LEARNING FROM THE BEST: A TEACHER-STUDENT MULTILINGUAL FRAMEWORK FOR LOW-RESOURCE LANGUAGES

Deblin Bagchi[*]

Dept. of Computer Science and Eng.
The Ohio State University
Columbus, Ohio
bagchi.16@osu.edu

William Hartmann[†]

Raytheon BBN Technologies
Cambridge, Massachusetts
william.hartmann@raytheon.com

ABSTRACT

The traditional method of pretraining neural acoustic models in low-resource languages consists of initializing the acoustic model parameters with a large, annotated multilingual corpus and can be a drain on time and resources. In an attempt to reuse TDNN-LSTMs already pre-trained using multilingual training, we have applied Teacher-Student (TS) learning as a method of pretraining to transfer knowledge from a multilingual TDNN-LSTM to a TDNN. The pretraining time is reduced by an order of magnitude with the use of language-specific data during the teacher-student training. Additionally, the TS architecture allows us to leverage untranscribed data, previously untouched during supervised training. The best student TDNN achieves a WER within 1% of the teacher TDNN-LSTM performance and shows consistent improvement in recognition over TDNNs trained using the traditional pipeline over all the evaluation languages. Switching to TDNN from TDNN-LSTM also allows sub-real time decoding.

Index Terms— Teacher-student learning, Low-resource speech, Multilingual training, Automatic speech recognition

1. INTRODUCTION

State-of-the-art speech recognition systems have achieved human-level performance—depending on the domain and measure of human performance—in the last couple years because of neural network-based acoustic models trained on huge amounts of annotated speech [1]. However, acquiring such annotations are expensive and thus, similar progress on low-resource settings is difficult to achieve. People have resorted to multilingual or cross-lingual training which transfers knowledge from a well-trained model to scenarios where transcriptions are limited [2, 3, 4, 5]. Multilingual training has been extremely popular in low-resource speech recognition settings, especially after the emergence of neural networks in speech recognition [6]. BBN has observed a gain in ASR performance by training neural network models on a 1560-hour multilingual training corpus and then fine-tuning this network with the target language after swapping out the output layer to monolingual mode [7, 8]. Given the large amount of multilingual data, sophisticated neural network models (like recurrent neural networks) can be used during multilingual training. However, during the fine-tuning stage, the limited amount of data in the target language can lead to an optimization problem. Attempts have been made to only update a reduced number of parameters during finetuning, but updating the entire network with a reduced learning rate or fewer epochs typically performs better [9]. While the gains from multilingual pretraining are often impressive, the pretraining stage takes significant time and computational resources. Given the cost of the pretraining, it is expensive to explore alternative architectures. Performance using small monolingual corpora do not necessarily correlate with using the model in a multilingual setting.

In this work, we use teacher-student learning to transfer knowledge from recurrent neural networks (pretrained on multilingual data and fine-tuned on monolingual data) to non-recurrent time-delay neural network (TDNN) models. We speed up the pretraining process by using only monolingual training data and record lower word error rates (WERs) compared to TDNNs trained using both the monolingual as well as multilingual pipeline. Since annotations are not a necessity for TS training, we have also leveraged untranscribed data to achieve a further gain in ASR performance in the target languages. Instead of the popular Knowledge Distillation approach, which uses cross entropy loss over the senone distribution, we use *L2-loss* to match the performance of a hidden layer activation.

The work presented here has surface similarities with [5] in that they have also used teacher-student learning for multilingual training to help in the low-resource setting. However, they have used the entire multilingual dataset to train their student networks. They have also used cross entropy on softmax

[*]The author performed the work during an internship at Raytheon BBN Technologies.

[†]This work was supported by the IARPA Material program.

outputs to compute their loss function, which is the widely-used approach for speech.

2. MULTILINGUAL SYSTEMS

2.1. Chain TDNN and Chain TDNN-LSTM

Our acoustic models use the "chain model" topology [10] and three frame subsampling of the input features. We use both the TDNN and TDNN-LSTM architectures. While the TDNN-LSTM acoustic model provides a significant reduction in WER, it does come with increased computational cost, both during training and decoding.

