Heuristic-free Knowledge Distillation for Streaming ASR via Multi-modal Training

Ji Won Yoon

Department of AI, Chung-Ang University, Seoul, South Korea.
jiwonyoon@cau.ac.kr

Abstract

Existing knowledge distillation (KD) studies for streaming automatic speech recognition (ASR) adopt a non-streaming model as the teacher and a streaming model as the student, respectively. Since the non-streaming teacher usually has less emission latency compared to the streaming student, the teacher's prediction is typically shifted by τ frames, where the parameter τ is selected heuristically. In this paper, we observe that this manual shifting is sub-optimal and propose a novel framework, namely *Heuristic-free KD*. Instead of leveraging knowledge from the non-streaming teacher model, we employ a self-distillation setup, distilling the knowledge within the streaming architecture itself. Since the teacher and student share the same streaming ASR backbone, the alignment mismatch issue can be effectively mitigated without requiring any time shifting by τ. Additionally, we incorporate full-context textual information as an auxiliary multi-modal input for the proposed teacher. Although the streaming architecture lacks future context, the additional linguistic input enables it to generate more accurate knowledge for self-distillation. We empirically demonstrate that the proposed KD approach significantly improves the performance of the streaming ASR model, outperforming conventional methods that rely on the offline teacher and heuristic parameter.

Introduction

In recent years, a streaming automatic speech recognition (ASR) framework has attracted significant attention for its ability to transcribe spoken language into text in real time. Such capability is crucial for enhancing interactions with various applications, including real-time captioning, on-the-fly language translation, voice command recognition, and dialogue systems, by providing immediate feedback and interactions. However, despite its potential, the streaming ASR model often underperforms compared to its non-streaming (a.k.a. offline or full-context) counterpart. This performance degradation stems from a significant constraint: streaming ASR lacks future context, a limitation designed to minimize the delay between spoken input and textual output. Improving performance has proven to be challenging in streaming ASR (He et al. 2019; Li et al. 2020; Sainath et al. 2020).

To bridge the gap between offline and online ASR models, there have been numerous efforts to adopt knowledge distillation (KD) (Hinton, Vinyals, and Dean 2014; Bucila, Caruana, and Niculescu-Mizil 2006). In the context of KD for streaming ASR, the non-streaming and streaming models are typically adopted as the teacher and student, respectively. Since non-streaming models benefit from having access to the entire speech utterance, they can offer more accurate hypotheses, thus producing better knowledge for the ASR task. By leveraging the knowledge from a powerful offline teacher during training, the online student can improve its performance compared to its baseline, which is trained solely with ground truth.

Despite the widespread adoption of KD for streaming ASR, most existing methods share a common limitation: they manually shift the output of the non-streaming teacher in a heuristic manner (Yang, Li, and Woodland 2022; Weninger et al. 2022; Yu et al. 2021b). Since the non-streaming model (teacher) has no incentive to delay its output, its predictions usually have lower latency compared to those of the streaming model (student). Due to significant misalignment between the teacher and student models, it has been confirmed that naive frame-to-frame KD can misguide the streaming student (Yu et al. 2021b; Inaguma and Kawahara 2021; Liang et al. 2023). Therefore, the teacher's predictions are typically required to be shifted by τ frames before being transferred to the student model. Here, the shifting parameter τ is determined heuristically. Setting a sensible value for τ is crucial for successful knowledge transfer, as the performance of the student model varies with different τ values (Yang, Li, and Woodland 2022).

In this paper, we observe that using a fixed parameter τ for KD is sub-optimal and introduce a novel framework for streaming ASR, namely *Heuristic-free KD*. Specifically, Heuristic-free KD encapsulates two key components. First, we employ a self-distillation setup for streaming ASR, which aims at distilling the knowledge within the streaming architecture itself without any extra non-streaming teacher. Rather than extracting knowledge from the non-streaming teacher, the proposed teacher and student share the same streaming ASR model parameters, ensuring that the frame-level alignment from the teacher is highly similar to that of the student. This allows the teacher's knowledge to be transferred without requiring any time shifting by τ. Second, we incorporate full-context textual information as an auxiliary multi-modal input for the proposed teacher. In par-

ticular, the student equipped with multi-modal fusion acts as the teacher, and the predictions benefiting from linguistic information are then transferred to the original streaming student. Even within the self-distillation framework using the streaming architecture, the proposed multi-modal training not only mitigates the streaming model's context scarcity but also generates a more accurate guidance for self-distillation. We implement cross-attention in multi-modal fusion, enabling the proposed teacher to learn the inter-modal dependencies between acoustic and linguistic features.

Extensive experiments are conducted on the LibriSpeech (Panayotov et al. 2015) benchmark, utilizing two ASR architectures: connectionist temporal classification (CTC) (Graves et al. 2006) and hybrid Transducer-CTC (Noroozi et al. 2024), since both CTC and Transducer (Graves 2012) models have gained favor in real-world streaming settings recently (Wang et al. 2023; Tian et al. 2023a,b; Yao et al. 2021; Zhang et al. 2022; Yu et al. 2021a). Compared to existing KD methods that rely on a powerful offline teacher and the shifting parameter τ, the proposed KD significantly improves the student's performance, achieving the best results in all configurations. It is important to note that Heuristic-free KD does not require any extra teacher model and time shifting by τ. In a detailed case study, we also confirm that the proposed teacher can generate more accurate knowledge for KD while preserving the alignment of the student.

Related Work

KD aims at transferring knowledge from a computationally intensive teacher model to a more efficient student model. By transferring the knowledge, the distilled student performs better than its baseline. It has been shown that the distillation framework is effective for compressing full-context ASR models (Li et al. 2014; Kim and Rush 2016; Kim et al. 2019; Senior et al. 2015; Takashima, Li, and Kawai 2018, 2019; Yoon et al. 2021, 2022).

In the context of KD for streaming ASR, the non-streaming and streaming models are typically adopted as the teacher and student, respectively. Due to the significant alignment mismatch problem between the two models, previous studies have suggested shifting the teacher model's knowledge by τ frames (Shim et al. 2023; Yang, Li, and Woodland 2022; Weninger et al. 2022), employing multi-stage training (Kurata and Saon 2020), and using CTC alignments (Inaguma and Kawahara 2021). Among these methods, the most widely-used approach is time shifting with τ.

