AN INVESTIGATION OF A KNOWLEDGE DISTILLATION METHOD FOR CTC ACOUSTIC MODELS

Ryoichi Takashima, Sheng Li and Hisashi Kawai

National Institute of Information and Communications Technology, Japan

ryoichi.takashima@nict.go.jp

ABSTRACT

End-to-end acoustic models, such as connectionist temporal classification (CTC) and the attention model, have been studied, and their speech recognition accuracies come close to those of conventional deep neural network (DNN)-hidden Markov models. However, most high-performance end-to-end models are not suitable for real-time (streaming) speech recognition because they are based on bidirectional recurrent neural networks (RNNs). In this study, to improve the performance of unidirectional RNN-based CTC, which is suitable for real-time processing, we investigate the knowledge distillation (KD)-based model compression method for training a CTC acoustic model. we evaluate a frame-level KD method and a sequence-level KD method for CTC model. The speech recognition experiments on Wall Street Journal tasks demonstrate that, the frame-level KD worsens the WERs of unidirectional CTC model, whereas sequence-level KD can improve the WERs of the model.

***Index Terms*—** Speech recognition, acoustic model, connectionist temporal classification, knowledge distillation, long short-term memory

1. INTRODUCTION

End-to-end acoustic models, such as connectionist temporal classification (CTC) [1, 2] and the attention model [3, 4] have been studied for automatic speech recognition (ASR) [5, 6, 7, 8, 9, 10, 11, 12, 13]. One advantage of the end-to-end model is its simplicity in the decoding process because they do not use the hidden Markov model (HMM). Miao et al. [7] proposed an end-to-end speech recognition framework that applied a CTC acoustic model to a weighted finite-state transducer (WFST) [14], and demonstrated a decoding speed that was over three times faster than a conventional deep neural network (DNN)-HMM-based WFST. The improvement of recognition accuracy has also been studied, and the state-of-the-art model outperforms a DNN-HMM baseline model in recognition accuracy without using any language models [13]. However, most high-performance end-to-end models are not suitable for real-time (streaming) ASR, despite their simple decoding process, because they are based on bidirectional recurrent neural networks (RNNs), which require a whole utterance to predict a frame output and thus cause high latency.

Unidirectional RNN-based CTC is one of the simplest implementations of the real-time end-to-end ASR; however, its recognition accuracy is worse than those of bidirectional high-performance models. There are some studies to improve that unidirectional model [15, 16]. In [15], they proposed a low-latency sequence-to-sequence model that can recognize a speech for every block of frames. In [16], they used RNN transducer (RNN-T) [17], which is an extension to the CTC, and showed promising results with their optimized RNN-T. In this paper, we use an easily implementable unidirectional CTC without modifying the model, and attempt to improve its performance by using the approach of knowledge distillation (KD).

KD [18, 19] is a model compression method for DNNs, and often used to bridge the gap of performance between a smaller model and a larger model. The KD method trains a smaller model (called *student model*) using the output of a larger model (called *teacher model*) as training labels so as to transfer the knowledge of the teacher model to the student model. The effectiveness of KD has been confirmed in speech recognition tasks [20, 21, 22, 23]. Fukuda et al. [23] showed that KD can improve the word error rate (WER) of a student convolutional neural network (CNN) using a VGG network [24] and a long short-term memory (LSTM) network [25] as teacher models. An interesting point of this work is that KD successfully transferred the sequential information of an LSTM to a CNN without recurrent structures. From that observation, we expect that KD can also transfer the knowledge of a bidirectional LSTM (bi-LSTM) to a unidirectional LSTM (uni-LSTM).

In previous studies, KD has been mainly applied through frame-level cross-entropy (CE) training, that is, for training CE-DNN-HMMs (we call it *frame-level KD*). There are also studies of KD technique for sequence training, such as attention model [26] and maximum mutual information (MMI)-based training [27] (we call it *sequence-level KD*). The frame-level KD has been applied to a CTC acoustic model in [28]; however, the performance degraded compared with the CTC model directly trained using correct labels. The sequence-

