Knowledge Distillation from Multilingual and Monolingual Teachers for End-to-End Multilingual Speech Recognition

Jingyi Xu, Junfeng Hou, Yan Song, Wu Guo, Lirong Dai
National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology in China, Hefei, China
E-mail: {sa6186, hjf176}@mail.ustc.edu.cn, {songy, guowu, lrdai}@ustc.edu.cn

***Abstract*—Attention-based encoder-decoder models significantly reduce the burden of developing multilingual speech recognition systems. By means of end-to-end modeling and parameters sharing, a single model can be efficiently trained and deployed for all languages. Although the single model benefits from jointly training across different languages, it should handle the variation and diversity of the languages at the same time. In this paper, we exploit knowledge distillation from multiple teachers to improve the recognition accuracy of the end-to-end multilingual model. Considering that teacher models learning from monolingual and multilingual data contain distinct knowledge of specific languages, we introduce multiple teachers including monolingual teachers of each language, and multilingual teacher to teach a same sized multilingual student model so that the multilingual student will learn various knowledge embedded in the data and intend to outperform multilingual teacher. Different from conventional knowledge distillation which usually relies on a linear interpolation for hard loss from true label and soft losses from teachers, a new random augmented training strategy is proposed to switch the optimization of the student model between hard or soft losses in random order. Our experiments on Wall Street Journal (English) and AISHELL-1 (Chinese) composed multilingual speech dataset show the proposed multiple teachers and distillation strategy boost the performance of the student significantly relative to the multilingual teacher.**

I. INTRODUCTION

Building a state-of-the-art speech recognition system requires a large amount of manual annotation of speech data, but it is expensive and time-consuming [1]. Especially, there are not enough well-annotated resources available in many languages in many cases. These problems have attracted a growing concern in multilingual and cross-lingual modeling, which allows for knowledge transfer across languages and thus relieves burdensome data requirements [2].

Conventional multilingual speech recognition model [3-5] needs language dependent components including pronunciation model (PM), acoustic model (AM), and language model (LM) which means that the model should know the language identity corresponding to each language during train and inference [2]. Moreover, errors are prone to accumulate from one component to next following components in a way that was not easily eliminated during training because of the independent optimization of AMs, PMs, LMs [2, 6]. On the contrary, attention based encoder-decoder model transcribes input speech sequence to output label sequences directly by integrating pronunciation models, acoustic models, language models into a unified structure, which simplifies the recognition process compared to conventional multilingual speech recognition model [6].

Knowledge distillation is first proposed by Hinton [7], and the idea is typically to transfer the knowledge of a high-capacity teacher with desired high performance to a more compact student. Although the student cannot match the teacher when trained directly on the same data, the distillation process brings the student closer to matching the predictive power of the teacher [9]. To improve the performance of student networks, more recent work has focused on multiple teacher models which combine the outputs of teacher networks to make the student learn this ensemble distribution so as to observe various "view" of the data [10].

However, knowledge distillation usually relies on a linear interpolation for each loss, which disables student model to directly access the individual complimentary teacher distributions and weakens the complementariness obtained by multiple teachers due to averaging them. To advance the distillation techniques, Fukuda [10] proposed an augmented-training strategy which uses each target loss sequentially to update the parameters of the student model per mini-batches. The strategy allows student model trained explicitly with a dedicated loss which is unaffected by other losses to enhance the versatility of capturing knowledge fields of teachers. A similar augmented-training strategy was used in multi-task learning [11], where the shared encoder of an encoder-decoder model in multiple tasks is updated sequentially during the optimization of each task loss. But they both almost ignored the influence that different orders of the losses will bring in.

Knowledge distillation has been applied on multilingual ASR and machine translation in previous works. In ref. [12], they investigated knowledge distillation as applied to different types of NN models and models trained with different input features. Ref. [13] proposed a distillation-based approach in neural machine translation where individual models are first trained and regarded as teachers and then the multilingual model is trained to fit the training data and match the outputs of individual models through knowledge distillation.