Recently, dual-mode training (Yu et al. 2021b; Liang et al. 2023; Liu et al. 2022) has been proposed, based on self-distillation and using the same model for both streaming and non-streaming scenarios. Despite its promise, there are significant limitations. Firstly, this approach requires many architectural modifications, including dual-mode convolution, dual-mode average pooling, dual-mode self-attention, and dual-mode normalization, making it difficult to reproduce. Additionally, the dual-mode framework still requires the heuristic parameter τ since it involves KD from non-streaming to streaming models. In contrast to the conventional dual-mode training, the Heuristic-free KD does not modify the original ASR model architecture and does not rely on both non-streaming teacher and time shifting with τ, providing a more general solution for KD. Furthermore, the primary focus of our proposed KD method is on improving the overall quality of the ASR student model rather than reducing latency.

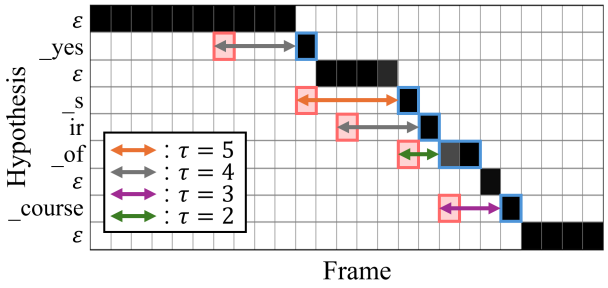

Figure 1: Frame-wise alignment example of streaming student baseline for utterance 7902-96594-0016 in LibriSpeech test-other dataset, where the target is 'YES SIR OF COURSE'. The red boxes highlight the non-blank predictions from the non-streaming teacher model. The two-way arrows in various colors indicate different optimal τ shifts. 'ϵ' represents the blank label in CTC.

Proposed Method

Motivation

When applying the conventional time shifting with τ, two primary challenges arise. Firstly, identifying the proper τ value requires additional parameter tuning, which involves further training sessions and subsequent performance evaluations. The ideal τ is variable and affected by factors such as model architecture, dataset characteristics, and streaming context. For example, Dual-mode ASR (Yu et al. 2021b) performs a small-scale hyper-parameter sweep, shifting from -2 to 2 frames. According to Yang, Li, and Woodland (2022), the best-performing τ is 7, while in our experiments, $\tau = 5$ yielded the best performance for competing KD methods on LibriSpeech. Secondly, our empirical observations suggest that adopting a single τ value for all alignments could be sub-optimal. As shown in Figure 1, to align the 's' label of the teacher model with the 's' label of the student model, the τ value should be adjusted to 5. Meanwhile, other labels might achieve better performance with τ values of 4, 3, or 2. This variation indicates that employing a single τ value for time shifting may not be optimal for all scenarios, potentially resulting in less effective KD for the streaming student. Given the importance of the role of streaming ASR models in various industries, it's surprising how little research has been devoted to eliminating the heuristic parameter τ. This motivated us to ask a simple question: how can we provide optimal guidance for a streaming student model without using a non-streaming teacher and manual shifting with τ?

Heuristic-free KD

In this paper, we introduce Heuristic-free KD for streaming ASR that does not require the manual shifting of τ frames during the distillation process.

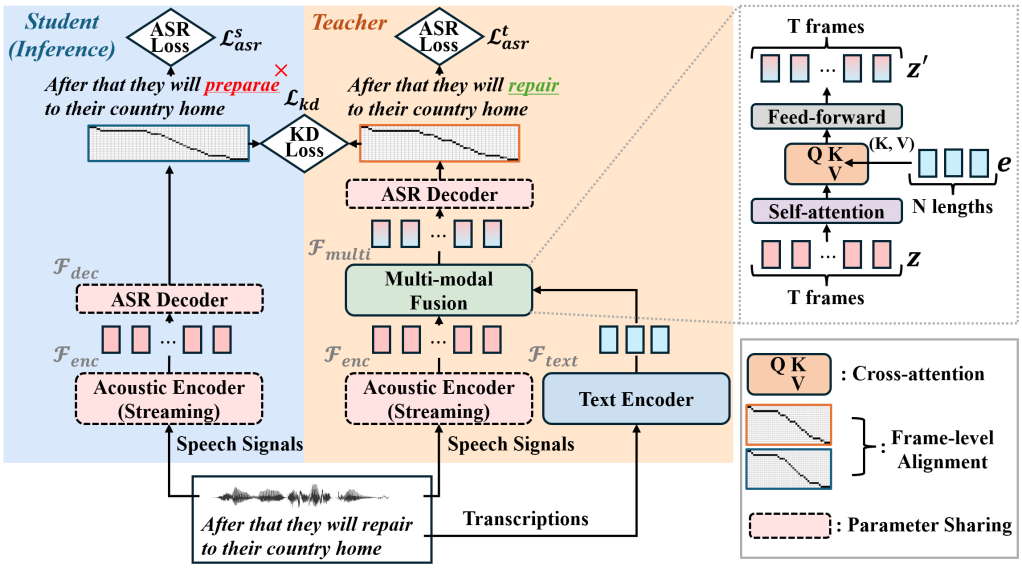

Figure 2: Overview of Heuristic-free KD framework for streaming ASR. In the proposed self-distillation framework, the streaming ASR model equipped with multi-modal fusion serves as the teacher mode. The teacher and student modes share the parameters of both $\mathcal{F}_{enc}$ and $\mathcal{F}_{dec}$. Consider that the student makes an erroneous prediction with "repair", predicting it as "preparae". The proposed teacher with multi-modal training benefits from the full-context target text and then provides more accurate knowledge for KD. During multi-modal fusion, cross-attention utilizes the acoustic representation as queries and the textual representation as key-value pairs.

