



# Conformer with dual-mode chunked attention for joint online and offline ASR

Felix Weninger, Marco Gaudesi, Md Akmal Haidar, Nicola Ferri, Jesús Andrés-Ferrer, Puming Zhan

Nuance Communications, Inc.

felix.weninger@nuance.com

## Abstract

In this paper, we present an in-depth study on online attention mechanisms and distillation techniques for dual-mode (i.e., joint online and offline) ASR using the Conformer Transducer. In the dual-mode Conformer Transducer model, layers can function in online or offline mode while sharing parameters, and in-place knowledge distillation from offline to online mode is applied in training to improve online accuracy. In our study, we first demonstrate accuracy improvements from using chunked attention in the Conformer encoder compared to autoregressive attention with and without lookahead. Furthermore, we explore the efficient KLD and 1-best KLD losses with different shifts between online and offline outputs in the knowledge distillation. Finally, we show that a simplified dual-mode Conformer that only has mode-specific self-attention performs equally well as the one also having mode-specific convolutions and normalization. Our experiments are based on two very different datasets: the Librispeech task and an internal corpus of medical conversations. Results show that the proposed dual-mode system using chunked attention yields 5% and 4% relative WER improvement on the Librispeech and medical tasks, compared to the dual-mode system using autoregressive attention with similar average lookahead.

**Index Terms:** online speech recognition, knowledge distillation

## 1. Introduction

End-to-end (E2E) systems have become dominant in automatic speech recognition (ASR) because of their simplicity and better performance. In addition to new advancements in model architectures [1–3], one of the major efforts is to make E2E ASR systems support online streaming applications with strict latency requirements. The Recurrent Neural Network Transducer (RNN-T) has been a favorite E2E model architecture for online streaming because of its time-synchronous processing of input audio and superior performance over the CTC model [4–6].

There have been significant improvements to RNN-T since it was proposed in [4], such as replacing the LSTM/BLSTM encoder with Transformer [7, 8], Conformer [2], and ContextNet [1]. A major difference between online streaming and offline batch-mode E2E model is that the former is subject to strict and often application-dependent latency constraints. To reduce the deterministic latency incurred during inference, an online ASR system is only allowed to access limited future context. Since many popular E2E ASR systems are based on bidirectional long-range context modeling (BLSTM, Transformer, etc.) in the encoder, this is the primary reason that online E2E ASR systems generally underperform their offline counterparts. The degradation in accuracy is largely determined by the accessed amount of future context. There has been extensive research in effectively utilizing future context with limited latency for improving online E2E model performance [7–12].

From a deployment efficiency point of view, it is beneficial to have a single model able to serve multiple different applications:

from offline batch-mode to online streaming under different latency requirements. Unfortunately, a model trained for the offline use case generally does not perform well in the online use case and vice versa. Therefore, there is a direction of research towards making a single model suitable for multiple use cases with different latency requirements [13–17].

**Contributions of our paper:** We extend the dual-mode ASR work in [16] in several aspects that were not covered there or in similar works [15, 17]: 1. Comprehensive evaluation of different online streaming approaches (i.e., autoregressive and chunked attention) based on Conformer Transducer in dual-mode training on two very different data sets. 2. Evaluation of different distillation approaches for offline-to-online distillation in dual-mode training and the importance of modeling the output shift between offline and online modes. 3. Propose a dual-mode model trained with shared convolution (i.e., causal convolution) and normalization layers across modes.

## 2. Methodology

### 2.1. Dual-mode Conformer Transducer

In our paper, we use end-to-end ASR systems based on the Conformer Transducer (Conf-T) architecture, which combines the concept of the recurrent neural network transducer (RNN-T) [4] with the Conformer encoder [2, 8]. Each Conformer encoder block consists of feedforward, multi-head self-attention (MHSA) [18], and convolution layers. Following the dual-mode approach [16], a single Conformer Transducer model can operate in both online and offline mode. In online mode, the outputs of convolutions and attention layers are calculated by masking the weights corresponding to future frames. Moreover, online and offline mode use different sets of (batch / layer) normalization parameters (running average statistics and scales / offsets). For the convolutions, the alternative approach proposed in our paper is to simply use causal (left) padding everywhere.