2.2. Multilingual training and knowledge transfer

The multilingual training and fine-tuning pipeline is similar to [7] with some minor revisions. The standard criterion used during multilingual training is LF-MMI with an output layer that is simply a combination of all phone states for all languages. Note that each phone is tagged with a language ID, so the acoustic states are not shared across languages. To tune the multilingual network to the target low resource language, the multilingual output layer is discarded and a new output layer for the target language is created. The entire network is then fine-tuned using the LF-MMI criterion. The data used in this stage is the transcribed data for the target language. Following this, the model parameters are further sharpened with sMBR training; 1 epoch for TDNN-LSTM and 4 epochs for TDNNs.

3. TEACHER-STUDENT TRAINING

Teacher-student (TS) training is a technique to train a low-complexity student model from a more cumbersome, high-complexity teacher model. For DNN-HMM based speech recognition models, most literature has performed TS training using a KL-divergence based loss between the softmax distributions of the teacher and student models [11]. Hinton et.al [12] introduced the temperature parameter to sharpen (or flatten) the softmax distribution and uses TS training as an auxillary function for regularization of supervised training. Generalized distillation (GD) [13, 14, 15] extends distillation methods by training a teacher network with separate clean data. A student network is trained on noisy data and, at the same time, guided by the soft-labels from a teacher which has access to synchronized clean speech. The generalized distillation methods showed improved performance on CHiME4 and Aurora2 corpora. A variant of teacher-student learning has also been used in speech enhancement for noise-robust speech recognition [16, 17].

However, the TDNN and TDNN-LSTM models we use as teacher models do not use a softmax output. The final output layer is treated as a pseudo log-likelihood instead. Hence, we opt for the TS training method used in [18] for model compression. At first, a bottleneck layer of dimensionality D is chosen in both the teacher and the student. A forward pass of the input $x(m,u)$ through both the teacher and the student obtains activations $f()$ and $g()$ respectively. An *L2-based* distance loss between the two activations is then calculated and the student model parameters are updated based on this loss.

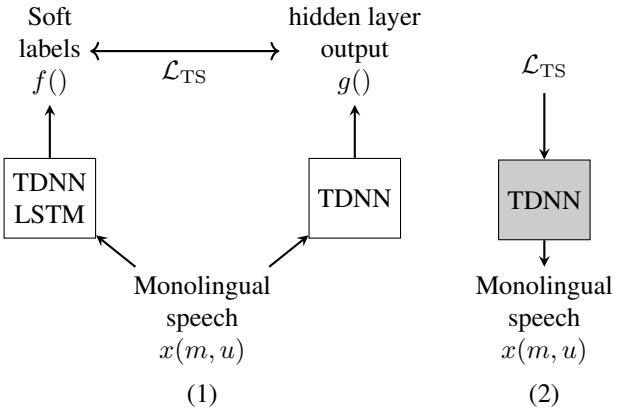

Fig. 1. This diagram depicts our teacher-student pretraining process from a teacher TDNN-LSTM to student TDNN. (1) Monolingual speech data is passed to both the teacher and student. The $L2-Loss$ is calculated according to Eqn 1. (2) The parameters of the student model are updated based on the loss $\mathcal{L}_{\text{TS}}$. The gray color portrays the update of the student TDNN parameters using the L2-loss all the way down the student TDNN.

$$\mathcal{L}_{\text{TS}}(x(m,u)) = \frac{1}{D}\sum_{d=1}^{D}(g(x(m,u)) - f(x(m,u)))^2 \quad (1)$$

where $x(m,u)$ refers to frame m of utterance u. We believe this approach offers several advantages. Since the softmax distribution is not used, this approach is more flexible and can be applied to more types of networks. More importantly, it is not hindered by the over confidence of the teacher model. In our previous experience, better, more confident teacher models can actually produce worse students. It should also be noted that we have only used language-specific data during the teacher-student training phase as opposed to multilingual data which significantly reduces the time taken for pretraining. A pictorial representation of the TS pretraining pipeline is shown in Fig 1.

4. EXPERIMENTS

The languages used to test our framework are Dari (20-hours of transcribed speech), Swahili (40 hours of transcribed speech) and Tagalog (80 hours of transcribed speech). We test our pipeline and tune parameters on Dari and apply the tuned architectures on Swahili and Tagalog.