Self-distillation As aforementioned, it has been proven that the predictions of the non-streaming teacher have a significant mismatch with those of the streaming student. The proposed framework refers to the idea that instead of using the knowledge of the non-streaming model, we rather employ a self-distillation setup, distilling the knowledge within the streaming architecture itself. In Heuristic-free KD, the teacher and student share the same streaming ASR backbone, ensuring that the frame-level alignment of the teacher is highly similar to that of the student. Therefore, the alignment mismatch issue can be effectively mitigated, allowing the teacher's knowledge to be transferred without requiring any time shifting by τ. Since the teacher and student utilize the same model parameters, the proposed KD can be considered as self-distillation, thus using the terms "teacher mode" and "student mode" rather than "teacher model" and "student model".

Multi-modal Training The limited context available to the conventional streaming model restricts its ability to generate informative knowledge for self-distillation. The key idea behind the proposed approach is to refine the output of the student model via multi-modal training and then use it as guidance for KD. Inspired by prior research (Shim, Choi, and Sung 2022) that revealed the significance of linguistic information for predicting ASR targets, our framework injects full-context textual information into the streaming model, which serves as the teacher mode (see Figure 2). Even though the streaming model lacks future context, the additional target text input provides full-context linguistic information. This enables the streaming architecture to generate more accurate knowledge for self-distillation.

Student Mode. Without loss of generality, we consider an ASR model that maps the speech input x to a latent representation z and predicts the transcription y with an acoustic encoder $\mathcal{F}_{enc}$ and a decoder $\mathcal{F}_{dec}$, as follows:

$$\mathcal{F}_{enc}(x) \rightarrow z, \quad \mathcal{F}_{dec}(z) \rightarrow y. \tag{1}$$

This sequence represents the forward pass of the student model in our KD framework. It is important to note that during inference, only the parameters of the student's encoder $\mathcal{F}_{enc}$ and decoder $\mathcal{F}_{dec}$ are utilized.

Teacher Mode. In the Heuristic-free KD framework, the student equipped with multi-modal fusion serves as the teacher. Before the modality fusion, the target transcription y is transformed by a text encoder $\mathcal{F}_{text}$, which is composed of an embedding layer followed by a single self-attention layer, to effectively incorporate y into the fusion process. Meanwhile, the speech input x is transformed by the acoustic encoder $\mathcal{F}_{enc}$. The process can be described as follows:

$$\mathcal{F}_{enc}(x) \rightarrow z, \quad \mathcal{F}_{text}(y) \rightarrow e. \tag{2}$$

The multi-modal fusion block $\mathcal{F}_{multi}$ is inserted before the ASR decoder $\mathcal{F}_{dec}$. Motivated by recent multi-modal studies (Radhakrishnan et al. 2023; Nadeem et al. 2024; Yoon et al. 2023a,b; Qian et al. 2023), we leverage a cross-attention mechanism within $\mathcal{F}_{multi}$ for fusion, enabling the teacher to effectively learn inter-modal dependencies between acoustic and textual features. Specifically, the acoustic embedding is used as the query, and the textual embedding is leveraged as key-value pairs. Though the acoustic

embedding from the streaming encoder $\mathcal{F}_{enc}$ lacks future context, this selective fusion approach not only facilitates the integration of auditory and linguistic information but also addresses the context scarcity of the streaming architecture, thus generating better guidance for self-distillation. This modification restructures the ASR framework as outlined below:

$$\mathcal{F}_{multi}(z, e) \rightarrow z', \quad \mathcal{F}_{dec}(z') \rightarrow y. \tag{3}$$

Since the teacher's forward pass shares the parameters $\mathcal{F}_{enc}$ and $\mathcal{F}_{dec}$ with the student, the additional parameters employed during this forwarding are only $\mathcal{F}_{multi}$ and $\mathcal{F}_{text}$.

Training Objective. Our self-distillation framework consists of three training objectives: the ASR loss for the student $\mathcal{L}^s_{asr}$, the ASR loss for the teacher $\mathcal{L}^t_{asr}$, and the KD loss $\mathcal{L}_{kd}$. Both $\mathcal{L}^s_{asr}$ and $\mathcal{L}^t_{asr}$ are original ASR losses. When training the CTC model, they represent CTC losses; for the hybrid CTC-Transducer model, they are hybrid CTC-Transducer losses. For $\mathcal{L}_{kd}$, we use the SKD loss (Yoon et al. 2021) to transfer the frame-level predictions from the teacher to the student. Therefore, the final objective is given as follows:

$$\mathcal{L}_{total} = \mathcal{L}^s_{asr} + \mathcal{L}^t_{asr} + \lambda \cdot \mathcal{L}_{kd} \tag{4}$$

where λ is a tunable parameter to balance the two ASR losses and the KD loss.

Comparison with Text Injection ASR. Our approach appears similar to recent text injection ASR approaches at first glance (Kang et al. 2022; Thomas et al. 2022b; Zhang et al. 2024; Yao et al. 2022; Yue et al. 2023; Chen et al. 2022; Thomas et al. 2022a; Yoon et al. 2023b,a). However, our motivation is different. The main purpose of multi-modal training in our framework is to provide full-context textual information to the streaming model, thereby enabling it to extract more accurate guidance for self-distillation. Unlike the offline ASR model that can consider the entire utterance, the streaming ASR model lacks future context, which makes it difficult to generate informative knowledge from the streaming teacher. Motivated by prior work (Shim, Choi, and Sung 2022) that underscored the significance of linguistic information for the ASR task, we use the target text as an auxiliary multi-modal input for the teacher. This approach overcomes the context scarcity in the streaming scenario and effectively improves self-distillation. Interestingly, although the target text directly provides informative context for the ASR task, it does not lead to a trivial solution where the model merely copies the target input as its output, as will be shown in the analysis section. This behavior is related to the inherent property of the CTC model. In the CTC framework, the relationship between the CTC alignment π and target y is many-to-one. Since it is extremely difficult to predict the "many" from the "one", the CTC model cannot directly copy the target input (e.g., {c,a,t}) into the output CTC alignment (e.g., {c,c,c,c,a,a,a,a,a,t,t,t}), even when the target is used as additional input.

