Continuous Soft Pseudo-Labeling in ASR

Tatiana Likhomanenko, Ronan Collobert, Navdeep Jaitly, Samy Bengio
Apple
{antares,collobert,njaitly,bengio}@apple.com

Abstract

Continuous pseudo-labeling (PL) algorithms such as slimIPL have recently emerged as a powerful strategy for semi-supervised learning in speech recognition. In contrast with earlier strategies that alternated between training a model and generating pseudo-labels (PLs) with it, here PLs are generated in end-to-end manner as training proceeds, improving training speed and the accuracy of the final model. PL shares a common theme with teacher-student models such as distillation in that a teacher model generates targets that need to be mimicked by the student model being trained. However, interestingly, PL strategies in general use hard-labels, whereas distillation uses the distribution over labels as the target to mimic. Inspired by distillation we expect that specifying the whole distribution (aka soft-labels) over sequences as the target for unlabeled data, instead of a single best pass pseudo-labeled transcript (hard-labels) should improve PL performance and convergence. Surprisingly and unexpectedly, we find that soft-labels targets can lead to training divergence, with the model collapsing to a degenerate token distribution per frame. We hypothesize that the reason this does not happen with hard-labels is that training loss on hard-labels imposes sequence-level consistency that keeps the model from collapsing to the degenerate solution. In this paper, we show several experiments that support this hypothesis, and experiment with several regularization approaches that can ameliorate the degenerate collapse when using soft-labels. These approaches can bring the accuracy of soft-labels closer to that of hard-labels, and while they are unable to outperform them yet, they serve as a useful framework for further improvements.

1 Introduction and Related Works

The Deep Learning community has recently made significant progress in semi-supervised learning via self-training or pseudo-labeling (PL) [1] in vision, speech and other domains [2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. In this paper we focus on self-training in the context of automatic speech recognition (ASR). ASR presents some unique challenges in self-training that are not seen in some other tasks stemming from their sequential nature. Inputs and targets have different lengths and targets are not produced with the same regularity, compared to the inputs. Further, ASR models are often coupled with language models, that can be additionally used to aid the pseudo-labeling process.

In semi-supervised learning [12] we have access to labeled and unlabeled data, where the unlabeled data is typically much larger, since it is often easier to find unlabeled data, compared to labeled data. PL, in its simplest formulation, employs a base model trained with labeled data which is later used as a "teacher" to generate predictions on the unlabeled data. The unlabeled data is converted to "pseudo-labeled" data through the addition of these teacher generated targets. A student model is then trained on the combination of labeled and pseudo-labeled data to yield a final model.

PL with ASR has recently been applied to three different types of end-to-end models – Connectionist Temporal Classification (CTC) [13], Transducer [14], and sequence-to-sequence [15]. To achieve state-of-the-art results, PLs are generated from the models using an external language model (LM)

I Can't Believe It's Not Better Workshop at NeurIPS 2022.

either with beam-search decoding [16, 6, 17] or through shallow fusion [18, 19, 20, 21]. Unlike in vision, here PLs are sequences of tokens, such as characters[1]. With this setting, however, acoustic models tend to overfit to the text training set of the LM used for PL [6, 20, 11].

Recent approaches on PL in ASR have attempted to rectify this problem through continuous training [22, 5], in which the model that is being trained is the same model that is used to generate PLs, and PLs are generated without using an LM. The main challenge for continuous PL is stability, which was reported independently by several authors [7, 9, 10, 11] and resolved partially by introducing either a caching mechanism [7] which kept an evolving cache of PL data or by using an Exponential Moving Average (EMA) of the model being trained to generate the PLs [8, 9, 10, 11]. This continuous training greatly simplifies the pipeline, improves convergence speed and achieves state-of-the-art compared to other semi-supervised approaches [7].

Interestingly, the above strategies in general use one (or a small number) of PLs from the teacher to train the student. By contrast distillation [23] and soft-labeling [3, 4] approaches use the predicted distribution over labels from the teacher as the target to mimic. Inspired by these works we expect that specifying the whole distribution (aka soft-labels) over sequences as the target for unlabeled data should improve performance and convergence. However, with CTC-based models for labeled data we find experimentally that training with soft-labels leads to models where the distribution converges to a degenerate solution, e.g. emitting only the blank token (this only happens in continuous pseudo-labeling, while teacher-student exhibit stable soft-labeling, see Appendix E.2). By contrast, hard-labels are more stable and collapse to a degenerate solution is not observed. We believe this is because structured CTC loss used for hard-labels imposes sequence-level consistency while loss on soft-labels (cross-entropy or L_2) does not have this property, could destroy the sequence-level representation and thus easily converge to the degenerate solution, quite bad for generalization. We apply different regularization approaches to try and prevent degenerate solutions such as entropy regularization, smoothing of the soft-labels, sampling from the teacher distribution [24, 25], blending the soft-labels with hard-labels and PLs filtering. These approaches can bring the accuracy of soft-labels closer to that of hard-labels, and while they are unable to outperform them yet, they serve as a useful framework for further improvements.

2 Background on Pseudo-Labeling and slimIPL

Let $L = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ and $U = \{\boldsymbol{x}_j\}$ be the labeled and unlabeled datasets respectively. We consider a PL approach where an acoustic model $\mathcal{A}(\boldsymbol{x}; \boldsymbol{\theta})$ with model parameters $\boldsymbol{\theta}$ is continuously trained on a combination of L and a pseudo-labelled set derived from U. The model is trained by minimizing a loss $\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_L(\boldsymbol{\theta}) + \lambda \mathcal{L}_U(\boldsymbol{\theta})$ where $\lambda \in \mathbb{R}^+$ is a tunable hyper-parameter controlling the importance of unlabeled data. The loss for labeled data is defined as $\mathcal{L}_L(\boldsymbol{\theta}) = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim L} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$, where $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$ is the conditional distribution defined by model $\mathcal{A}(\boldsymbol{x}; \boldsymbol{\theta})$. The loss for unlabeled data is defined as $\mathcal{L}_U(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{x}\sim U} \log p_{\boldsymbol{\theta}}(\hat{\boldsymbol{y}}|\boldsymbol{x})$, where $\hat{\boldsymbol{y}}$ is the PL transcription for a data point generated using the model being trained and a prior LM. Specifically,

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} \ \log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) + \alpha \ \log p^{lm}(\boldsymbol{y}), \ \boldsymbol{x} \in U, \tag{1}$$

where $p^{lm}(\boldsymbol{y})$ is an LM, and $\alpha \geq 0$ is a hyper-parameter controlling the reliance on the LM. The above minimization is usually approximated with a beam-search decoding algorithm [19]. In recent works it was found that $\alpha > 0$ gives stable training but can lead to over-fitting to the language model used [6] while $\alpha = 0$ can lead to training instability and divergence [7].

To resolve these issues an LM free ($\alpha = 0$) continuous PL approach called slimIPL was recently proposed for ASR [7]. It also optimizes further the speed of the PL generation: instead of Eq. (1) with $\alpha = 0$ which defines hard-transcriptions, slimIPL uses hard-alignment defined in Eq. (2). Models trained with the CTC loss output distributions over token set $\boldsymbol{w}$ (including an auxiliary blank token) for every output frame, or timestamp, $t = \overline{1, T}$ as $p^t_{\boldsymbol{\theta}}(w|\boldsymbol{x}), w \in \boldsymbol{w}$, where T is the total number of frames in the sample. The CTC loss maximizes the probability $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$ of a transcription $\boldsymbol{y}$ across all possible alignments[2] $\boldsymbol{a} = (a_1, \ldots, a_T)$ which lead to $\boldsymbol{y}$ assuming

[1] We refer to them as hard-labels because only some sequences are used and not the whole distribution over all sequences.

[2] E.g. for $T = 12$ one of the alignments for $\boldsymbol{y}$ 'cat' is 'cc###aatttt#', where # is a CTC blank token.

factorization $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\boldsymbol{a}\to\boldsymbol{y}} \prod_{t=1}^{T} p_{\boldsymbol{\theta}}^{t}(a_t|\boldsymbol{x})$. In order to get PLs generated by Eq. (1) ($\alpha = 0$) we would need to compute the total probability of the transcription $\boldsymbol{y}$ by summing probabilities of every alignment $\boldsymbol{a} \to \boldsymbol{y}$ and then taking the most probable $\boldsymbol{y}$. This is intractable but beam search could be used in this case in order to approximate the argmax search. With slimIPL, the authors decided to use a simpler approach – hard-alignment (or hard-path), obtained with:

$$\hat{\boldsymbol{y}} = \mathcal{G}(\hat{\boldsymbol{a}}), \qquad \hat{\boldsymbol{a}} = (\hat{a}_1, \ldots, \hat{a}_T), \qquad \hat{a}_i = \operatorname{argmax}_w p_{\boldsymbol{\theta}}^{i}(w|\boldsymbol{x}), \tag{2}$$

where for every frame the most probable token is $\hat{a}_i$. From there, the final pseudo-label transcription $\hat{\boldsymbol{y}}$ is obtained through the transformation denoted by $\mathcal{G}$, which removes repeated tokens as well as the CTC blank token[3]. This is equivalent to minimizing Eq. (1) with a (lexicon-free) beam search constrained to a beam size of 1.