level KD has not been applied to the CTC model to the best of our knowledge. In this paper, we explore efficient methods to apply the KD technique to a CTC acoustic model by comparing the frame-level KD and the sequence-level KD. We use a uni-LSTM-based CTC model (uni-LSTM-CTC) as a student model and a bi-LSTM-based CTC model (bi-LSTM-CTC) as a teacher model. In the frame-level KD framework, similar to the previous work of [28], we train a student uni-LSTM under the CE criteria using frame-level outputs of the teacher CTC model. In the sequence-level KD framework, following the method proposed in [26], we extract the hypotheses of the label sequence and their posterior probability estimated by the teacher CTC model, and a student CTC model is trained using the hypotheses under sequence-level CE criteria. We evaluate these KD methods on Wall Street Journal (WSJ) large vocabulary continuous speech recognition (LVCSR) tasks.

2. CONNECTIONIST TEMPORAL CLASSIFICATION

For general speech recognition, we need to map a sequence of label estimated for each frame (called 'path' and denoted as π) into a label sequence (denoted as $\mathbf{l}$) of length equal to or less than the number of frames. In the CTC framework [1], a path is converted into a label sequence by introducing the deletion of repeated labels and insertion of blank labels (i.e., "no label"). We call this conversion "CTC mapping" with function $\mathcal{B}$, where $\mathbf{l} = \mathcal{B}(\pi)$. Because there are multiple possible paths mapped into an identical label sequence, the conditional probability of the label sequence $\mathbf{l}$ given the input sequence $\mathbf{x}$ is defined as the sum of the probabilities of all possible corresponding paths:

$$p(\mathbf{l}|\mathbf{x}) = \sum_{\pi \in \mathcal{B}^{-1}(\mathbf{l})} p(\pi|\mathbf{x}). \tag{1}$$

The conditional probability of path π is calculated as follows:

$$p(\pi|\mathbf{x}) = \prod_{t=1}^{T_{\mathbf{x}}} p(k = \pi_t|\mathbf{x}). \tag{2}$$

Here, the posterior probability $p(k = \pi_t|\mathbf{x})$ of label π_t of t-th frame given input sequence $\mathbf{x}$ is modeled with an RNN. $T_{\mathbf{x}}$ denotes the number of frames. The CTC model is trained by maximizing the likelihood, that is, minimizing the loss function $\mathcal{L}_{\mathrm{CTC}}$ defined as

$$\mathcal{L}_{\mathrm{CTC}} = -\sum_{(\mathbf{x},\mathbf{l}) \in Z} \ln p(\mathbf{l}|\mathbf{x}) = \sum_{(\mathbf{x},\mathbf{l}) \in Z} \mathcal{F}_{\mathrm{CTC}}(\mathbf{l}|\mathbf{x}), \tag{3}$$

where Z denotes the training dataset, and $\mathcal{F}_{\mathrm{CTC}}(\mathbf{l}|\mathbf{x}) = -\ln p(\mathbf{l}|\mathbf{x})$ is the local loss defined for explanation in Section 4.2. The local loss $\mathcal{F}_{\mathrm{CTC}}(\mathbf{l}|\mathbf{x})$ is efficiently computed using the forward-backward algorithm.

3. KNOWLEDGE DISTILLATION

KD [19] is a model compression method for DNNs. The main idea of KD is to train a smaller student model using the output of a larger teacher model as training labels (often called soft labels) such that the student model works like the teacher model. In the KD framework, first, a teacher model is trained using the correct label. Then, the student model is trained using the outputs of the teacher model that corresponds to training data under the CE criteria as follows:

$$\mathcal{L}_{\mathrm{KD}} = -\sum_{l} p_{\mathrm{tea}}(l|x) \ln p_{\mathrm{stu}}(l|x), \tag{4}$$

where $p_{\mathrm{tea}}(l|x)$ denotes the posterior probability of label l given input x estimated by the teacher model (i. e. the output of the teacher model) and $p_{\mathrm{stu}}(l|x)$ is that estimated by the student model.

Because the CE of the original KD is defined as training samples (i.e., frames in the ASR task) are independent, we call the original KD frame-level KD. There are also studies of KD technique for sequence training, such as attention model [26] and MMI-based training [27]. In these methods, the student model is trained by minimizing the CE between the probability distributions of the label sequences on a teacher model and a student model. We call this KD approach sequence-level KD.