Experiments

Experimental Setup

Dataset and Metrics. We evaluated the performance of models using the LibriSpeech (Panayotov et al. 2015) benchmark, the most widely used ASR dataset, which is freely available under the CC BY 4.0 license. During the training, we employed 'train-clean-100', 'train-clean-360', and 'train-other-500'. For evaluations, we utilized 'dev-clean', 'dev-other', 'test-clean', and 'test-other'. The experimental results on the Common Voice 7.0 Spanish dataset (Ardila et al. 2020) can be found in the extended Appendix. Two widely-used metrics were used to evaluate performance: word error rate (WER) and relative error rate reduction (RERR). WER is a standard metric for assessing speech recognition accuracy, while RERR quantifies the proportionate reduction in WER compared to a baseline.

Implementation. Our experiments were mainly conducted with the NeMo (Kuchaiev et al. 2019) toolkit, and greedy decoding was applied to compare WERs. We implemented two ASR architectures for comparison: CTC and hybrid Transducer-CTC. For the CTC, we adopted cache-aware streaming FastConformer (Noroozi et al. 2024) as the student baseline. Two streaming settings were considered: look-aheads of 1040 ms and 480 ms, with corresponding right context sizes of 13 and 6, respectively. Although no look-ahead will decrease latency, we kindly note that using a look-ahead framework is generally accepted in the streaming ASR scenario (Noroozi et al. 2024; Chen et al. 2024; Kumar et al. 2024; Zhang et al. 2020). To reproduce conventional KD methods, the FastConformer (Rekesh et al. 2023) was employed as the non-streaming CTC teacher model. Guided CTC (Kurata and Audhkhasi 2019) and SKD (Yoon et al. 2021) were adopted as competing approaches. Since Dual-mode ASR (Yu et al. 2021b) required many architectural modifications, it was difficult to fairly compare the performance of KD methods while preserving the original ASR model structure. Given that Dual-mode ASR employed the conventional KD approach using the shifting parameter τ, we believe our experiments with various KD methods and settings sufficiently demonstrated the effectiveness of our approach. In the case of the hybrid Transducer-CTC, cache-aware streaming FastConformer was utilized as both the teacher and student baseline since we conducted a streaming-to-streaming KD scenario. For all conventional teacher models, we used the pre-trained checkpoints provided by the NeMo. The training details will be provided in Appendix.

Experimental Results

Table 1 reports the WER and RERR results on the LibriSpeech benchmark. We adopted the cache-aware streaming FastConformer as the student baseline, which had 12.8 M parameters. For competing KD methods, including Guided CTC training (Kurata and Audhkhasi 2019) and SKD (Yoon et al. 2021), the large size (about 115 M parameters) of FastConformer-CTC was utilized as the non-streaming teacher model. According to the original papers of Guided CTC and SKD, it has been confirmed that both KD

Method	τ	dev-clean		dev-other		test-clean		test-other	
		WER	**RERR**	**WER**	**RERR**	**WER**	**RERR**	**WER**	**RERR**
Teacher (Offline)	–	1.87 %	–	4.22 %	–	2.08 %	–	4.22 %	–
Student (Online)	–	7.46 %	–	17.10 %	–	7.42 %	–	17.49 %	–
SKD	0	8.77 %	-17.56 %	18.37 %	-7.42 %	8.99 %	-21.16 %	19.00 %	-8.63 %
	3	6.79 %	8.98 %	16.42 %	4.04 %	6.97 %	6.06 %	16.07 %	8.12 %
	5	6.60 %	11.53 %	16.12 %	5.73 %	6.78 %	8.63 %	15.79 %	9.71 %
	7	7.08 %	5.09 %	16.19 %	5.32 %	7.21 %	2.83 %	16.16 %	7.61 %
Guided CTC	0	10.58 %	-41.82 %	20.49 %	-19.82 %	10.49 %	-41.37 %	20.84 %	-19.08 %
	3	7.58 %	-1.61 %	17.80 %	-4.09 %	7.51 %	-1.21 %	17.63 %	-0.80 %
	5	7.03 %	5.76 %	17.13 %	-0.18 %	7.16 %	3.50 %	17.18 %	1.77 %
	7	7.88 %	-5.63 %	17.78 %	-3.98 %	8.10 %	-9.16 %	17.41 %	0.46 %
Ours, Heuristic-free KD	–	**5.95** %	**20.24** %	**14.89** %	**12.92** %	**6.24** %	**15.90** %	**15.05** %	**13.94** %

Table 1: Streaming CTC student with a look-ahead of 1040 ms: comparison of WER and RERR on the LibriSpeech benchmark. Note that the proposed method did not require an additional teacher model. The offline teacher model was only used for conventional KD methods. Results that represent superior performance are highlighted in bold.

Method	τ	dev-clean		dev-other		test-clean		test-other	
		WER	**RERR**	**WER**	**RERR**	**WER**	**RERR**	**WER**	**RERR**
Teacher (Offline)	–	1.87 %	–	4.22 %	–	2.08 %	–	4.22 %	–
Student (Online)	–	5.45 %	–	13.89 %	–	5.59 %	–	13.89 %	–
Guided CTC	5	5.34 %	2.02 %	13.82 %	0.50 %	5.52 %	1.25 %	13.93 %	-0.29 %
SKD	5	5.36 %	1.65 %	13.85 %	0.29 %	5.78 %	-3.40 %	13.67 %	1.58 %
Ours, Heuristic-free KD	–	**4.68** %	**14.13** %	**13.24** %	**4.68** %	**4.96** %	**11.27** %	**12.80** %	**7.85** %

Table 2: Streaming CTC student with a look-ahead of 480 ms: comparison of WER and RERR on the LibriSpeech benchmark. Results that represent superior performance are highlighted in bold.

frameworks were effective in minimizing the alignment mismatch between non-streaming CTC models, without using manual frame shifting. However, as reported in the results, both SKD and Guided CTC with $\tau = 0$ failed to mitigate the misalignment between the non-streaming teacher and streaming student, resulting in the distilled student performing worse than its baseline. This finding suggests that the issue at hand was particularly challenging. While varying the value of τ, we found that $\tau = 5$ performed well with conventional KD methods compared to other settings. The results in Table 1 show that the proposed framework achieved the best performance in all configurations. It yielded an RERR of 20.24 % on dev-clean and 12.92 % on dev-other, respectively. Considering that the best results of SKD and Guided-CTC were RERR 11.53 % and RERR 5.76 % on dev-clean, the Heuristic-free KD demonstrated a significant improvement over these conventional methods. Note that the proposed method did not require any pre-trained teacher and the shifting parameter τ. While the other approaches used the extra teacher that had about 115 M parameters, Heuristic-free KD required only 1.3 M additional parameters for KD.