A schematic overview of slimIPL is presented in Appendix A Algorithm 1. slimIPL exploits a dynamic cache, which cleverly facilitates access to PLs from previous model states (this can be viewed as model ensemble for PL generation). This dynamic cache was proposed in order to stabilize the optimization and to avoid sudden model divergence due to collapsing to empty transcription predictions. slimIPL has several hyper-parameters: (i) λ, (ii) M, the training step at which PL generation begins, (iii) C, the dynamic cache size and (iv) p, the probability of updating the cache.

Our goal in the next sections is to investigate how "soft"-labeling can further improve slimIPL as we believe it may contain richer information compared to hard-labels. Interestingly, "soft"-labeling is computationally cheap, similar to the cost implied by hard-paths obtained from Eq. (2), and much cheaper than with pseudo-labels which would be obtained by minimizing Eq. (1) (via a beam search) but requires more memory. We start our analysis from the best models found in slimIPL work [7].

3 Experimental Setup and Methods

Inspired by results on soft-labeling in vision [3, 4] and knowledge distillation [23] our empirical hypothesis is the following: exploiting the whole distribution over sequences should be more powerful for PL and bring benefit in either faster convergence and / or lower error at the end of training. In this section we describe the data, the training setup and methods used to test this hypothesis.

Data We use the LibriSpeech dataset [26] and the *train-10h* subset of Libri-Light [27]. We use the *train-clean-360* and *train-other-500* standard LibriSpeech subsets as unlabeled data and consider either 10h (*train-10h*) or 100h (*train-clean-100*) as the labeled data. The standard sets *dev-clean* and *dev-other* are used to tune all hyper-parameters and to select the best models. The sets *test-clean* and *test-other* are used to report final token (TER) and word (WER) error rates.

Training setup We closely follow slimIPL experiments [7] with the same hyper-parameters, including transformer model, CTC loss and character tokens (see Appendix B). We make only few changes: a) to speed up training and decrease memory usage we get rid of relative positional embedding [28] and use CAPE [29] instead; b) we train with larger batch per GPU but on 8 GPUs. In our experiments we only vary C, p and the type of PL (soft-labels vs hard-alignment), while keeping everything else fixed. We run experiments with 3 seeds and we report mean and standard deviation as the final performance.

Soft-labeling slimIPL Even though speech recognition involves a sequence-to-sequence alignment problem, we are going to perform "soft"-labeling on a per-frame basis. In that respect, the soft-label strategy is similar than what is commonly used for non-sequential data. Formally, let's denote logits as $z_w^t = \log p_{\boldsymbol{\theta}}^t(w|\boldsymbol{x})$ and PL logits $\hat{z}_w^t = \log p_{\boldsymbol{\theta}'}^t(w|\boldsymbol{x})$ for $\boldsymbol{x} \in U$. We also denote as $q_w^t(\boldsymbol{x};\tau) = \frac{e^{z_w^t/\tau}}{\sum_{w'} e^{z_{w'}^t/\tau}}$ and $\hat{q}_w^t(\boldsymbol{x};\tau) = \frac{e^{\hat{z}_w^t/\tau}}{\sum_{w'} e^{\hat{z}_{w'}^t/\tau}}$. Then the loss function on unlabeled data $\boldsymbol{x} \in U$ is defined in a similar way than the distillation loss [23]: either as cross-entropy (CE) with temperature $\tau > 0$ or as a L_2 regression (equivalent to CE when $\tau = \infty$, see [23]):

$$\mathcal{L}_{U,ce}(\boldsymbol{\theta}) = -\beta \mathbb{E}_{\boldsymbol{x}\sim U} \sum_{t,w} q_w^t \log \hat{q}_w^t; \qquad \mathcal{L}_{U,L_2}(\boldsymbol{\theta}) = -\beta \mathbb{E}_{\boldsymbol{x}\sim U} \sum_{t,w} (z_w^t - \hat{z}_w^t)^2. \tag{3}$$

[3] This can be viewed as hard-alignment, obtained from per-frame hard-labels. In contrast, CTC decoding considers all alignments leading to the same transcription to obtain this hard-alignment.

Table 1: Comparison between different PL generation ways with their best hyper-parameters: "hard-path" denotes Eq. (2) PLs, "hard-beam" denotes approximation of Eq. (1) ($\alpha = 0$) via beam search, "soft" denotes PLs from Sec. 3. Upper and lower bounds denote training on labeled data only.

PL type	Cache	PL vs golden label		PL vs hard-alignment		WER %			
	C	TER %	WER %	TER %	WER %	dev-clean	dev-other	test-clean	test-other
lower bound (sup. LS)	-	-	-	-	-	2.6 ± 0.1	6.9 ± 0.1	2.7 ± 0.1	6.9 ± 0.1
100h, hard-path	1k	1.5 ± 0.1	4.8 ± 0.1	0.0	0.0	3.9 ± 0.1	8.6 ± 0.1	4.1 ± 0.1	9.2 ± 0.2
100h, hard-path	100	1.3 ± 0.1	4.2 ± 0.1	0.0	0.0	3.7 ± 0.1	8.0 ± 0.1	3.9 ± 0.1	8.2 ± 0.1
100h, hard-beam (beam 10)	100	1.3 ± 0.1	4.1 ± 0.1	0.007 ± 0.001	0.04 ± 0.01	3.7 ± 0.1	7.9 ± 0.1	3.9 ± 0.1	8.3 ± 0.1
100h, sampling ($\tau = 0.4$)	100	1.4 ± 0.1	4.2 ± 0.1	0.04 ± 0.01	0.2 ± 0.1	3.7 ± 0.1	8.0 ± 0.2	3.9 ± 0.1	8.3 ± 0.2
100h, soft ($\tau = 10, \beta = 1$)	1k	1.6 ± 0.1	5.3 ± 0.3	-	-	4.7 ± 0.4	9.0 ± 0.5	4.8 ± 0.4	9.5 ± 0.5
+hard ($\delta = 0.1$)	100	1.4 ± 0.1	4.5 ± 0.1	-	-	4.1 ± 0.1	8.6 ± 0.2	4.2 ± 0.1	9.2 ± 0.1
+hard ($\tau = 1, \beta = 0.01, \delta = 0.1$)	100	1.4 ± 0.1	4.4 ± 0.1	-	-	4.0 ± 0.1	8.3 ± 0.1	4.2 ± 0.1	8.7 ± 0.2
10h, hard-path	1k	4.3 ± 0.1	14.7 ± 0.3	-	-	14.4 ± 0.3	18.8 ± 0.4	15.1 ± 0.4	19.3 ± 0.3
10h, hard-beam (beam 100)	1k	3.8 ± 0.1	13.1 ± 0.1	0.008 ± 0.001	0.05 ± 0.01	13.0 ± 0.1	16.8 ± 0.2	13.5 ± 0.1	17.3 ± 0.1
10h, sampling ($\tau = 0.15$)	1k	4.3 ± 0.3	14.7 ± 1.0	0.01 ± 0.02	0.08 ± 0.02	14.2 ± 0.8	18.5 ± 1.2	15.1 ± 0.9	19.1 ± 1.1
10h, soft ($\tau = 1, \beta = 0.01$)	1k-3k	we are not able to train, blowing up							
+hard ($\delta = 0.1$)	1k	4.6 ± 0.2	15.5 ± 0.2	-	-	14.9 ± 0.4	19.8 ± 0.2	15.7 ± 0.3	20.4 ± 0.5
upper bound (sup. 100h)	-	-	-	-	-	5.9 ± 0.1	17.9 ± 0.1	6.2 ± 0.1	18.1 ± 0.1
upper bound (sup. 10h)	-	-	-	-	-	37.1 ± 0.1	58.4 ± 0.1	37.7 ± 0.3	58.4 ± 0.2

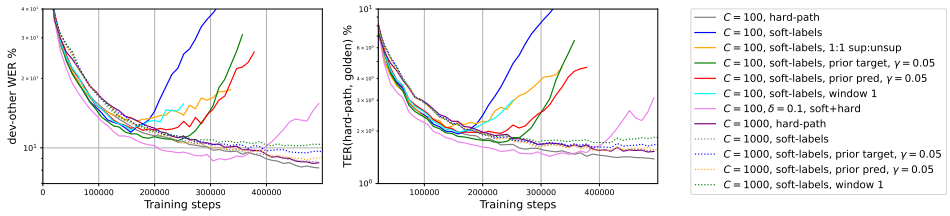

Figure 1: Comparison between different soft-labels for 100h supervision where additional regularizations are used, default parameters are $p = 0.1, \tau = 10, \beta = 1, N_U = 3$.