Offline and online Transducer outputs are calculated as  $z_{\text{on}} = M_{\text{on}}^{\theta^{\text{on}}}(x)$ ,  $z_{\text{off}} = M_{\text{off}}^{\theta^{\text{off}}}(x) \in [0, 1]^{T \times U \times K}$ , where  $M$  is the model,  $\theta^{\text{on}} = [\theta; \nu^{\text{on}}]$ ,  $\theta^{\text{off}} = [\theta; \nu^{\text{off}}]$  are the online and offline model weights,  $\nu^{\text{on}}$ ,  $\nu^{\text{off}}$  are the corresponding normalization parameters, and  $T$ ,  $U$ ,  $K$  denote # frames, # tokens, and vocabulary size. In dual-mode training [16], the offline and online mode are trained jointly while knowledge transfer is done via in-place distillation from the offline to the online mode. More precisely, the following loss is minimized:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{trd}}(y^*, z_{\text{on}}) + \beta \mathcal{L}_{\text{trd}}(y^*, z_{\text{off}}) + \gamma \mathcal{L}_{\text{dist}}(z_{\text{off}}, z_{\text{on}}), \quad (1)$$

where  $\mathcal{L}_{\text{trd}}$  is the transducer loss [4],  $\mathcal{L}_{\text{dist}}$  is a distillation loss (cf. Section 2.3),  $y^*$  are the training labels, and  $\alpha, \beta, \gamma \geq 0$  are hyperparameters.

### 2.2. Chunked Attention

To adapt Transformer-like architectures for the streaming use case, the key part to consider is the MHSA block. The MHSA

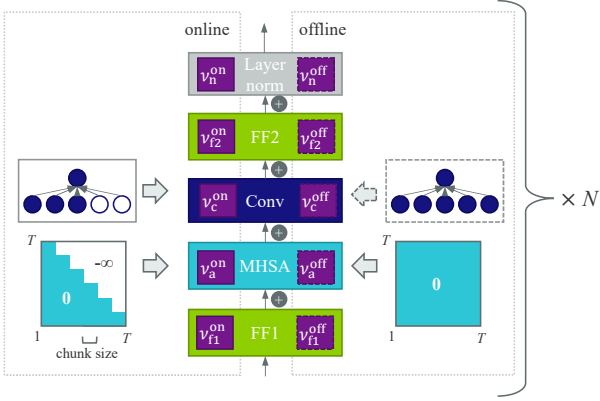


Figure 1: Schema of the dual-mode Conformer encoder with chunked/global attention mask (added to logits), causal/non-causal convolutions and dual-mode normalization with parameters  $v_{(\cdot)}^{(\cdot)}$ . Dashed lines indicate optional components.

can be made strictly online by using autoregressive attention [18], i.e., every frame in the encoder can only attend to previous frames. This constraint is efficiently implemented by adding  $-\infty$  (or a large negative value) to the attention logits at the ‘invalid’ positions. There are several ways to modify this under specified latency requirements, such as truncated lookahead [8] or contextual lookahead [12]. While the first approach builds lookahead that accumulates through layers into a larger overall lookahead of the encoder, the latter increases the computational cost at inference by overlapping the input audio chunks. In this work, we focus on the chunked attention approach [11], where we divide the input audio into non-overlapping chunks. For each encoder input chunk, the MHSA query at each position uses as memory (key and values) all the other positions that belong to the same chunk or previous chunks. Figure 1 shows the dual-mode Conformer encoder with chunked attention mask.

### 2.3. Distillation

In addition to architectural advancements and various approaches for effectively leveraging limited future context, knowledge distillation from offline to online model is another way of improving online model performance [19]. We compare the approaches proposed in [20, 21] in the context of in-place distillation in dual-mode training.

Knowledge distillation by using the Kullback-Leibler divergence (KLD) loss [22, 23] directly is inefficient for Transducers due to the large size of the output lattice. The efficient KLD [20] and 1-best distillation [21] losses address this issue by restricting the calculation to a reduced output lattice. Moreover, in order to make these approaches work for offline-to-online distillation, we consider the potential emission delay between online and offline model via a tunable shift parameter  $\tau$ , similar to [16, 21].

