

Label-Context-Dependent Internal Language Model Estimation for CTC

Zijian Yang^{*1}, Minh-Nghia Phan^{*1}, Ralf Schlüter^{1,2}, Hermann Ney^{1,2}

¹Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University, Germany

²AppTek GmbH, Germany

{zyang, schluter, ney}@ml.rwth-aachen.de, minh.nghia.phan@rwth-aachen.de

Abstract

Although connectionist temporal classification (CTC) has the label context independence assumption, it can still implicitly learn a context-dependent internal language model (ILM) due to modern powerful encoders. In this work, we investigate the implicit context dependency modeled in the ILM of CTC. To this end, we propose novel context-dependent ILM estimation methods for CTC based on knowledge distillation (KD) with theoretical justifications. Furthermore, we introduce two regularization methods for KD. We conduct experiments on Librispeech and TED-LIUM Release 2 datasets for in-domain and cross-domain evaluation, respectively. Experimental results show that context-dependent ILMs outperform the context-independent priors in cross-domain evaluation, indicating that CTC learns a context-dependent ILM. The proposed label-level KD with smoothing method surpasses other ILM estimation approaches, with more than 13% relative improvement in word error rate compared to shallow fusion.

Index Terms: speech recognition, CTC, internal language model, knowledge distillation

1. Introduction & Related Work

In automatic speech recognition, sequence-to-sequence (seq2seq) models have drawn more and more attention in recent years, due to their simplified training/decoding pipeline and strong performance. The most famous architectures include attention-based encoder-decoder models (AED) [1], recurrent neural network transducers (RNN-T) [2] and connectionist temporal classification (CTC) [3]. Although seq2seq models can be used as standalone models for decoding, integrating an external language model (ELM) trained on large-scale text data often leads to better performance for both in- and cross-domain tasks. Among various integration methods, Shallow Fusion (SF) is a widely used approach, combining the ELM and acoustic model (AM) probabilities in a log-linear manner.

Since a seq2seq model provides posterior probabilities for sequences, integrating it with an ELM during decoding requires compensation for the prior/internal language model (ILM) of the ASR model according to Bayes rule. While ILM estimation methods have been extensively studied for models with explicit label context dependency, such as AED and RNN-T [4, 5, 6, 7], due to the encoder-only architecture and the label context independence assumption, estimating the ILM of CTC is inherently challenging. Most existing ILM estimation approaches for CTC rely on heuristics. Das et al. proposed a heuristic ILM estimation method based on masking out the acoustic input and accumulating log-posteriors for only the masked time frames [8]. Zhao et al. proposed to apply different masking rates for the

acoustic input to detect the strength of the ILM of CTC [9]. Other works apply the marginal frame-level prior as the ILM based on the context-independent assumption [10, 11]. However, due to powerful architectures applied in modern models, a label-context-dependent ILM can be learned in the encoder implicitly. In [12], the authors showed that the bigram ILM outperformed the unigram ILM in search, indicating that the ILM of CTC is context-dependent. Nevertheless, [12] estimates the ILM using transcription text, which is decorrelated from the CTC output.

In this work, we aim to address the implicit context dependence of the CTC ILM. Since CTC does not have a language-model-like module, we propose novel label-context-dependent ILM estimation methods for CTC based on label-level and sequence-level knowledge distillation (KD) with theoretical justifications. More precisely, we employ the CTC model as the teacher model, and a small ILM estimator as the student model, to distill CTC probabilities to the ILM estimator on label and sequence levels. To the best of our knowledge, this is the first context-dependent ILM estimation method based on CTC outputs. For the label-level distillation, we propose to compute the label posterior via CTC prefix probabilities. Moreover, we introduce two regularization methods for KD, i.e. smoothing and masking. Systematic comparisons are conducted for both in-domain and cross-domain evaluation on Librispeech (LBS) [13] and TEDLIUM Release 2 (Tedlium2) [14] datasets, respectively. Experimental results on Tedlium2 demonstrate that ILMs with context dependence work better than the context-independent priors in cross-domain evaluation, indicating a context-dependent ILM of CTC. Furthermore, in cross-domain evaluation, our proposed methods outperform previous ILM estimation methods with over 13% relative improvement in word error rate compared to shallow fusion (SF). We also investigate the effect of combining frame-level prior with other ILM estimation approaches, as well as the impact of different context lengths on ILM compensation performance in decoding.