To further validate the effectiveness of Heuristic-free KD, we considered a different setting: the look-ahead was set to 480 ms, which is suitable for low-latency streaming applications. Since reduced right context may lead to performance degradation, a larger student baseline (about 20.4 M parameters) was leveraged to improve the WER performance. Note that this configuration did not make the results incomparable across different look-aheads. In KD, enhancing an already strong student model is generally more difficult than improving a weaker one. By using a larger student model with a 480 ms look-ahead, we mitigated the degradation issue caused by the reduced right context, ensuring a fair but more challenging comparison within the KD framework. We set $\tau = 5$ for competing methods, as this setting performed well in previous experiments. As presented in Table 2, the Heuristic-free KD outperformed the conventional methods. Interestingly, we found that the distilled students using SKD and Guided CTC did not always perform better than the student baseline; in some cases, their performance was even worse. In configurations where conventional methods did improve the student's performance, the RERR values were minimal, indicating that the low-latency scenario with the look-ahead of 480 ms was more challenging than the previous one. In contrast, the Heuristic-free KD showed significant WER improvements in all cases, achieving RERR of 11.27 % on test-clean and 7.85 % on test-other, respectively.

We proceeded to verify how well the proposed KD approach can distill knowledge using the hybrid Transducer-CTC architecture, a recent version of the Transducer with

Method	dev-clean		dev-other		test-clean		test-other	
	WER	RERR	WER	RERR	WER	RERR	WER	RERR
Teacher (Online)	2.20 %	–	5.60 %	–	2.33 %	–	5.54 %	–
Student (Online)	5.28 %	–	13.32 %	–	5.53 %	–	13.68 %	–
SKD	5.20 %	1.52 %	13.21 %	0.83 %	5.39 %	2.53 %	13.49 %	1.39 %
Ours, Heuristic-free KD	**5.00** %	**5.30** %	**13.06** %	**1.95** %	**5.14** %	**7.05** %	**12.87** %	**5.93** %

Table 3: Streaming hybrid Transducer-CTC with a look-ahead of 1040 ms: comparison of WER and RERR on the LibriSpeech benchmark. Results that represent superior performance are highlighted in bold.

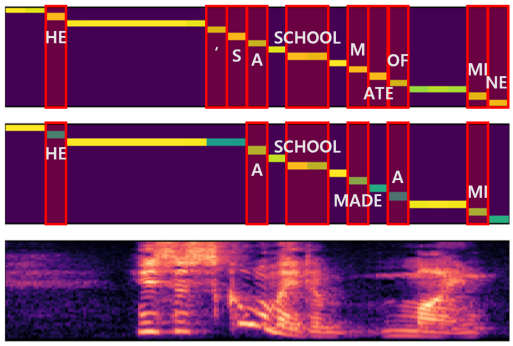

Figure 3: From top to bottom: frame-wise alignments generated with the teacher with multi-modal fusion, the distilled student in the proposed framework, and the corresponding spectrogram, where the target transcription is "HE'S A SCHOOLMATE OF MINE". For the frame-level alignment, the x-axis and y-axis represent frames and hypotheses, respectively. The red boxes highlight the non-blank predictions.

an added CTC layer on top of the encoder. Since we experimentally found that KD using the CTC outputs was more effective than using those of the Transducer, we used the SKD approach with CTC outputs as a comparative method. For fair comparison, the $\mathcal{L}_{kd}$ for Heuristic-KD also minimized the distance between the CTC outputs. In this case, $\mathcal{F}_{enc}$ and $\mathcal{F}_{dec}$ equated to the encoder and CTC decoder, respectively. Unlike previous experiments focused on KD from non-streaming to streaming models, this experiment considered KD from streaming to streaming models, which was an unexplored setting so far. Since both the teacher and student had the same look-ahead setting of 1040 ms, the prediction from the streaming teacher did not need to be shifted using parameter τ. Table 3 summarizes the WER and RERR results on LibriSpeech. Compared to SKD, Heuristic-free KD achieved better WER performance in all configurations, providing RERR of 7.05 % and 5.93 % on test-clean and test-other, respectively. We confirmed that the proposed framework still performed well with the Transducer-based model.

Analysis

Frame-level Alignments in Heuristic-free KD. As shown in Figure 3, we contrasted frame-wise alignments of the distilled student and teacher in the proposed framework. The distilled student produced the prediction "HE A SCHOOL MADE A MI", struggling to predict the correct sentence except for the words "HE" and "SCHOOL". However, the proposed teacher with multi-modal training offered an accurate prediction "HE'S A SCHOOLMATE OF MINE" while maintaining a similar frame-wise alignment to that of the student. As shown in Figure 3, the proposed teacher not only minimized the alignment mismatch issue but also successfully refined the output of the student, converting "MADE A" to "MATE OF" and "MI" to "MINE", respectively. This implies that it can provide more optimal and accurate knowledge for the streaming student, eliminating the need for the shifting parameter τ. Though the conventional streaming model struggled to generate accurate knowledge due to limited context, Heuristic-free KD could provide informative knowledge for the student within the streaming architecture.

Teacher Model	Params.	clean	other
CTC (Offline)	115 M	2.08 %	4.22 %
Proposed Teacher	14 M	1.92 %	5.04 %
	22 M	1.52 %	4.09 %
Transducer-CTC (Online)	114 M	2.33 %	5.54 %
Proposed Teacher	16 M	1.28 %	3.21 %

Table 4: Teacher performance comparison on the LibriSpeech test datasets. CTC (Offline) refers to the non-streaming CTC teacher model in Table 1, and Transducer-CTC (Online) refers to the streaming hybrid Transducer-CTC teacher model in Table 3.