In this paper, we exploit knowledge distillation from multiple teachers to improve the performance of the student model —

end-to-end multilingual speech recognition model. In our work, several same sized encoder-decoder models are selected as our multiple teacher models including multilingual teacher and monolingual teachers of each language. For monolingual teachers, models are trained with corresponding monolingual data, while multilingual teacher model is trained with all the language data. By this mean, monolingual teachers focus on each single language, and multilingual teacher takes care of all languages. The complementary knowledge fields allow multilingual student to learn various knowledge embedded in the data and intend to perform better than multilingual teacher. What's more, observing that the performance of student depends largely on the last update loss in augmented-training strategy (see the discussion in IV-C), we propose a random augmented-training strategy to avoid always fixing the last update loss which will be demonstrated to help the student model to escape the local minimum. The proposed multiple teachers and random augmented-training strategy boost the performance of the student with about 2.5% word error rate (WER) reduction in EN and 1.5% character error rate (CER) reduction in CHN relative to the multilingual teacher.

The remainder of this paper is organized as follows: in Section II, we briefly introduce teacher-student distillation framework and our multiple teacher models. In Section III, different training strategies including interpolated-training, augmented-training and random augmented-training strategy are described. Experimental results and analyses are provided in Section IV. Finally, Section V concludes the paper.

II. KNOWLEDGE DISTILLATION IN MULTILINGUAL SPEECH RECOGNITION

A. Knowledge distillation

In knowledge distillation, firstly train a teacher model to get the frame-level output distribution. Then train a student model using criteria that minimize the distance between the probability distribution of the teacher and the student [7-18]. The distance is so-called soft loss, defined as below:

$$L_{KD} = -\sum_{t=1}^{T}\sum_{k=1}^{\boldsymbol{\pi}} q(y_t = k|\boldsymbol{Y_{t-1}}, \boldsymbol{X}) log p(y_t = k|\boldsymbol{Y_{t-1}}, \boldsymbol{X}).$$

Where $\mathbf{X}$ is the input vector, y_t is output label in time step t, and $\boldsymbol{\pi}$ is a set of possible classes. $q(y_t = k|\boldsymbol{Y_{t-1}}, \boldsymbol{X})$ is the probability distribution of teacher model which works as soft label and $p(y_t = k|\boldsymbol{Y_{t-1}}, \boldsymbol{X})$ is the probability distribution of student model. Soft label is no longer a one-hot vector, instead, the competing classes have small but nonzero posterior probabilities for each training example [17]. Hinton [7] suggested that the small posterior probabilities are valuable information that encodes correlations among different classes which makes soft label superior to original hard label.

B. Multiple teachers in multilingual speech recognition

Various works have been proposed in multiple teacher models where multiple teachers offer multiple streams of information to help student to observe various "view" of the data. In our multiple teacher models, multiple individual models serve as monolingual teachers, each handling a special language, while the multilingual teacher handles all the languages in a single model. We use original hard label and soft labels from multiple teachers to help multilingual student to learn various knowledge embedded in the data and intend to outperform multilingual teacher.

III. ENSEMBLES OF MULTIPLE TEACHERS

A. Interpolated-training strategy

Algorithm 1 interpolated-training strategy (IP)

for all mini-batches in training data **do**
  pick mini-batch i ;
  use multilingual teacher model and corresponding monolingual teacher models to obtain soft losses ;
  combine soft losses and hard loss with preassigned weights for each loss;
  update neural network model with mini-batch i;
end for

Interpolated-training strategy (Algorithm 1) is the conventional method to use multiple teachers, which relies on a weighted average loss, making student model not directly optimized for each loss and weakens complementariness obtained by multiple models due to averaging them. Considering that dissimilarities between knowledge fields of teachers should be more explicitly maintained/leveraged to make student contain various characteristics, augmented-training strategy [10] was proposed to explicitly use multiple teachers for more effective parameter updating.

B. Augmented-training strategy

Augmented-training strategy was proposed to update the network by each loss sequentially per mini-batches, and thus augment the training data by creating multiple copies of the original data that reflect knowledge field of each teacher. Besides, augmented-training strategy simplifies the manual work with no need to tune the weight of each loss by grid search. The augmented-training strategy is depicted as below:

Algorithm 2 augmented-training strategy (AU)

set a fixed update order of each loss (soft losses and hard loss)
for all mini-batches in training data **do**
  pick mini-batch i ;
  use multilingual teacher model and corresponding monolingual teacher models to obtain soft losses ;
  update neural network model with mini-batch i sequentially according to the order;
end for

C. Random augmented-training strategy

We observe that the performance of student model depends largely on the last update loss in augmented-training strategy, while other losses are not as important as the last one. As soft labels are superior to hard label, the last loss should be soft to achieve better model performance. When trapped in the local minimum of soft loss, the model can escape the local minimum if we set the last loss from soft to hard (discussed in IV-C).