2. CTC

Given an input sequence $X = x_1, x_2, \dots, x_{T'}$ and a label vocabulary \mathcal{V} , the probability of the label sequence a_1^S with $a_s \in \mathcal{V}$ is modeled by CTC as follows:

$$\begin{aligned} P(a_1^S | X) &= \sum_{y_1^T: \mathcal{B}(y_1^T) = a_1^S} P(y_1^T | h_1^T) \\ &= \sum_{y_1^T: \mathcal{B}(y_1^T) = a_1^S} \prod_{t=1}^T P(y_t | h_t), \end{aligned} \quad (1)$$

where y_1^T is the blank-augmented alignment sequence, \mathcal{B} is the collapse function and h_1^T is the sequence of encoder output vectors with potential subsampling. As shown in the decompo-

* denotes equal contribution

sition in Eq. (1), the output of CTC has no explicit label context dependency. However, as a seq2seq model, CTC models the posterior probability of the label sequence. Therefore, when integrating an external language model (ELM), the correction for the internal language model (ILM) should be considered based on Bayes' theorem.

$$X \rightarrow a_1^{S^*} = \arg \max_{S, a_1^S} P(a_1^S | X) \frac{P_{\text{ELM}}^{\lambda_1}(a_1^S)}{P_{\text{ILM}}^{\lambda_2}(a_1^S)} \quad (2)$$

Here, $P_{\text{ELM}}(a_1^S)$ is the ELM probability, and $P_{\text{ILM}}(a_1^S)$ is the ILM probability of CTC. λ_1 and λ_2 are scales for the ELM and ILM, respectively. In practice, Viterbi approximation is usually applied for efficiency, which takes the best path during decoding. The ILM of CTC is defined as:

$$P_{\text{ILM}}(a_1^S) := \sum_X Pr(X) P(a_1^S | X), \quad (3)$$

where $Pr(X)$ denotes the true marginal distribution of the data. Since the true distribution is unknown, and computing the summation for all label sequences a_1^S is intractable, in practice, P_{ILM} is estimated by various approaches.

3. Internal Language Model Estimation for CTC

3.1. Frame-Level Unigram Prior Estimation

Label prior is a simple estimation of the ILM. Manohar et al. proposed to compute the label prior by marginalizing the model posterior over all acoustic inputs [11]. For CTC, this approach provides a context-independent frame-level prior (FP) estimation.

$$P_{\text{FP}}(y) = \frac{1}{\sum_{n=1}^N T_n} \sum_{n=1}^N \sum_{t=1}^{T_n} P(y | h_{n,t}) \quad (4)$$

Here, N denotes the number of training sequences, T_n denotes the encoder output length of sequence n , and $h_{n,t}$ denotes the hidden state of sequence n at time frame t . Since this unigram is estimated on frame level, it is applied to the CTC output for each frame during decoding, rather than only applied when a new label is emitted. That is, we use the following $q(y_t, h_t)$ with a prior scale λ_3 to replace the original CTC output $p(y_t | h_t)$ when integrating with an during in decoding.

$$q(y_t, h_t) = \frac{P(y_t | h_t)}{P_{\text{FP}}^{\lambda_3}(y_t)}$$

3.2. ILM Estimation with Knowledge Distillation

We propose to utilize knowledge distillation (KD) to estimate a context-dependent ILM of CTC. More precisely, we employ CTC as the teacher model and a standard autoregressive LM $q_\theta(a_1^S)$ (ILM estimator) as the student model, allowing the LM to learn the ILM probability distribution from CTC. Since CTC does not explicitly model context dependencies, we assume that its ILM is relatively weak. Therefore, we employ a small LSTM language model as the ILM estimator in practice.

3.2.1. Label-Level Knowledge Distillation for ILM estimation

The label posterior of ILM can be computed as follows:

$$\begin{aligned} P_{\text{ILM}}(a_s | a_1^{s-1}) &= \sum_X P(a_s, X | a_1^{s-1}) \\ &\approx \sum_X P(a_s | a_1^{s-1}, X) Pr(X | a_1^{s-1}) \\ &= \sum_X P(a_s | a_1^{s-1}, X) \frac{Pr(X, a_1^{s-1})}{Pr(a_1^{s-1})} \end{aligned} \quad (5)$$

Although CTC does not model the label posterior probability $P(a_s | a_1^{s-1}, X)$ directly, it can be computed as follows:

$$P(a_s | a_1^{s-1}, X) = \frac{P(a_1^s, \dots | X)}{P(a_1^{s-1}, \dots | X)}, \quad (6)$$

where $P(a_1^s, \dots | X)$ is the CTC prefix probability defined as the marginalized probability of all label sequences that have a_1^s as their prefix [15, 16].

$$P(a_1^s, \dots | X) = \sum_{v \in (\cup_{i=1}^{\infty} \mathcal{V}^i \cup \emptyset)} P(a_1^s \cdot v | X)$$