WER Performance of Teachers. We compared the proposed teacher with the non-streaming and streaming teacher models used for competing KD methods. As mentioned earlier, we used the checkpoints provided by the NeMo (Kuchaiev et al. 2019) for the conventional teacher models, as they required a significant amount of GPU resources and training time. Table 4 reports the WER performance of the teacher models. Unlike the conventional teacher models, the proposed teacher required fewer additional parameters for KD. For example, the proposed teacher for CTC had about 14 M parameters, with 12.8 M for the ASR student model ($\mathcal{F}_{enc}$ and $\mathcal{F}_{dec}$) and 1.3 M for the multi-modal fusion ($\mathcal{F}_{text}$ and $\mathcal{F}_{multi}$). For hybrid Transducer-CTC framework, the proposed teacher had about 16 M parameters, with

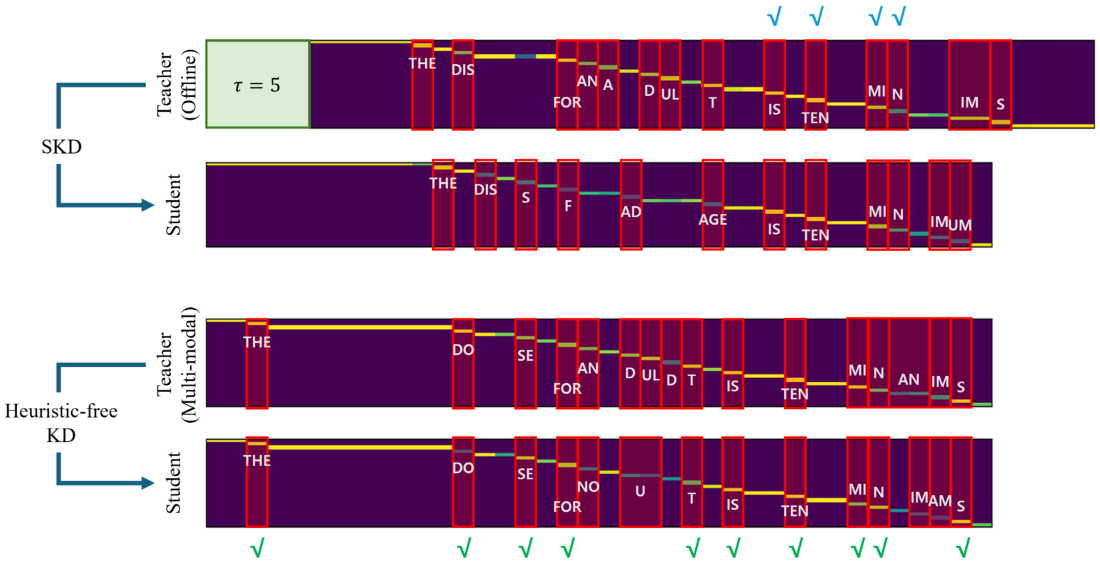

Figure 4: Frame-level alignment examples in the test-other dataset, where the target is "THE DOSE FOR AN ADULT IS TEN MINIMS". The x-axis and y-axis represent acoustic frames and hypotheses, respectively. The red boxes highlight the non-blank predictions.

the multi-modal fusion requiring 1.2 M. From the results, we confirmed that the WER results of the proposed teacher in Heuristic-free KD were more accurate than those of the conventional teacher models in most configurations, though not zero. If the proposed framework copied the target text input as its output, the WER performance would be zero, indicating that the model perfectly predicted the target and resulted in a trivial solution during training. It is verified that the Heuristic-KD not only performed better in training the streaming student but also prevented the trivial solution while using the target text input.

Alignment Comparison with Conventional KD. As illustrated in Figure 4, we presented four frame-level alignments generated by the offline teacher, the distilled student using SKD, the proposed teacher, and the distilled student in the Heuristic-free KD, respectively. Since $\tau = 5$ yielded the best performance for SKD, we shifted the prediction of the non-streaming teacher model by 5 frames. In Figure 4, the check marks indicate frames where the prediction was correct and the frame-level alignment between the teacher and student matched. While the distilled student with SKD only matched four frames with its teacher, the student in the proposed framework aligned with its teacher on most of the frames. This means that the Heuristic-free KD minimized the gap between the teacher and student more effectively than conventional KD.

Ablation Study

When training the proposed framework, we considered one tunable parameter λ in Eq. (4). As shown in Figure 5, we evaluated the WER performance while varying the parameter λ from $\{0.500, 0.250, 0.667, 0.125\}$ From the results, it is verified that the best WER performance on LibriSpeech was obtained when $\lambda = 0.250$.

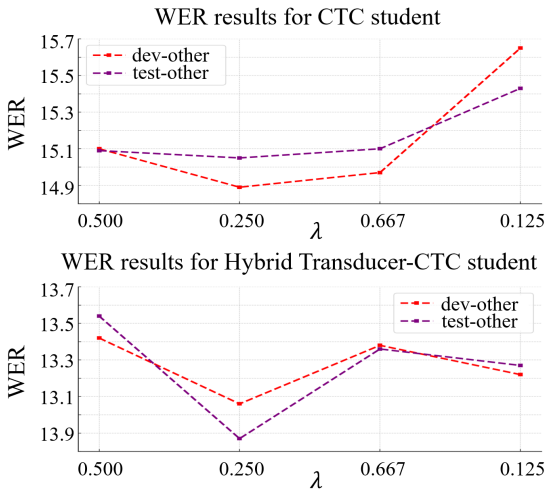

Figure 5: WER performance on LibriSpeech. The evaluation is conducted with varying values of λ.