4. KD FOR THE CTC ACOUSTIC MODEL

Before training a student CTC (uni-LSTM-CTC) model, we train a teacher CTC (bi-LSTM-CTC) model under the conventional CTC training criteria (i.e. Eq. (3)) on the correct label sequence. Then, using the output of the teacher CTC that corresponds to the training data, we train the student CTC model in frame-level or sequence-level KD frameworks.

4.1. Frame-level KD

The frame-level KD for the CTC model is the same framework used in previous works based on the original KD technique [20, 21, 22, 23], except that, for the CTC model, the posterior probability of the blank label is also considered. Figure 1 shows the overview. We extract the output of the teacher CTC model given the training data for each frame. Then, using the teacher's output as soft labels, we train the neural network (i.e., uni-LSTM in this work) of the student CTC model under the CE criteria:

$$\mathcal{L}_{\mathrm{CTC-KD_{frame}}} = -\sum_{\mathbf{x} \in Z} \sum_{t=1}^{T_{\mathbf{x}}} \sum_{k \in K} p_{\mathrm{tea}}(k|x_t) \ln p_{\mathrm{stu}}(k|x_t). \tag{5}$$

x_t denotes the t-th frame sample in an input sequence $\mathbf{x}$ of length $T_{\mathbf{x}}$. k denotes a label in the CTC label set $K = L \cup \{blank\}$, where L denotes the original label set. $p_{\mathrm{tea}}(k|x_t)$

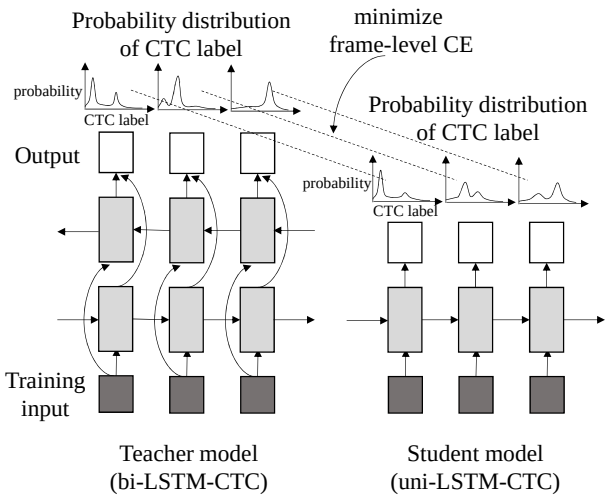

Fig. 1. Overview of the frame-level KD for the CTC acoustic model.

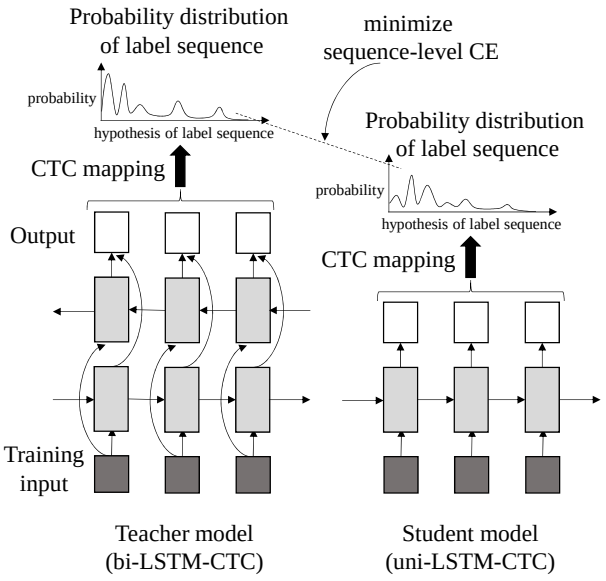

Fig. 2. Overview of the sequence-level KD for the CTC acoustic model.

is the posterior probability of label k estimated by the teacher CTC model, and $p_{\mathrm{stu}}(k|x_t)$ is the posterior probability estimated by the neural network of the student CTC model.

The frame-level KD has been applied to the CTC model in [28]. In that report, however, the performance degraded compared with the CTC model directly trained using correct labels. We guess the reason is that; as reported in previous studies [1, 7], the probability of the blank label estimated by the CTC model is nearly 1.0 in the majority of frames. Therefore, when we use the frame-level criteria, only few frames that have a higher probability of non-blank labels might be used effectively for training the student model.

