INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Bridging the gap between streaming and non-streaming ASR systems by distilling ensembles of CTC and RNN-T models

Thibault Doutre, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Olivier Siohan, Liangliang Cao

Google Inc., USA

doutre@google.com, llcao@google.com

Abstract

Streaming end-to-end automatic speech recognition (ASR) systems are widely used in everyday applications that require transcribing speech to text in real-time. Their minimal latency makes them suitable for such tasks. Unlike their non-streaming counterparts, streaming models are constrained to be causal with no future context and suffer from higher word error rates (WER). To improve streaming models, a recent study [1] proposed to distill a non-streaming teacher model on unsupervised utterances, and then train a streaming student using the teachers' predictions. However, the performance gap between teacher and student WERs remains high. In this paper, we aim to close this gap by using a diversified set of non-streaming teacher models and combining them using Recognizer Output Voting Error Reduction (ROVER). In particular, we show that, despite being weaker than RNN-T models, CTC models are remarkable teachers. Further, by fusing RNN-T and CTC models together, we build the strongest teachers. The resulting student models drastically improve upon streaming models of previous work [1]: the WER decreases by 41% on Spanish, 27% on Portuguese, and 13% on French.

Index Terms: speech recognition, streaming ASR, model distillation, model ensemble, CTC, RNN-T

1. Introduction

ASR systems are used to transcribe speech to text in many daily applications. They can be embedded on devices like smart home devices or smartphones and can also be used in cloud services. These systems can be designed to be either steaming or non-streaming. Non-streaming models can take advantage of the full sequence of audio when transcribing speech [2, 3]. However, this requires that the full speech sequence be sent to the system before having access to a transcript. Streaming end-to-end ASR systems [4, 5, 6, 7, 8, 9, 10] are developed for real-time recognition tasks such as smart assistants and real-time captioning. Due to their streaming constraint, such systems cannot use the full context and generally perform less than non-streaming systems. This paper addresses this performance gap between streaming and non-streaming models.

Recent research has focused on improving streaming ASR by trading off latency [11, 3, 12]. For example, [13] shows that allowing streaming models to add a small amount of future information is effective in reducing the WER of RNN-T models. Some work attempts on improving streaming ASR using model distillation [14, 15, 16, 17]. More recently, [1] extends this approach by distilling a non-streaming teacher model on large-scale unsupervised YouTube dataset. By scaling the unsupervised data set, streaming models are trained on a large amount of training data and can outperform streaming models trained on a limited amount of supervised data. However, there is still a significant gap between teacher and student WERs.

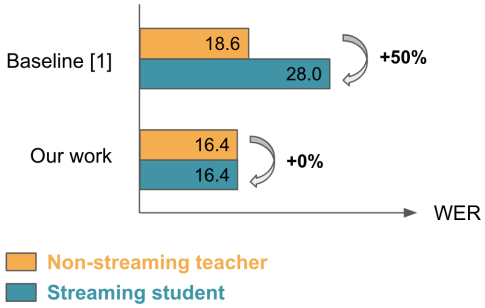


Figure 1: *Comparing WER between streaming students and their respective teachers (Spanish). The gap in performance between teachers and students WERs significantly improves over previous work [1].*

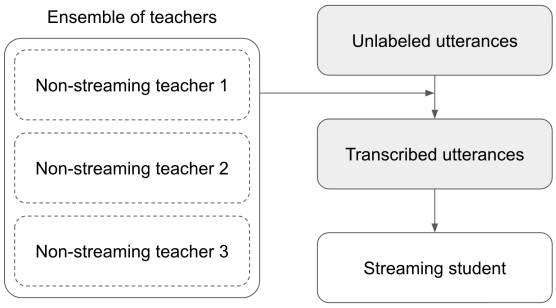


Figure 2: *Teacher-student framework. The student model is trained on an arbitrarily large set of utterances, transcribed by an ensemble of pre-trained teacher models.*

In this paper, we aim to further reduce the gap between non-streaming teachers and streaming students. We propose to improve this approach by (1) using a variety of non-streaming teacher models (2) combining teacher models to improve the quality of transcription (3) using stronger student models. In particular, we find Connectionist Temporal Classification (CTC) models [18] better teacher models than RNN-T models. In previous literature, RNN-T was considered better than CTC [5], as CTC suffers from the frame-independence assumption and often relies on an additional language model to capture the context information. In this work, however, RNN-T student models appear to generalize better when trained from CTC teachers. In addition, we find it advantageous to combine CTC with RNN-T, as the ensemble brings significant gains: the diversity of such models creates stronger teachers than any single model.