While the end-of-sentence (<EOS>) token is crucial for an autoregressive LM to ensure proper normalization, CTC does not explicitly model the probability of <EOS>. However, CTC implicitly accounts for the <EOS> probability when computing the probability of the whole sequence. Therefore, the posterior probability of <EOS> can be computed via:

$$P(<EOS> | a_1^s, X) = \frac{P(a_1^s | X)}{P(a_1^s, \dots | X)}$$

During training, We reuse the CTC training data to estimate the ILM. Let $(X_n, a_{n,1}^{S_n})$ be the n -th pair in the training data, and N be the number of training pairs. The training criterion based on Kullback–Leibler divergence is defined as follows:

$$\begin{aligned} F(\theta) &= \frac{1}{N} \sum_{n=1}^N \sum_{s=1}^{S_n+1} \sum_{a \in \mathcal{V}^+} P(a | a_{n,1}^{s-1}, X_n) \log \frac{P(a | a_{n,1}^{s-1}, X_n)}{q_\theta(a | a_{n,1}^{s-1})} \\ &= \sum_{X, S, a_1^S} \tilde{Pr}(X, a_1^S) \sum_{s=1}^{S+1} \sum_{a \in \mathcal{V}^+} P(a | a_1^{s-1}, X) \log \frac{P(a | a_1^{s-1}, X)}{q_\theta(a | a_1^{s-1})}, \end{aligned} \quad (7)$$

(8)

where \mathcal{V}^+ is the <EOS>-augmented vocabulary. The label posterior $P(a | a_{n,1}^{s-1}, X)$ is derived from the CTC output via Eq. (6) and serves as the fixed teacher model. Meanwhile, $q_\theta(a | a_{n,1}^{s-1})$ is the output of a standard LM with learnable parameters θ , applied as the ILM estimator for the ILM of CTC. Eq. (8) is derived from Eq. (7) by representing the training data with the empirical distribution \tilde{Pr} . The motivation of this training criterion is: if there is sufficient amount of training data, i.e. $\tilde{Pr}(X, a_1^S) \approx Pr(X, a_1^S)$, it can be proven that the label-level ILM approximation in Eq. (5) is exactly the global optimum \hat{q} of the training criterion (8), i.e. $\hat{q}(a | a_1^{s-1}) \approx P_{\text{ILM}}(a | a_1^{s-1})$. A detailed proof is derived in Appendix A.

In practice, the available training data is limited, and achieving a global optimum is not guaranteed. Since CTC training data is used for ILM estimation, the CTC model may become overconfident on previously seen data. We observe that the posterior probability is usually close to 1 for the ground truth label in training data and 0 for other labels, which makes this criterion close to the standard cross-entropy criterion for LM training on transcriptions. To mitigate this issue, we propose two regularization methods for ILM estimator training: KD with smoothing and masking.

Label-level KD with smoothing: Due to the limited amount of training data, the empirical distribution \tilde{Pr} is typically sparse. Inspired by some count-based LM smoothing methods like absolute discounting and Jelinek–Mercer smoothing, we propose to smooth the empirical distribution $\tilde{Pr}(X, a_1^S)$ in Eq. (8) by interpolating the marginal distributions $\tilde{Pr}(X)$ and $\tilde{Pr}(a_1^S)$:

$$\tilde{Pr}(X, a_1^S) = \alpha \tilde{Pr}(X, a_1^S) + (1 - \alpha) \tilde{Pr}(X) \tilde{Pr}(a_1^S),$$

where α is an interpolation factor. The latter smoothing term can be regarded as dropping the interdependencies between X and a_1^S . The training criterion is then defined as:

$$\begin{aligned}
F_{\text{smoothing}}(\theta) &= \sum_{X, S, a_1^S} \overline{Pr}(X, a_1^S) \sum_{s=1}^{S+1} \sum_{a \in \mathcal{V}^+} P(a|a_1^{s-1}, X) \log \frac{P(a|a_1^{s-1}, X)}{q_\theta(a|a_1^{s-1})} \\
&= \sum_{n=1}^N \sum_{n'=1}^N \frac{\beta(n, n')}{N} \sum_{s=1}^{S_n+1} \sum_{a \in \mathcal{V}^+} P(a|a_{n,1}^{s-1}, X_{n'}) \log \frac{P(a|a_{n,1}^{s-1}, X_{n'})}{q_\theta(a|a_{n,1}^{s-1})}
\end{aligned} \tag{9}$$

$$\tag{10}$$

Here, $\beta(n, n')$ is defined by Kronecker delta δ :

$$\beta(n, n') = \delta(n, n')\alpha + \frac{1-\alpha}{N},$$

The equivalence between Eq. (9) and Eq. (10) is derived in Appendix B. In general, for a label sequence $a_{n,1}^{S_n}$ in the training data, the smoothing criterion computes an averaged probability over all inputs, rather than only considering the pairwise input X_n like in Eq. (7). In practice, since the smoothing over the whole training data is infeasible, we conduct the smoothing within a mini-batch.