4.2. Sequence-level KD

Figure 2 shows the overview of the sequence-level KD for the CTC model. The main approach is the same as that proposed in [26], and we apply that approach to the CTC's training framework. We extract the hypotheses of the label sequence and their posterior probabilities estimated by the teacher CTC model. Then, using the hypotheses and posterior probabilities, we train a student CTC model under sequence-level CE criteria:

$$\mathcal{L}_{\mathrm{CTC-KD_{seq}}} = -\sum_{\mathbf{x}\in Z}\sum_{\mathbf{h}\in\mathcal{H}} p_{\mathrm{tea}}(\mathbf{h}|\mathbf{x}) \ln p_{\mathrm{stu}}(\mathbf{h}|\mathbf{x}). \quad (6)$$

$\mathbf{h}$ denotes a hypothesis of the label sequence in the set of all possible hypotheses $\mathcal{H}$. $p_{\mathrm{tea}}(\mathbf{h}|\mathbf{x})$ and $p_{\mathrm{stu}}(\mathbf{h}|\mathbf{x})$ are the posterior probabilities of hypothesis $\mathbf{h}$ estimated by the teacher CTC model and the student CTC model, respectively. Since Eq. (6) can be expressed as

$$\mathcal{L}_{\mathrm{CTC-KD_{seq}}} = \sum_{\mathbf{x}\in Z}\sum_{\mathbf{h}\in\mathcal{H}} p_{\mathrm{tea}}(\mathbf{h}|\mathbf{x}) \mathcal{F}_{\mathrm{CTC}}(\mathbf{h}|\mathbf{x}), \quad (7)$$

the sequence-level KD criteria for CTC model can be summarized as the weighted mean of the original CTC loss regarding each hypothesis of the label sequence.

Since extracting $p_{\mathrm{tea}}(\mathbf{h}|\mathbf{x})$ for all possible hypotheses is unrealistic, we approximate them using N-best hypotheses as follows:

$$\tilde{\mathcal{L}}_{\mathrm{CTC-KD_{seq}}} = \sum_{\mathbf{x}\in Z}\sum_{n=1}^{N} \tilde{p}_{\mathrm{tea}}(\mathbf{h}_n|\mathbf{x}) \mathcal{F}_{\mathrm{CTC}}(\mathbf{h}_n|\mathbf{x}), \quad (8)$$

where $\mathbf{h}_n$ denotes the n-th hypothesis in the N-best hypotheses, and $\tilde{p}_{\mathrm{tea}}(\mathbf{h}_n|\mathbf{x})$ denotes the posterior probability approximated as

$$\tilde{p}_{\mathrm{tea}}(\mathbf{h}_n|\mathbf{x}) = \frac{p_{\mathrm{tea}}(\mathbf{h}_n|\mathbf{x})}{\sum_{n=1}^{N} p_{\mathrm{tea}}(\mathbf{h}_n|\mathbf{x})}. \quad (9)$$

In our experiments, we used ($N = 10$)-best hypotheses.

Depending on the approximate precision, the use of Eq. (8) has a risk that it trains a student model with inadequate labels. In an extreme case, when the Eq. (8) is used on $N = 1$, the student model will be trained using hard labels, which might include incorrect labels. To avoid this problem, we use the interpolation between Eq. (8) and the original CTC loss function using correct labels as follows:

$$\tilde{\mathcal{L}}'_{\mathrm{CTC-KD_{seq}}} = (1-q)\mathcal{L}_{\mathrm{CTC}} + q\tilde{\mathcal{L}}_{\mathrm{CTC-KD_{seq}}}, \quad (10)$$

where $q \in (0, 1]$ is a tunable parameter.

Table 1. WERs [%] on "eval92" (trained on "train_si84").

Acoustic Model (train_si84, 15hours)	q	WER
bi-LSTM-CTC	-	10.35
uni-LSTM-CTC	-	11.77
uni-LSTM-CTC with frame-level KD	-	16.04
uni-LSTM-CTC	1.0	11.54
with sequence-level KD (10-best)	0.9	11.25
	0.8	11.06
	0.7	**10.83**
	0.6	11.20

Table 2. WERs [%] on "eval92" (trained on "train_si284").

