Interspeech 2018
2-6 September 2018, Hyderabad

Knowledge Distillation for Sequence Model

Mingkun Huang[1], *Yongbin You*[2], *Zhehuai Chen*[1], *Yanmin Qian*[1], *Kai Yu*[1]

[1]Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
[2]AISpeech Ltd.

{mingkunhuang,chenzhehuai,yanminqian,kai.yu}@sjtu.edu.cn, yongbin.you@aispeech.com

Abstract

Knowledge distillation, or *teacher-student training*, has been effectively used to improve the performance of a relatively simpler deep learning model (the student) using a more complex model (the teacher). It is usually done by minimizing the Kullback-Leibler divergence (KLD) between the output distributions of the student and the teacher at each frame. However, the gain from *frame-level knowledge distillation* is limited for sequence models such as Connectionist Temporal Classification (CTC), due to the mismatch between the sequence-level criterion used in teacher model training and the frame-level criterion used in distillation. In this paper, *sequence-level knowledge distillation* is proposed to achieve better distillation performance. Instead of calculating a teacher posterior distribution given the feature vector of the current frame, sequence training criterion is employed to calculate the posterior distribution given the whole utterance and the teacher model. Experiments are conducted on both English Switchboard corpus and a large Chinese corpus. The proposed approach achieves significant and consistent improvements over the traditional frame-level knowledge distillation using both labeled and unlabeled data.

Index Terms: Knowledge Distillation, Connectionist Temporal Classification, Kullback-Leibler Divergence.

1. Introduction

Although deep neural networks (DNNs) have achieved state-of-the-art performance in automatic speech recognition (ASR) [1], the large amount of parameters take up considerable memory for storage. The complex models also require much time and power to evaluate [2]. These factors impede the deployment of such models on resource limited systems. Thus researches have been conducted on transferring the learned models to low-resource platforms [3, 4, 5, 6].

Transfer learning (knowledge distillation) [7], or teacher-student (TS) training [2], is a machine learning paradigm that shows potential in model compression. Namely, it utilizes knowledge learned by teacher models to help the student model converge faster or with better predictions [8]. In this way, the knowledge in a conventionally trained DNN can be distilled into a narrower and shallower model with fewer parameters and comparable system performance. In [2], teacher models predict the soft targets as the supervision of much smaller student model. Teacher models can be larger sizes, different structures or trained by different criteria including cross entropy (CE) and sequence discriminative training. In [9], the knowledge distillation can be combined with newly proposed parameter-efficient neural network structures.

All these methods operate at frame level. Namely, KullbackLeibler divergence (KLD) between posterior outputs of the student model and teacher models at each frame is minimized. The disadvantages include: i) separately optimizing at frame level while ASR is inherently a sequence labeling problem. ii) treating un-transcribed and transcribed data equally and not utilizing transcription of the latter. Moreover, researches haven't been conducted carefully in the context of end-to-end sequence model such as connectionist temporal classification (CTC).

In this work, knowledge distillation for CTC is investigated for the first time. *Sequence-level* knowledge distillation is proposed. Namely, the posterior probability of the student model is still optimized at sequence level. Nevertheless, the posterior probability of the label sequence given the feature sequence is obtained from the teacher model using CTC criterion, which can be extended to sequence discriminative training criteria. Experiments are conducted on both Switchboard corpus and a larger Chinese corpus. The proposed method achieves consistent improvement versus traditional frame-level knowledge distillation. In Section 2.2, knowledge distillation and its application in ASR are briefly reviewed. Sequence-level knowledge distillation is proposed in Section 2.3 and compared with prior works in Section 3. In Section 4, experiments are conducted. Finally we present our conclusions in Section 5.

2. Knowledge distillation for CTC

In this section, both traditional frame-level knowledge distillation and sequence-level knowledge distillation are introduced to CTC [1].

2.1. Connectionist temporal classification

The CTC criterion [10] was introduced to map the input speech frames into an output label sequence. To handle the issue that the length of output labels is smaller than that of input frames, CTC allows the repetition of labels and introduces a blank label (denoted as ϕ) to map the label sequence into a CTC path, which makes the input and output sequence having the same length T.

Denote the phoneme label sequence as $\mathbf{l}$, the corresponding input frame sequence as $\mathbf{x}$, and $\mathcal{B}^{-1}(\mathbf{l})$ represents all the CTC paths mapped from $\mathbf{l}$. The CTC objective function $\mathcal{L}_{\text{CTC}}$ is defined as the negative log conditional probability of the ground

The corresponding author is Kai Yu. This work has been supported by the National Key Research and Development Program of China under Grant No.2017YFB1002102, the China NSFC project (No. U1736202) and the JiangSu NSF project (BK20161244). Experiments were carried out on the PI supercomputer at Shanghai Jiao Tong University. AISpeech also provide facility support.