Label-level KD with Masking To mitigate the overconfidence of CTC, we propose to mask the acoustic evidence of the label based on alignments. Let t_1^S be (sub-)word boundaries provided by a GMM alignment, where t_s is the end time frame of label s . Each label position s has a probability p_{mask} to mask out the corresponding acoustic input $x_{t_{s-1}^S+1}^{t_s^S}$. For training efficiency, we mask out multiple positions for each training pair. Let the set of masked positions for the n -th sequence be $\mathcal{M}_n \subseteq \{1, 2, \dots, S_n\}$, the training criterion is defined as:

$$\begin{aligned}
F_{\text{masking}}(\theta) &= \frac{1}{N} \sum_{n=1}^N \sum_{s \in \mathcal{M}_n} \sum_{a \in \mathcal{V}^+} P(a|a_{n,1}^{s-1}, \overline{X}_n) \log \frac{P(a|a_{n,1}^{s-1}, \overline{X}_n)}{q_\theta(a|a_{n,1}^{s-1})}
\end{aligned}$$

where \overline{X}_n is the masked input sequence.

3.2.2. Sequence-Level Knowledge Distillation for ILM estimation

The KD from CTC to ILM estimator can also be done on the sequence level. Following the idea of the smoothing method in Sec. 3.2.1, we smooth the empirical distribution for sequence-level KD:

$$F_{\text{seq}}(\theta) = \sum_{n=1}^N \sum_{n'=1}^N \frac{\beta(n, n')}{N} P(a_{n,1}^{S_n} | X_{n'}) \log \frac{P(a_{n,1}^{S_n} | X_{n'})}{q_\theta(a_{n,1}^{S_n})} \tag{11}$$

The main difference between the sequence-level KD criterion (11) and the label-level KD criterion (10) is that (11) transfers sequence probabilities to the ILM estimator, while (10) transfers label posterior probabilities.

4. Experiments

4.1. Setup

Experiments are done on the 960 hours Librispeech (LBS) dataset [13] for in-domain evaluation and TED-LIUM Release 2 (Tedlium2) dataset [14] for cross-domain evaluation using the RETURNN framework [17] based on Pytorch [18]. We use byte-pair-encoding (BPE) [19] as output labels with a vocabulary size of 10k.

We use a 12-layer conformer [20] CTC model as our baseline acoustic model. The hidden dimension of the conformer

block is 512, and the feed-forward dimension is 2048. We employ 80-dimensional log Mel features as the input to CTC. SpecAugment is applied for data augmentation in CTC training. A downsampling factor of 6 is applied. The model is trained for 100 full epochs on LBS with AdamW optimizer [21] and one-cycle learning rate schedule [22, 23], with a peak learning rate of 1e-3. This CTC model is also applied as the teacher model in KD for ILM estimation. We assume that the ILM of CTC is relatively weak. Therefore, we employ a one-layer LSTM LM as the ILM estimator. The embedding dimension is 128, and the LSTM hidden dimension is 1000. ILM estimators are trained via criteria proposed in Sec. 3.2. When computing teacher probabilities, SpecAugment is disabled to reflect the exact output of CTC. We train ILM estimators on the LBS transcriptions for 5 full epochs. We also train a standard LM with the same architecture as the ILM estimator on LBS transcriptions for 30 full epochs, referred to as the transcription LM.

During decoding, we employ a 24-layer transformer LM [24] as the ELM for in-domain evaluation and a 4-layer LSTM LM as the ELM for cross-domain evaluation, with perplexity (PPL) 37 and 48 on dev sets, respectively. We apply the time-synchronous Viterbi search implemented in RETURNN. All the scales applied in search are optimized on the dev sets. Since we observed that the ILM PPL is not correlated with the word error rate (WER), we select the ILM estimator checkpoints based on evaluation results on dev sets. The smoothing factor α and the masking rate p_{mask} are also tuned based on recognition results on dev sets. We use $\alpha = 0.5$ for smoothing and $p_{\text{mask}} = 0.4$ for masking in our experiments. Our code can be found online.¹

4.2. In-Domain and Cross-Domain Evaluation

Table 1 shows evaluation results for the in-domain Librispeech and cross-domain Tedlium2 Corpora. The unigram ILM is computed via renormalizing FP without the blank label and applied on the label level. Similar to [12], the transcription LM served as an approximation of the ILM of CTC. As expected, integrating the ELM brings significant improvements over the standalone CTC, and all ILM correction approaches further improve the performance compared to SF. Moreover, different ILM estimation methods perform comparably on in-domain tasks, but differ more on cross-domain tasks. We assume that this is because ILM correction has two functions: boosting label emission and rebalancing the label distribution, consistent with discussions on ILM in [7, 25]. For in-domain tasks, since the ILM of CTC is relatively weak, ILM correction mainly serves to boost the label emission, while precise ILM estimation for the label distribution reshaping is unnecessary. Therefore, all ILM estimation methods perform similarly. However, for cross-domain tasks, because of the mismatch between the underlying distribution of the source and target domain, proper ILM estimation for label distribution rebalancing is important.