When comparing our work to [1], we drastically improve

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-637

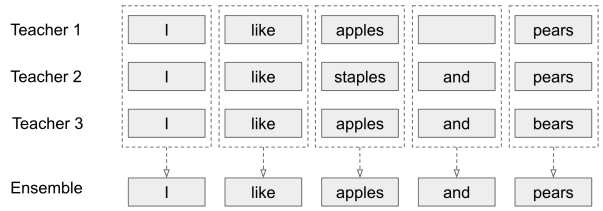


Figure 3: *Transcripts of multiple teacher models combined in a transition network by aligning them iteratively. The resulting transcript is obtained with majority voting.*

the performance of streaming models using a similar amount of labeled data in multiple languages (later shown in Table 5). According to [1], the WER increase of student models ranges from 24% to 50% relative to their respective teachers, depending on the language. Figure 1 shows that the performance gap between streaming students and non-streaming teachers is closed in Spanish. In section 3, we will show that this behavior is consistent in all languages.

2. Method

2.1. The teacher-student framework

In [1], the authors describe a method to improve streaming ASR by using non-streaming models. Given a pre-trained non-streaming teacher model, a set of randomly segmented YouTube audio data is transcribed. The resulting data set is then augmented with pseudo labels and used to train streaming models. The authors showed that the teacher's performance is an important factor in the student's performance. By scaling this method to millions of hours of unlabeled data, streaming models improve up to 39% compared to a baseline RNN-T model.

2.2. Ensemble of teachers

To improve the teacher-student framework, we propose to train several teacher models and combine them with ROVER [19]. The data is first transcribed with multiple teachers. The resulting transcripts are iteratively aligned to form a word transition network (WTN). Then the edges of the WTN are merged by majority voting, as described in Figure 3. Since the resulting transcripts come from the predictions of several teachers, they can be viewed as the transcription of an ensemble of teacher models. Should the ensemble of teachers perform better than any single teacher, the resulting transcripts may also be better to learn from.

2.3. Combining RNN-T and CTC models

ROVER is known to perform best when using various models, so it can exploit differences in the categories of errors generated by the different ASR systems [20, 21]. In this paper, we choose to combine multiple RNN-T [4] and CTC [18] models. For all teacher models, we adopt a Conformer [22] based encoder which has shown improvements over alternatives by taking advantage of both convolution neural networks (CNN) and transformers' architectures to capture local and global context.

In all languages, we consider three teacher models. The first non-streaming teacher is the **MD-RNNT** model [22]. This model stacks 17 macro layers in the encoder. Each macro layer consists of 8 attention heads, 1-D convolutions with a kernel size of 15 and a relative positional embedding of dimension 512. We use a right context of size 42 and a left context of size 43. The encoder dimension is 512. The decoder network has a uni-directional LSTM with 1024 hidden units. The joint network has 512 hidden units and the final output uses a 4k word piece model. The total number of parameters for this model is 141 million. The second teacher model we are considering is the **YT-RNNT** model [23]. The encoder starts with 2 layers of 3x3 2D convolution layers at the bottom with 4x time reduction, with a channel size of 640. There are a total of 16 conformer blocks, that are similar to the macro layers of MD-RNNT, except for the convolutions which have a kernel size of 32 and a positional embedding of size 128. The encoder dimension is 640. The decoder layer has 1 LSTM with 2048 cells and a projection layer of 640 outputs. This makes up a total of 175 million parameters. For both MD-RNNT and YT-RNNT models, we follow [22] to set the front-end and SpecAugment hyper-parameters. The third teacher model used in this study is a CTC model: **YT-CTC**. It has the same encoder as the YT-RNNT model, but its decoder only has 1 layer of 4096 hidden units with a projection layer of 640 outputs. Due to the simplified nature of the encoder, this model has 167 million parameters which is fewer than the YT-RNNT model. Both YT-CTC and YT-RNNT models are trained from YouTube data, whereas MD-RNNT is trained on multi-domain data [24]. The data include audio from YouTube, but also other data sources that come from user requests made on various devices.