Conclusions

In this paper, we introduced a novel framework for streaming ASR, termed Heuristic-free KD. By leveraging the self-distillation setup, we effectively mitigated the alignment mismatch during KD, eliminating the requirement for non-streaming teacher and shifting parameter τ. Additionally, we enhanced the overall quality of knowledge within the streaming architecture by incorporating linguistic information as an auxiliary multi-modal input. Our empirical results demonstrated that the proposed KD approach significantly improved the performance of the streaming model, surpassing conventional methods that used the extra non-streaming teacher model and heuristic parameter.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2021-0-01341,Artificial Intelligence Graduate School Program(Chung-Ang University)).

References

Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2020. Common voice: a massively-multilingual speech corpus. In *Proc. LREC*.

Bucila, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model compression. In *Proc. ACM SIGKDD*, 535–541.

Chen, Z.; Huang, H.; Hrinchuk, O.; Puvvada, K. C.; Koluguri, N. R.; Żelasko, P.; Balam, J.; and Ginsburg, B. 2024. Bestow: Efficient and streamable speech language model with the best of two worlds in gpt and t5. *arXiv preprint arXiv:2406.19954v1*.

Chen, Z.; Zhang, Y.; Rosenberg, A.; Ramabhadran, B.; Moreno, P.; Bapna, A.; and Zen, H. 2022. Maestro: matched speech text representations through modality matching. In *Proc. INTERSPEECH*.

Graves, A. 2012. Sequence transduction with recurrent neural networks. In *Proc. ICML Workshop on Representation Learning*.

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML*, 369–376.

He, Y.; Sainath, T. N.; Prabhavalkar, R.; McGraw, I.; Alvarez, R.; Zhao, D.; Rybach, D.; Kannan, A.; Wu, Y.; Pang, R.; Liang, Q.; Bhatia, D.; Shangguan, Y.; Li, B.; Pundak, G.; Sim, K. C.; Bagby, T.; Chang, S.; Rao, K.; and Gruenstein, A. 2019. Streaming end-to-end speech recognition for mobile devices. In *Proc. ICASSP*.

Hinton, G.; Vinyals, O.; and Dean, J. 2014. Distilling the knowledge in a neural network. In *Proc. NIPS Workshop Deep Learn.*

Inaguma, H.; and Kawahara, T. 2021. Alignment knowledge distillation for online streaming attention-based speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Kang, Y.; Liu, T.; Li, H.; Hao, Y.; and Ding, W. 2022. Self-supervised audio-and-text pre-training with extremely low-resource parallel data. In *Proc. AAAI*.

Kim, H. G.; Na, H.; Lee, H.; Lee, J.; Kang, T. G.; Lee, M. J.; and Choi, Y. S. 2019. Knowledge distillation using output errors for self-attention end-to-end models. In *Proc. ICASSP*.

Kim, Y.; and Rush, A. 2016. Sequence-level knowledge distillation. In *Proc. EMNLP*.

Kuchaiev, O.; Li, J.; Nguyen, H.; Hrinchuk, O.; Leary, R.; Ginsburg, B.; Kriman, S.; Beliaev, S.; Lavrukhin, V.; Cook, J.; Castonguay, P.; Popova, M.; Huang, J.; and Cohen, J. M. 2019. Nemo: a toolkit for building AI applications using neural modules. *arXiv preprint arXiv:1909.09577*.

Kumar, S.; Madikeri, S.; Gomez, J. Z.; Tello, E. V.; Nigmatulina, I.; Motlicek, P.; E, M.; and Ganapathiraju, A. 2024. Xlsr-transducer: streaming asr for self-supervised pretrained models. *arXiv preprint arXiv:2407.04439v1*.

Kurata, G.; and Audhkhasi, K. 2019. Guiding ctc posterior spike timings for improved posterior fusion and knowledge distillation. In *Proc. INTERSPEECH*.

Kurata, G.; and Saon, G. 2020. Knowledge distillation from offline to streaming rnn transducer for end-to-End speech recognition. In *Proc. INTERSPEECH*.

Li, B.; Chang, S.; Sainath, T. N.; Pang, R.; He, Y.; Strohman, T.; and Wu, Y. 2020. Towards fast and accurate streaming end-to-end asr. In *Proc. ICASSP*.

Li, J.; Zhao, R.; Huang, T. J.; and Gong, Y. 2014. Learning small-size DNN with output-distribution-based criteria. In *Proc. INTERSPEECH*.

Liang, C.; Zhang, X. L.; Zhang, B.; Wu, D.; Li, S.; Song, X.; Peng, Z.; and Pan, F. 2023. Fast-u2++: Fast and accurate end-to-end speech recognition in joint ctc/attention frames. In *Proc. ICASSP*.

Liu, C.; Shangguan, Y.; Yang, H.; Shi, Y.; Krishnamoorthi, R.; and Kalinli, O. 2022. Learning a dual-mode speech recognition model via self-pruning. In *Proc. IEEE SLT*.

Nadeem, A.; Hilton, A.; Dawes, R.; Thomas, G.; and Mustafa, R. 2024. Cad-contextual multi-modal alignment for dynamic avqa. In *Proc. WACV*.

Noroozi, V.; Majumdar, S.; Kumar, A.; Balam, J.; and Ginsburg, B. 2024. Stateful conformer with cache-based inference for streaming automatic speech recognition. In *Proc. ICASSP*.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *Proc. ICASSP*, 5206–5210.

Qian, X.; Wang, Z.; Wang, J.; Guan, G.; and Li, H. 2023. Audio-visual cross-attention network for robotic speaker tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 550 – 562.

Radhakrishnan, S.; Yang, C.; Khan, S.; Kumar, R.; Kiani, N.; Cabrero, D.; and Tegnér, J. 2023. Whispering llama: a cross-modal generative error correction framework for speech recognition. In *Proc. EMNLP*.

Rekesh, D.; Koluguri, N.; Kriman, S.; Majumdar, S.; Noroozi, V.; Huang, H.; Hrinchuk, O.; Puvvada, K.; Kumar, A.; Balam, J.; and Ginsburg, B. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. In *Proc. ASRU*.