Therefore, we propose a random augmented-training strategy that does not always fix the last update loss. Different from augmented-training strategy where the update order of each loss is fixed, in random augmented-training strategy, firstly set two fixed update orders that the last losses are soft and hard separately, then randomly select the update order by sampling from a Bernoulli distribution per mini-batches. Considering that soft loss should mainly guide the model to converge, we set a higher sampling probability for the order that the last loss is soft in our experiments. The random augmented-training strategy is depicted as below:

Algorithm 3 random augmented-training strategy (RAU)

set two fixed update orders that the last losses are soft and hard separately
for all mini-batches in training data **do**
  pick mini-batch i;
  use multilingual teacher model and corresponding monolingual teacher models to obtain soft losses ;
  randomly select the order in pool of orders by sampling from a Bernoulli distribution ;
  update neural network model with mini-batch i sequentially according to the order;
end for

IV. EXPERIMENT

A. Data

We conduct our experiments on data from Wall Street Journal (English), AISHELL-1(Chinese). AISHELL-1 and WSJ have 150 and 80 hours in training data respectively. We train two monolingual teacher models independently on data for each language. As with the multilingual training set, the classes of the multilingual teacher model are also a union of language-specific classes. Because of limited computing power and strict time budget, we randomly select 20k sentences in each language as the subset of the whole multilingual training data, yielding totally 66h speech (the training data of English is about 42h and Chinese is about 24h). Multilingual student models trained with different teachers and training strategies are evaluated on the subset to sufficiently explore the best teachers and training strategy. Afterwards, trained with the best configuration which integrates proposed multiple teachers and random augmented-training strategy, our student model is verified on the whole dataset and show that our work successfully makes student outperform multilingual teacher under the same training data and architecture.

B. Model configuration

The input feature is 40 mel-scale filter bank features together with the energy in each frame, and first and second temporal differences. We tune the hyperparameters on English teacher model and reuse the optimal configuration to the remaining models including Chinese teacher model, multilingual teacher model and multilingual student models. The best configuration for WSJ is a 4 layer encoder comprised of 256 biGRU cells (i.e. 256 cells in forward layer and 256 cells in backward layer) with 1/2 downsampling in the last two layers which reduces the time resolution by $4 = 2^2$, and a 1 layer decoder containing 256 GRU cells. Gradient norm clipping is set to 1, together with Gaussian weight noise with $N(0,0.075)$ and L2 weight decay with 1e-5. We decay the learning rate from 1e-3 to 1e-4 when no improvement is found on validation set. Label smoothing [19] is applied with correct class probability set to 0.9.

C. Results

Firstly, Table I lists the results of multilingual teacher model, two monolingual teacher models on the whole language data and multilingual student model without knowledge distillation (which is trained the same as multilingual teacher model but on the subset). "Multi" means multilingual teacher, "Mono" means monolingual teacher. The performance of the student model is far worse than the teachers' because of less training data. We also list prior work of attention-based end-to-end model with similar architectures to convince our baseline.

Secondly, we report the results of the student models with knowledge distillation from a single teacher (only one teacher for each language, different from multiple teachers which offer multiple teachers for each language) using interpolated-training strategy and augmented-training strategy in Table II. "Baseline" is the student model without knowledge distillation (same as "Multi-student" in Table I). "IP" means interpolated-training strategy (the best preassigned weights for soft and hard loss is 0.8 and 0.2). "AU" means augmented-training strategy. As is shown in Table II, knowledge distillation improves the performance impressively, and augmented-training strategy with the update order from hard to soft performs better than interpolated-training strategy. We can initially conclude that the performance of student depends largely on the last update loss. More experimental demonstrations can be found in Table III and Fig. 1 under multiple teachers condition, where student model is trained with six possible update orders of losses respectively using augmented-training strategy.

TABLE I
The results of the teacher models and student model.