Table 1: *Architecture of the different teacher models.*

	Encoder	Decoder	Data
MD-RNNT	17 blocks	1 LSTM	Multi-domain
YT-RNNT	16 blocks	1 LSTM	YouTube
YT-CTC	16 blocks	1 layer	YouTube

As for our student model, we use the same streaming RNN-T model with the Conformer encoder across all languages. This model is similar to the MD-RNNT, with the exception that attention layers and convolutions no longer allow full context to ensure its streaming quality.

3. Experiments on YouTube data

We confirm our method on YouTube data in three different languages: Spanish, Portuguese, and French. For each language, we are given three non-streaming teacher models, later combined with ROVER. To ensure a fair comparison, we use the same streaming RNN-T student model in all experiments.

3.1. Evaluation set

We evaluate our models on *YT-long*, a set of utterances generated by sampling and hand-transcribing popular YouTube videos based on their view counts. The length of utterances varies between 40 seconds and 30 minutes. When decoding an utterance, no segmentation is applied: the whole utterance is being processed at once.

3.2. Training data

As in [1], we train our teacher models on *Confisland* data [25]. The data is semi-supervised and has been acquired from a set of videos with transcripts uploaded by YouTube video owners. Then, as described in [25], only part of the audio where the transcript matches the audio with a certain confidence is being

Table 2: *Number of hours of the Confisland and YT-segments data sets for different languages.*

	Confisland	*YT-segments*
Spanish	13,000	41,000
French	10,000	29,000
Portugese	2,500	5,000

Table 3: *WERs of a streaming Conformer student model trained on YT-segments, distilled from non-streaming teacher models.*

	Teacher model	Teacher WER on *YT-long*	Student WER on *YT-long*
Spanish	MD-RNNT	16.4	33.4
	YT-RNNT	18.6	23.4
	YT-CTC	20.2	16.9
	Teacher ensemble	18.1	16.4
Portuguese	MD-RNNT	29.1	31.9
	YT-RNNT	22.8	26.7
	YT-CTC	24.8	23.0
	Teacher ensemble	21.9	20.5
French	MD-RNNT	31.9	42.8
	YT-RNNT	18.8	23.6
	YT-CTC	21.0	16.6
	Teacher ensemble	18.5	16.7

kept for training. Approximately 30% to 50% of the audio is kept, depending on the language.

As for our unsupervised set *YT-segments*, we randomly segment the audio of the original set of *Confisland* into utterances of lengths varying uniformly between 5 and 15 seconds. Since *YT-segments* are unlabeled, we can use all the data, even if the user uploaded transcripts are missing or incorrect. The size of each data set is described in Table 2.

3.3. Results

In Table 3 we compare how well the same streaming RNN-T model performs when trained from different teachers and from the ensemble of teachers. We note that, despite the higher WER of CTC models, they are better teachers than their non-streaming RNN-T counterparts. In particular, RNN-T students trained from CTC teachers surpass their teachers on *YT-long*. When using the model ensemble, teacher performance not only improves, but also its corresponding student. The gap between the best streaming student and the best non-streaming teacher is significantly reduced.

Specifically, Table 4 and Table 5 show the improvement over the results in [1]: by using an ensemble of teachers and a RNN-T student model, we outperform baselines trained from single non-streaming RNN-T teachers on similar data. The relative improvement in WER varies between 27% for Portuguese and 42% for Spanish.