4.1. Multilingual experiment specifications

The multilingual models are trained using 1560-hour Conversational Telephonic Speech (CTS) taken from a set of 11 different languages[1]. The training corpus used in this work is the same as the one used in [8] and [7]. It was trained with the BBN Sage toolkit [19], specifically using the integrated Kaldi [20] portion of the toolkit. The input features used were 40 dimensional high-resolution MFCC features for the input frame and the frames surrounding it, and 100-dimension i-vectors, for a total input feature vector of size 220. The structure of the model is the standard chain model as described in [10], and the data is subsampled to examine 1 out of every 3 frames. The TDNN architecture consists of 7 hidden layers with 576 neurons in each layer. The splicing configuration for the TDNN is as follows: {0}, {-1,0,1},{-1,0,1},{-1,0,1},{-3,0,3},{-3,0,3},{-3,0,3},{-3,0,3}, {0}. This can be read as each layer containing the splicing of the layers at a time relative to 0, with 0 being the current frame. So the first hidden layer uses solely the concatenated input feature vector, the second hidden layer uses a concatenation of the hidden layer outputs at the previous time-step, the current time-step, and the next time-step, and so on.The TDNN-LSTM has an architecture of tdnn-lstm-tdnn-lstm-tdnn-lstm (three alternating layers of TDNN and LSTM), where the splicing configurations of the TDNN layers is {-3,0,3}. There are 1024 neurons in the TDNN layer and 256 neurons in the LSTM layer.

Multilingual training was performed on the TDNN and TDNN-LSTM models using the LF-MMI objective function, with an initial learning rate of 1×10^{-3}, which gradually went down to 1×10^{-4}. The output layer consists of $\sim$16000 phone targets from all the member languages in the multilingual dataset.

After multilingual training, the TDNN and TDNN-LSTM models are fine-tuned for a single epoch using LF-MMI with an initial learning rate of 3×10^{-3}, which gradually goes down to 3×10^{-4}. We also perform sMBR (1-4 epochs based on the model) after LF-MMI. The learning rate of sMBR training is fixed at 5×10^{-6}. In Table 1, we list the Word Error Rates (WER) of models trained using multilingual training and fine-tuning on Dari, Swahili and Tagalog. Although Swahili and Tagalog corpora are both included in the multilingual training corpus, membership in the multilingual corpus does not seem to affect knowledge transfer optimization [7].

We compare models trained using multilingual training with models trained monolingually where only the target-language data is passed through randomly initialized neural networks with the same architecture as above. Parameters are updated first using LF-MMI and then sMBR using the same recipe as above. From the table, it can be concluded that there is a universal improvement in word recognition performance for models trained multilingually compared to monolingual models.

[1]The languages and amounts of data for the multilingual corpus are as follows: English (380hrs), Mandarin (250hrs), Spanish (245hrs), Cantonese (110hrs), Pashto (98hrs), Tagalog (90hrs), Vietnamese (90hrs), French (85hrs), Turkish (83hrs), Haitian (80hrs), Swahili (50hrs).

Language	Baseline Experiment	WER
Dari	Mono TDNN	49.1
	Multi TDNN	45.5
	Multi TDNN-LSTM	43.1
Swahili	Mono TDNN	53.5
	Multi TDNN	50.3
	Multi TDNN-LSTM	45.5
Tagalog	Mono TDNN	54.9
	Multi TDNN	51.8
	Multi TDNN-LSTM	47.6

Table 1. *Monolingual and multilingual baseline Word Error Rates (WER) for Dari, Swahili and Tagalog.*

4.2. Teacher-student experiments

Our main focus is to use TDNNs without recurrent connections to emulate the performance of TDNN-LSTMs. We use the TDNN-LSTM obtained after the fine-tuning step of multilingual training as teacher. Owing to the small size of Dari, we test the correctness of the teacher-student pipeline and finalize the student TDNN architecture on this language. We investigate student TDNNs with different configurations and compare performance to the teacher TDNN-LSTM. The results are listed in Table 2. It should be observed that when a student TDNN is trained using a teacher TDNN, the recognition performance of the student TDNN is similar to the teacher. We see no difference in performance by increasing the number of hidden layers or widening the layers. However, the situation changes when the teacher is a TDNN-LSTM. The recognition performance of a student TDNN improves with the inclusion of more hidden layers with a large number of neurons per layer. The improvement in student TDNN performance may be attributed to the fact that increasing the number of hidden layers in a TDNN also increases the length of context observed by the model. The best performing student TDNN has 12 layers and 1280 neurons in each layer. This increases the context from 30 frames in the baseline TDNN to 54 frames. With this model, we achieve WERs within 1% of the TDNN-LSTM teacher in Dari. Even though this student TDNN has more parameters than the teacher TDNN-LSTM, the decoding time is still faster than the teacher due to the absence of recurrent connections.

