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

Self-Adaptive Distillation for Multilingual Speech Recognition: Leveraging Student Independence

Isabel Leal, Neeraj Gaur, Parisa Haghani, Brian Farris[*], *Pedro J. Moreno*[*], *Manasa Prasad*[*], *Bhuvana Ramabhadran*[*], *Yun Zhu*[*]

Google, USA

{isabelleal, neerajgaur, parisah}@google.com

Abstract

With a large population of the world speaking more than one language, multilingual automatic speech recognition (ASR) has gained popularity in the recent years. While lower resource languages can benefit from quality improvements in a multilingual ASR system, including unrelated or higher resource languages in the mix often results in performance degradation. In this paper, we propose distilling from multiple teachers, with each language using its best teacher during training, to tackle this problem. We introduce *self-adaptive* distillation, a novel technique for automatic weighting of the distillation loss that uses the student/teachers confidences. We analyze the effectiveness of the proposed techniques on two real world use-cases and show that the performance of the multilingual ASR models can be improved by up to 11.5% without any increase in model capacity. Furthermore, we show that when our methods are combined with increase in model capacity, we can achieve quality gains of up to 20.7%.

Index Terms: speech recognition, knowledge distillation, multilingual, RNN-T

1. Introduction

Multilingual automatic speech recognition (ASR) models that can transcribe speech in different languages are appealing from different angles. First, the same model can be used for multilingual users who can speak different languages. Second, they can improve the performance of lower-resource languages by learning shared representations from data available from other languages [1, 2, 3, 4, 5]. Finally, they can simplify deployment, as one model, instead of N, is used to provide transcription for N languages. With the advent and progress of end-to-end speech recognition models [6] in the recent years, multilingual modeling has been further simplified, replacing conventional ASR components such as the acoustic, language, and pronunciation models, with a single unified model to provide multilingual recognition.

While multilingual models are able to achieve higher performance for low-resource languages than their monolingual counterparts, high-resource and unrelated languages often see a negative impact in performance [7]. In order to achieve high performance for linguistically diverse languages with varying amounts of data, recent work has focused on building higher-capacity models [8, 9, 10, 7] that often introduce structure to the network by pre-assigning parts of the network to be language-specific. However, using larger models is expensive during inference and can hinder using these models for serving traffic.

[*]: Authors sorted alphabetically.

In this paper we propose to use knowledge distillation [11] with the goal of arriving at a multilingual model with performance as good as or superior to similar capacity/architecture monolingual counterparts. Our method is inspired by [12] which showed that for NLP tasks, building a multi-task student by distilling from multiple single-task models can lead to high-quality multi-task models capable of outperforming both single-task and multi-task baselines.

A key challenge in knowledge distillation consists of effectively balancing knowledge distillation and training from ground-truth, i.e., effectively leveraging "student independence" from the teacher(s). In [12], a teacher annealing method is used to decrease distillation over time. In [13], the authors train the student model by first simultaneously matching the outputs of the teachers and the ground-truth, and then switching to ground-truth only. In [14], the authors use a random augmented training strategy that switches between ground-truth and distillation in random order. The challenge of balancing distillation with learning from ground-truth, and more generally of improving objective functions via careful weighting, has been a topic of several works in speech recognition and machine learning at broad [15, 16, 17, 18, 19].

We make the following novel contributions:

- We propose a novel method, called *self-adaptive distillation*, that combines ground-truth and distillation losses in a non-linear fashion that leverages student independence during training by automatically balancing learning from the teachers and ground-truth.
- We use self-adaptive distillation to train a multilingual student model that learns from ground-truth and multiple teachers. The teachers are either monolingual or multilingual models capable of recognizing fewer languages than the student.
- We further study the interaction between distillation and model capacity.

We study these methods in the context of streaming end-to-end multilingual RNN-T models. Our experimental results show that with self-adaptive distillation, the student not only surpasses the performance of the baseline multilingual model trained without distillation, but it also demonstrates stronger performance compared to individual teachers. Additionally, we find that gains from distillation are additive to gains from increasing the capacity of the student model.

2. Knowledge distillation for RNN-T

2.1. Background

Knowledge distillation [11] consists of transferring knowledge from one or more *teacher models* to a *student model*, and has a

http://dx.doi.org/10.21437/Interspeech.2021-614

wide range of applications in speech recognition, such as model adaptation and model compression [15, 20, 21]. The idea behind it is to learn from the predictions of the teacher by making the output of the student similar to the output of the teacher.