3.4. CTC vs RNN-T teachers

Figure 4 and Table 3 show that student models trained from CTC and RNN-T teachers behave differently. When CTC is used, the corresponding student models outperform their teachers on *YT-long*. However, when RNN-T teachers are used, students preform worse than their teachers. Looking at several teacher predictions on the French *YT-segments*, we found that

Table 4: *Relative change of streaming students WERs, compared to their respective non-streaming teachers. The lower the increase the better. When multiple teachers are involved, the WER of the best teacher is used.*

	Baseline [1]	Our streaming student
Spanish	+50%	+0%
Portugese	+24%	-6%
French	+34%	-11%

Table 5: *Comparison of the WER of streaming models in this paper compared with streaming baselines [1] trained on similar data.*

	Spanish	Portuguese	French
Streaming RNN-T on *Confisland* [1]	35.9	30.8	34.5
Baseline streaming student [1]	28.0	28.3	25.0
Our streaming student	16.4	20.5	16.7
Relative improvement relative to the baseline streaming student	41%	27%	13%

CTC transcripts are generally worse than their RNN-T counterparts. Table 6 shows a typical utterance from *YT-segments*, transcribed using multiple teachers. The CTC transcript suffers from grammatical problems and made-up words. This is likely a result of the local nature of the CTC decoder, not being able to form a coherent sentence from word pieces.

Despite generating worse labels, the CTC model appears to be a better teacher to learn from. To determine whether this behavior is tied to *YT-long*, we look at the WER on a *Confisland* dev set: across all languages and for all models, students always perform worse than their teachers. Table 7 shows such gap between teachers and students when evaluated on *Confisland* data. The main difference between *Confisland* and *YT-long* being the length of utterances, it seems that student models trained from CTC teachers suffer less from long form errors.

Finally, we look at the students' training losses, normalized per token. Figure 5 shows that students trained from CTC teachers have a higher loss than those trained from RNN-T teachers. This indicates that RNN-T transcripts are easier to fit than CTC transcripts. The high loss of students trained from CTC teachers

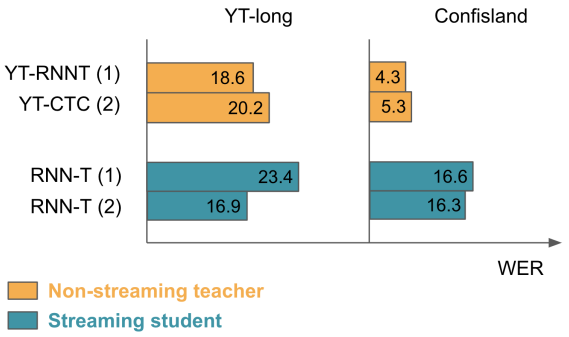


Figure 4: *Comparing a RNN-T streaming student model trained from CTC and RNN-T non-streaming teachers. The CTC model is a better teacher, despite having a higher WER.*

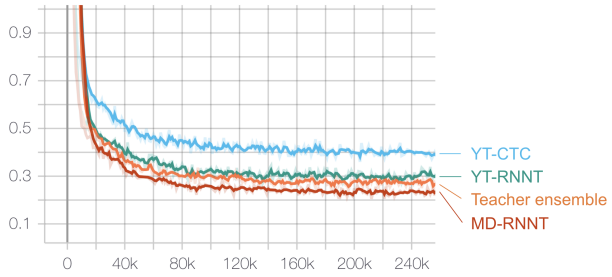


Figure 5: *Loss of student models trained from different teachers on the Spanish Confisland dev set, normalized per token.*

Table 6: *Comparing different teachers' predictions on French YT-segments. Bold words are not in the French vocabulary and underlined words are not in the ground truth.*

Teacher model	Prediction
MD-RNNT	euh qui achète pour revendre sur les vide-greniers sans s'acquitter des taxes et obligations
YT-RNNT	gens qui achètent pour revendre sur les vide-greniers sans s'acquitter des taxes et obligations ou
YT-CTC	qui **achèèteent** pour revendre sur le **videe-** grenier sans accquitter des taxes et obligations ou
Teacher ensemble	gens qui pour revendre sur le vide-greniers sans s'acquitter des taxes et obligations ou
Ground truth	qui achètent pour revendre sur les vide-greniers sans s'acquitter des taxes et obligations

and their low WERs on *YT-long* suggests that students trained from CTC teachers generalize better on long utterances. The student-teacher framework could be more robust when teachers and students are of different nature. Such behavior has already been observed in the experiments of [26], where the authors compare two different student models for the same ContextNet teacher [27] in a noisy student setting [28]. The Conformer student has better results than the ContextNet student on Librispeech [29].