After tuning our models on Dari, we tested whether our observations held on Swahili and Tagalog. Hence, we chose one student architecture with 7 hidden layers and 576 neurons and pitted it against the best performing student model in Dari (12 hiden layers and 1280 neurons). The results are listed in Table 3. We see that a larger context is a positive influence across all languages. The best performing student TDNN out-

Teacher	Student	WER
Multi TDNN	(7, 576)	45.3
	(9, 1280)	45.4
Multi TDNN-LSTM	(7, 576)	45.9
	(7, 1024)	45.1
	(7, 1280)	45.1
	(9, 1280)	44.3
	(12, 1280)	43.9

Table 2. *Word Error Rates (WER) for teacher-student pretraining in Dari with TDNN and TDNN-LSTM teachers. The students are TDNNs represented as (number of layers, number of neurons in each layer)*

Language	Teacher	Student	WER
Swahili	TDNN	(7, 576)	50.6
		(12, 1280)	50.1
	TDNN-LSTM	(7, 576)	50.5
		(12, 1280)	47.9
Tagalog	TDNN	(7, 576)	52.7
		(12, 1280)	51.4
	TDNN-LSTM	(7, 576)	52.7
		(12, 1280)	50.4

Table 3. *Word Error Rates (WER) for teacher-student pretraining in Swahili and Tagalog with TDNN and TDNN-LSTM teachers. The students are TDNNs represented as (number of layers, number of neurons in each layer)*

performs a multilingual TDNN by 1.6%, 2.4% and 1.4% and a monolingual TDNN by 5.5%,5.6% and 4.5% absolute (refer to Table 1) for Dari, Swahili and Tagalog respectively.

4.3. Leveraging untranscribed data

The TDNN student WER for Dari was within 1% of the WER of the TDNN-LSTM teacher. However, in Table 3, we find that is not the case for Swahili and Tagalog. We suspect this is due to poor generalization on the part of the student TDNN models. The transcribed training data of Swahili and Tagalog is made up of conversational telephone speech (CTS). However, the evaluation is done over a mixture of CTS, news broadcast (NB) and topical broadcast (TB), i.e. part of the evaluation is from an unseen domain. Our suspicions are confirmed in Table 4 where we look at the WER breakdown for CTS, NB and TB respectively. The student model performance seems to have degraded over broadcast data.

The easiest fix would be to incorporate broadcast data to the training pipeline. Even though Swahili and Tagalog has 70 and 80 hours of broadcast data respectively, the data is not transcribed. The typical approach to leveraging untranscribed data in acoustic model training is to decode the data and treat the hypothesized transcripts as truth. However, with teacher-student learning, it is simple to combine the unlabelled data with the transcribed data and pass it through the teacher because labels are not required to train the student TDNNs during TS training.

Language	Model	All	CTS	NB	TB
Swahili	Before	47.9	34.4	48.5	53.2
	After	46.6	34.3	46.6	51.6
	TDNN-LSTM	45.5	32.8	45.3	50.8
Tagalog	Before	50.4	37.3	49.5	52.1
	After	48.8	37.5	47.5	50.4
	TDNN-LSTM	47.6	37.0	46.2	49.2

Table 4. *Swahili and Tagalog teacher-student trained TDNN WERs before and after leveraging untranscribed data, compared with multilingual TDNN-LSTM model.*

After using the combination of transcribed and untranscribed data in the teacher-student learning stage, we find that the performance of student TDNNs improve. The results are in Table 4. It can be observed that even though the recognition performance remains almost same for CTS, performance on NB improves by 1.9% and 2% absolute and performance on TB improves by 1.6% and 1.7% absolute for Swahili and Tagalog respectively.

5. CONCLUSION

In this work, we replace the traditional multilingual training phase for low-resource speech recognition with teacher-student training. The proposed training mechanism facilitates knowledge transfer from a TDNN-LSTM, trained and fine-tuned using multilingual training to a TDNN. The student TDNN models perform better than TDNNs trained using supervised training approaches (multilingual and monolingual training). The improvement is consistent across all the three languages we evaluate on, namely, Dari, Swahili and Tagalog. Traditionally, multilingual training is a drain on time and resources due to the bulkiness of the corpus. We reduce training time by using monolingual data during teacher-student training. The best student TDNN architecture has 12 hidden layers and 1280 neurons in each layer and achieves a WER within 1% of TDNN-LSTM performance for all test languages. Since increase in number of TDNN layers means subsequent increase in context, we understand that context plays an important role in improving the performance of student TDNNs. Further, teacher-student pretraining enables us to leverage large amounts of unlabelled data in a fairly easy way, enhancing the performance of the student TDNNs. It is also worth mentioning that in terms of decoding time, the best performing student TDNN model is still faster than the TDNN-LSTM due to the absence of recurrent connections.