Several previous works have focused on distilling from teachers with different architectures [22] and non-streaming models [23, 24]. In this paper, we focus on the case where teachers and students are streaming end-to-end RNN-T models [6]. RNN-Ts are sequence to sequence models which are popular in streaming ASR applications. One of the primary reasons for their popularity is that they do not need alignments between the input and output sequences. Instead, in the RNN-T loss, the probability of the output sequence is calculated by marginalizing over all possible alignments of the input and output sequences [6]. For each training example, the model produces $U \times T$ predictions over K symbols, where U is the length of the output sequence and T is the length of the input sequence. A natural way to define a distillation loss for such models is then

$$\mathcal{L}_{\text{full distill},\mathcal{T}} = \sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{k=1}^{K} P_{\mathcal{T}}(k|\mathbf{x}) \ln \frac{P_{\mathcal{T}}(k|\mathbf{x})}{P_{\mathcal{S}}(k|\mathbf{x})} \quad (1)$$

where $\mathcal{T}$ is the teacher, $\mathcal{S}$ the student, and $\mathbf{x}$ the input sequence. To reduce the computation cost of Equation (1), [20] came up with a method to simplify the loss by replacing the distribution over K by a categorical distribution over $S_{u,t} = \{y_{u,t}, \phi, r_{u,t}\}$ where $y_{u,t}$ is the u-th symbol in the output sequence emitted after consuming t input frames, ϕ is the blank symbol, and $r_{u,t}$ is a symbol that accumulates the probability over all the remaining $K-2$ symbols. The distillation loss can then be defined as:

$$\mathcal{L}_{\text{distill},\mathcal{T}} = \sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{k\in S_{u,t}} P_{\mathcal{T}}(k|\mathbf{x}) \ln \frac{P_{\mathcal{T}}(k|\mathbf{x})}{P_{\mathcal{S}}(k|\mathbf{x})} \quad (2)$$

2.2. Multi-teacher distillation for multilingual ASR

Recent works have shown knowledge distillation to be effective in the context of multilingual machine translation [13] and multilingual speech recognition [14, 22]. In this work, we use a method inspired by the single-task to multi-task distillation developed in [12] to improve multilingual ASR. We use as teachers multiple monolingual or multilingual models capable of recognizing fewer languages than the student.

More precisely, for each teacher $\mathcal{T}$, we modify $\mathcal{L}_{\text{distill},\mathcal{T}}$ in Equation (2) by multiplying it by a term which is 1 when the training example belongs to a language that can be recognized by $\mathcal{T}$, and 0 otherwise. The total loss is then defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RNNT}} + \sum_{\mathcal{T}\in\text{teachers}} \mathcal{W}_{\mathcal{T}}\mathcal{L}_{\text{distill},\mathcal{T}} \quad (3)$$

where $\mathcal{W}_{\mathcal{T}}$ is the weight of the distillation loss $\mathcal{L}_{\text{distill},\mathcal{T}}$. The way in which $\mathcal{W}_{\mathcal{T}}$ is defined determines the type of distillation, as it will be discussed in Section 3.

3. Self-adaptive distillation

The weights $\mathcal{W}_{\mathcal{T}}$ in Equation (3) determine how much the student learns from the teachers and from ground-truth. In this section we discuss different ways of defining $\mathcal{W}_{\mathcal{T}}$, leading up to the introduction of the novel self-adaptive distillation, in which $\mathcal{W}_{\mathcal{T}}$ depends on both the teacher and student confidences.

3.1. Vanilla distillation

We call vanilla distillation the simplest type of distillation, in which $\mathcal{W}_{\mathcal{T}}$ is a constant α. In this case, the total loss is simply a linear combination of the student's RNN-T loss and the per-teacher distillation losses.

3.2. Distillation weight schedule

Leveraging student independence was shown to be important in [12], which proposed an annealing method to combine teacher predictions with ground-truth in a way to decrease distillation over time. Based on this idea, we experiment using a schedule for $\mathcal{W}_{\mathcal{T}}$. Instead of setting $\mathcal{W}_{\mathcal{T}}$ to a constant α, we let it be a decreasing function of the number of training steps, $\alpha(t)$.