The efficient KLD loss [20] collapses the probability distribution of the tokens as follows:

$$\mathcal{L}_{\text{dist}}^{\text{eff}} = \sum_{t,u} \sum_{l \in \{y, \emptyset, r\}} p_{\text{off}}(l|t, u) \log \frac{p_{\text{off}}(l|t, u)}{p_{\text{on}}(l|t - \tau, u)}, \quad (2)$$

where  $y$ ,  $\emptyset$  and  $r$  denote the correct, the blank, and all other labels,  $p_{(\cdot)}$  denotes the probability obtained from the Transducer output  $z_{(\cdot)}$ , and  $t$  and  $u$  denote time frame and token indices.

Table 1: Single-mode baselines on Librispeech 100h (LA: lookahead).

Mode	Attention	Test WER [%]	
		cln	other
Online	Autoregressive	9.6	26.9
Online	Autoreg. LA	8.4	24.8
Online	Chunked	7.9	23.4
Offline (causal conv)	Full context	6.3	18.4
(non-causal conv)	Full context	6.3	18.3
Offline [27]	Full context	6.8	18.9

Conversely, the 1-best distillation loss [21] takes into account the full probability distribution, but only along the 1-best path in the teacher lattice. We extend this approach to in-place distillation by regenerating the 1-best path of the offline model on-the-fly in each training step. For consistency, we also use KLD, not cross-entropy as in [21]:

$$\mathcal{L}_{\text{dist}}^{1\text{-best}} = \sum_{(t,u) \in 1\text{-best}} \sum_{k=1}^K p_{\text{off}}(k|t, u) \log \frac{p_{\text{off}}(k|t, u)}{p_{\text{on}}(k|t - \tau, u)}, \quad (3)$$

where  $k$  is the index of a symbol in the vocabulary.

## 3. Experiments and Results

### 3.1. Librispeech Data

#### 3.1.1. Training recipe

We first perform a comparative evaluation using the 100 hour training subset of the Librispeech [24] corpus. Speed perturbation [25] with factors 0.9, 1.0 and 1.1 and SpecAugment [26] are applied to improve generalization. The topology of the Conformer Transducer and the training recipe are similar to [27]. The encoder consists of a feature frontend that extracts 80-dimensional log-Mel features, two convolutional layers that perform downsampling on the time axis by a factor of 4, and 18 Conformer blocks with hidden dimension 256 and feed-forward dimension 1024. The prediction network has a single LSTM layer with 256 hidden units, and the joint network has 256 units. The vocabulary contains 30 characters. Models are trained for 300 epochs. The training hyperparameters (especially learning rate schedule) were tuned for the offline model using a limited grid search on the clean development set of Librispeech, then applied to all other models (online and dual-mode) without further tuning. For dual-mode training, the online and offline losses are weighted equally ( $\alpha = \beta = 0.5$ ) and the distillation weight is set to  $\gamma = 0$  (no distillation) or  $\gamma = 0.01$ . We measure the word error rate (WER) on the ‘clean’ and ‘other’ test set of Librispeech. Decoding is done by beam search with beam size 8, without using an external language model.

#### 3.1.2. Online and offline baselines

The results of our single-mode baselines are shown in Table 1. Our offline Conformer Transducer system outperforms the reference result obtained by ESPnet [27]. We also investigated the usage of causal (left padded) 1-D depthwise convolutions in the Conformer blocks in the offline model. The WER was similar to the standard non-causal (centered) convolutions. Hence, we chose to apply causal convolutions for offline mode as well, thereby simplifying the implementation compared to the original dual-mode Conf-T [16].