[1] Our recent work shows that the proposed approach can be extended to other sequence discriminative training criteria.

10.21437/Interspeech.2018-1589

truth labels of all training sequences (using one sequence as an example)

$$\mathcal{L}_{\text{CTC}} = -\ln(P(\mathbf{l}|\mathbf{x})) \tag{1}$$

The gradient of (1) with respect to (w.r.t.) the output at each frame can be calculated as

$$\frac{\partial \mathcal{L}_{\text{CTC}}}{\partial y_k^t} = -\frac{1}{P(\mathbf{l}|\mathbf{x})}\frac{\partial P(\mathbf{l}|\mathbf{x})}{\partial y_k^t} \tag{2}$$

where y_k^t denotes the softmax output of k-th phoneme at frame t. The probability of label sequence $\mathbf{l}$ can be calculated by summing over the probabilities of all possible paths:

$$P(\mathbf{l}|\mathbf{x}) = \sum_{\boldsymbol{\pi}} P(\boldsymbol{\pi}|\mathbf{x}) = \sum_{\boldsymbol{\pi}} \prod_{t=1}^{T} y_{\pi_t}^t \tag{3}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T) \in \mathcal{B}^{-1}(\mathbf{l})$, and we assume that labels at each time step are conditionally independent. The required gradient w.r.t the softmax activation a_k^t is

$$\frac{\partial \mathcal{L}_{\text{CTC}}}{\partial a_k^t} = \sum_{k'} \frac{\partial \mathcal{L}_{\text{CTC}}}{\partial y_{k'}^t}\frac{\partial y_{k'}^t}{\partial a_k^t} = y_k^t - \frac{\sum_{\pi_t=k} P(\boldsymbol{\pi}|\mathbf{x})}{P(\mathbf{l}|\mathbf{x})} \tag{4}$$

Denote $\sigma_{\text{CTC}}(k,t)$ as $\frac{\sum_{\pi_t=k} P(\boldsymbol{\pi}|\mathbf{x})}{P(\mathbf{l}|\mathbf{x})}$, which is the posterior probability of the k-th phoneme at frame t, and can be efficiently calculated using the forward-backward algorithm [10].

2.2. Frame-level knowledge distillation

To train the student model, the Kullback-Leibler divergence (KLD) between the frame posterior distributions of the teacher and student models is minimized as below:

$$\begin{aligned} KLD(p,q) &= \sum_i p_i \log \frac{p_i}{q_i} \\ &= \sum_i p_i \log p_i - \sum_i p_i \log q_i \end{aligned} \tag{5}$$

where p and q are the teacher and student distributions respectively. Because the first term in Equation (5) is not related to the student model optimization, only the second term (cross entropy) is used. Comparing Equation (5) with CE criterion, the hard label is replaced by the posterior distribution inferred from the teacher model using source data at each frame. In ASR, the optimization of knowledge distillation is conducted on the frame level KLD in each sequence (using one training sequence as an example):

$$\mathcal{L}_{\text{KLD}} = \sum_{t=1}^{T}\sum_{k} h_k^t \log y_k^t \tag{6}$$

where h_k^t and y_k^t are the k-th phoneme's posterior probability of teacher and student models at frame t respectively.

Researches have been conducted on different types of teacher and student models. In [2], teacher models can be larger sizes, different structures or trained by different criteria including cross entropy and sequence discriminative training. In [9], the knowledge distillation can be combined with newly proposed parameter-efficient neural network structures. In [11], the KLD optimization can be combined with permutation invariant training in multi-outputs single channel overlapped speech recognition.

The frame-level knowledge distillation discussed above can be extended to CTC (F-CTC) as below,

$$\frac{\partial \mathcal{L}_{\text{F-CTC}}}{\partial a_k^t} = y_k^t - h_k^t \tag{7}$$

y_k^t and h_k^t are the k-th phoneme outputs at frame t of student and teacher models respectively.

The training procedure is similar to [2]: i) Train a large-size CTC teacher model with the standard procedure. ii) For each mini-batch, do forward propagation of both teacher and student models to obtain y_k^t and h_k^t. iii) Calculate the error signal as Equation (7) and do back-propagation only for the student model. iv) Repeat step ii to iv until convergence.

Disadvantages of the current framework are summarized:

- **Frame level optimization**. ASR is inherently a sequence labeling problem. Sequence level criteria, CTC and sequence discriminative training, are proposed and achieved significant and consistent improvement [12]. Nevertheless, the traditional knowledge distillation operates at frame level, although teacher models can be sequence discriminative trained [2].
- **Missing transcription**. The traditional knowledge distillation treats un-transcribed and transcribed data equally as the input of both teacher and student models. Although large amount of un-transcribed data can be helpful [13], discarding transcription can hamper the performance and requires additional steps to utilize the transcription during [14] or after knowledge distillation [9, 11].