Acoustic Model (train_si284, 81hours)	q	WER
bi-LSTM-CTC	-	8.70
uni-LSTM-CTC	-	10.37
uni-LSTM-CTC with frame-level KD	-	12.71
uni-LSTM-CTC		
with sequence-level KD (10-best)	0.7	**9.57**

5. EXPERIMENTS

5.1. Experimental conditions

We evaluated the KD approaches for the CTC acoustic model on WSJ LVCSR tasks. The experiments were conducted on two training datasets: (1) only "WSJ0 (LDC93S6B) [29]" (15 hours, known as "train_si84" in the Kaldi recipe [30]); and (2) "WSJ0" and "WSJ1 (LDC94S13B) [31]" (81 hours, known as "train_si284" in the Kaldi recipe). For both experiments, we used a dataset called "eval92" [32] for evaluation.

We extracted 40-dimensional mel-filterbank features with their first and second-order derivatives (FBANK+Δ+$\Delta\Delta$, 120 dimensions in total) as acoustic features, and the target labels were defined to include 69 phonemes, two noise marks, and a blank (72 labels in total). The teacher model and the student model were a bi-LSTM-CTC and a uni-LSTM-CTC, respectively. Both networks had three hidden layers and 512 memory cells in each hidden layer. We used the CNTK toolkit [33] to train models and optimized the model parameters using Adam algorithm [34] with an initial learning rate of 0.0001.

For decoding, we used EESEN software [7], which integrates the CTC acoustic model, lexicon, and language model on the WFST framework. We used the CMU dictionary as the lexicon and the 20,000-word vocabulary WSJ pruned language model, known as "lm_tgpr" in the Kaldi recipe, as the language model.

In the sequence-level KD method, we also used EESEN software to extract the ($N = 10$)-best hypotheses. The 10-best hypotheses were extracted using WFST-based beam-search. In this process, we used only the token WFST (defined as T.fst in EESEN software), which maps a sequence of frame-level CTC labels to a single lexicon unit (i.e., a phoneme in this experiment). Note that we did not use dictionary and language model in this process.

5.2. Results

Table 1 shows the WERs on "train_si84" training dataset. As shown in this table, the WER of uni-LSTM-CTC without KD was 1.42% higher than that of bi-LSTM-CTC. When we applied the frame-level KD to the student uni-LSTM-CTC, the performance worsened further as reported in [28].[1] When we applied the sequence-level KD, the WERs of student uni-LSTM-CTC were improved. We observed the best performance when the interpolating parameter q (See Eq. (10)) was set to 0.7, and the WER of the student uni-LSTM-CTC was improved by 0.94%. That means the sequence-level KD reduced the performance difference between the student model and the teacher bi-LSTM-CTC by 66.2% relatively. In the following experiment, we fixed $q = 0.7$.[2]

Table 2 shows the WERs on "train_si284" training dataset. The tendency of the results was similar to that on "train_si84". The WER of uni-LSTM-CTC without KD was 1.67% higher than that of bi-LSTM-CTC, and the frame-level KD worsened the performance further. By applying the sequence-level KD, the WER of the student uni-LSTM-CTC were improved by 0.80%, and that means the performance difference between the student model and the teacher model was reduced by 47.9% relatively.

6. CONCLUSION

To improve the recognition performance of the unidirectional CTC acoustic model for real-time end-to-end speech recognition, we explored KD methods for training CTC models. We considered the application of both frame-level KD and sequence-level KD. Our experiments on the LVCSR tasks using WSJ dataset show that, the frame-level KD worsened the WERs of the student uni-LSTM-CTC model, whereas sequence-level KD improved the WERs of the student model.

We used the 10-best hypotheses because of the limitation of computational resources. In future work, to analyze the relationship between the number of hypotheses and the performance, we will investigate the method to efficiently perform sequence-level KD with more hypotheses. Additionally, the recognition accuracy of a student unidirectional CTC model might be improved further using higher-performance teacher models. Therefore, we will evaluate the performance of the KD on various teacher models.

[1] We also tried using modified criteria, that is, a naive interpolation between the original CTC loss function (Eq. (3)) and the frame-level KD (Eq. (5)), but the training failed to converge.

[2] We used 5% of training data as a development set, and we also observed the best performance on the development set when we set $q = 0.7$.