Table 2: *Librispeech 100h task: WER obtained by dual-mode systems in online and offline inference with and without efficient KLD distillation (loss weight  $\gamma$ , shift  $\tau$ ).*

Online att.	$\gamma$	$\tau$	Test WER [%]			
			Online		Offline	
			cln	other	cln	other
Autoreg.	0.0	-	9.0	25.2	7.2	21.8
Autoreg.	0.01	0	9.0	25.8	7.2	21.2
Autoreg.	0.01	-6	8.4	24.2	7.0	20.6
Autoreg. LA	0.0	-	7.7	23.1	6.8	19.9
Autoreg. LA	0.01	0	7.7	22.6	7.0	19.8
Autoreg. LA	0.01	-6	7.5	22.2	6.7	19.8
Chunked	0.0	-	7.4	22.0	6.4	19.2
Chunked	0.01	0	7.7	22.4	6.4	19.2
Chunked	0.01	-6	<b>7.1</b>	21.5	<b>6.1</b>	18.9

For the online systems, we compare autoregressive attention, autoregressive attention with 12 frames ( $\approx 0.5$  seconds) lookahead in the 9th encoder layer<sup>1</sup>, and chunked attention (see Section 2.2) with a chunk size of 25 frames ( $\approx 1$  second). Using autoregressive attention leads to a drastic WER increase compared to the offline model (52% relative). However, the relative WER increase is still much smaller than the one reported in [16], suggesting that our online baseline is competitive. As expected, the lookahead reduces the gap between online and offline WER significantly. Furthermore, despite having the same average lookahead of about 0.5 seconds, the chunked attention performs better than the autoregressive attention with lookahead (6% WER reduction (WERR)).

### 3.1.3. Dual-mode systems

Table 2 shows the results obtained by dual-mode training for various types of online attention. Compared to the online baselines in Table 1, the WER in online mode is improved by dual-mode training in all cases (6%, 8% and 6% WERR on test\_clean for autoregressive, autoregressive with lookahead and chunked attention, respectively). Furthermore, we observe that there is a consistent gain in online performance from using distillation with shift  $\tau = -6$ , but no gain from distillation without shift. Still, the gain from distillation is diminished when lookahead is used, likely because this brings the online performance closer to the offline model and reduces the benefit of knowledge distillation. The dual-mode system using autoregressive attention in the online mode improves on the WER of the corresponding single-mode online system by 12% relative. Conversely, the offline performance is degraded by 11% relative. The trend is similar for the autoregressive attention with lookahead, despite overall better performance. In contrast, using chunked attention in the online mode avoids the degradation in offline mode, and the corresponding dual-mode system performs better than the single-mode baselines in both offline and online mode, achieving 10% and 3% relative WERR, respectively. This is likely because the lookahead for a given frame is not constant in chunked attention, which makes the online prediction task more similar to the offline one and thus facilitates joint training.

We also investigate the impact of the shift  $\tau$  between offline teacher and online student in the in-place distillation with both efficient distillation and 1-best distillation. As can be seen from

<sup>1</sup>We did not observe significant performance differences when putting the lookahead in another encoder layer or distributing it across multiple encoder layers.

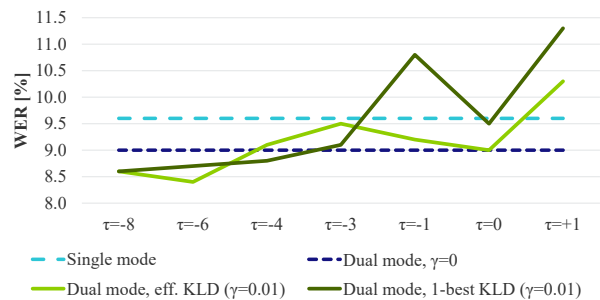


Figure 2: *Dual-mode WER on Librispeech 100h task with efficient and 1-best distillation, varying the shift parameter  $\tau$ .*

Table 3: *Single-mode baselines on medical conversation data*

Mode	Attention	WER [%]
Online	Autoreg. LA	14.7
Online	Chunked	14.4
Offline (causal conv)	Full context	13.1
(non-causal conv)	Full context	13.2

Figure 2, gains from distillation can be achieved only with fairly large shifts (e.g.,  $\tau = -6$  corresponds to a  $\approx 240$  ms emission delay), which is consistent with the findings in [21], while results are unstable for small shifts (in fact, the experiments with  $\tau = -2$  diverged). The best result with efficient distillation is achieved at  $\tau = -6$  (8.4% WER), whereas the 1-best distillation performs best with  $\tau = -8$ .