In order to keep the same λ (ratio between labeled and unlabeled data losses, introduced in Section 2) for both hard-labels and soft-labels experiments, we add a re-scaling factor $\beta > 0$ in Eq. (3), which provides a way to keep gradients computed for $\mathcal{L}_U$ and $\mathcal{L}_{U,ce}/\mathcal{L}_{U,L_2}$ at the same scale[4].

4 Empirical Study and Discussion

Training with soft-labels is unstable First experiments with both soft-labels losses reveal some training instability, when using the same cache size C and probability p than with hard-labels: training diverges, with WER increasing on validation sets as well. The WER between PLs and their corresponding golden transcription also diverges. By increasing the cache size by 10x (from 100 to 1k) and lowering its probability (0.1) the training is however stabilized (Figure 1). We performed similar experiments with EMA [8, 11] and observed the same trend: the EMA decay factor needs to be increased (from 0.999 to 0.9999/0.99995) compared to the one used with hard-alignments. Previously it was reported that EMA and slimIPL are robust to hyper-parameters and that using a small cache or a larger value of p would still lead to stable training, and similar WER results. All our experiments demonstrate that this is not the case with soft-labeling. Moreover, the best soft-labeling performing model ($\tau = 10$ and $\beta = 1$) has worse WER performance than when using hard-alignment, as shown in Table 1[5]. The same 1k cache for hard-alignment gives worse WER performance than with 100 cache (as PLs are then more slowly updated), but it is still better than with soft-labels.

Soft-labeling easily collapses to blank To understand instability issues in soft-labeling, we track how the PLs of a subset of the unlabeled samples evolve over the training with $C = 100, p = 0.1$,

[4]As we use adaptive optimizer, we tried also to use two different optimizers for different losses as their dynamics are different. However, experiments with two optimizers were significantly worse, not to mention it brings complexity in practice.

[5]Temperatures τ and scaling factor β were searched in the range $[0.0001, 100]$. Different cache size were tried (10, 100, 1k, 10k). Cache probability was picked in (0.1, 0.5, 1).

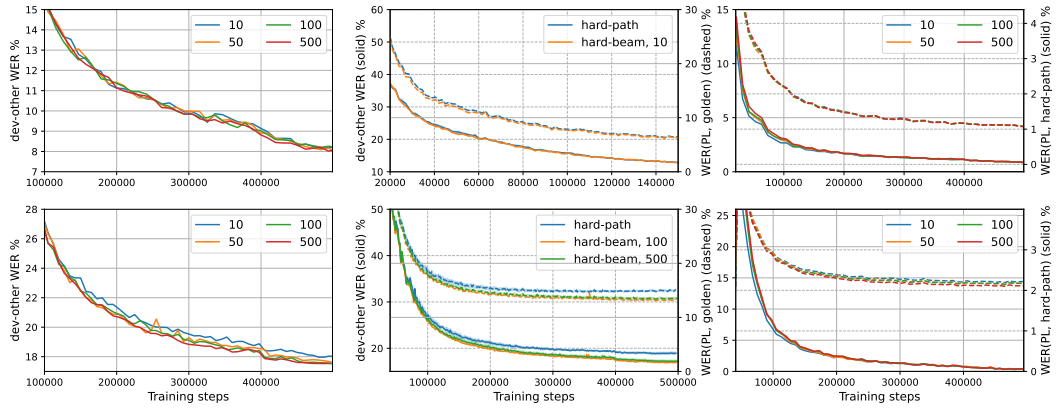

Figure 2: Comparison between hard-transcription approximated with a beam (hard-beam) of different size and hard-path for 100h (top) and 10h (bottom) of supervision. Left: *dev-other* WER for different beams; middle: best hard-beam vs hard-path across 3 different seeds on *dev-other* (solid) and their PL quality (dashed). Right: PL quality (dashed) and WER(PL, hard-path) (solid) for different beams.

see Appendix C Figure 8 (top). Interestingly, only several (2-3) tokens dominate every frame during the training. When the model diverges, e.g. the blank token probability increases over all frames, ultimately taking over the whole distribution mass of soft-labels.

Hard-labels frame-level transitions are noisy. Sequence-level ones are smooth. Compared to soft-labels, hard-path dynamics (Appendix C Figure 6) exhibit non-smooth transition between alignments. Instead of looking at each sample individually, we summarize the evolution of the same subset of unlabeled samples in Figure 4 (for hard-labels) and Figure 5 (for soft-labels). We track the average distance between the PLs generated for the same sample, for two consecutive iterations. We report the frame-level Levenshtein distance of these consecutive PLs, after converting them into hard-alignments $\hat{\boldsymbol{a}}$ (whether the PLs were soft or hard). We also report the Levenshtein distance at the sequence level $\hat{\boldsymbol{y}}$. When tracking individual samples, we observe more transition noise in hard-labels, than for soft-labels. However, at the sequence level, hard-labels PLs transitions are significantly smoother as the training goes.

Sequence-level PLs may lead to better WER We now consider Eq. (1) with beam sizes from 10 to 500 to understand if a richer sequence-level representation can be helpful, by turning off the LM ($\alpha = 0$). Figure 2 top shows that beam and hard-path perform similarly. Moreover, the WER between PLs generated from beam search and PLs generated from hard-path becomes smaller than 2% (right plot) very quickly (after 30k updates). However, Figure 2 bottom, shows that if we switch to using only 10h of supervision, we observe a 1.5-2% absolute decrease in WER from increasing the beam size to 50-100, while the PLs from beam and hard-path are still similar (WER is 1-3%).

Based on the above observations we decided to investigate several ways to improve soft-labeling: regularizing the soft-label distribution in the hope to avoid blank collapses, sampling as a way to regularize soft-labels transitions, and finally combining soft and hard-labels.

4.1 CTC blank regularization does not solve instability issue

To resolve the issue of collapsing to blank tokens we start with different regularization schemes to constrain the predictive distribution to be close to a specified prior distribution. All variants we tried slow down divergence, but were unable to improve on our best slimIPL results (Figure 1). The different prior distribution we tried are as follows:
Entropy regularization: This prior attempts to maximize the entropy of the predictive distribution.
Sequence-level prior: The average predicted probability vector across all time steps for a sample is regularized to be close to (using KL-divergence) either (i) a uniform prior across tokens, or (ii) a token distribution computed from *train-clean-100* text. Since text does not have a blank token, the prior for the blank token is estimated by a supervised model trained on *train-clean-100*.
Frame-level prior: Each frame is regularized to either match a target blank probability, or the predictions and/or targets as label-smoothed by averaging with a uniform distribution with weight γ,

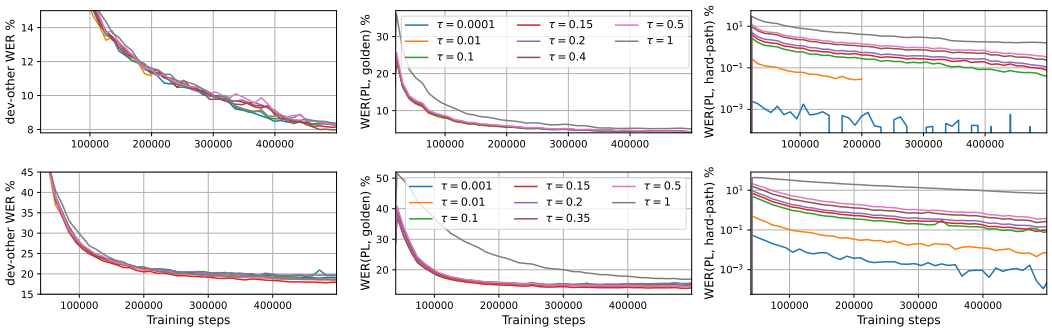

Figure 3: Sampling for 100h (top) and 10h (bottom) of supervision. WER on *dev-other* (left) depending on temperature τ and WER between transcriptions obtained with sampling (and different temperature τ) and golden transcription (left) / hard-alignment (middle).

which is varied from 0.01 to 0.1.
Logits averaging: Logits from neighboring frames are smoothed with window $w = 1 \cdots 10$ frames.