2.3. Sequence-level knowledge distillation

In traditional deep learning models trained by the cross entropy criterion, the quality of the supervision is always crucial to the performance [1]. Several works comparing traditional systems with CTC based systems also show that the quality of $\sigma_{\text{CTC}}(k,t)$ is the bottleneck of the convergence speed and performance [15].

Frame-level knowledge distillation allows transfer of these frame distributions directly from the softmax output of teacher model. Ideally however, we would like to fully utilize the sequence training criterion. From Equation (3) (4), we know that

$$\frac{\partial \mathcal{L}_{\text{CTC}}}{\partial a_k^t} = y_k^t - \sigma_{\text{CTC}}(k,t) \tag{8}$$

where $\sigma_{\text{CTC}}(k,t)$ is the posterior probability of the k-th phoneme at frame t. Since $\sum_k \sigma_{\text{CTC}}(k,t) = 1$ and $\sigma_{\text{CTC}}(k,t) >= 0$, we can treat $\sigma_{\text{CTC}}(k,t)$ as the soft label used in traditional teacher student training. According to the derivative in Equation (4), the loss function can be equally written as below:

$$\mathcal{L}_{\text{CTC}} = -\sum_{t=1}^{T}\sum_{k} \sigma_{\text{CTC}}(k,t) \log y_k^t \tag{9}$$

To distill the knowledge, we do the forward-backward calculation on teacher model to obtain $\sigma_{\text{CTC}}(k,t)$, which can be extended to other sequence discriminative training criteria. Therefore, the sequence-level knowledge distillation training objective function $\mathcal{L}_{\text{S-CTC}}$ can be derived as below:

$$\frac{\partial \mathcal{L}_{\text{S-CTC}}}{\partial a_k^t} = y_k^t - \sigma'_{\text{CTC}}(k,t) \tag{10}$$

where $\sigma'_{\text{CTC}}(k,t)$ is calculated on teacher model. Equation 10 is similar to Equation 7 for both of them calculate frame-wise KLD. Essentially, we compute the conditional probability by marginalizing all possible alignments and, at each frame t, force the student model to focus on the correct labels, which thoroughly utilizes the label sequence.

The training procedure is summarized: i) Train a large-size CTC teacher model with the standard procedure. ii) For each mini-batch, do forward propagation of both teacher and student models to obtain y_k^t and $\sigma'_{\text{CTC}}(k,t)$. iii) Do forward-backward calculation on each sequence at the teacher model and obtain $\sigma'_{\text{CTC}}(k,t)$ as Equation (4). iv) Calculate the error signal as Equation (10) and do back-propagation for the student model. v) Repeat step ii to iv until convergence.

Compared with frame-level knowledge distillation, as shown in Figure 1, the key difference is that the transcription is utilized. In Equation (7), the student model is optimized to imitate the frame posterior output of the teacher model, no matter whether it is correct or wrong compared with the transcription. Nevertheless, S-CTC forces the student model to only learn a correct phoneme alignment through the state occupation probability obtained from forward-backward calculation on the transcription using the teacher model.

For data without transcriptions, we can first utilize it using frame-level knowledge distillation, then fine-tune the student model using sequence-level knowledge distillation with labeled data. Details of this effect will be discussed in Section 4.3.

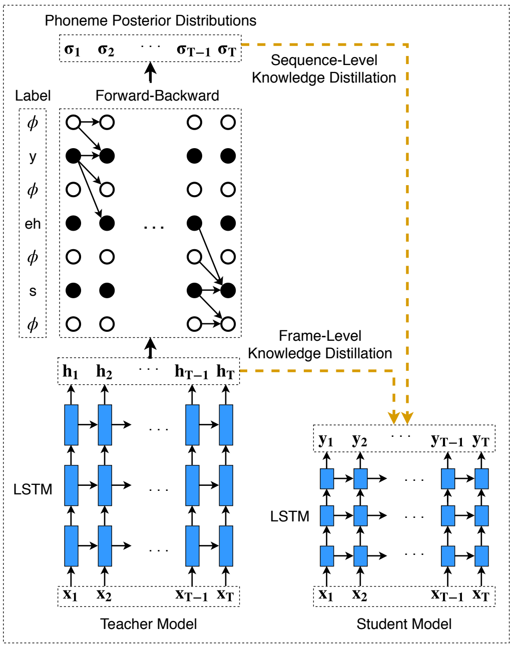


Figure 1: *Frame and sequence level knowledge distillation.*

3. Relation to prior work

In this work, knowledge distillation based model compression [8] for CTC is investigated for the first time. The student model is still trained by CE criterion, but the supervision, the frame posterior probability, is obtained from the teacher model using sequence training criterion.