7. REFERENCES

[1] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML2006*. ACM, 2006, pp. 369–376.

[2] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks.," in *ICML2014*, 2014, vol. 14, pp. 1764–1772.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR2015*. IEEE, 2015.

[4] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition," in *ICASSP2016*. IEEE, 2016, pp. 4945–4949.

[5] Haşim Sak, Andrew Senior, Kanishka Rao, and Françoise Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," in *Interspeech*. ISCA, 2015, pp. 1468–1472.

[6] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *ICML2016*, 2016, pp. 173–182.

[7] Yajie Miao, Mohammad Gowayyed, and Florian Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *ASRU2015*. IEEE, 2015, pp. 167–174.

[8] Naoyuki Kanda, Xugang Lu, and Hisashi Kawai, "Maximum-a-posteriori-based decoding for end-to-end acoustic models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1023–1034, 2017.

[9] Naoyuki Kanda, Xugang Lu, and Hisashi Kawai, "Mininum bayes risk training of ctc acoustic models in maximum a posteriori based decoding framework," in *ICASSP2017*. IEEE, 2017, pp. 4855–4859.

[10] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *ICASSP2017*. IEEE, 2017, pp. 4835–4839.

[11] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP2016*. IEEE, 2016, pp. 4960–4964.

[12] Yu Zhang, William Chan, and Navdeep Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *ICASSP2017*. IEEE, 2017, pp. 4845–4849.

[13] Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," in *Interspeech*. ISCA, 2017, pp. 949–953.

[14] Mehryar Mohri, Fernando Pereira, and Michael Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.

[15] Navdeep Jaitly, David Sussillo, Quoc V. Le, Oriol Vinyals, Ilya Sutskever, and Samy Bengio, "An online sequence-to-sequence model using partial conditioning," in *NIPS2016*, pp. 5067–5075. 2016.

[16] Kanishka Rao, Haşim Sak, and Rohit Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer," in *ASRU2017*. IEEE, 2017, pp. 193–199.

[17] Alex Graves, "Sequence transduction with recurrent neural networks," in *ICML 2012 Workshop on Representation Learning*, 2012.

[18] Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong, "Learning small-size dnn with output-distribution-based criteria," in *Interspeech*, 2014, pp. 1911–1914.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2014.

[20] Yevgen Chebotar and Austin Waters, "Distilling knowledge from ensembles of neural networks for speech recognition," in *Interspeech*, 2016, pp. 3439–3443.

[21] Liang Lu, Michelle Guo, and Steve Renals, "Knowledge distillation for small-footprint highway networks," in *ICASSP2017*, 2017, pp. 4820–4824.

[22] Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R. Hershey, "Student-teacher network learning with enhanced features," in *ICASSP2017*, 2017, pp. 5275–5279.

[23] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers," in *Interspeech*, 2017, pp. 3697–3701.

[24] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR2015*, 2015.

[25] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] Yoon Kim and Alexander M. Rush, "Sequence-level knowledge distillation," in *EMNLP2016*, 2016.

[27] Jeremy H. M. Wong and Mark J. F. Gales, "Sequence student-teacher training of deep neural networks," in *Interspeech*, 2016.

[28] Andrew Senior, Haşim Sak, Félix de Chaumont Quitry, Tara N. Sainath, and Kanishka Rao, "Acoustic modelling with cd-ctc-smbr lstm rnns," in *ASRU2015*. IEEE, 2015, pp. 604–609.

[29] John Garofolo et al., "CSR-I (WSJ0) Complete LDC93S6A," DVD. Philadelphia: Linguistic Data Consortium, 1993.

[30] 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 *ASRU2011*. IEEE Signal Processing Society, 2011.

[31] John Garofolo et al., "CSR-II (WSJ1) Complete LDC94S13A," DVD. Philadelphia: Linguistic Data Consortium, 1994.

[32] Francis Kubala et al., "The hub and spoke paradigm for csr evaluation," in *ARPA Human Language Technology Workshop*, 1994, pp. 37–42.

[33] "CNTK," https://github.com/Microsoft/CNTK.

[34] Diederik P. Kingma and Jimmy Lei Ba, "Adam: A method for stochastic optimization," in *ICLR2015*, 2015.