4.2 Sampling as stable soft-labeling

Next, we investigate how sampling [24, 25] from the model performs because sampling encapsulates the uncertainty in the predictions. We generate $\hat{\boldsymbol{y}} = \mathcal{G}(\hat{a}_1, \ldots, \hat{a}_T)$ where $\hat{a}_t \sim p^t_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{x})$ for $t = \overline{1, T}$. A temperature τ is varied from 0.0001 to 10 to change the peakiness of the distribution, and CTC loss is used on PL data. For $\tau \to 0$ sampling corresponds to the hard-path while $\tau \to \infty$ corresponds to the uniform distribution over tokens[6]. We can see from Figure 3 that sampling converges to hard-path in the end (top) while in limited supervision (bottom) only $\tau = 1$ does not. As with hard-transcription we observe that sampling is similar to hard-path for 100h and is better for 10h of supervision. Unlike soft-labels, sampled PLs are stable like the hard-path PLs.

4.3 Blending soft-labels with hard-labels as regularization

We observed and discussed above, when models are trained with hard-alignments, hard-transcriptions or sampling, their frame-level predictions vary a lot from step to step during training. In contrast soft-labels do not display such rapid changes in the predictive distribution (compare Appendix C Figure 4 vs Figure 5). We hypothesize that soft-labeling loss is trivial to optimize to degenerate solutions and the model over-fits to it easily while memorizing the entire labeled set (see Appendix D). Meanwhile, hard PL versions provide sequence level targets which might prevent a collapse. So we combine hard-path PL and soft-PL using $\delta\mathcal{L}_U + (1-\delta)\mathcal{L}_{U,ce}$, $\delta = \{0.01, 0.1, 0.5, 0.9, 0.99\}$. This is the most effective way to stabilize training: e.g. see Appendix E.3 Figure 15 with $\tau = 1, C = 100$. Further, with this blend models can be trained with 10h of labeled data which was not possible with soft-labels alone (see Table 1). The strategy also works better than the other soft-PL variants (Table 1) while converging 1.5-2x faster (see Figure 1). However some issues still remain as it slowly starts to diverge at the very end of training, and the results are slightly worse than hard-path.

5 Conclusion

At first glance, a change from hard-labels to soft-labels for continuous pseudo-labeling in ASR seemed simple and obvious, but our experiments revealed entirely unexpected results. Soft-labeling was found to quickly drive the PLs that are generated continuously to trivial, degenerate solutions while memorizing the entire labeled data at the same time. We found that different regularization approaches help to improve stability of soft-label training but are unable to match the accuracy of hard-label trained models. Hard-transcription and sampling from the distribution on the other hand show stability and improved results in lower supervision setting. We found that combining both hard-alignment PLs with CTC loss and soft-label PLs with cross-entropy loss on unsupervised data leads to improved results. This almost matches hard-alignment and speeds up convergence by 1.5-2x.

[6]With $\tau > 1$ a distribution is closer to uniform and models do not train.

References

[1] H. Scudder, "Probability of error of some adaptive pattern-recognition machines," *IEEE Transactions on Information Theory*, vol. 11, no. 3, pp. 363–371, 1965.

[2] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[3] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[4] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," *arXiv preprint arXiv:1911.09785*, 2019.

[5] Y. Chen, W. Wang, and C. Wang, "Semi-supervised asr by end-to-end self-training," *Proc. Interspeech 2020*, pp. 2787–2791, 2020.

[6] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative pseudo-labeling for speech recognition," *Proc. Interspeech 2020*, pp. 1006–1010, 2020.

[7] T. Likhomanenko, Q. Xu, J. Kahn, G. Synnaeve, and R. Collobert, "slimipl: Language-model-free iterative pseudo-labeling," *Proc. Interspeech*, 2021.

[8] V. Manohar, T. Likhomanenko, Q. Xu, W.-N. Hsu, R. Collobert, Y. Saraf, G. Zweig, and A. Mohamed, "Kaizen: Continuously improving teacher using exponential moving average for semi-supervised speech recognition," *arXiv preprint arXiv:2106.07759*, 2021.

[9] Y. Higuchi, N. Moritz, J. L. Roux, and T. Hori, "Momentum pseudo-labeling for semi-supervised speech recognition," *Proc. Interspeech*, 2021.

[10] Y. Higuchi, N. Moritz, J. Le Roux, and T. Hori, "Advancing momentum pseudo-labeling with conformer and initialization strategy," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7672–7676.

[11] ——, "Momentum pseudo-labeling: Semi-supervised asr with continuously improving pseudo-labels," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–14, 2022.

[12] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-supervised Learning*. Mit Press, 2010.

[13] A. Graves, S. Fernández, F. Gomez, and J. 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.

[14] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[15] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*. IEEE, 2016, pp. 4960–4964.

[16] W.-N. Hsu, A. Lee, G. Synnaeve, and A. Hannun, "Semi-supervised speech recognition via local prior matching," *arXiv preprint arXiv:2002.10336*, 2020.

[17] Q. Xu, A. Baevski, T. Likhomanenko, P. Tomasello, A. Conneau, R. Collobert, G. Synnaeve, and M. Auli, "Self-training and pre-training are complementary for speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3030–3034.

[18] J. Kahn, A. Lee, and 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, pp. 7084–7088.

[19] G. Synnaeve, Q. Xu, J. Kahn, T. Likhomanenko, E. Grave, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end asr: from supervised to semi-supervised learning with modern architectures," in *Workshop on Self-supervision in Audio and Speech, ICML*, 2020.

[20] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," *Proc. Interspeech 2020*, pp. 2817–2821, 2020.

[21] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2010.10504*, 2020.

[22] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013.

[23] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[24] K. Imamura, A. Fujita, and E. Sumita, "Enhancement of encoder and attention using target monolingual corpora in neural machine translation," in *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, 2018, pp. 55–63.

[25] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 489–500.

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

[27] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[28] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 464–468.

[29] T. Likhomanenko, Q. Xu, G. Synnaeve, R. Collobert, and A. Rogozhnikov, "Cape: Encoding relative positions with continuous augmented positional embeddings," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[30] T. Likhomanenko, Q. Xu, V. Pratap, P. Tomasello, J. Kahn, G. Avidov, R. Collobert, and G. Synnaeve, "Rethinking evaluation in asr: Are our models robust enough?" *Proc. Interspeech*, 2021.

[31] 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," *Proc. Interspeech 2019*, pp. 2613–2617, 2019.

[32] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of machine learning research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

A slimIPL Algorithm Overview

Algorithm 1: slimIPL algorithm where arbitrary algorithm for PL generation is used

Data: labeled $L = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ and unlabeled $U = \{\boldsymbol{x}_j\}$ data, $\tilde{\boldsymbol{x}} = augmentation(\boldsymbol{x})$, initialization $\boldsymbol{\theta}^0$ and cache $\mathcal{C} = \{\}$, learning rate η_k, losses $\mathcal{L}_L$ and $\mathcal{L}_U$.