## 3.2. Medical Data

Additionally, we conduct experiments on an internal data set which consists of conversational speech data in the medical domain (doctor-patient conversations). The experiments are based on a training set of 1 k hours manually end-pointed and transcribed speech covering various medical specialties. We measure WER on a speaker-independent test set consisting of 263 k words.

The model topology and training recipe are similar to the one used for Librispeech. In the encoder, we use 16 Conformer blocks with hidden dimension 512 and feed-forward dimension 1024 after the frontend. The prediction network consists of a single Transformer layer with the same dimensions, and the joint network has 512 units. The vocabulary contains 2 k word-pieces.

Table 3 shows the single-mode baselines. For the online systems, we compare autoregressive attention with lookahead (12 frames) and chunked attention (24 frames). Unlike on Librispeech, a pure autoregressive model (without any lookahead) did not yield satisfactory performance. The chunked attention improves the WER of the online system by 2% relative compared to the autoregressive attention with 12 frames lookahead. Still, there remains a gap of about 9% relative WER difference between the online and the offline system.

Table 4 shows the results obtained by dual-mode training. We use the efficient KLD loss in case of  $\gamma > 0$ . As in the Librispeech scenario, using chunked attention in the online mode helps improving both online and offline performance. The dual-mode system with chunked attention obtains 3.7% / 3.1% relative WERR compared to the one using autoregressive attention with lookahead, and 5.4% / 1.6% with respect to the corresponding single-mode online / offline system. However, unlike on

Table 4: *Medical conversation data: WER obtained by dual-mode systems in online and offline inference with and without efficient distillation (loss weight  $\gamma$ , shift  $\tau$ ).*

Online att.	$\gamma$	$\tau$	WER [%]	
			Online	Offline
Autoreg. LA	0.0	-	14.2	13.3
Autoreg. LA	0.01	0	14.2	13.3
Autoreg. LA	0.01	-6	14.1	13.2
Chunked	0.0	-	<b>13.7</b>	<b>12.9</b>
Chunked	0.01	0	<b>13.7</b>	<b>12.9</b>
Chunked	0.01	-6	13.8	13.0

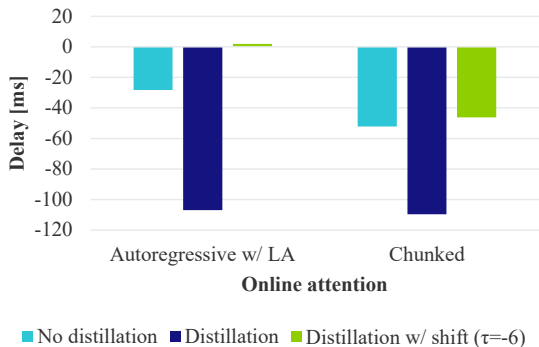


Figure 3: *Emission delay of dual-mode systems vs. single-mode online reference measured on medical data (lower means earlier).*

Librispeech, we do not observe any gain from distillation, even with  $\tau = -6$ . One possible reason is that the performance difference between the online and offline models on the medical data set is small (about 9% relative, see Table 1) compared to that on the Librispeech data set (23% relative, see Table 1).

### 3.3. Emission Timing

Figure 3 shows the time delay of transcriptions produced by our dual-mode models, with respect to a single-mode reference. We compute this delay as the average time difference of all the matching couples of correct words. For each word we consider the time of the emitting frame of its last word-piece in the RNN-T output alignment. If the encoder is chunked, times are rounded up to the end of their corresponding encoder chunks. The absolute emission delay is similar between autoregressive attention with lookahead and chunked attention models.

As can be seen in Figure 3, dual-mode models typically emit faster than the single-mode ones. When the dual-mode model is trained without distillation, we observe a slightly lower delay in the chunked configuration compared with the autoregressive one, while a significant improvement in terms of emission delay ( $\approx 110$  ms) is evident in both cases when the distillation is enabled. However, the latency gain from distillation vanishes when the teacher targets are shifted with a negative  $\tau$  value, because this configuration encourages later emission.