6. REFERENCES

[1] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, et al., "English conversational telephone speech recognition by humans and machines," *arXiv preprint arXiv:1703.02136*, 2017.

[2] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky, "Multilingual mlp features for low-resource lvcsr systems," in *Proceedings of IEEE ICASSP*, 2012, pp. 4269–4272.

[3] Kate M Knill, Mark JF Gales, Shakti P Rath, Philip C Woodland, Chao Zhang, and S-X Zhang, "Investigation of multilingual deep neural networks for spoken term detection," in *Proceedings of IEEE ASRU*, 2013, pp. 138–143.

[4] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proceedings of IEEE ICASSP*, 2013, pp. 7304–7308.

[5] Jia Cui, Brian Kingsbury, Bhuvana Ramabhadran, Abhinav Sethy, Kartik Audhkhasi, Xiaodong Cui, Ellen Kislal, Lidia Mangu, Markus Nussbaum-Thom, Michael Picheny, et al., "Multilingual representations for low resource speech recognition and keyword search," in *Proceedings of IEEE ASRU*, 2015, pp. 259–266.

[6] Ngoc Thang Vu, Wojtek Breiter, Florian Metze, and Tanja Schultz, "An investigation on initialization schemes for multilayer perceptron training using multilingual data and their effect on asr performance," in *Proceedings of Interspeech*, 2012, pp. 2586–2589.

[7] Jeff Ma, Francis Keith, Tim Ng, Man-hung Siu, and Owen Kimball, "Improving deliverable speech-to-text systems with multilingual knowledge transfer," in *Proceedings of Interspeech*, 2017, pp. 127–131.

[8] Francis Keith, William Hartmann, Man-Hung Siu, Jeff Ma, and Owen Kimball, "Optimizing multilingual knowledge transfer for time-delay neural networks with low-rank factorization," in *Proceedings of IEEE ICASSP*, 2018, pp. 4924–4928.

[9] Khe Chai Sim, Arun Narayanan, Ananya Misra, Anshuman Tripathi, Golan Pundak, Tara Sainath, Parisa Haghani, Bo Li, and Michiel Bacchiani, "Domain adaptation using factorized hidden layer for robust automatic speech recognition," in *Proceedings of Interspeech*, 2018, pp. 892–896.

[10] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi.," in *Proceedings of Interspeech*, 2016, pp. 2751–2755.

[11] Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong, "Learning small-size dnn with output-distribution-based criteria," in *Proceedings of Interspeech*, 2014.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[13] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik, "Unifying distillation and privileged information," *arXiv preprint arXiv:1511.03643*, 2015.

[14] Konstantin Markov and Tomoko Matsui, "Robust speech recognition using generalized distillation framework.," in *Proceedings of Interspeech*, 2016, pp. 2364–2368.

[15] Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R Hershey, "Student-teacher network learning with enhanced features," in *Proceedings of IEEE ICASSP*, 2017, pp. 5275–5279.

[16] Deblin Bagchi, Peter Plantinga, Adam Stiff, and Eric Fosler-Lussier, "Spectral feature mapping with mimic loss for robust speech recognition," *arXiv preprint arXiv:1803.09816*, 2018.

[17] Peter Plantinga, Deblin Bagchi, and Eric Fosler-Lussier, "An exploration of mimic architectures for residual network based spectral mapping," *arXiv preprint arXiv:1809.09756*, 2018.

[18] Jimmy Ba and Rich Caruana, "Do deep nets really need to be deep?," in *Advances in neural information processing systems*, 2014, pp. 2654–2662.

[19] Roger Hsiao, Ralf Meermeier, Tim Ng, Zhongqiang Huang, Maxwell Jordan, Enoch Kan, Tanel Alumäe, Jan Silovskỳ, William Hartmann, Francis Keith, et al., "Sage: The new bbn speech processing platform.," in *Proceedings of Interspeech*, 2016, pp. 3022–3026.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *Proceedings of IEEE ASRU*, 2011.