



Knowledge Distillation from Non-streaming to Streaming ASR Encoder using Auxiliary Non-streaming Layer

Kyuhong Shim, Jinkyu Lee, Simyoung Chang, and Kyuwoong Hwang

Qualcomm AI Research[†], Qualcomm Korea YH, Seoul, Republic of Korea

{kshim, jinkyu, simychan, kyuwoong}@qti.qualcomm.com

Abstract

Streaming automatic speech recognition (ASR) models are restricted from accessing future context, which results in worse performance compared to the non-streaming models. To improve the performance of streaming ASR, knowledge distillation (KD) from the non-streaming to streaming model has been studied, mainly focusing on aligning the output token probabilities. In this paper, we propose a layer-to-layer KD from the teacher encoder to the student encoder. To ensure that features are extracted using the same context, we insert auxiliary non-streaming branches to the student and perform KD from the non-streaming teacher layer to the non-streaming auxiliary layer. We design a special KD loss that leverages the autoregressive predictive coding (APC) mechanism to encourage the streaming model to predict unseen future contexts. Experimental results show that the proposed method can significantly reduce the word error rate compared to previous token probability distillation methods.

Index Terms: streaming speech recognition, non-streaming speech recognition, knowledge distillation

1. Introduction

Streaming automatic speech recognition (ASR) is a fundamental technology used in various speech-related applications, such as live captioning, voice command recognition, and voice chat. However, streaming ASR models face significant limitations because of their typical target systems' requirements for on-device and real-time speech recognition. The primary challenge is that the model cannot access the future context in order to minimize the delay from input speech to output text. Moreover, the size of the model cannot be too large as the inference speed and hardware resource usage are critical considerations for on-device use cases. Due to these constraints, streaming models often exhibit lower recognition accuracy compared to non-streaming (a.k.a., full-context) models.

Knowledge distillation (KD) [1, 2] has proved its effectiveness in reducing the performance gap between streaming and non-streaming models. Specifically, KD-based methods leverage the non-streaming model as a teacher to transfer the learned knowledge to the streaming student model [3]. The simple and frequently used approach is to minimize Kullback-Leibler divergence (KLD) between the output token emission probabilities (i.e., posterior distributions) of the teacher and student. However, several studies have reported that direct frame-to-frame distillation could misguide the streaming model [4, 5, 6]. This is because the audio-text frame alignments are different between the non-streaming and streaming models; the former generally

[†]Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

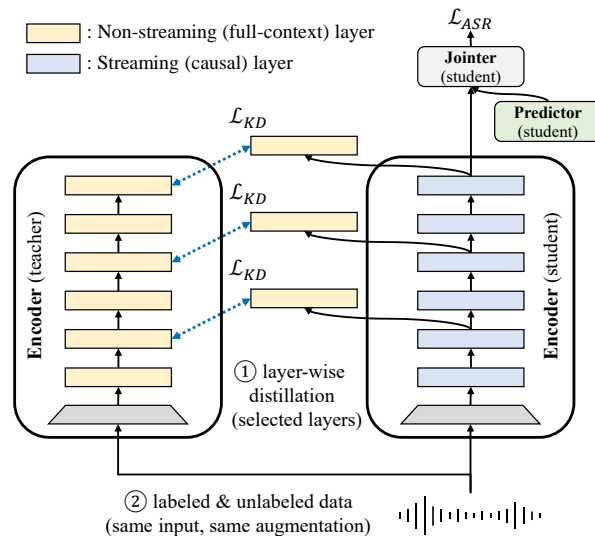


Figure 1: Overview of the proposed knowledge distillation (KD) from non-streaming ASR teacher to streaming ASR student. Auxiliary non-streaming (yellow) layers are attached for KD.

emit tokens faster than the latter. On the other hand, recent studies have introduced dual-mode [4], a type of self-KD training, which shares the same model for both streaming and non-streaming cases. While this approach improves streaming ASR performance, there exists a critical limitation that the student model size should be the same as the teacher [7].