### 3.4. Effect of Dual-mode Normalization and Joint Training

In Table 5, we assess the importance of the dual-mode normalization layers in the Conformer blocks proposed by [16] vs. simply sharing the normalization layers between online and offline mode. On Librispeech (using chunked attention and efficient

Table 5: *WER obtained by dual-mode systems in online and offline inference, using dual normalization layers (one for online and one for offline) or single normalization layers (shared between online and offline mode).*

Norm. layers	WER [%]	
	Online	Offline
<i>Librispeech 100h (clean / other)</i>		
dual	7.1 / 21.5	6.1 / 18.9
single	7.1 / 21.3	6.3 / 19.0
<i>Medical conversation task</i>		
dual	13.7	12.9
single	13.7	13.0

distillation with  $\tau = -6$ ), the online performance is very similar between dual and single-mode normalization, while there is a small degradation in offline WER. The medical conversation task (using chunked attention but no distillation) shows a similar picture. Since we use causal convolutions for both online and offline mode as in the previous experiments, using a single set of normalization layers means that convolutional and feedforward components are identical to the single-mode Conformer, and the MHSA layers vary only the attention mask. Thus, single-mode normalization considerably simplifies the implementation while yielding similar performance.

Motivated by these results, we also investigated a further simplification of the dual-mode training where the attention mask for all MHSA layers is randomly chosen as the global (offline) or chunked (online) one for each line in the current mini-batch, instead of training both modes on the entire batch (joint training, cf. Eq. (1)). This is similar in spirit to the sampling techniques in [15–17]. The advantage is that only one model (online or offline) is computed for each utterance, thus saving approximately 50% of computation and memory requirement. We found such dual-mode training to yield a single model for both online and offline mode that performed similar to the dedicated single-mode models (14.2% / 13.3% WER on the medical task). However, unlike joint training, it did not result in a sizable WER gain compared to the single-mode baselines.

## 4. Conclusions

In this paper, we presented an in-depth study on the performance of dual-mode training for online Conformer Transducer architectures. We could obtain significant WER improvements in online mode on both the Librispeech and a medical conversational speech task, even without in-place distillation, and match the performance of dedicated offline models. Best results in online mode were obtained using chunked attention. Our results also shed light on the importance of modeling emission delay when doing offline-to-online knowledge distillation: we found that distillation without shift is helpful for reducing latency, while distillation with shift can reduce the WER at the expense of emission delay. The latter could potentially be mitigated by techniques such as FastEmit [28]. In general, the gain from distillation depends on the online configuration (especially the lookahead) and the data set. Furthermore, we explored several modifications to the original training approach, and found a simplified version, where only the attention mask is exchanged between online and offline modes, to perform equally well as the original proposal [16]. In future work, we will apply our findings to multi-mode ASR [17] for improving robustness of the online model in multiple latency requirements.