In cross-domain evaluation results, we observe that all the context-dependent ILM estimation methods outperform the context-independent priors (unigram and FP), indicating that CTC learns a context-dependent ILM implicitly due to the powerful encoder. The label-level KD without any regularization performs similar to the transcription LM, which verifies the statement in Sec. 3.2 that without regularization, the estimated ILM would be close to a transcription LM. Moreover, our proposed label-level KD methods with regularization (smoothing and masking) surpass other methods, indicating better ILM estimation and the necessity of regularization. Meanwhile, the

¹<https://github.com/rwth-i6/returnn-experiments/tree/master/2025-label-dependent-ILM-ctc>

Table 1: Recognition results for the in-domain Librispeech and cross-domain Tedlium2 Corpora. The ILM regularization is done during the ILM estimator training. The ILM PPLs are computed on the dev sets. The unigram ILM is computed via renormalizing frame-level prior (FP) without the blank label. Trans LM refers to applying transcription LM as an approximation of the ILM. Label and seq KD refer to label- and sequence-level KD methods for ILM estimation proposed in Sec. 3.2.

ELM	ILM	ILM reg- ularization	Librispeech						Tedlium2		
			ILM PPL	WER[%]				ILM PPL	WER[%]		
				dev		test			dev	test	
				clean	other	clean	other				
no	no		-	3.0	6.8	3.2	7.2	-	17.7	18.7	
yes			-	2.2	4.8	2.2	5.3	-	14.3	15.9	
	FP		-	2.1	4.4	2.1	4.9	-	13.2	14.7	
	unigram	-	1223	2.1	4.4	2.1	4.9	1307	12.9	14.9	
	trans LM		145	2.1	4.4	2.1	4.9	278	12.4	14.5	
			140	2.1	4.4	2.1	4.9	255	12.4	14.4	
	label KD	masking	169	2.0	4.3	2.0	4.8	286	12.2	14.0	
			191	2.0	4.4	2.0	4.8	297	11.9	13.8	
seq KD	smoothing	293	2.1	4.5	2.1	5.0	395	12.5	14.3		

Table 2: WERs [%] on Tedlium2 dev and test dataset for different ILM estimation methods combined with frame-level prior. The ELM is used in decoding. The ILM regularization is done during the ILM estimator training. Trans LM refers to applying transcription LM as an approximation of the ILM. Label and seq KD refer to label- and sequence-level KD methods for ILM estimation proposed in Sec. 3.2.

ILM	ILM regularization	dev		test	
		w/ FP	w/o FP	w/ FP	w/o FP
no	-	13.2	14.3	14.7	15.9
trans LM		12.0	12.4	13.9	14.5
		12.1	12.4	13.8	14.4
label KD	masking	12.1	12.2	13.9	14.0
		11.8	11.9	13.7	13.8
seq KD	smoothing	11.9	12.5	14.1	14.3

sequence-level KD is worse than the label-level KD. This may be because label-level KD provides more information about the CTC output distribution, as it considers the probabilities of all labels in the vocabulary for a given prefix in the training data, whereas sequence-level distillation only accounts for the probabilities of sequences present in the training data. Overall, our proposed label-level KD with smoothing achieves the best results in cross-domain evaluation, yielding a relative improvement of over 13% compared to SF. We also compute PPLs of the estimated ILMs. Unlike the known correlation between the PPL of the ELM and WER [26, 27], we do not observe a correlation between ILM PPLs and WERs. This suggests that ILM estimator checkpoints should not be selected based on the PPL, and new metrics need to be explored to better evaluate ILMs.

4.3. Ablation Study

4.3.1. Using Frame-Level Prior and ILM Jointly

As discussed in Sec. 3.1, the frame-level prior is applied at each time frame. Strictly speaking, this is not a standard ILM correction method, as ILM correction should be defined on label sequences. Therefore, in this section, we examine the effect of combining the ILM defined on labels with frame-level prior correction. Table 2 shows the comparison between various ILM estimation methods with and without frame-level prior. It is observed the transcription LM achieves notable improvement when combined with the frame-level prior. This is because the transcription LM models only the distribution of text

data and is decorrelated from the CTC output. Consequently, it benefits from the frame-level prior estimated from CTC outputs. Sequence-level KD also benefits from the local distribution of the frame-level prior, as it learns the probabilities of entire sequences from CTC, lacking local information. In contrast, label-level KD with smoothing and masking gains little from the frame-level prior. This indicates that ILM correction has a similar effect to frame-level prior correction, but our method provides a more accurate estimation of the ILM.