In this work, we propose a novel KD method that applies layer-wise distillation on the encoder part of the model. The proposed method offers several advantages over existing methods, including faster training, better robustness to frame misalignment, and the ability to utilize unlabeled data. First, we insert auxiliary non-streaming layers into the streaming model (see Figure 1). These layers allow both the source and target of KD to observe the entire sequence, thereby minimizing discrepancies in terms of accessible context and making the training easier. Second, we design a special KD loss function to enhance the feature extraction process (see Figure 2). Our KD loss is composed of three losses targeting different objectives: 1) reducing the difference in the extracted feature of each frame, 2) matching the frame-to-frame relationship between all frames, and 3) predicting the future from the past context. Especially, the third loss is inspired by autoregressive predictive coding (APC), which was originally suggested for self-supervised learning tasks. We empirically discover that this *future prediction* task is very helpful for streaming models.

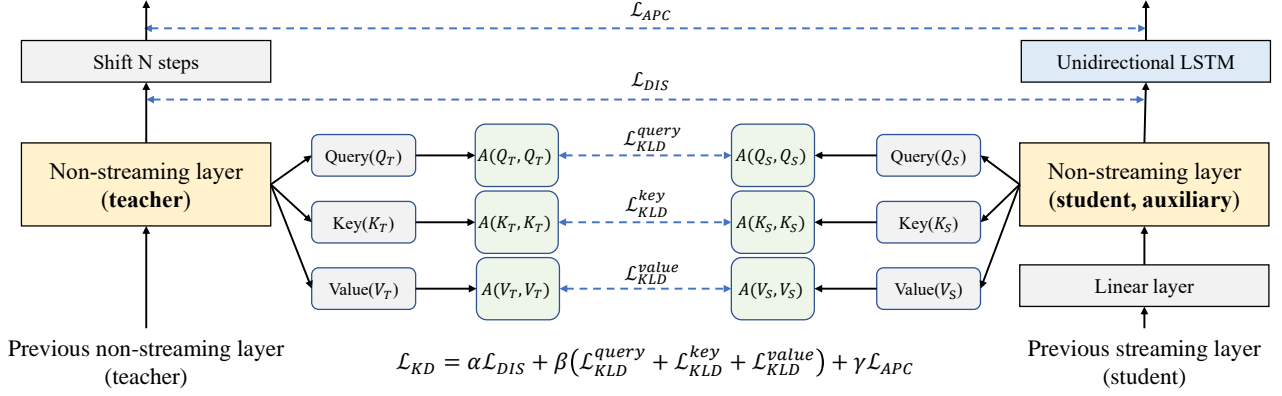


Figure 2: Illustration of three distillation loss functions: feature distance error (DIS), Kullback–Leibler divergence (KLD) for query, key, value relationships, and auto-regressive predictive coding (APC) error. α , β , and γ are the weighting coefficients.

Experimental results show that the proposed KD method considerably improves the streaming ASR performance. Compared to the baseline, our method achieves 16% relative WER reduction on LibriSpeech *test-other* dataset. By incorporating the large unlabeled data as an additional training resource, we could additionally reduce the WER by 3%.

2. Related Work

2.1. Distillation for Improving Streaming ASR

KD has been widely used for improving the streaming ASR model by setting the non-streaming ASR model as a teacher. The most popular approach is to distill the output token probabilities [2, 5, 8]. To mitigate the alignment mismatch problem, previous studies suggested manual shifting [9, 10], multi-stage alignment matching [8], or exploiting CTC alignments [5].

Another line of research called dual-mode, where the non-streaming teacher and streaming student share the same parameter set, has been proposed [4]. Dual-mode can provide a considerably better performance, but this may be not the best fit for low-resource streaming ASR: 1) the teacher and the student should be the same (large) architecture, and 2) it has not been tested for chunk-wise streaming setting. The problems of dual-mode have partially solved by joint training and student pruning together [7], but this method exploits block sparsity that has limitations on most hardware platforms.

Recently, a layer-wise KD method has been introduced [11]. However, this work is very distinct from ours; the main difference is the use of auxiliary non-streaming layers. The method directly minimizes the frame-wise feature differences between the streaming and non-streaming layer, and this method requires the teacher and student to use the same feature dimension. In addition, our KD loss can transfer diverse characteristics using three different loss objectives.