## 5. References

- [1] W. Han, Z. Zhang, Y. Zhang, J. Yu, C.-C. Chiu, J. Qin, A. Gulati, R. Pang, and Y. Wu, "ContextNet: Improving convolutional neural networks for automatic speech recognition with global context," in *Proc. Interspeech 2020*, 2020, pp. 3610–3614.
- [2] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented Transformer for speech recognition," in *Proc. of INTERSPEECH*. Shanghai, China: ISCA, 2020, pp. 5036–5040.
- [3] F. Weninger, M. Gaudesi, R. Leibold, R. Gemello, and P. Zhan, "Dual-encoder architecture with encoder selection for joint close-talk and far-talk speech recognition," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Cartagena, Colombia: IEEE, 2021, pp. 534–540.
- [4] A. Graves, "Sequence transduction with recurrent neural networks," in *Proc. of ICML Workshop on Representation Learning*. Edinburgh, UK: PMLR, 2012.
- [5] E. Battenberg, J. Chen, R. Child, A. Coates, Y. G. Y. Li, H. Liu, S. Satheesh, A. Sriram, and Z. Zhu, "Exploring neural transducers for end-to-end speech recognition," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Okinawa, Japan: IEEE, 2017, pp. 206–213.
- [6] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, "A comparison of sequence-to-sequence models for speech recognition," in *Proc. Interspeech 2017*, 2017, pp. 939–943.
- [7] C.-F. Yeh, J. Mahadeokar, K. Kalgaonkar, Y. Wang, D. Le, M. Jain, K. Schubert, C. Fuegen, and M. L. Seltzer, "Transformer-Transducer: End-to-end speech recognition with self-attention," *ArXiv*, vol. abs/1910.12977, 2019.
- [8] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott *et al.*, "Transformer Transducer: A streamable speech recognition model with Transformer encoders and RNN-T loss," in *Proc. of ICASSP*. Barcelona, Spain: IEEE, 2020, pp. 7829–7833.
- [9] J. Li, Y. Wu, Y. Gaur, C. Wang, R. Zhao, and S. Liu, "On the comparison of popular end-to-end models for large scale speech recognition," in *Proc. Interspeech 2020*, 2020, pp. 1–5.
- [10] B. Li, A. Gulati, J. Yu, T. N. Sainath, C.-C. Chiu, A. Narayanan, S.-Y. Chang, R. Pang, Y. He, J. Qin, W. Han, Q. Liang, Y. Zhang, T. Strohman, and Y. Wu, "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)*, 2021, pp. 5634–5638.
- [11] 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)*, 2021, pp. 5904–5908.
- [12] Y. Shi, Y. Wang, C. Wu, C.-F. Yeh, J. Chan, F. Zhang, D. Le, and M. Seltzer, "Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6783–6787.
- [13] A. Tripathi, J. Kim, Q. Zhang, H. Lu, and H. Sak, "Transformer Transducer: One model unifying streaming and non-streaming speech recognition," *arXiv:2010.03192*, 2020.
- [14] Z. Gao, S. Zhang, M. Lei, and I. McLoughlin, "Universal ASR: unifying streaming and non-streaming ASR using a single encoder-decoder model," *CoRR*, vol. abs/2010.14099, 2020.
- [15] K. Audhkhasi, T. Chen, B. Ramabhadran, and P. J. Moreno, "Mixture model attention: Flexible streaming and non-streaming automatic speech recognition," in *Proc. Interspeech 2021*, 2021, pp. 1812–1816.
- [16] 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 *ICLR 2021*, 2021.
- [17] K. Kim, F. Wu, P. Sridhar, K. J. Han, and S. Watanabe, "Multi-Mode Transformer Transducer with Stochastic Future Context," in *Proc. Interspeech 2021*, 2021, pp. 1827–1831.
- [18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.
- [19] G. Kurata and G. Saon, "Knowledge distillation from offline to streaming RNN transducer for end-to-end speech recognition," in *Proc. Interspeech 2020*, 2020, pp. 2117–2121.
- [20] 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)*, 2021, pp. 5639–5643.
- [21] X. Yang, Q. Li, and P. C. Woodland, "Knowledge distillation for neural transducers from large self-supervised pre-trained models," *arXiv:2110.03334*, 2021.
- [22] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. of ICASSP*. Vancouver, Canada: IEEE, 2013, pp. 7893–7897.
- [23] F. Weninger, J. Andrés-Ferrer, X. Li, and P. Zhan, "Listen, Attend, Spell and Adapt: Speaker adapted sequence-to-sequence ASR," in *Proc. of INTERSPEECH*. Graz, Austria: ISCA, 2019, pp. 3805–3809.
- [24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. of ICASSP*. Brisbane, Australia: IEEE, 2015, pp. 5206–5210.
- [25] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech 2015*, 2015, pp. 3586–3589.
- [26] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. of INTERSPEECH*. Graz, Austria: ISCA, 2019, pp. 2613–2617.
- [27] Y. Higuchi, N. Chen, Y. Fujita, H. Inaguma, T. Komatsu, J. Lee, J. Nozaki, T. Wang, and S. Watanabe, "A comparative study on non-autoregressive modelings for speech-to-text generation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 47–54.
- [28] J. Yu, C.-C. Chiu, B. Li, S.-Y. Chang, T. N. Sainath, Y. R. He, A. Narayanan, W. Han, A. Gulati, Y. Wu, and R. Pang, "FastEmit: Low-latency streaming ASR with sequence-level emission regularization," in *ICASSP 2021*, 2021.