3.3. Adaptive distillation

Previous works [15, 16, 17] have found different ways to improve distillation by taking into account teacher confidence when computing the total loss. Sharing a similar high-level idea, we define an *adaptive distillation* in which the total loss combines the RNN-T loss and the distillation losses linearly, but the weight of the distillation losses are adjusted on-the-fly by taking into account the teachers' RNN-T losses. More precisely, $\mathcal{W}_{\mathcal{T}}$ is in this case given by

$$\mathcal{W}_{\mathcal{T}} = g(\mathcal{L}_{\text{RNNT},\mathcal{T}}), \quad (4)$$

where $g(x)$ is a non-constant decreasing function and $\mathcal{L}_{\text{RNNT},\mathcal{T}}$ the teacher's RNN-T loss. This way, when the teacher is not confident, distillation is down-weighted. In our experiments, g is chosen to be of the form $g(x) = \frac{\alpha}{(1+x)}$.

3.4. Self-adaptive distillation

Self-adaptive distillation modifies adaptive distillation by making $\mathcal{W}_{\mathcal{T}}$ also dependent on the student's RNN-T loss:

$$\mathcal{W}_{\mathcal{T}} = f(\mathcal{L}_{\text{RNNT}},\ \mathcal{L}_{\text{RNNT},\mathcal{T}}), \quad (5)$$

where $f(x, y)$ is increasing on x and decreasing on y. We assume $f(x, \cdot)$ is not constant. This way, when the student is not confident (i.e., $\mathcal{L}_{\text{RNNT}}$ is high), distillation is up-weighted, and when the teacher is not confident (i.e., $\mathcal{L}_{\text{RNNT},\mathcal{T}}$ is high), distillation is down-weighted. In our experiments, f is chosen to be of the form $f(x, y) = \alpha\frac{x}{1+y}$.

Self-adaptive distillation effectively leverages student independence by letting the model decide on its own how heavily to weight distillation. Observe that the total loss becomes a non-linear combination of the RNN-T loss and the distillation losses. During back-propagation, $\mathcal{L}_{\text{RNNT},\mathcal{T}}$ is treated as a constant, but $\mathcal{L}_{\text{RNNT}}$ is not, and this is shown to be important for performance.

4. Experimental setup

4.1. Languages and Data

We conduct our experiments on two groups of languages. The first group, which we call *supercluster*, is a combination of a Nordic/Germanic cluster of languages (Danish, Finnish, Norwegian, Swedish, and Dutch) with a high-resource language (English). The second group, which we call *interspersed*, is a combination of unrelated high-resource languages, consisting of English, Brazilian Portuguese, Russian, and Turkish.

The amounts of training data vary per language (see Table 1). For all of these languages, the training and test data are anonymized and transcribed by humans. As in [10], we use two forms of data augmentation to mitigate overfitting: noise and reverberation based augmentation [25] and SpecAugment [26].

Table 1: *Training data sizes (in number of utterances) per language for the supercluster and interspersed language groups.*

Supercluster			Interspersed		
Language		**Train (M)**	**Language**		**Train (M)**
en-us	*(English)*	35.3	en-us	*(English)*	35.3
da-dk	*(Danish)*	3.5	pt-br	*(Br. Portuguese)*	18.1
fi-fi	*(Finnish)*	3.6	ru-ru	*(Russian)*	25.4
nb-no	*(Norwegian)*	5.4	tr-tr	*(Turkish)*	19.3
nl-nl	*(Dutch)*	8.4			
sv-se	*(Swedish)*	7.5			

4.2. Models

The input acoustic features to the model are 80-dimensional log-Mel features stacked over three frames as described in [27]. Our model follows the encoder-decoder RNN-T streaming architecture detailed in [28] with changes to the encoder, which in our case is formed by 17 conformer layers as described in [29]. All of our models (baselines, students, and teachers) share the same architecture up to model capacity. We use two types of models with respect to capacity: the standard capacity models (in which the encoder model dimension is set to 512, leading to a total of 137M parameters), and the high-capacity models, which have around twice as many parameters (the model capacity is increased by increasing the encoder model dimension to 768, leading to a total of 278M parameters). All models share the same target vocabulary, a wordpiece model containing 4096 wordpieces and including all languages. All models are trained with an effective batch size of 4096.