2.2. Distillation for ASR Model Compression

ASR models have adopted KD for model compression as KD’s original purpose [1]. In particular, Transducer-based end-to-end ASR models [12] have been actively studied for compression [13, 14, 15, 16, 17] because the framework is suitable for low-resource scenarios: on-device and real-time inference. In addition, large self-supervised models have also tried KD to squeeze their model size [18].

2.3. Cascaded Non-streaming and Streaming Layers

Cascade encoders are a special encoder design that stack several non-streaming layers on top of the streaming layers [19, 20]. By doing so, the model can be jointly trained for non-streaming and streaming purposes, and the model can dynamically select which path to pass through [21]. Our work also attaches additional non-streaming layers, but they are only used for KD purpose and then removed after training is finished.

3. Distillation from Non-streaming Teacher to Streaming Student

3.1. Preliminaries on ASR Encoder

Given the input speech frames $\mathbf{x} = \{x_1, x_2, \dots, x_T\}$ and the target text tokens $\mathbf{y} = \{y_1, y_2, \dots, y_U\}$, the goal of ASR model is to maximize the likelihood $P(\mathbf{y}|\mathbf{x})$. The input \mathbf{x} first passes through an encoder, which is usually a stack of identical layers, to extract contextualized features for each frame. Let the output feature of the ℓ -th layer $\mathbf{h}^\ell = \{h_1^\ell, h_2^\ell, \dots, h_T^\ell\}$, and then the next layer’s t -th frame output can be represented as: $h_t^{\ell+1} = f^{\ell+1}(h_{t-LC}^\ell, \dots, h_{t+RC}^\ell)$ and $f^{\ell+1}(h_1^\ell, \dots, h_T^\ell)$ for streaming and non-streaming models, respectively. LC and RC are the left/right-context size, where the right-context size is often set to zero in low-latency use cases. For clarity, we will use \mathbf{h}^ℓ for the teacher and \mathbf{s}^ℓ for the student. In this work, we employ Transducer [12] architecture; the encoder output is jointly combined with the predictor output and generates a token emission probability matrix of size $T \times U$.

3.2. Layer-wise Distillation using Auxiliary Layer

Instead of the output token probability distillation, we propose a layer-wise KD from the teacher encoder to the student encoder. We conjecture that the naive frame-wise feature mapping from h_t^ℓ to s_t^ℓ would be unnecessarily difficult for the student model because of the very different context information teacher and student use. The observation that non-streaming and streaming models produce different audio-text alignments can be strong evidence of our hypothesis.

In order to solve this context mismatch, we insert auxiliary non-streaming layers at selected layers of the student (see Figure 1). Formally, if an auxiliary layer is attached at ℓ -th layer, the output frame features of auxiliary layer $z_t^\ell = g^\ell(h_1^\ell, \dots, h_T^\ell)$ can observe the entire sequence. Then, we perform KD between

the teacher layer \mathbf{h}^ℓ and the auxiliary layer \mathbf{z}^ℓ . By doing so, we can remove the fundamental context difference between the non-streaming source and streaming target, thus expecting the student to more focus on improving the quality of features.

The auxiliary layer consists of three parts (see Figure 2). The feature first passes through a simple projection layer to expand the feature dimension the same as the teacher. Then, it passes through a single Transformer layer and finally passes through a unidirectional LSTM layer. Note that the auxiliary layers are jointly trained with the student encoder, while the teacher is fixed during training.

3.3. Distillation Loss

We propose a novel loss function designed for KD from the non-streaming ASR teacher to the streaming ASR student.

3.3.1. Feature similarity loss

First, we minimize the distance between the teacher/student features of the same frame to guide the feature extraction. We adopt the distance function introduced in DistilHuBERT [18].

$$L_{\text{DIS}} = \sum_{t=1}^T \left[\frac{1}{D} \|h_t - z_t\|_1 - \log \sigma(\cos(h_t, z_t)) \right] \quad (1)$$

where D is the feature vector dimension, σ is sigmoid activation function, and $\cos(\cdot, \cdot)$ is cosine similarity.