3.5. Ablation studies

In this subsection, we aim to better understand how RNN-T and CTC teachers work differently with student models.

3.5.1. A non-streaming student

First, we train a non-streaming version of the student model on the Spanish *YT-segments* distilled from the ensemble of teachers. This non-streaming student has the same architecture as MD-RNNT and improves the WER by 10% compared to its streaming counterpart (Table 8). This underlines that distillation itself contributes to improving non-streaming RNN-T models.

3.5.2. An ensemble of RNN-T teachers

In this subsection, we examine the importance of CTC in the ensemble of teachers. Specifically, we replace the YT-CTC teacher with another version of MD-RNNT. The WER of the

Table 7: *WERs of a streaming Conformer student model and their teachers on a Spanish Confisland dev set.*

Teacher model	Teacher WER on *Confisland*	Student WER on *Confisland*
MD-RNNT	6.8	17.8
YT-RNNT	4.3	16.6
YT-CTC	5.3	16.3
Teacher ensemble	4.1	16.2

Table 8: *WERs of teachers and their respective students in different ablation studies for the Spanish YT-long test set.*

Teacher ensemble	RNN-T Student	Teacher WER	Student WER
RNN-T + CTC	Streaming	18.1	16.4
RNN-T + CTC	Non-streaming	18.1	14.7
RNN-T only	Streaming	17.5	20.1
CTC only	Streaming	19.6	18.0

streaming student trained from this new ensemble of RNN-T models increases by 19% compared to when using the diverse set of teachers (Table 8). This highlights the fact that at least one CTC model as a teacher is critical to the student's performance.

3.5.3. An ensemble of CTC teachers

Lastly, we provide evidence that a diverse set of teachers is beneficial for student models. We train student models from an ensemble of CTC teachers made from five different checkpoints of YT-CTC. The performance of such teachers ranges from 20.2 to 21.0. The resulting ensemble improved the WER to 19.6. Table 8 shows, however, that the student model that is trained from this resulting ensemble of CTC teachers does not perform as well as with the diverse group of teachers: the WER increases from 16.4 to 18.0.

4. Conclusions

In this paper, we develop a robust teacher-student learning method. In particular, we observe that CTC teachers, despite having a higher WER than their RNN-T counterpart, are a better teacher for streaming RNN-T models. By combining CTC and RNN-T models into a teacher ensemble, we can train a new streaming student model that not only drastically improves the results in [1], but also close the gap between single non-streaming teachers and streaming students. Future work will include improving the CTC teacher by complementing it with a language model, expanding the amount of unsupervised data and adding more teachers to the ensemble.

5. Acknowledgements

We are very thankful for our colleagues Yu Zhang, Basi García, Min Ma, Zhiyun Lu, Yongqiang Wang, Anmol Gulati, Françoise Beaufays, and Trevor Strohman for their help and suggestions.

6. References

[1] T. Doutre, W. Han, M. Ma, Z. Lu, C.-C. Chiu, R. Pang, A. Narayanan, A. Misra, Y. Zhang, and L. Cao, "Improving streaming asr with non-streaming model distillation on unsupervised data," in *ICASSP 2021*, 2021.

[2] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.

[3] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[4] A. Graves, "Sequence Transduction with Recurrent Neural Networks," *arXiv e-prints*, p. arXiv:1211.3711, Nov. 2012.

[5] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *Proc. ICASSP*. IEEE, 2019, pp. 6381–6385.

[6] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7829–7833.