4.3.2. ILM with Different Context Lengths

We also investigate the effect of different context lengths for the ILM. We employ two-layer feedforward networks for limited-context ILM estimators. For transcription LMs, since the distribution is independent of the CTC output, no clear correlation is observed between context length and performance. For label-level ILM, we observe that ILM with context-6 performs slightly better than context-1 and context-10 but worse than full-context. This may be due to the feedforward network’s limited ability to model long contexts (e.g., context-10), making context-10 less effective than context-6. Overall, the optimal context length for CTC ILM requires further investigation.

Table 3: WERs [%] on Tedlium2 dev and test dataset for transcription LM and label-level ILM estimation with sampling across different context lengths. ELM is applied in recognition, while frame-level prior is not applied. Transcription LM is applied as approximated ILM.

ILM	Context	WER	
		dev	test
-	-	14.3	15.9
transcription LM	1	12.4	14.3
	6	12.7	14.6
	10	12.8	14.4
	full context	12.4	14.5
label-level KD w/ smoothing	1	12.4	14.2
	6	12.2	14.2
	10	12.2	14.4
	full context	11.9	13.8

5. Conclusion

In this work, we investigated the implicit context dependence of the internal language model (ILM) of connectionist temporal classification (CTC). To this end, we proposed novel ILM estimation methods for CTC based on knowledge distillation (KD). We provided a solid theoretical background for label- and sequence-level KD in the probability space. Moreover, two regularization methods were introduced to improve the performance of the estimated ILM. We conducted our experiments on Librispeech for in-domain evaluation and TEDLIUM Release 2 for cross-domain evaluation. Experimental results showed that in cross-domain tasks, context-dependent ILMs outperformed the context-independent priors, indicating that CTC inherently learned a context-dependent ILM. Moreover, our proposed label-level KD with sampling method achieved the best performance, demonstrating a relative improvement of over 13% in word error rate compared to shallow fusion. Additionally, we investigated the impact of context length and the effect of combining frame-level prior and ILMs for label sequences, providing further insights into ILM estimation.

Acknowledgments

This work was partially supported by the project RESCALE within the program *AI Lighthouse Projects for the Environment, Climate, Nature and Resources* funded by the Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV), funding ID: 67K132006A.