3.3.2. Self-Attention loss

Second, we also distill the self-attention distribution to not only mimic the teacher frame-wise but also follow the inter-frame relationships. Specifically, we transfer three relationships (Query-Query, Key-Key, and Value-Value) individually as suggested in MiniLM-v2 [22]. For simplicity, we only present the Query-Query relationship below, but the equation is easily expandable for Key and Value cases.

$$L_{\text{KLD}}^{\text{query}} = \frac{1}{A} \sum_{a=1}^A \sum_{t=1}^T \text{KLD}(R_{T(a,t)}^{\text{query}} \| R_{S(a,t)}^{\text{query}}) \quad (2)$$

where A is the number of attention heads. $R_{T(a,t)}, R_{S(a,t)}$ indicate teacher/student attention probability distribution of the t -th frame computed at a -th attention head.

$$R_{(a,t)}^{\text{query}} = \text{Softmax}_k \left(q_{(a,t)} q_{(a,k)}^T / \sqrt{d_A} \right) \quad (3)$$

where $d_A = D/A$ is the attention head dimension and $q_{(a,t)}$ is a query vector of t -th frame. The self-attention loss is the sum of losses computed from Query, Key, and Value as:

$$L_{\text{KLD}} = L_{\text{KLD}}^{\text{query}} + L_{\text{KLD}}^{\text{key}} + L_{\text{KLD}}^{\text{value}} \quad (4)$$

3.3.3. Future prediction loss

Finally, we employ the future prediction loss, inspired by the autoregressive predictive coding (APC) [23, 24]. The objective of APC is to predict the future feature (after several frames) using previous history. To do so, APC includes a unidirectional RNNs and train those layers to minimize the difference. Although APC was originally proposed for self-supervised speech pre-training, we discover that APC can be especially helpful for streaming ASR model distillation. Let the output features of

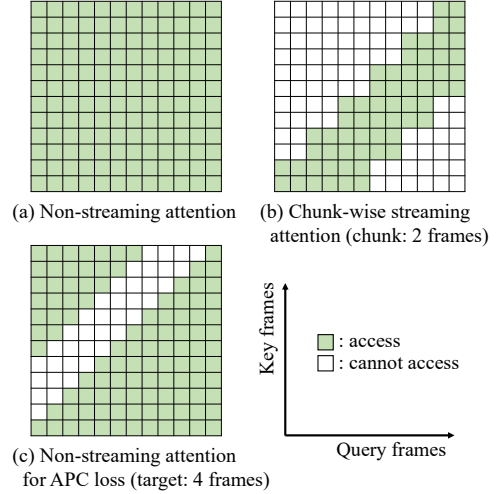


Figure 3: Illustration of self-attention masks in (a) non-streaming layer in teacher, (b) streaming layer in student, and (c) non-streaming layer in auxiliary branches. Note that the number of frames in the figure is chosen for clear understanding.

the unidirectional LSTM layer $r_t = \text{LSTM}(z_1, z_2, \dots, z_t)$ (see Figure 2), then our APC loss is computed as:

$$L_{\text{APC}} = \sum_{t=1}^T \left[\frac{1}{D} \|h_{t+N} - r_t\|_1 - \log \sigma(\cos(h_{t+N}, r_t)) \right] \quad (5)$$

where N indicates how far the target frame is located. Note that the equation is almost the same to the feature similarity loss, but the target is shifted by N frames. APC loss encourages the model to predict future, trying to follow the teacher’s behavior as if the student can exploit the future context.

To maximize the efficiency of APC loss, we modify the Transformer attention mask applied for auxiliary non-streaming layers (see Figure 3(c)). In particular, we mask future N frames $[t+1, t+N]$ so that each frame cannot directly observe frames between the current one and the target of APC loss. By doing so, we can prevent the model from cheating (a.k.a., ‘information leakage’) by directly taking the target from the input sequence.

3.3.4. Total loss

Overall, the final loss function is combined as follows:

$$L_{\text{total}} = \frac{1}{N_s} \sum_{i \in \{N_s\}} L_{\text{ASR}}(\mathbf{x}_i) + \frac{1}{N_s + N_u} \sum_{i \in \{N_s, N_u\}} (\alpha L_{\text{DIS}} + \beta L_{\text{KLD}} + \gamma L_{\text{APC}})(\mathbf{x}_i) \quad (6)$$

where N_s, N_u are the number of labeled and unlabeled samples, respectively. We set $\alpha = 0.01, \beta = 0.0005$, and $\gamma = 0.005$ for the experiments.