Model	EN(WER)	CHN(CER)
Multi-teacher	14.61	12.12
Mono-teacher-EN	11.84	-
Mono-teacher-CHN	-	11
Multi-student	18.55	32.94
Seq2seq-WSJ[19]	14.76	-
Seq2seq-WSJ[20]	12.9	-
Seq2seq-AISHELL[21]	-	19.8
Seq2seq-AISHELL[22]	-	10.56

TABLE II
The results of the student models with knowledge distillation from single teacher using interpolated-training strategy and augmented-training strategy.

Model	EN(WER)	CHN(CER)
Baseline	18.55	32.94
Multi-IP	16.44	28.21
Multi-AU(from soft to hard)	17.40	27.68
Multi-AU(from hard to soft)	**16.15**	**24.87**
Mono-IP	16.57	27.31
Mono-AU(from soft to hard)	16.78	29.37
Mono-AU(from hard to soft)	**15.64**	**25.70**

TABLE III
The results of student models with six possible update orders of losses in multiple teacher models using augmented-training strategy.

Order	EN(WER)	CHN(CER)
mono → multi → hard	19.01	31.55
multi → mono → hard	18.23	32.35
mono → hard → multi	15.70	25.93
hard → mono → multi	15.13	23.10
multi → hard → mono	**15.12**	**22.27**
hard → multi → mono	17.03	25.65

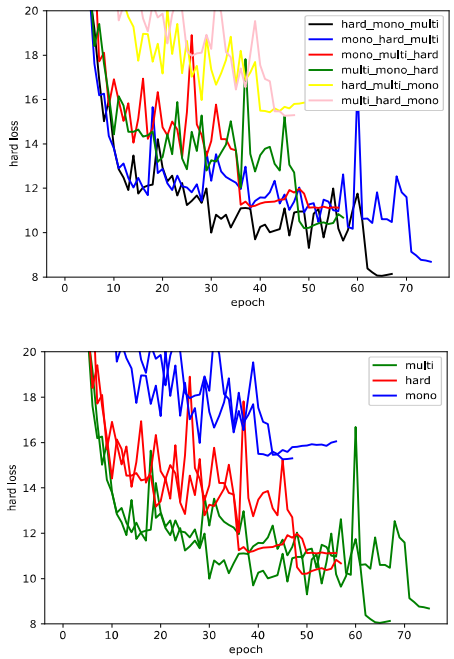

Fig. 1 The hard losses of student models with six possible update orders of losses in multiple teacher models using augmented-training strategy per epoch. The figure below is the same as the figure above only with the difference by setting the same color for the same last loss.

In Table III and Fig.1, "mono" represents soft loss from corresponding monolingual teacher (two monolingual teachers, one monolingual teacher for one language), "multi" represents soft loss from multilingual teacher, "hard" represent hard loss. We can observe that the error rates (in Table III) and hard losses (in Fig. 1) of the student models are similar when the last update losses are same, which indicates that the performance of student depends largely on the last update loss. And the best order of augmented-training strategy in multiple teacher models is "multi →hard → mono", which will be our AU baselines in following experiments.

Thirdly, comparing the best results obtained by knowledge distillation from single teacher in Table II and from multiple teachers in Table III, the WERs down from 15.64% to 15.12% in EN and the CERs down from 25.70% to 22.27% in CHN suggest that our multiple teacher models bring a significant reduction in error rate.

Fourthly, we compare the performance of students using random augmented-training strategy (RAU) with students using augmented-training strategy under single teacher model (in Table IV) and multiple teacher model (in Table V). For single teacher model, RAU works with sampling probability 0.2 for update order from mono to hard and 0.8 for order from hard to mono. For multiple teacher models, the sampling probability is 0.2 for update order from multi to mono to hard and 0.8 for order from multi to hard to mono. We can see that RAU gains a considerable improvement on the performance of the student model relative to AU, which will be demonstrated to help the model to escape the local minimum when trapped in soft loss, as discussed as following.

TABLE IV
Comparing the performance of random augmented-training strategy with augmented-training strategy under single teacher model.

Training-strategy	EN(WER)	CHN(CER)
AU(from hard to mono)	15.64	25.70
RAU	**14.55**	**22.68**

TABLE V
Comparing the performance of random augmented-training strategy with augmented-training strategy under multiple teacher models.