Result: Acoustic model $\mathcal{A}(\boldsymbol{x}; \boldsymbol{\theta})$

1. Train $\mathcal{A}$ on $(\boldsymbol{x}, \boldsymbol{y}) \in L$ for M steps: $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \eta_k \nabla \mathcal{L}_L(\mathcal{A}(\tilde{\boldsymbol{x}}; \boldsymbol{\theta}^k), \boldsymbol{y}), k = \overline{1, M}$;
2. Decrease model's $\mathcal{A}(\boldsymbol{x}, \boldsymbol{\theta})$ dropout;
3. **for** $k = \overline{M+1, M+C}$ **do**
   - For random $\boldsymbol{x} \in U$ generate $\hat{\boldsymbol{y}} = PL(\mathcal{A}_{inference}(\boldsymbol{x}; \boldsymbol{\theta}^k))$ and $\mathcal{C} \leftarrow \mathcal{C} \bigcup \{\boldsymbol{x}, \hat{\boldsymbol{y}}\}$;
   - $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \eta_k \nabla \mathcal{L}_L(\mathcal{A}(\tilde{\boldsymbol{x}}; \boldsymbol{\theta}^k), \boldsymbol{y}), (\boldsymbol{x}, \boldsymbol{y}) \in L$;

end

repeat
- **if** $rand(0, 1) < N_L/(N_L + N_U)$ **then**
  - Take random $(\boldsymbol{x}, \boldsymbol{y}) \in L$ and $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \eta_k \nabla \mathcal{L}_L(\mathcal{A}(\tilde{\boldsymbol{x}}; \boldsymbol{\theta}^k), \boldsymbol{y})$
- **else**
  - Take random $b = (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{C}$ and $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \eta_k \nabla \mathcal{L}_U(\mathcal{A}(\tilde{\boldsymbol{x}}; \boldsymbol{\theta}^k), \boldsymbol{y})$;
  - **if** $rand(0, 1) < p$ **then**
    - For random $\boldsymbol{x}' \in U$ generate $\hat{\boldsymbol{y}}' = PL(\mathcal{A}_{inference}(\boldsymbol{x}; \boldsymbol{\theta}^k))$ and $\mathcal{C} \leftarrow \mathcal{C} \setminus b \bigcup \{\boldsymbol{x}', \hat{\boldsymbol{y}}'\}$
  - **end**
- **end**
- $k \leftarrow k + 1$;

until *convergence or maximum iterations are reached*;

Table 2: Comparison between slimIPL and fully supervised prior works: the same model architecture (transformer with learnable relative positional embedding), loss and token set is used. This gives lower and upper bounds on the WER one can expect with semi-supervised approaches.

	Cache C	WER %			
		dev-clean	dev-other	test-clean	test-other
lower bound (sup. 960h) [30]	-	2.6	7.0	2.7	6.8
slimIPL 100h, hard-path [7]	100	3.7	7.3	3.8	7.5
upper bound (sup. 100h) [7]	-	6.2	16.8	6.2	16.8
slimIPL 10h, hard-path [7]	1000	11.4	14.0	11.4	14.7
upper bound (sup. 10h) [7]	-	31.9	52.3	32.6	52.4

B Empirical Setup

We keep the original 16kHz sampling rate from LibriSpeech, and compute log-mel filterbanks with 80 coefficients for a 25ms sliding window, strided by 10ms. All features are normalized to have zero mean and unit variance per input sequence before feeding them into the acoustic model. We closely follow the architecture and training pipeline from slimIPL, including same hyper-parameters, CTC loss, transformer-based architecture, token set (26 English alphabet letters augmented with the apostrophe and a word boundary token), SpecAugment [31] parameters, Adagrad optimizer [32] and dynamic batching. We make only few changes: a) to speed up the training and decrease memory usage, we replace relative positional embedding [28] by CAPE positional embedding [29] (only a global shift of 30s is used); b) we train all models on 8 A100 40GB GPUs with dynamic batching of ~ 290 audio per GPU, and FP32 tensor core computations for up to 500k updates. In all experiments we only vary the PL type and few slimIPL hyper-parameters (cache size C and cache probability update p), while keeping everything else fixed. Most of experiments are run with 3 different seeds, and we report mean and standard deviation as the final performance. We start with models where $C = 100, M = 20,000, N_U = 3, p = 0.1$ for 100h and $C = 1000, M = 60,000, N_U = 10, p = 0.1$ for 10h of supervision.

C Frame-Level and Sequence-Level Evolution in PLs

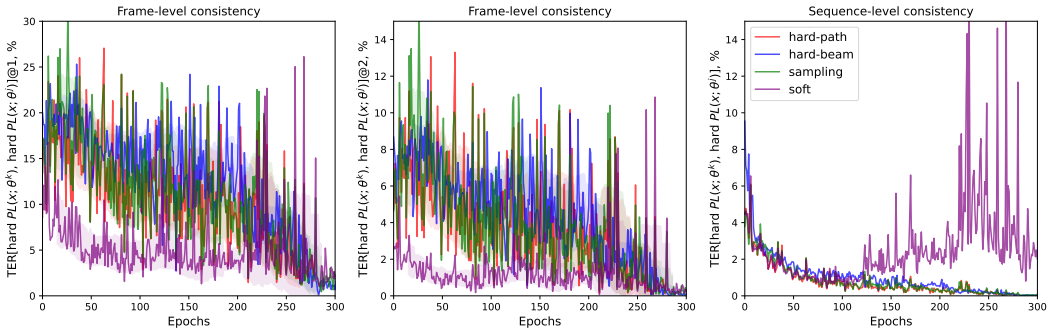

Figure 4: Levenshtein distance at the frame level (first and second columns) and sequence level (third column) between PLs for the same subset of unlabeled data for two consecutive steps for hard-path hard-beam, sampling and soft PLs.

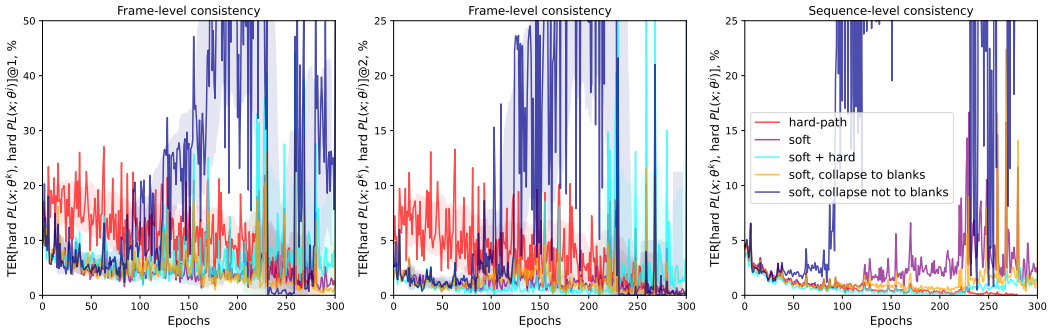

Figure 5: Levenshtein distance at the frame level (first and second columns) and sequence level (third column) between PLs for the same subset of unlabeled data for two consecutive steps for hard-path, soft, soft blended with hard ("soft + hard") soft blended with hard diverged to blank only ("soft, collapse to blanks"), soft with prior on target diverged to uniform distribution ("soft, collapse not to blanks") PLs.

To further analyse convergence to a degenerate solution for soft-labels, we keep tracking of PLs for the same subset of unlabeled samples as training goes (200 samples). We then compute the average distance between the PLs generated for the same sample, for two consecutive iterations. We report the frame-level Levenshtein distance of these consecutive PLs, after converting them into hard-alignments $\hat{\boldsymbol{a}}$ (whether the PLs were soft or hard). Here we consider also top-2 distance: we check whether $\hat{\boldsymbol{a}}_t^j$ is in top-2 tokens set of $p_{\boldsymbol{\theta}^k}^t(w|\boldsymbol{x})$, where k and j are two consecutive iterations, and $k < j$. We also report the Levenshtein distance at the sequence level $\hat{\boldsymbol{y}}$. Results (for cache $C = 100$ and $p = 0.1$) are shown for hard-path, beam and sampling in Figure 4, and for different variants of soft-labels with regularization in Figure 5. When tracking individual samples, we observe more transition noise in hard-labels, than for soft-labels. However, at the sequence level, hard-labels PLs transitions are significantly smoother as the training goes. Moreover, for models converged to a degenerate solution (Figure 5) we observe first sequence-level and later also frame-level consistency destruction between PLs for two consecutive iterations.

Below in Figures 6,7,8,9,10,11 we show detailed per-frame token distribution evolution during training with 100h labeled data, $C = 100, p = 0.1$. Top part of these Figures displays evolution in training of output token distribution per frame for one particular random sample from unlabeled set U. Every sub-figure is a time frame t of the sample. For every sub-figure: x-axis is a training step when we trained on the sample, y-axis is the probability of a token in range $[0, 1]$. CTC blank token is denoted by #. Bottom left part shows total probability of top-3 tokens per frame for set of unlabeled samples averaged across training steps, while bottom right shows percentage of CTC blank token in hard-path PLs (solid) and percentage of empty hard-alignment PLs (dashed) during training.