4. Experimental Results

4.1. Setup

Dataset. We use LibriSpeech-960h [25] dataset for supervised ASR training. Additionally, we exploit 6K hours of unlabeled data from the same domain (i.e., audio book reading) [26] to better transfer knowledge from the teacher encoder.

Table 1: Word error rate (%) for three streaming student models (S1-S3) and two small models (NS and S0) trained from scratch. The teacher model is a non-streaming model.

Model	dev-		test-	
	clean	other	clean	other
Teacher (106M)	2.2	5.3	2.5	5.4
From scratch (30M)				
- NS (non-streaming)	2.4	5.7	2.7	5.9
- S0 (streaming)	4.9	12.2	4.7	12.8
Distillation (30M)				
- S1 (prob. KLD)	4.8	12.0	4.4	13.3
- S2 (KD w/o aux. layer)	4.0	11.1	4.4	12.3
- S3 (KD w/ aux. layer)	3.8	10.1	4.1	10.7

Table 2: Ablation study on the knowledge distillation losses. N is the relative target distance in APC loss.

Loss ablation			dev-		test-	
DIS	KLD	APC	clean	other	clean	other
✓			4.0	10.8	4.3	11.6
✓	✓		4.0	10.7	4.2	11.2
✓	✓	✓($N=2$)	3.9	10.4	4.1	10.9
✓	✓	✓($N=4$)	3.8	10.1	4.1	10.7
✓	✓	✓($N=8$)	4.0	10.5	4.2	11.2

Model Architecture. In this work, we employ the Conformer-Transducer [27] ASR models. The teacher encoder is a stack of 16 Conformer layers with 512-dim. feature and 8 attention heads. The student encoder is a stack of 16 Conformer layers with 256-dim. feature and 4 attention heads. For the streaming student, the convolution layer is set to be causal and the attention mask is restricted to only access chunks within the left context limit. For the auxiliary layers, we use a single Transformer layer (self-attention + feed-forward). We apply KD from the 4, 8, 12, 16-th layers of the teacher to 4, 8, 12, 16-th layers of the student, respectively.

Training Details. We basically follow the training setup of the original Conformer paper [27] for student training, using a batch size of 1024 and peak learning rate of $1.25e^{-3}$. The streaming student uses a chunk-wise training and inference [28], where the chunk size is 160ms and the left/right context size is 640ms and 0ms, respectively (see Figure 3(b)). When unlabeled data is used, we mix the labeled and unlabeled data to a 1:1 ratio to construct an epoch.

4.2. ASR Performance

Table 1 compares WER of different training methods on LibriSpeech evaluation datasets. Note that the external language model is not used during decoding. S1 is the model trained with token probability distillation, and S2 is the case that directly applies L_{DIS} for layer-wise KD without the proposed auxiliary layer. We can observe that the proposed training method (S3) achieves the lowest WER. Interestingly, the conventional method (S1) only achieves a marginal improvement compared to the model without KD (from scratch, S0). We would like to mention that we have tested various hyper-parameter configurations for S1 to stabilize the training and improve performance. We are cautiously assuming that the method would not be the best fit for our target setup, small-size model and chunk-wise streaming.

Table 3: Effect of the size of unlabeled data for the distillation.

#Unlabeled data (hrs)	dev-		test-	
	clean	other	clean	other
0 (none)	3.8	10.1	4.1	10.7
+600 (small)	3.8	9.8	4.2	10.4
+6,000 (medium)	3.8	9.7	4.4	10.3

4.3. Ablation Study

Table 2 shows that all three KD losses contribute to the performance. For APC, we test three N configurations (see Section 3.3.3) and empirically find that the best results are produced with $N=4$, which corresponds to 160ms of speech. Furthermore, by comparing S2 (Table 1) and the first row (Table 2) that use the same loss, we can verify that the auxiliary non-streaming layer indeed plays an important role.