Sainath, T. N.; He, Y.; Li, B.; Narayanan, A.; Pang, R.; Bruguier, A.; Chang, S.; Li, W.; Alvarez, R.; Chen, Z.; Chiu, C.; Garcia, D.; Gruenstein, A.; Hu, K.; Jin, M.; Kannan, A.; Liang, Q.; McGraw, I.; Peyser, C.; Prabhavalkar, R.; Pundak, G.; Rybach, D.; Shangguan, Y.; Sheth, Y.; Strohman, T.; Visontai, M.; Wu, Y.; Zhang, Y.; and Zhao, D. 2020. A streaming on-device end-to-end model surpassing server-side conventional model quality and latency. In *Proc. ICASSP*.

Senior, A.; Sak, H.; C. Quitry, F.; Sainath, T.; Rao, K.; et al. 2015. Acoustic modelling with cd-ctc-smbr lstm rnns. In *Proc. ASRU*, 604–609.

Shim, K.; Choi, J.; and Sung, W. 2022. Understanding the role of self attention for efficient speech recognition. In *Proc. ICLR*.

Shim, K.; Lee, J.; Chang, S.; and Hwang, K. 2023. Knowledge distillation from non-streaming to streaming asr encoder using auxiliary non-streaming layer. In *Proc. INTERSPEECH*.

Takashima, R.; Li, S.; and Kawai, H. 2018. An investigation of a knowledge distillation method for ctc acoustic models. In *Proc. ICASSP*, 5809–5813.

Takashima, R.; Li, S.; and Kawai, H. 2019. Investigation of sequence-level knowledge distillation methods for ctc acoustic models. In *Proc. ICASSP*, 6156–6160.

Thomas, S.; Kingsbury, B.; Saon, G.; and Kuo, H. J. 2022a. Integrating text inputs for training and adapting rnn transducer asr models. In *Proc. ICASSP*.

Thomas, S.; Kuo, H.-K. J.; Kingsbury, B.; and Saon, G. 2022b. Towards reducing the need for speech training data to build spoken language understanding systems. In *Proc. ICASSP*.

Tian, Z.; Xiang, H.; Li, M.; Lin, F.; Ding, K.; and Wan, G. 2023a. Building accurate low latency asr for streaming voice search in e-commerce. In *Proc. ACL: Industry Track*.

Tian, Z.; Xiang, H.; Li, M.; Lin, F.; Ding, K.; and Wan, G. 2023b. Peak-first ctc: reducing the peak latency of ctc models by applying peak-first regularization. In *Proc. ICASSP*.

Wang, Y.; Chen, Z.; Zheng, C.; Zhang, Y.; Han, W.; and Haghani, P. 2023. Accelerating rnn-t training and inference using ctc guidance. In *Proc. ICASSP*.

Weninger, F.; Gaudesi, M.; Haidar, M. A.; Ferri, N.; Andres-Ferrer, J.; and Zhan, P. 2022. Conformer with dual-mode chunked attention for joint online and offline asr. In *Proc. INTERSPEECH*.

Yang, X.; Li, Q.; and Woodland, P. C. 2022. Knowledge distillation for neural transducers from large self-supervised pre-trained models. In *Proc. ICASSP*.

Yao, Z.; Ren, S.; Chen, S.; Ma, Z.; Guo, P.; and Xie, L. 2022. Tessp: text-enhanced self-supervised speech pre-training. *arXiv preprint arXiv:2211.13443v1*.

Yao, Z.; Wu, D.; Wang, X.; Zhang, B.; Yu, F.; Yang, C.; Peng, Z.; Chen, X.; Xie, L.; and Lei, X. 2021. Wenet: production oriented streaming and non-streaming end-to-end speech recognition toolkit. In *Proc. INTERSPEECH*.

Yoon, J. W.; Ahn, S.; Lee, H.; Kim, M.; Kim, S.; and Kim, N. S. 2023a. EM-network: oracle guided self-distillation for sequence learning. In *Proc. ICML*.

Yoon, J. W.; Kim, H. Y.; Lee, H.; Ahn, S.; and Kim, N. S. 2023b. Oracle teacher: towards better knowledge distillation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 2974–2987.

Yoon, J. W.; Lee, H.; Kim, H. Y.; Cho, W. I.; and Kim, N. S. 2021. Tutornet: towards flexible knowledge distillation for end-to-end speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 1626–1638.

Yoon, J. W.; Woo, B. J.; Ahn, S.; Lee, H.; and Kim, N. S. 2022. Inter-kd: intermediate knowledge distillation for ctc-based automatic speech recognition. In *Proc. SLT*.

Yu, J.; Chiu, C.-C.; Li, B.; Chang, S.; Sainath, T. N.; He, Y.; Narayanan, A.; Han, W.; Gulati, A.; Wu, Y.; and Pang, R. 2021a. Fastemit: low-latency streaming asr with sequence-level emission regularization. In *Proc. ICASSP*.

Yu, J.; Han, W.; Gulati, A.; Chiu, C.; Li, B.; Sainath, T.; Wu, Y.; and Pang, R. 2021b. Dual-mode asr: unify and improve streaming asr with full-context modeling. In *Proc. ICLR*.

Yue, X.; Ao, J.; Gao, X.; and Li, H. 2023. Token2vec: a Joint self-supervised pre-training framework using unpaired speech and text. In *Proc. ICASSP*.

Zhang, B.; Wu, D.; Peng, Z.; Song, X.; Yao, Z.; Lv, H.; Xie, L.; Yang, C.; Pan, F.; and Niu, J. 2022. Wenet 2.0: more productive end-to-end speech recognition toolkit. *arXiv preprint arXiv:2203.15455v2*.

Zhang, Q.; Lu, H.; Sak, H.; Tripathi, A.; McDermott, E.; Koo, S.; and Kumar, S. 2020. Transformer transducer: a streamable speech recognition model with transformer encoders and rnn-t loss. In *Proc. ICASSP*.

Zhang, Z.; Chen, S.; Zhou, L.; Wu, Y.; Ren, S.; Liu, S.; Yao, Z.; Gong, X.; Dai, L.; Li, J.; and Wei, F. 2024. Speechlm: enhanced speech pre-training with unpaired textual data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32.