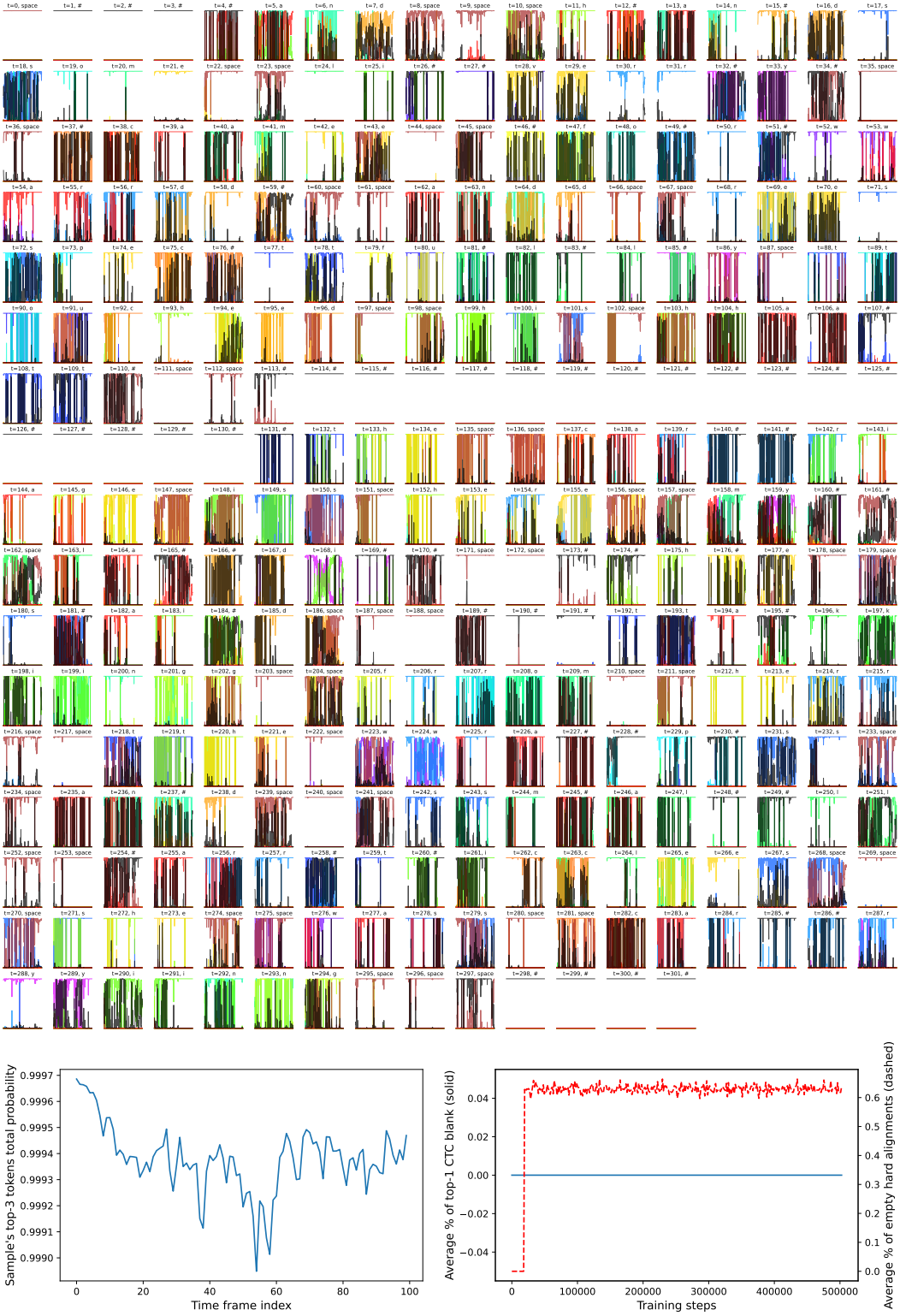

Figure 6: Training dynamics for hard-path with 100h labeled data, $C = 100, p = 0.1$.

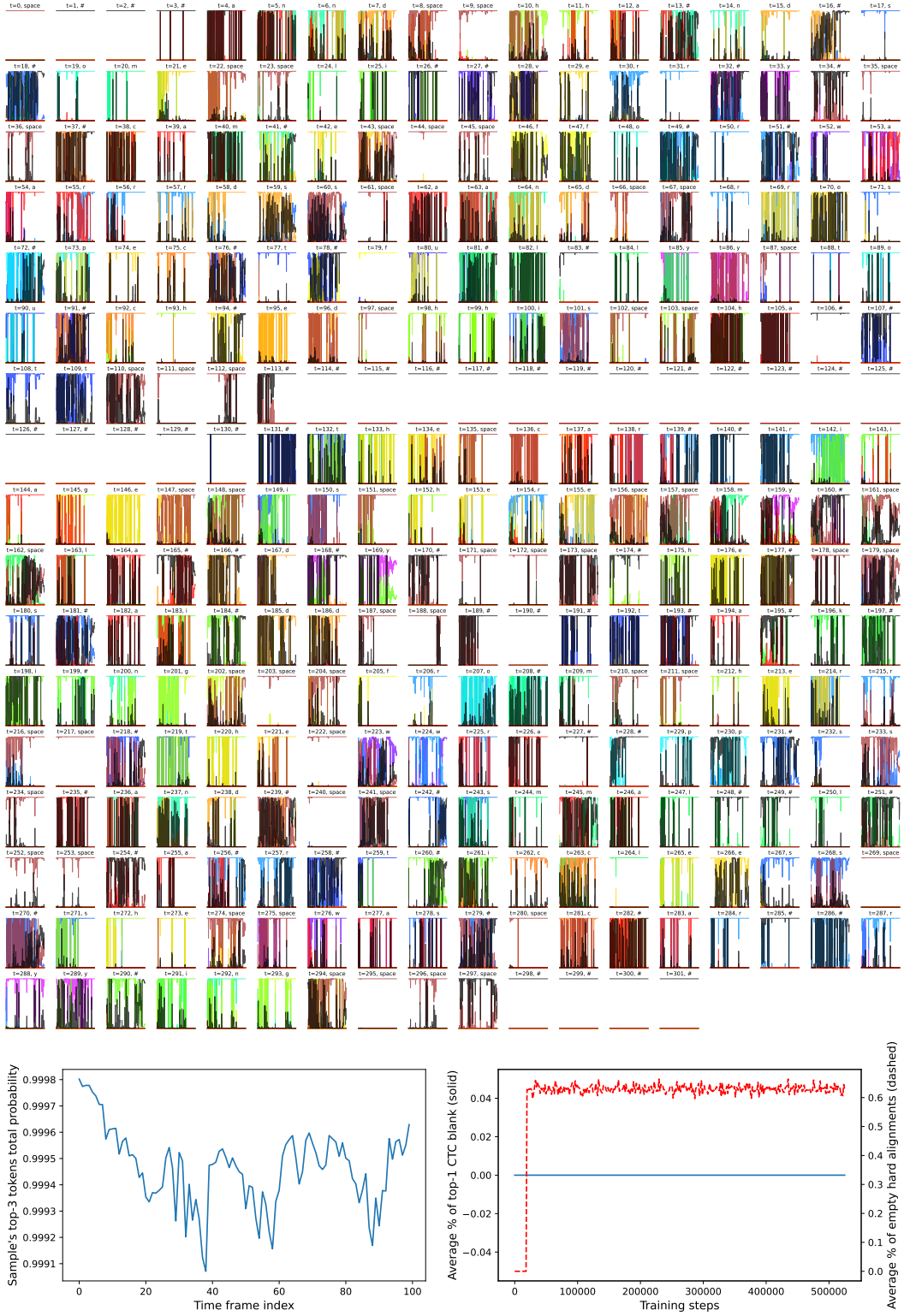

Figure 7: Training dynamics for sampling with 100h labeled data, $C = 100, p = 0.1, \tau = 0.4$.

Figure 8: Training dynamics for soft-labels with 100h labeled data, $C = 100, p = 0.1, \tau = 10, \beta = 1$.