4.4. Additional Unsupervised Training

Table 3 presents performance improvements when using 600 (small) and 6,000 (medium) hours of unlabeled data in LibriLight dataset [26]. When using 6,000 hours of unlabeled data, we observe a clear WER reduction on the challenging LibriSpeech *dev-other* and *test-other* datasets. Nevertheless, we observe a slight WER increase in *dev-clean* and *test-clean*; we believe this is due to the imbalance of the number of labeled and unlabeled samples. Because such large-scale data training takes much longer time, the models in Table 3 started from the best model (S3) which is trained only with supervised data.

5. Discussion

5.1. Teacher-to-Student Layer Connection

Although we connected the layers of the same position from both teacher and student, we expect more optimization can be possible if we could find the optimal connection between layers. One approach is to identify the layers of similar behaviors and connect them. Indeed, recent studies show that the information encoded in each layer can be examined, for both ASR models [29, 30] and self-supervised speech models [31, 32]. On the other hand, it would be an interesting research direction to reduce the number of layers in the student model for additional compression [18].

5.2. Relationship to Pseudo Label Generation

To utilize unlabeled data, several studies have suggested generating pseudo labels for such data using the out-of-the-shelf large ASR model [33, 9]. Previous work tested this idea for improving streaming ASR performance [34] and achieved considerable WER reduction, but only tested a large streaming model with more than 120M parameters. We believe this approach can be combined with ours as they are orthogonal to each other.

6. Conclusion

In this paper, we proposed a novel KD method for improving the small streaming student using the large non-streaming teacher. The key idea is to insert non-streaming auxiliary layers on the student, which aims at matching the context of the target features for knowledge distillation. We also designed a special layer-wise KD loss function, which includes future prediction task using APC loss. Experiments showed that our method improves the streaming ASR performance by a clear margin.