5. Results

In this section we present the results for our experiments. For our self-adaptive distillation experiments, we let f in Equation (5) be of the form $f(x, y) = \alpha\frac{x}{1+y}$, and for our adaptive distillation experiments, we let g in Equation (4) be of the form $g(x) = \frac{\alpha}{(1+x)}$. The weight $\mathcal{W}_\mathcal{T}$ for the distillation loss is then $\alpha\frac{\mathcal{L}_{\text{RNNT}}}{1+\mathcal{L}_{\text{RNNT},\mathcal{T}}}$ for self-adaptive, $\alpha\frac{1}{1+\mathcal{L}_{\text{RNNT},\mathcal{T}}}$ for adaptive, and α for vanilla distillation. For our distillation weight schedule experiments, we start with a distillation weight α and decrease it linearly to zero during training. In each case we try different values of α in the range $[10^{-4}, 10^{-2}]$ and report the best.

5.1. Supercluster

In this section we compare baselines, teachers, and the different types of distillation for the supercluster language group consisting of English and Nordic/Germanic languages. The teachers are a monolingual English model and a Nordic/Germanic multilingual model. In this Section, we show:

- Self-adaptive distillation outperforms baselines, all other types of distillation, and teacher models.
- Gains from self-adaptive distillation and gains from increase in model capacity are additive.

5.1.1. Teachers and baselines

Our teachers are a monolingual English model, and a multilingual model trained with data from Danish, Finnish, Norwegian, Swedish, and Dutch. Both are standard capacity models as described in Section 4.2. WERs for the teachers have been congregated in the "teachers" column of Table 2.

We train two types of baselines for our experiments – standard capacity baselines, and high-capacity baselines, as described in Section 4.2. Both types are trained on the same multilingual data as the student models but without distillation. We find that the standard capacity baseline see regressions of up to 10.6% relative to the teachers. Increasing capacity effectively eliminates these regressions. Results for the standard capacity and high-capacity baselines are shown in the "base" and "base768" columns of Table 2.

5.1.2. Students

Our best standard capacity students are the ones trained using self-adaptive distillation (column "selfadap" in Table 2). The student trained with self-adaptive distillation outperforms the baseline "base" by up to 11.5% relative, and manages to outperform the *teachers* by up to 3.5% relative.

When combined with capacity increase, self-adaptive distillation outperforms the baseline "base" by up to 18.3% relative, and outperforms the same-sized baseline "base768" by up to 8.2% relative. Additionally, this model outperforms the *teachers* by up to 11.1% relative. WERs for this high-capacity student are reported in the "selfadap768" column of Table 2.

The WERs for vanilla distillation, distillation with a weight schedule, and adaptive distillation are presented in the "vanilla", "schedule", and "adap" columns of Table 2. The distillation weight schedule can benefit some languages, but the improvements over "vanilla" are neither consistent nor large. Adaptive distillation brings improvements over vanilla distillation on some languages, but regressions on several others.

The non-linearity in the total loss combination of self-adaptive distillation is important; we also experimented training a model in which the distillation weight is defined in the same way, i.e. $\mathcal{W}_\mathcal{T} = \alpha\frac{\mathcal{L}_{\text{RNNT}}}{1+\mathcal{L}_{\text{RNNT},\mathcal{T}}}$, but is treated as a constant during back-propagation. This model (column "stopgrad" of Table 2) underperforms self-adaptive distillation, showing the importance of non-linearity in the gradient computation.

5.2. Interspersed

In order to study scaling distillation to a larger number of teachers, and to understand the effectiveness of distillation when regressions of the baseline over the teachers are large, we apply distillation to the 4-lingual interspersed language group consisting of English, Brazilian Portuguese, Russian, and Turkish. These languages were chosen due to their linguistic dissimilarities and the fact that there is a large amount of training data available for each one of them (see Table 1). Despite optimizations such as balanced training data and augmentation, as we increase the number of high-resource, unrelated languages, we see the baseline ASR quality degrade, since the languages start competing for capacity [7, 9]. See Figure 1 for the progression of en-us WER as the number of languages increases on models trained without distillation.

5.2.1. Teachers and baselines

When training 4-lingual (en-us, pt-br, ru-ru, tr-tr) interspersed students, we use four different teachers, one monolingual standard capacity model (as described in Section 4.2) for each language. WERs for these models have been congregated in the "teachers" column of Table 3.