6. References

- [1] Z. Tüske, G. Saon, K. Audhkhasi, and B. Kingsbury, “Single Headed Attention Based Sequence-to-Sequence Model for State-of-the-Art Results on Switchboard-300,” in *INTERSPEECH*, 2020.
- [2] A. Graves, “Sequence Transduction with Recurrent Neural Networks,” *arXiv preprint arXiv:1211.3711*, 2012.
- [3] 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.
- [4] Z. Meng, N. Kanda, Y. Gaur, S. Parthasarathy, E. Sun, L. Lu, X. Chen, J. Li, and Y. Gong, “Internal Language Model Training for Domain-adaptive End-to-End Speech Recognition,” in *ICASSP*. IEEE, 2021, pp. 7338–7342.
- [5] A. Zeyer, A. Merboldt, W. Michel, R. Schlüter, and H. Ney, “Librispeech transducer model with internal language model prior correction,” *arXiv preprint arXiv:2104.03006*, 2021.
- [6] M. Zeineldeen, A. Glushko, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, “Investigating Methods to Improve Language Model Integration for Attention-based Encoder-Decoder ASR Models,” in *INTERSPEECH*, 2021, pp. 2856–2860.
- [7] W. Zhou, Z. Zheng, R. Schlüter, and H. Ney, “On Language Model Integration for RNN Transducer based Speech Recognition,” in *ICASSP*. IEEE, 2022, pp. 8407–8411.
- [8] N. Das, M. Sunkara, S. Bodapati, J. Cai, D. Kulshreshtha, J. Faris, and K. Kirchhoff, “Mask the bias: Improving domain-adaptive generalization of ctc-based asr with internal language model estimation,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [9] Z. Zhao and P. Bell, “Regarding the existence of the internal language model in ctc-based e2e asr.” [Online]. Available: <https://zhaozeyu1995.github.io/pdf/icassp2025.pdf>
- [10] Y. Miao, M. Gowayyed, and F. Metze, “Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding,” in *2015 IEEE workshop on automatic speech recognition and understanding (ASRU)*. IEEE, 2015, pp. 167–174.
- [11] V. Manohar, D. Povey, and S. Khudanpur, “Semi-supervised maximum mutual information training of deep neural network acoustic models,” in *Interspeech*, 2015, pp. 2630–2634.
- [12] N. Kanda, X. Lu, and H. Kawai, “Maximum a posteriori based decoding for ctc acoustic models,” in *Interspeech*, 2016, pp. 1868–1872.
- [13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR Corpus based on public Domain Audio Books,” in *ICASSP*. IEEE, 2015, pp. 5206–5210.
- [14] A. Rousseau, P. Deléglise, Y. Esteve *et al.*, “Enhancing the ted-lum corpus with selected data for language modeling and more ted talks,” in *LREC*, 2014, pp. 3935–3939.
- [15] A. Graves, “Supervised sequence labelling with recurrent neural networks,” Ph.D. dissertation, Technical University Munich, 2008.
- [16] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, “Hybrid ctc/attention architecture for end-to-end speech recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.
- [17] A. Zeyer, T. Alkhoul, and H. Ney, “RETURNN as a Generic Flexible Neural Toolkit with Application to Translation and Speech Recognition,” 2018, pp. 128–133.
- [18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimeshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.
- [19] R. Sennrich, “Neural machine translation of rare words with subword units,” *arXiv preprint arXiv:1508.07909*, 2015.
- [20] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented transformer for speech recognition,” in *Interspeech 2020*, 2020, pp. 5036–5040.
- [21] I. Loshchilov, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [22] G. Saon, Z. Tüske, D. Bolanos, and B. Kingsbury, “Advancing rnn transducer technology for speech recognition,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5654–5658.
- [23] J. Xu, W. Zhou, Z. Yang, E. Beck, and R. Schlüter, “Dynamic encoder size based on data-driven layer-wise pruning for speech recognition,” in *Interspeech 2024*, 2024, pp. 4563–4567.
- [24] K. Irie, A. Zeyer, R. Schlüter, and H. Ney, “Language Modeling with Deep Transformers,” in *INTERSPEECH*, 2019.
- [25] Z. Yang, W. Zhou, R. Schlüter, and H. Ney, “On the relation between internal language model and sequence discriminative training for neural transducers,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 627–12 631.
- [26] D. Klakow and J. Peters, “Testing the correlation of word error rate and perplexity,” *Speech Communication*, vol. 38, no. 1-2, pp. 19–28, 2002.
- [27] K. Irie, “Advancing neural language modeling in automatic speech recognition.” Ph.D. dissertation, RWTH Aachen University, Germany, 2020.

Appendix

A. The Global Optimum of Criterion (8)

In this section, we derive the global optimum of the training criterion (8), and show that this is equivalent to the ILM definition in Eq. (5). In order to distinguish a complete sequence and a prefix, we write out the $\langle \text{EOS} \rangle$ label explicitly. To simplify the discussion, we assume that the maximum sequence length is finite. Therefore, all summations are done over a finite set, and exchanging the order of summations is always possible. Assume an unlimited amount of training data so that empirical distributions are the same as true distributions, i.e. $\tilde{Pr}(\cdot) = Pr(\cdot)$. We have:

$$\begin{aligned}
 F(\theta) &= \\
 &= \sum_{S, X, a_1^S} Pr(X, a_1^S \cdot \langle \text{EOS} \rangle) \sum_{s=1}^{S+1} \sum_{a \in \mathcal{V}^+} P(a|a_1^{s-1}, X) \log \frac{P(a|a_1^{s-1}, X)}{q_\theta(a|a_1^{s-1})} \\
 &= \sum_S \sum_{s=1}^{S+1} \sum_{X, a_1^S} Pr(X, a_1^S \cdot \langle \text{EOS} \rangle) \sum_{a \in \mathcal{V}^+} P(a|a_1^{s-1}, X) \log \frac{P(a|a_1^{s-1}, X)}{q_\theta(a|a_1^{s-1})} \\
 &= \sum_S \sum_{s=1}^{S+1} \sum_{X, a_1^{s-1}} Pr(X, a_1^{s-1}) \cdot \sum_{a_s^S} Pr(a_s^S \cdot \langle \text{EOS} \rangle | a_1^{s-1}, X) \\
 &\quad \cdot \sum_{a \in \mathcal{V}^+} P(a|a_1^{s-1}, X) \log \frac{P(a|a_1^{s-1}, X)}{q_\theta(a|a_1^{s-1})} \\
 &= \sum_{X, s, a_1^{s-1}} Pr(X, a_1^{s-1}) \sum_{a \in \mathcal{V}^+} P(a|a_1^{s-1}, X) \log \frac{P(a|a_1^{s-1}, X)}{q_\theta(a|a_1^{s-1})} \\
 &\quad \cdot \underbrace{\sum_{S \geq s-1, a_s^S} Pr(a_s^S \cdot \langle \text{EOS} \rangle | a_1^{s-1}, X)}_{= 1 \text{ c.f. the normalization of } Pr(\cdot | a_1^{s-1}, X), \text{ } a_s^{s-1} \text{ refers to an empty sequence}} \\
 &= \sum_{X, s, a_1^{s-1}} Pr(X, a_1^{s-1}) \sum_{a \in \mathcal{V}^+} P(a|a_1^{s-1}, X) \log \frac{P(a|a_1^{s-1}, X)}{q_\theta(a|a_1^{s-1})} \\
 &= \sum_{s, a_1^{s-1}} Pr(a_1^{s-1}) \sum_{a \in \mathcal{V}^+} \sum_X P(a|a_1^{s-1}, X) Pr(X|a_1^{s-1}) \log \frac{P(a|a_1^{s-1}, X)}{q_\theta(a|a_1^{s-1})} \tag{12}
 \end{aligned}$$