7. References

- [1] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
- [2] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, “Learning small-size dnn with output-distribution-based criteria,” in *Fifteenth annual conference of the international speech communication association*, 2014.
- [3] S. Kim, M. L. Seltzer, J. Li, and R. Zhao, “Improved training for online end-to-end speech recognition systems,” *arXiv preprint arXiv:1711.02212*, 2017.
- [4] J. Yu, W. Han, A. Gulati, C.-C. Chiu, B. Li, T. N. Sainath, Y. Wu, and R. Pang, “Dual-mode asr: Unify and improve streaming asr with full-context modeling,” in *International Conference on Learning Representations*, 2021.
- [5] H. Inaguma and T. Kawahara, “Alignment knowledge distillation for online streaming attention-based speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.
- [6] C. Liang, X.-L. Zhang, B. Zhang, D. Wu, S. Li, X. Song, Z. Peng, and F. Pan, “Fast-u2++: Fast and accurate end-to-end speech recognition in joint ctc/attention frames,” *arXiv preprint arXiv:2211.00941*, 2022.
- [7] C. Liu, Y. Shangguan, H. Yang, Y. Shi, R. Krishnamoorthi, and O. Kalinli, “Learning a dual-mode speech recognition model via self-pruning,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 273–279.
- [8] G. Kurata and G. Saon, “Knowledge distillation from offline to streaming rnn transducer for end-to-end speech recognition,” in *Interspeech*, 2020, pp. 2117–2121.
- [9] X. Yang, Q. Li, and P. C. Woodland, “Knowledge distillation for neural transducers from large self-supervised pre-trained models,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8527–8531.
- [10] F. Weninger, M. Gaudesi, M. A. Haidar, N. Ferri, J. Andrés-Ferrer, and P. Zhan, “Conformer with dual-mode chunked attention for joint online and offline asr,” *arXiv preprint arXiv:2206.11157*, 2022.
- [11] A. Kojima, “Knowledge Distillation for Streaming Transducer–Transducer,” in *Proc. Interspeech 2021*, 2021, pp. 2841–2845.
- [12] A. Graves, “Sequence transduction with recurrent neural networks,” *arXiv preprint arXiv:1211.3711*, 2012.
- [13] K. Zhao, H. Nguyen, A. Jain, N. Susanj, A. Mouchtaris, L. Gupta, and M. Zhao, “Knowledge Distillation via Module Replacing for Automatic Speech Recognition with Recurrent Neural Network Transducer,” in *Proc. Interspeech 2022*, 2022, pp. 4436–4440.
- [14] J. Rathod, N. Dawalatabad, S. SINGH, and D. Gowda, “Multi-stage Progressive Compression of Conformer Transducer for On-device Speech Recognition,” in *Proc. Interspeech 2022*, 2022, pp. 1691–1695.
- [15] J. Park, S. Jin, J. Park, S. Kim, D. Sandhyana, C. Lee, M. Han, J. Lee, S. Jung, C. Han *et al.*, “Conformer-based on-device streaming speech recognition with kd compression and two-pass architecture,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 92–99.
- [16] R. V. Swaminathan, B. King, G. P. Strimel, J. Droppo, and A. Mouchtaris, “Codert: Distilling encoder representations with co-learning for transducer-based speech recognition,” *arXiv preprint arXiv:2106.07734*, 2021.
- [17] S. Panchapagesan, D. S. Park, C.-C. Chiu, Y. Shangguan, Q. Liang, and A. Gruenstein, “Efficient knowledge distillation for rnn-transducer models,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5639–5643.
- [18] H.-J. Chang, S.-w. Yang, and H.-y. Lee, “Distillhubert: Speech representation learning by layer-wise distillation of hidden-unit bert,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7087–7091.
- [19] B. Li, A. Gulati, J. Yu, T. N. Sainath, C.-C. Chiu, A. Narayanan, S.-Y. Chang, R. Pang, Y. He, J. Qin *et al.*, “A better and faster end-to-end model for streaming asr,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5634–5638.
- [20] T. N. Sainath, Y. He, A. Narayanan, R. Botros, W. Wang, D. Qiu, C.-C. Chiu, R. Prabhavalkar, A. Gruenstein, A. Gulati *et al.*, “Improving the latency and quality of cascaded encoders,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8112–8116.
- [21] S. Ding, W. Wang, D. Zhao, T. N. Sainath, Y. He, R. David, R. Botros, X. Wang, R. Panigrahy, Q. Liang *et al.*, “A unified cascaded encoder asr model for dynamic model sizes,” *arXiv preprint arXiv:2204.06164*, 2022.
- [22] W. Wang, H. Bao, S. Huang, L. Dong, and F. Wei, “Minilmv2: Multi-head self-attention relation distillation for compressing pre-trained transformers,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2140–2151.
- [23] Y.-A. Chung and J. Glass, “Generative pre-training for speech with autoregressive predictive coding,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3497–3501.
- [24] Y.-A. Chung, H. Tang, and J. Glass, “Vector-Quantized Autoregressive Predictive Coding,” in *Proc. Interspeech 2020*, 2020, pp. 3760–3764.
- [25] 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.
- [26] J. Kahn, M. Riviere, 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.
- [27] 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 *Proc. Interspeech 2020*, 2020, pp. 5036–5040.
- [28] X. Chen, Y. Wu, Z. Wang, S. Liu, and J. Li, “Developing real-time streaming transformer transducer for speech recognition on large-scale dataset,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5904–5908.
- [29] K. Shim, J. Choi, and W. Sung, “Understanding the role of self attention for efficient speech recognition,” in *International Conference on Learning Representations*, 2022.
- [30] K. Shim and W. Sung, “Similarity and Content-based Phonetic Self Attention for Speech Recognition,” in *Proc. Interspeech 2022*, 2022, pp. 4118–4122.
- [31] S. wen Yang, A. T. Liu, and H. yi Lee, “Understanding Self-Attention of Self-Supervised Audio Transformers,” in *Proc. Interspeech 2020*, 2020, pp. 3785–3789.
- [32] A. Pasad, J.-C. Chou, and K. Livescu, “Layer-wise analysis of a self-supervised speech representation model,” in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.
- [33] D. Hwang, K. C. Sim, Z. Huo, and T. Strohmaier, “Pseudo label is better than human label,” *arXiv preprint arXiv:2203.12668*, 2022.
- [34] T. Doutré, W. Han, M. Ma, Z. Lu, C.-C. Chiu, R. Pang, A. Narayanan, A. Misra, Y. Zhang, and L. Cao, “Improving streaming automatic speech recognition with non-streaming model distillation on unsupervised data,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6558–6562.