Training-strategy	EN(WER)	CHN(CER)
AU(from multi to hard to mono)	15.12	22.27
RAU	**14.09**	**18.55**

In order to study why RAU works better than AU, we firstly conduct the following experiment: The student model is trained with AU from beginning, and AU is replaced with RAU for next following epochs when no improvement is found for AU training strategy. After a lower error rate is reached, the training strategy switches back to AU. As is shown in Fig. 2, when using augmented-training strategy and converged, the soft loss (the red line in Fig. 2) of the model escapes the local minimum soon after trained the epoch with RAU (the epoch indicated by the green line in Fig. 2). To further demonstrate this phenomenon, we offer experimental results of two variants of RAU in Table VI: "Variant 1" is the strategy that 1 mini-batch is trained with the order from mono to hard (last loss of this mini-batch is hard loss) after every 4 mini-batches have been trained with the order from hard to mono (last losses of these 4 mini-batches are soft loss); and "Variant 2" is the strategy that 2 mini-batches are trained with the order from mono to hard after every 8 mini-batches are trained with the order from hard to mono. Similar to RAU, the last loss of a mini-batch in these two variant strategies could be

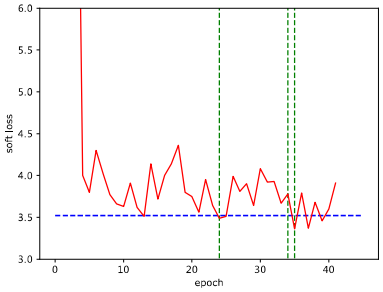

Fig. 2 The soft loss of the model trained by AU and trained by RAU when the model converges. The red line is the epochs trained by AU and the green line points the epochs trained by RAU. The blue line is the lowest loss before the model is trained by RAU.

TABLE VI
The performance of two variants of RAU.

Training-strategy	EN(WER)	CHN(CER)
Variant 1	17.46	26.45
Variant 2	14.61	23.72

Table VII
Comparing the performance of our best student with multilingual model on the same data and architecture.

model	EN(WER)	CHN(CER)
Multilingual-model (subset)	18.55	32.94
Best student(subset)	14.09	18.55
Multilingual-model(whole)	14.61	12.12
Best student(whole)	**11.98**	**10.83**

hard or soft loss, which is not fixed and the occurrence frequencies of hard and soft last update loss are nearly the same with RAU. Variant 1 performs worse than Variant 2 as that it is probably not enough to learn from hard loss only once when trapped in soft loss (Please note: only one hard loss after every 4 soft losses in Variant 1 vs. 2 hard loss after every 8 soft losses in Variant 2), otherwise their performance should be similar. In summary, Fig. 2 and Table VI indicate that RAU may help the model to escape the local minimum by learning from hard loss when trapped in soft loss.

Finally, we apply the whole multilingual language data on the best student which integrates the proposed multiple teachers and random augmented-training strategy, achieving the best WER of 14.09% in EN and the best CER of 18.55% in CHN on the subset in Table V. We compare the performance of multilingual model with our best student on the subset and the whole language data in Table VII. The multilingual models on subset and the whole language data are the baselines same as Multi-student and Multi-teacher in Table I. Table VII indicates that the best student integrating the proposed multiple teachers and random augmented-training strategy successfully reduces the error rate relative to multilingual model under the same language data and architecture. As multilingual teacher and multilingual student are a single model with no need to know language identity during inference, which differ from monolingual teachers and ensembles of multiple teachers in beam search during inference, we only compare the multilingual student with the multilingual teacher.

V. CONCLUSION

In this paper we propose multiple teacher models where the complementary knowledge fields of expertise of multilingual and monolingual teachers allow multilingual student to learn various knowledge embedded in the data. Besides, observing that the performance of student depends largely on the last update loss in augmented-training strategy, we propose the random augmented-training strategy that does not always fix the last update loss, which has been demonstrated to help the model to escape the local minimum. Comparison with multilingual teacher shows the proposed multiple teachers and distillation strategy improve the performance of the model significantly. In future work, we will apply the whole multilingual language data on AU training strategy to give a more obvious and concise contrast. What's more, the experiments on a larger dataset like AISHELL-2 (Chinese) and Librispeech (English) will be conducted to further study the upper limit of our method. Our work suggests that multilingual model can be better optimized with better knowledge distillation techniques. Therefore exploring more efficient distillation strategies will also be our future work.