We train two types of baselines – standard capacity baselines, and high-capacity baselines, as explained in Section 4.2, both trained on the same 4-lingual data as the student but with-

Table 2: *WER comparison for different types of supercluster distillation models, baselines, and teachers.*

		Baselines		Other distillation				Self-adaptive distillation	
	teachers	**base**	**base768**	**vanilla**	**schedule**	**adap**	**stopgrad**	**selfadap**	**selfadap768**
da-dk	9.4	10.4	9.1	9.7	9.5	9.5	9.5	9.2	8.5
fi-fi	15.9	17.4	16.0	16.2	16.1	16.4	16.2	16.0	15.0
nb-no	11.6	11.9	11.1	11.2	11.3	11.1	11.2	11.1	10.7
nl-nl	10.6	11.4	10.5	10.7	10.7	10.9	10.6	10.4	9.7
sv-se	12.2	13.0	11.8	11.8	11.8	12.0	11.9	11.8	10.9
en-us	6.3	6.5	6.1	6.0	6.2	5.9	6.1	6.1	5.6

Table 3: *WER comparison for 4-lingual interspersed distillation models, baselines, and teachers.*

		Baselines		Distillation		
	teachers	**base**	**base768**	**vanilla**	**selfadap**	**selfadap768**
en-us	6.3	7.2	6.5	6.9	6.9	6.2
pt-br	6.6	7.7	7.0	7.2	7.0	6.4
ru-ru	7.3	8.7	7.8	8.1	7.8	6.9
tr-tr	7.0	8.5	7.7	8.1	7.7	7.0

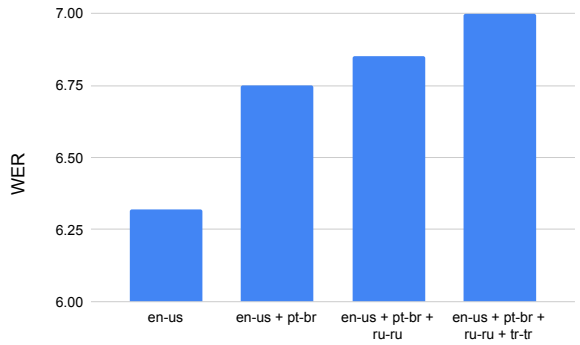

Figure 1: *Progression of en-us WER as more high-resource, unrelated languages are included in the training data.*

out distillation. Due to the choice of languages, the standard capacity baseline sees large regressions, up to 21.4% relative, when compared to the teacher models. Increasing capacity is effective in reducing regressions, but the high-capacity 4-lingual baseline still underperforms the teachers by up to 9.4% relative. WERs for standard and high-capacity baselines are shown in columns "base" and "base768" of Table 3, respectively.

5.2.2. Students

Consistently with what was seen for the supercluster language group in Section 5.1, the best standard capacity student is the one trained with self-adaptive distillation (see column "selfadap" in Table 3). This student achieves up to 10.3% improvement over the baseline "base". When combined with increase in capacity, self-adaptive distillation (column "selfadap768") outperforms the high-capacity baseline "base768" by up to 12.0% and the standard capacity baseline by up to 20.7%. This student is also able to outperform the teachers by up to 5.5%.

The WER for vanilla distillation is reported in the "vanilla" column of Table 3. We see that self-adaptive distillation outperforms vanilla distillation by up to 4.9%.

5.3. Self-adaptive distillation for model compression

To understand whether the effectiveness of self-adaptive distillation is dependent on the multi-teacher approach, we have also experimented using self-adaptive distillation in the classical setting of model compression. We find that self-adaptive distillation continues to be effective in this setting, outperforming the baseline and vanilla distillation. Results are reported in Table 4.

We use a high-capacity supercluster model as the only teacher (column "teacher" in Table 4), and train standard capacity self-adaptive and vanilla students (columns "selfadap" and "vanilla"). Self-adaptive distillation outperforms vanilla distillation by up to 4.1% and the baseline "base" by up to 9.6%.

Table 4: *WER comparison for model compression.*

		Baseline	Distillation	
	teacher	**base**	**vanilla**	**selfadap**
da-dk	9.1	10.4	9.8	9.4
fi-fi	16.0	17.4	16.4	16.2
nb-no	11.1	11.9	11.4	11.3
nl-nl	10.5	11.4	10.7	10.4
sv-se	11.8	13.0	12.1	11.8
en-us	6.1	6.5	6.1	6.1

6. Conclusions

This paper presents a novel knowledge distillation technique, called self-adaptive distillation, and a method for distilling from multiple teachers that, applied across several multilingual ASR systems for different language groups, brings significant improvements in WER of up to 11.5% over baselines trained on the same data but with no distillation. When the student's model capacity is increased, these gains are amplified to up to 20.7%. We also apply self-adaptive distillation to model compression and verify its effectiveness in that case as well.