Since $\sum_X P(a|a_1^{s-1}, X) Pr(X|a_1^{s-1})$ is a properly normalized distribution over $a \in \mathcal{V}^+$, according to Gibbs' inequality, the global optimum of Eq. (12) is obtained via:

$$\hat{q}_\theta(a|a_1^{s-1}) = \sum_X P(a|a_1^{s-1}, X) Pr(X|a_1^{s-1}), \forall a \in \mathcal{V}^+$$

Namely, equivalent to the definition Eq. (5)

B. Equivalence between Eq. (9) and (10)

Here, we derive that Eq. (9) is equivalent to Eq. (10). By definition, the empirical distributions are computed via:

$$\begin{aligned}
 \tilde{Pr}(X, a_1^S) &= \frac{1}{N} \sum_{n=1}^N \delta(X, X_n) \delta(a_1^S, a_{n,1}^{S_n}), \\
 \tilde{Pr}(X) &= \frac{1}{N} \sum_{n=1}^N \delta(X, X_n), \quad \tilde{Pr}(a_1^S) = \frac{1}{N} \sum_{n=1}^N \delta(a_1^S, a_{n,1}^{S_n})
 \end{aligned}$$

To simplify notations, we define

$$G(P, q|X, a_1^S) := \sum_{s=1}^{S+1} \sum_{a \in \mathcal{V}^+} P(a|a_1^{s-1}, X) \log \frac{P(a|a_1^{s-1}, X)}{q_\theta(a|a_1^{s-1})}$$

Therefore, Eq. (9) can be rewritten as:

$$\begin{aligned}
 F(\theta) &= \sum_{X, S, a_1^S} \tilde{Pr}(X, a_1^S) G(P, q|X, a_1^S) \\
 &= \sum_{S, X, a_1^S} \frac{\alpha}{N} \sum_{n=1}^N \delta(a_1^S, a_{n,1}^{S_n}) \delta(X, X_n) G(P, q|X, a_1^S) \\
 &\quad + \sum_{S, X, a_1^S} \frac{1-\alpha}{N} \sum_{n=1}^N \delta(a_1^S, a_{n,1}^{S_n}) \frac{1}{N} \sum_{n'=1}^N \delta(X, X_{n'}) G(P, q|X, a_1^S) \\
 &= \frac{\alpha}{N} \sum_{n=1}^N \sum_{(X, a_1^S): X=X_n, a_1^S=a_{n,1}^{S_n}} G(P, q|X, a_1^S) \\
 &\quad + \frac{1-\alpha}{N} \sum_{n=1}^N \sum_{a_1^S: a_1^S=a_{n,1}^{S_n}} \frac{1}{N} \sum_{n'=1}^N \sum_{X: X=X_{n'}} G(P, q|X, a_1^S) \\
 &= \frac{\alpha}{N} \sum_{n=1}^N G(P, q|X_n, a_{n,1}^{S_n}) + \frac{1}{N} \sum_{n=1}^N \sum_{n'=1}^N \frac{1-\alpha}{N} G(P, q|X_{n'}, a_{n,1}^{S_n}) \\
 &= \frac{1}{N} \sum_{n=1}^N \sum_{n'=1}^N \beta(n, n') G(P, q|X_{n'}, a_{n,1}^{S_n}) \\
 &= \sum_{n=1}^N \sum_{n'=1}^N \frac{\beta(n, n')}{N} \sum_{s=1}^{S_n+1} \sum_{a \in \mathcal{V}^+} P(a|a_{n,1}^{s-1}, X_{n'}) \log \frac{P(a|a_{n,1}^{s-1}, X_{n'})}{q_\theta(a|a_{n,1}^{s-1})}
 \end{aligned}$$

The derivation is done and we have shown that Eq. (9) and (10) are equivalent.