[7] C.-F. Yeh, J. Mahadeokar, K. Kalgaonkar, Y. Wang *et al.*, "Transformer-Transducer: End-to-End Speech Recognition with Self-Attention," *arXiv preprint arXiv:1910.12977*, 2019.

[8] E. Tsunoo, Y. Kashiwagi, T. Kumakura, and S. Watanabe, "Transformer asr with contextual block processing," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 427–433.

[9] N. Moritz, T. Hori, and J. Le, "Streaming automatic speech recognition with the transformer model," in *Proc. ICASSP*. IEEE, 2020, pp. 6074–6078.

[10] B. Li, S.-y. Chang, T. N. Sainath, R. Pang *et al.*, "Towards fast and accurate streaming end-to-end ASR," in *Proc. ICASSP*. IEEE, 2020, pp. 6069–6073.

[11] T. N. Sainath, R. Pang, D. Rybach, Y. He, R. Prabhavalkar, W. Li, M. Visontai, Q. Liang, T. Strohman, Y. Wu, I. McGraw, and C.-C. Chiu, "Two-Pass End-to-End Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2773–2777.

[12] T. N. Sainath, Y. He, B. Li, A. Narayanan, R. Pang, A. Bruguier, S. y. Chang, W. Li, R. Alvarez, Z. Chen, C. C. Chiu, D. Garcia, A. Gruenstein, K. Hu, A. Kannan, Q. Liang, I. McGraw, C. Peyser, R. Prabhavalkar, G. Pundak, D. Rybach, Y. Shangguan, Y. Sheth, T. Strohman, M. Visontai, Y. Wu, Y. Zhang, and D. Zhao, "A streaming on-device end-to-end model surpassing server-side conventional model quality and latency," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6059–6063.

[13] J. Li, R. Zhao, H. Hu, and Y. Gong, "Improving rnn transducer modeling for end-to-end speech recognition," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 114–121, 2019.

[14] S. Kim, M. L. Seltzer, J. Li, and R. Zhao, "Improved training for online end-to-end speech recognition systems," in *Interspeech 2018*. ISCA, September 2018.

[15] R. Takashima, S. Li, and H. Kawai, "An investigation of a knowledge distillation method for ctc acoustic models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5809–5813.

[16] G. Kurata and K. Audhkhasi, "Improved knowledge distillation from bi-directional to uni-directional lstm ctc for end-to-end speech recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 411–417.

[17] G. Kurata and G. Saon, "Knowledge Distillation from Offline to Streaming RNN Transducer for End-to-end Speech Recognition," in *Proc. Interspeech*, 2020.

[18] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376.

[19] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 08 2000.

[20] X. Haihua, Z. Jie, and G. Wu, "An efficient multistage rover method for automatic speech recognition," in *2009 IEEE International Conference on Multimedia and Expo*, 2009, pp. 894–897.

[21] H. Schwenk and J. Gauvain, "Combining multiple speech recognizers using voting and language model information," in *INTERSPEECH*, 2000.

[22] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and et al., "Conformer: Convolution-augmented transformer for speech recognition," *Interspeech 2020*, Oct 2020.

[23] C.-C. Chiu, A. Narayanan, W. Han, R. Prabhavalkar, Y. Zhang, N. Jaitly, R. Pang, T. Sainath, P. Nguyen, L. Cao, and Y. Wu, "Rnn-t models fail to generalize to out-of-domain audio: Causes and solutions," *ArXiv*, vol. abs/2005.03271, 2020.

[24] A. Narayanan, R. Prabhavalkar, C. Chiu, D. Rybach, T. N. Sainath, and T. Strohman, "Recognizing long-form speech using streaming end-to-end models," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 920–927.

[25] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *ASRU*, 2013.

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

[27] 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," *Interspeech 2020*, Oct 2020. [Online]. Available: http://dx.doi.org/10.21437/interspeech.2020-2059

[28] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," *Interspeech 2020*, Oct 2020. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1470

[29] 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)*, 2015, pp. 5206–5210.