7. References

[1] A. Waibel, H. Soltau, T. Schultz, T. Schaaf, and F. Metze, "Multilingual speech recognition," in *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, 2000, pp. 33–45.

[2] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7304–7308.

[3] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7319–7323.

[4] M. J. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: BABEL project research at CUED," in *Fourth International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2014)*. International Speech Communication Association (ISCA), 2014, pp. 16–23.

[5] Z. Tüske, P. Golik, D. Nolden, R. Schlüter, and H. Ney, "Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[6] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[7] N. Gaur, B. Farris, P. Haghani, I. Leal, P. J. M. Mengibar, M. Prasad, B. Ramabhadran, and Y. Zhu, "Mixture of Informed Experts for multilingual speech recognition," in *ICASSP 2021, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.

[8] A. Kannan, A. Datta, T. N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee, "Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model," *arXiv preprint arXiv:1909.05330*, 2019.

[9] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters," in *Proc. Interspeech 2020*, 2020, pp. 4751–4755. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2831

[10] Y. Zhu, P. Haghani, A. Tripathi, B. Ramabhadran, B. Farris, H. Xu, H. Lu, H. Sak, I. Leal, N. Gaur, P. J. Moreno, and Q. Zhang, "Multilingual Speech Recognition with Self-Attention Structured Parameterization," in *Proc. Interspeech 2020*, 2020, pp. 4741–4745. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2847

[11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *arXiv preprint arXiv:1503.02531*, 2015.

[12] K. Clark, M.-T. Luong, U. Khandelwal, C. D. Manning, and Q. V. Le, "BAM! Born-Again Multi-Task Networks for Natural Language Understanding," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5931–5937. [Online]. Available: https://www.aclweb.org/anthology/P19-1595

[13] X. Tan, Y. Ren, D. He, T. Qin, and T.-Y. Liu, "Multilingual Neural Machine Translation with Knowledge Distillation," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=S1gUsoR9YX

[14] J. Xu, J. Hou, Y. Song, W. Guo, and L. Dai, "Knowledge Distillation from Multilingual and Monolingual Teachers for End-to-End Multilingual Speech Recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 844–849.

[15] Z. Meng, J. Li, Y. Gaur, and Y. Gong, "Domain Adaptation via Teacher-Student Learning for End-to-End Speech Recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 268–275.

[16] Z. Meng, J. Li, Y. Zhao, and Y. Gong, "Conditional Teacher-student Learning," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6445–6449.

[17] S. Tang, L. Feng, W. Shao, Z. Kuang, W. Zhang, and Y. Chen, "Learning Efficient Detector with Semi-supervised Adaptive Distillation," *arXiv preprint arXiv:1901.00366*, 2019.

[18] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4057–4060.

[19] D. Povey and B. Kingsbury, "Evaluation of proposed modifications to mpe for large scale discriminative training," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 4, 2007, pp. IV–321–IV–324.

[20] S. Panchapagesan, D. S. Park, C.-C. Chiu, Y. Shangguan, Q. Liang, and A. Gruenstein, "Efficient Knowledge Distillation for RNN-Transducer Models," *arXiv preprint arXiv:2011.06110*, 2020.

[21] Y. Chebotar and A. Waters, "Distilling Knowledge from Ensembles of Neural Networks for Speech Recognition," in *Interspeech 2016*, 2016, pp. 3439–3443. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1190

[22] J. Cui, B. Kingsbury, B. Ramabhadran, G. Saon, T. Sercu, K. Audhkhasi, A. Sethy, M. Nussbaum-Thom, and A. Rosenberg, "Knowledge distillation across ensembles of multilingual models for low-resource languages," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4825–4829.

[23] T. Doutre, 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," *arXiv preprint arXiv:2010.12096*, 2021.

[24] 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. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2442

[25] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. N. Sainath, and M. Bacchiani, "Generation of Large-Scale Simulated Utterances in Virtual Rooms to Train Deep-Neural Networks for Far-Field Speech Recognition in Google Home," in *Proc. Interspeech 2017*, 2017, pp. 379–383. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1510

[26] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[27] A. Waters, N. Gaur, P. Haghani, P. Moreno, and Z. Qu, "Leveraging Language ID in Multilingual End-to-End Speech Recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 928–935.

[28] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S. yiin Chang, K. Rao, and A. Gruenstein, "Streaming End-to-end Speech Recognition For Mobile Devices," 2018.

[29] 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. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-3015