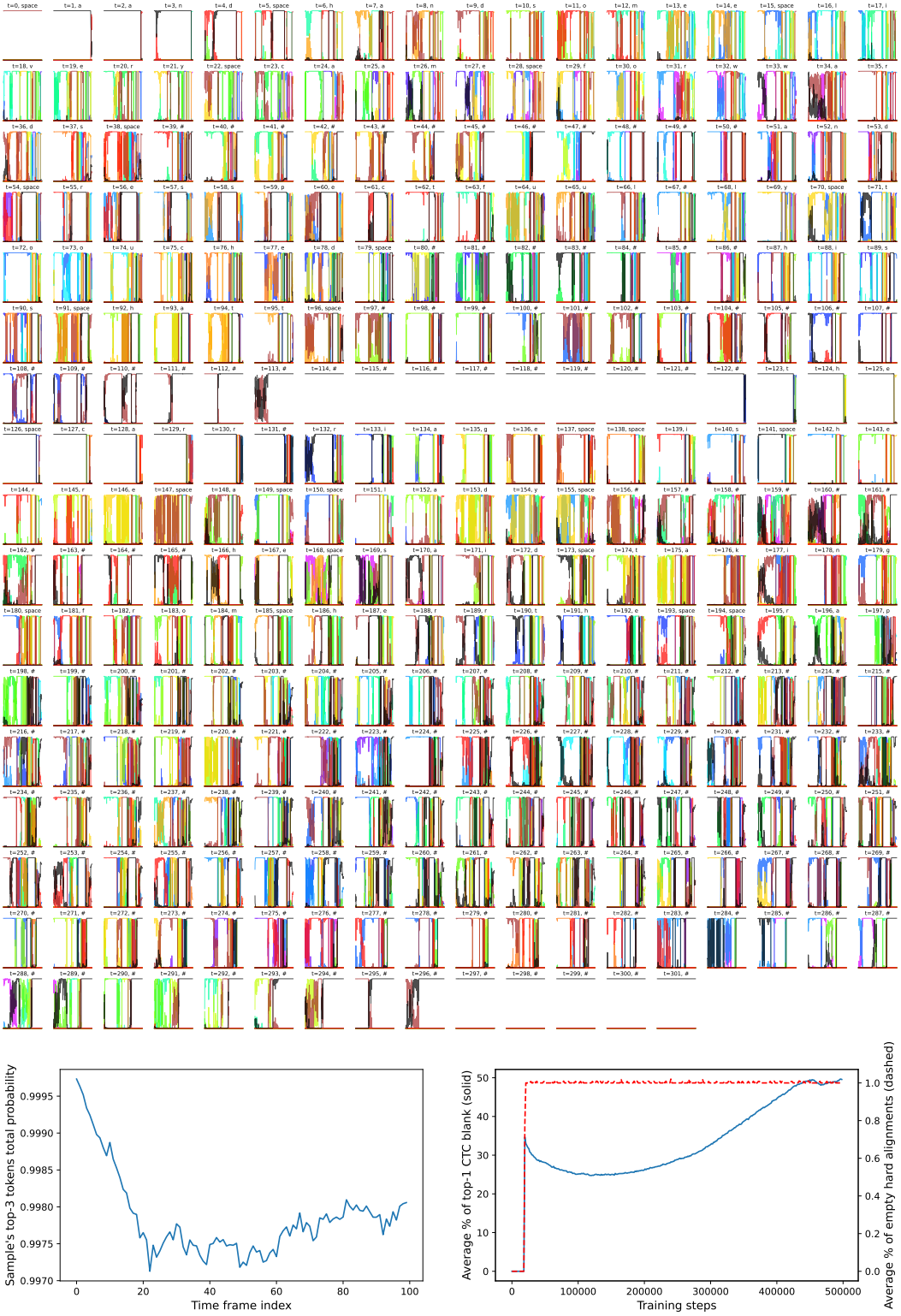

Figure 9: Training dynamics for best model with blending hard and soft-labels. We use 100h labeled data, $C = 100, p = 0.1, \tau = 10, \beta = 1, \delta = 0.1$.

Figure 10: Training dynamics for blending hard and soft-labels together when divergence to blank distribution happens. We use 100h labeled data, $C = 100, p = 0.1, \tau = 10, \beta = 1, \delta = 0.1$.

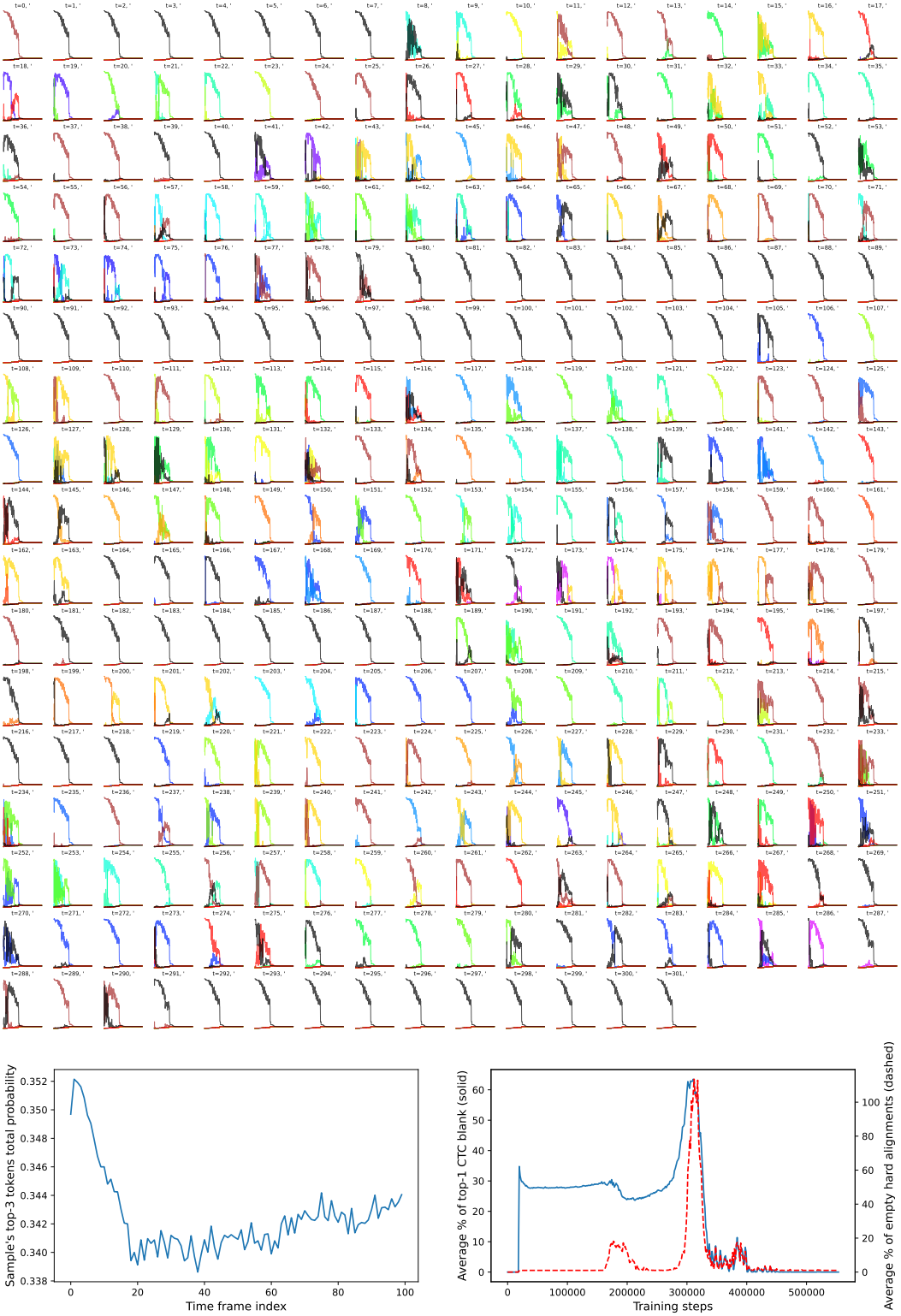

Figure 11: Training dynamics with soft-labels and target prior when divergence to non-blank distribution happens. We use 100h labeled data, $C = 100, p = 0.1, \tau = 10, \beta = 1, \gamma = 0.05$.

D Collapsing to a Degenerate Solution and Labeled Set Memorization

For some models where the WER between hard-alignment and golden transcription increases, we keep training for a very long time. In all cases, while we observe different dynamics, all models converge to some degenerate solution, while memorizing the labeled training set. Convergence to a degenerate solution can be very slow and may take a huge number of training steps. For all experiments we observe in some cases tendency to converge to blank prediction only (Figure 10) or to e.g. uniform distribution (Figure 11). In case of convergence to these trivial solutions we observe that the training loss on pseudo-labeled data becomes zero while the loss on labeled data is non-zero (and decreasing), as shown in Figure 12. When we look at the model's generated transcriptions, we see empty transcriptions for unlabeled data and close to ground truth (WER is a few percentages) for labeled data. This indicates that the model is in the regime of labeled data memorization.