VI. ACKNOWLEDGEMENT

This work was supported by National Key R&D Program of China (Grant No.2017YFB1002202) and the Key Science and Technology Project of Anhui Province (Grant No. 18030901016).

VII. REFERENCE

[1] S. Dalmia, R. Sanabria, F. Metze, and A. W Black, "Sequence-based multi-lingual low resource speech recognition," in *Proc. ICASSP*, 2018, pp. 4909–4913.

[2] S. Toshniwal, T. Sainath, R. Weiss, B. Li, P. Moreno, et al. "Multilingual speech recognition with a single end-to-end model," in *Proc. ICASSP*, 2018, pp. 4904-4908.

[3] T. Schultz and A.Waibel, "Fast bootstrapping of lvcsr systems with multilingual phoneme sets," in *Proc. Eurospeech*, 1997, pp. 371-374.

[4] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, 2001.

[5] T. Niesler, "Language-dependent state clustering for multilingual acoustic modelling," *Speech Communication*, vol. 49, no. 6, 2007.

[6] H. Seki, S. Watanabe, T. Hori, J. Le Roux, et al, "An end-to-end language-tracking speech recognizer for mixed-language speech," in *Proc. ICASSP*, 2018, pp. 4919–4923.

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv: 1503.02531*.

[8] M. Huang, Y. You, Z. Chen, Y. Qian, and K. Yu, "Knowledge distillation for sequence model," in *Proc. Interspeech*, 2018, pp. 3703-3707.

[9] T. Furlanello, Z. Lipton, C. Tschannen, M. Itti, and A. Anandkumar, "Born again neural networks," *arXiv preprint arXiv: 1805.04770*.

[10] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, et al, "Efficient knowledge distillation from an ensemble of teachers," in *Proc. Interspeech*, 2017, pp. 3697–3701.

[11] T. Moriya , S. Ueno , Y. Shinohara , M. Delcroix , Y. Yamaguchi , et al, "Multi-task learning with augmentation strategy for acoustic-to-word attention-based encoder-decoder speech recognition", in *Proc, Interspeech*, 2018, pp. 2399-2403.

[12] J. Cui, B. Kingsbury, B. Ramabhadran, G. Saon, T. Sercu, et al, "Knowledge distillation across ensembles of multilingual models for low-resource languages," in *Proc. ICASSP*, 2017, pp. 4825–4829.

[13] T.Xu, R.Yi, H.Di, Q.Tao, Z.Zhou and L.Tie-Yan, "Multilingual Neural Machine Translation with knowledge distillation". In Proc, ICLR, 2019.
[14] Y. Chebotar and A.Waters, "Distilling knowledge from ensembles of neural networks for speech recognition," in *Proc. Interspeech*, 2016, pp. 3439-3443.
[15] K. Markov and T. Matsui, "Robust speech recognition using generalized distillation framework," in *Proc. Interspeech*, 2016, pp. 2364–2368.
[16] W. Chan, N. R. Ke, and I. Lane, "Transferring knowledge from a RNN to a DNN," in *Proc. Interspeech*, 2015, pp. 3264–3268.
[17] L. Lu, M. Guo, and S. Renals, "Knowledge distillation for smallfootprint highway networks," in *Proc. ICASSP*, 2017, pp. 4820–4824.
[18] C.Szegedy, V.Vanhoucke, and S.Ioffe, "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, 2016, pp. 2818-2826.
[19] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *Proc. ICASSP*, 2017, pp. 4845-4849.
[20] W. Chan, Y. Zhang, Q. Le, and N. Jaitly, "Latent sequence decompositions," *arXiv preprint arXiv: 1610.03035*.
[21] M. Li and M. Liu. "End-to-end speech recognition with adaptive computation steps," in *Proc, ICASSP*, 2019, pp. 6246-6250.
[22] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, et al, "Component fusion: learning replaceable language model component for end-to-end speech recognition system," in *Proc. ICASSP*, 2019, pp. 5361-5635.