environment where the domains of the initial dataset and unlabeled dataset are significantly different; *AIAsst* (AI Assistant) and *MapNavi* (Map Navigation Voice Control). To this end, we configure an initial dataset (443h) from AIAsst, an unlabeled dataset (366h), and a test dataset (33.2h) from MapNavi. They have a lot of different vocabulary. For example, MapNavi contains many nouns for addresses not included in AIAsst.

As shown in table 2, the initial model (Initial) could not handle MapNavi samples since it was only exposed to the samples from a very different domain, AIAsst. However, supported by small HLS and our proposed training pipeline, the model restored its performance close to that of the full budget where all samples from MapNavi were used. The used training pipeline was +CR-SA-τ, which applying the preliminary filtering with $\tau = -0.5$ to +CR-SA by considering the harshness of heterogeneous domains.

5.5. Frequency of Pseudo-Labeling

The pseudo-labeling procedure takes a large portion of total training time since it contains the computationally intensive beam-search decoding. Therefore, we consider the periodic pseudo-labeling, which reduces the number of beam-search decoding. Figure 4 shows that the dynamics of CERs on test dataset with 1/10 labeling budget on several periods (Δ) (1, 3, and 5 epochs) and we can confirm that the degradation with a longer period is marginal (0.2%p difference between $\Delta = 1$ and $\Delta = 5$) as the training epochs proceeds even though we use the strongly constrained labeling budget. So, we can efficiently tradeoff freshness of pseudo-labels with training time when the amount of unlabeled samples is too large.

Table 2: Measured CER(%) on test dataset from MapNavi when the initial and the unlabeled dataset consist of samples from AIAsst and MapNavi, respectively.

Labeling Budget	Initial	+CR-SA-τ	Full Budget
38.6h (1/10)	49.25	10.62	7.55
57.0h (1/7)		9.27	
77.0h (1/5)		8.58	
137.0h (1/3)		7.74	

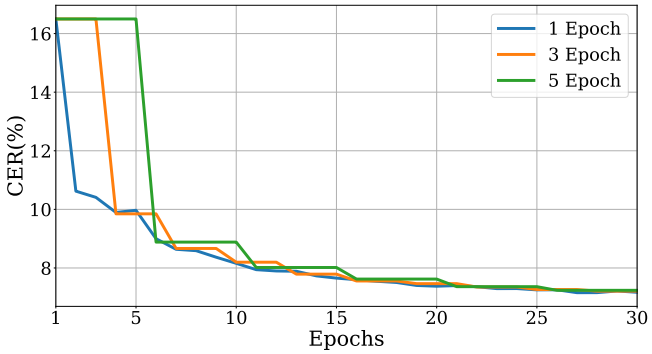


Fig. 4: The error of pseudo-labels (P-CER) over periods (Δ) for pseudo-labeling using +CR-SA and 1/10 labeling budget.

6. CONCLUSIONS

In this paper, we proposed the training pipeline boosting active learning under constrained labeling budget by incorporating semi-supervised learning with pseudo-labeling and consistency regularization. We showed that consistency regularization with well-configured augmentations effectively exploited unlabeled samples, which are not considered in active learning, by regulating the side-effects caused by noisy pseudo-labels. Our proposed training pipeline (+CR-SA) improved CERs by 12.76% and 4.02% compared to active learning (HLS) when labeling budgets cover 1/3 and 1/10 of total unlabeled samples, respectively. Moreover, we achieve the competitive performance (0.82%p worse) with 1/3 amount of samples (137 vs 386 hours). We highlight that this is the first work adopting the consistency regularization into ASR task and the results present the potential to remarkably reduce the performance degradation with insufficient labeling budget.

7. REFERENCES

[1] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[2] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[3] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[4] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al., "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.

[5] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[6] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[7] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv preprint arXiv:2001.07685*, 2020.

[8] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 5050–5060.

[9] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10687–10698.

[10] David D Lewis and Jason Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Machine learning proceedings 1994*, pp. 148–156. Elsevier, 1994.

[11] Dong Yu, Balakrishnan Varadarajan, Li Deng, and Alex Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech & Language*, vol. 24, no. 3, pp. 433–444, 2010.

[12] Gokhan Tur, Dilek Hakkani-Tür, and Robert E Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, no. 2, pp. 171–186, 2005.

[13] Donggeun Yoo and In So Kweon, "Learning loss for active learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 93–102.

[14] Yarin Gal and Zoubin Ghahramani, "Bayesian convolutional neural networks with bernoulli approximate variational inference," *arXiv preprint arXiv:1506.02158*, 2015.

[15] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler, "The power of ensembles for active learning in image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9368–9377.

[16] Ozan Sener and Silvio Savarese, "Active learning for convolutional neural networks: A core-set approach," *arXiv preprint arXiv:1708.00489*, 2017.

[17] Hieu T Nguyen and Arnold Smeulders, "Active learning using pre-clustering," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 79.

[18] Jiaji Huang, Rewon Child, Vinay Rao, Hairong Liu, Sanjeev Satheesh, and Adam Coates, "Active learning for speech recognition: the power of gradients," *arXiv preprint arXiv:1612.03226*, 2016.

[19] Y. Yuan, S. Chung, and H. Kang, "Gradient-based active learning query strategy for end-to-end speech recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2832–2836.

[20] Karan Malhotra, Shubham Bansal, and Sriram Ganapathy, "Active Learning Methods for Low Resource End-to-End Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2215–2219.

[21] Jacob Kahn, Ann Lee, and Awni 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, pp. 7084–7088.

[22] Jesper E Van Engelen and Holger H Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[23] Dong-Hyun Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, 2013, vol. 3, p. 2.

[24] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," *arXiv preprint arXiv:1908.02983*, 2019.

[25] Thomas Drugman, Janne Pylkkonen, and Reinhard Kneser, "Active and semi-supervised learning in asr: Benefits on the acoustic and language models," *arXiv preprint arXiv:1903.02852*, 2019.

[26] Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Edouard Grave, Tatiana Likhomanenko, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert, "End-to-end asr: from supervised to semi-supervised learning with modern architectures," *arXiv preprint arXiv:1911.08460*, 2019.

[27] Yang Chen, Weiran Wang, and Chao Wang, "Semi-supervised asr by end-to-end self-training," *arXiv preprint arXiv:2001.09128*, 2020.

[28] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel, "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," *arXiv preprint arXiv:1911.09785*, 2019.

[29] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Advances in neural information processing systems*, 2016, pp. 1163–1171.

[30] Samuli Laine and Timo Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[31] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[32] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[33] Jan Schlüter and Thomas Grill, "Exploring data augmentation for improved singing voice detection with neural networks.," in *ISMIR*, 2015, pp. 121–126.

[34] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[35] Terrance DeVries and Graham W Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[36] Jung-Woo Ha, Kihyun Nam, Jin Gu Kang, Sang-Woo Lee, Sohee Yang, Hyunhoon Jung, Eunmi Kim, Hyeji Kim, Soojin Kim, Hyun Ah Kim, et al., "Clovacall: Korean goal-oriented dialog speech corpus for automatic speech recognition of contact centers," *arXiv preprint arXiv:2004.09367*, 2020.

[37] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.