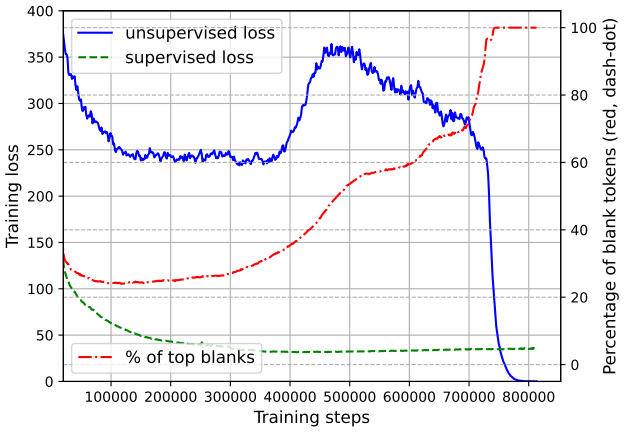

Figure 12: Example of a model trained with soft-labels that diverged during very long training to blank only token distribution (the same model as in Figure 10). By the time of this convergence to a trivial solution the model has perfect zero loss on unlabeled data (solid, blue) while having non-zero and still decreasing training loss on labeled data (dashed, green) what indicates memorization of labeled data. The percentage of top blank tokens is presented too (dashed-dot, red).

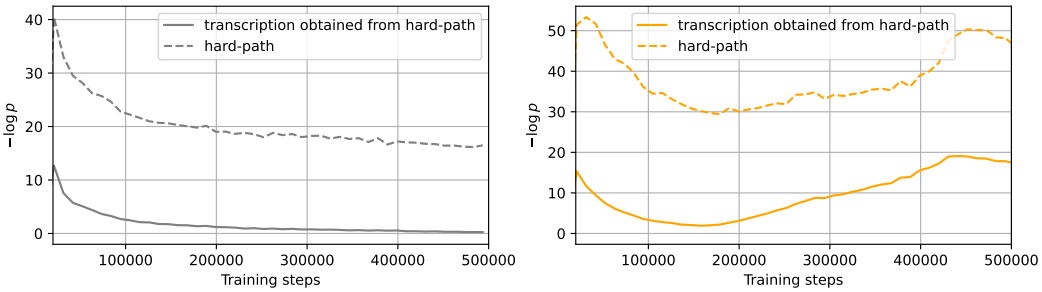

Figure 13: Comparison between hard-PL (left) and soft-PL (right) for 100h supervision, $C = 100, p = 0.1, N_U = 3, \tau = 10, \beta = 1$. We generate hard-path pseudo-label for a sample and compute this hard-path probability and transcription (to which this hard-path leads) probability from the logits of the current model state. We can see that one-path itself never dominates and there are other alignments which lead to the same transcription and have significant probability mass.

E Ablations

E.1 Labeled Data Exclusion from PL Phase

To confirm experimentally that there is no optimization problem (leading to trivial solutions) coming from the joint usage of different losses (CTC for labeled data, cross-entropy for unlabeled data), we

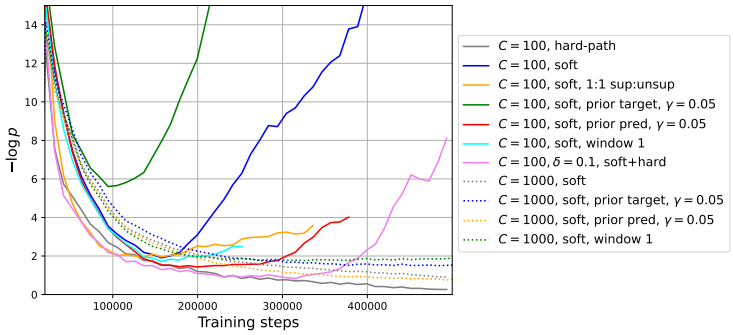

Figure 14: Comparison between different pseudo-labels for 100h supervision, $p = 0.1, N_U = 3, \tau = 10, \beta = 1$ for models from Figure 1. We generate hard-path from a pseudo-label for a sample and compute this hard-path probability from the logits of the current model state. We can see that best one-path probability decreasing ($-\log p$ is also growing) for unstable models.

stop to train on labeled data as soon as the cache is filled. With a cache size set to 100, we again observe model divergences with soft-labels, while cache size of 1000 is able to stabilize training. Both cases have worse performance than their hard-path counterpart with the same hyper-parameters.

E.2 Teacher-Student Training as a Stable Soft-Labeling

We tried also to disentangle continuous PL from soft-labels, and performed ablations with teacher-student PL and soft-labels. We first train a supervised only model on 100h of LibriSpeech (gets 18% WER on *dev-other*), use it to generate soft-labels for all unlabeled data and then train a new model from scratch on joint labeled and pseudo-labeled data. This model reduces WER on *dev-other* to 14.1-14.3% WER for different values of τ and β while being very stable in training. We perform then the second round of teacher-student training using the latter model as a new teacher. Training a new model from scratch gives on *dev-other* 12.8-13.4% WER being stable again. Hence, we conclude that unstable training with soft-labels may come from continuous training dynamics. It is still an open question how to make continuous pseudo-labeling stable with soft-labels.

E.3 Regularization: Blending Soft-Labels with Hard-Labels

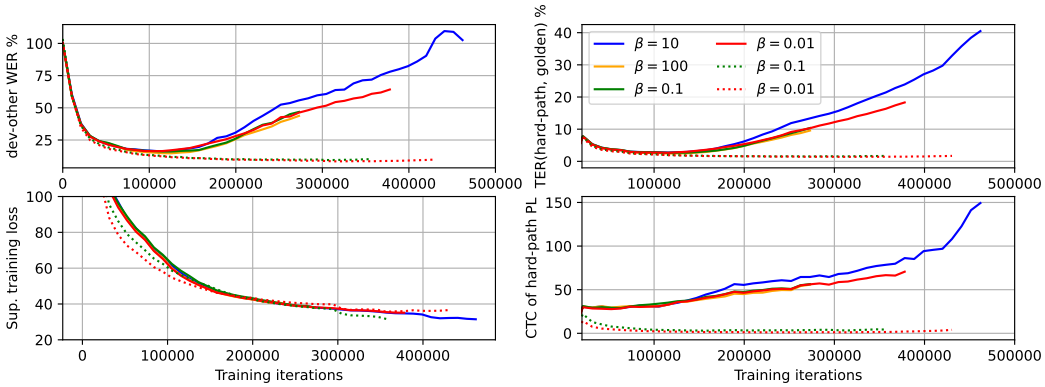

Figure 15: Comparison between soft-PL (solid) and blending soft-PL with hard-labels with $\delta = 0.1$ (dashed). Training is done with temperature $\tau = 1$ and 100h of labeled data, with cache parameters $C = 100, p = 0.1$.

E.4 Sequence-Level Consistency: Pseudo-Labels Filtering

As we discovered issues with sequence-level consistency for soft-labels (Figure 4 vs Figure 5) for continuous training we tried regularization based on simple filtering pseudo-labeled data which

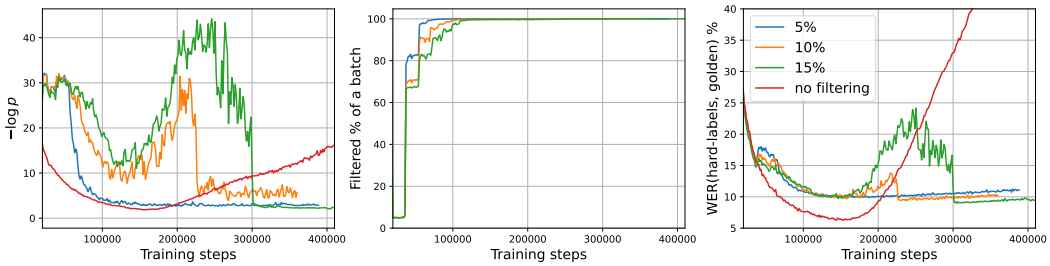

Figure 16: Training behaviour for PL filtering for sequence level consistency. We report three different thresholds on token distance between PLs across two consecutive training steps (5%, 10%, 15%) and the baseline soft-labeling without any filtering (red) for $C = 100, p = 0.1, \tau = 10, \beta = 1$. Left: log probability of hard-path transcription of PLs, middle: percentage of samples filtered from the batch, right: WER between PL hard-path and golden transcription.

breaks this consistency: we measure Levenshtein distance (on the token level) between hard-path transcriptions of current pseudo-label for a sample $\boldsymbol{x} \in U$ and *of previously used pseudo-label to make a training step* of the same sample; if the distance is larger than some threshold we filter a sample from the unlabeled batch for the current update step. Our preliminary experiments show that this filtering is able to stabilize the training with a proper threshold, see Figure 16, however it reaches quite bad performance yet.