A key difference versus the traditional frame-level knowledge distillation is that the proposed knowledge distillation is conducted on sequence level. [2, 9] propose to transfer the sequence discriminative trained teacher model by KLD based frame-level training. Sequence discriminative training can be conducted after transfer learning, which obtains further improvement [11]. We believe it is the evidence of the importance of sequence training. Another difference is that the transcription is utilized, which is different with hypothesis sequence-level knowledge distillation in [16]. The proposed method forces the student model to only learn a correct phoneme alignment through the state occupation probability obtained from doing forward-backward calculation on the transcription at the teacher model.

The closest related work is the sequence student-teacher training proposed for DNN-HMM in [17], where the student model is trained to emulate the hypothesis posterior distribution of the teacher model. The differences include: i) different models and criterion. [17] introduces sequence-level knowledge distillation to DNN-HMM trained by sequence discriminative criteria, while this work is based on LSTM trained by CTC. ii) hypothesis modeling. [17] constrains teacher hypotheses by n-best list from beam search, which can hinder the improvement. Nevertheless, this work computes the conditional probability by marginalizing over all possible alignments of the transcription and shows that sequence-level knowledge distillation can be combined with unsupervised frame-level knowledge distillation, which brings further improvement.

4. Experiments

Experiments were performed on both Switchboard corpus and a large Chinese corpus. The CTC teacher model is a 5-layer LSTM, each with 1024 memory cells and a recurrent projection layer of 256 units. For the student model, we used 3 LSTM layers of 400 cells, each with a recurrent projection layer of 128 units. For comparison purpose, a baseline hybrid system was trained by CE criterion and with the same structure as teacher model except the last layer, which is tri-phone states with 8K clusters. The CTC model was initialized by the baseline hybrid system above [18]. The weights in all the networks were randomly initialized with a uniform (-0.02, 0.02) distribution. We clipped gradients to [-5, 5]. A learning rate annealing and early stopping strategies as in [19] were used. All LSTM RNN models were trained using KALDI [20] and EESEN [21].

4.1. Experiments on Switchboard corpus

Switchboard [22] is a 310-hour English dataset with 4870 channels. 36-dimensional log-mel filterbank over 25 ms frames every 10 ms from the input speech signal was extracted. 45 monophones and a blank were predicted by the output layer of the neural network. Evaluation was carried out on the Switchboard (swbd) and Callhome (callhm) subset of the NIST 2000 CTS test set. The waveforms were segmented according to the NIST partitioned evaluation map (PEM) file. A 30k-vocabulary language model trained from transcription of the Switchboard corpus and interpolated with the Fisher corpus was used for decoding. Word error rate (WER) was taken as the metric.

Table 1 shows the performance comparison on Switchboard. Line 1 and 2 are the baseline systems of the teacher and student model respectively. With 10 times more parameters, the teacher model obtains 20% reduction in WER versus that of the student model, which is similar to the observation in [23].

Table 1: *Performance comparison of CTC based knowledge distillation on Switchboard corpus.*

Model	Criterion	WER (%)	
		swbd	callhm
Teacher	CTC	15.9	29.6
Student		20.0	34.3
Student	F-CTC	17.8	32.6
	S-CTC	**17.4**	**31.5**

In our preliminary experiments, after the knowledge distillation procedure described in Section 2.2 and 2.3, utilizing CTC criterion to fine-tune the model always brings about further slight improvement, which is the same to the observation in sequence discriminative training [9, 11]. Thus all numbers shown in the table are obtained after CTC fine-tuning. Compared with the student model, F-CTC can obtain 11% and 5% WER reduction in `swbd` and `callhm`, respectively. S-CTC further shows slight but consistent reduction versus F-CTC.

4.2. Experiments on Chinese corpus

CTC always need more data to achieve competitive performance versus hybrid systems [18, 24]. We used a 2000 hours hand-transcribed Chinese corpus to evaluate the proposed CTC based knowledge distillation paradigm. All the utterances were extracted from an online speech recognition service. Our training set consists of 2.5 million utterances with average duration of 3 seconds. The input of the LSTM RNNs was 40-dimensional log-mel filterbank energy features computed every 10ms using the Chinese corpus. The input layer frame skipping [25] was adopted to reduce computation. A stochastic data sweeping scheme [26] was used to accelerate the training procedure. 121 mono-phones and a blank were predicted by the output layer. A 3-gram language model was applied in evaluation. The evaluation set were also extracted from the online service without speaker duplication. The test set consists of 6500 utterances.

Table 2: *Performance comparison of CTC based knowledge distillation on Chinese corpus.*

Model	Method	CER (%)
Teacher	CTC	15.7
Student		20.5
Student	F-CTC	19.7
	S-CTC	**18.9**

Table 2 shows performance comparison on the large Chinese corpus. The observations from first two lines are consistent with that of Table 1. Comparing line 3 and 4 with the student baseline system in line 2, both knowledge distillation methods can significantly reduce the WER and S-CTC shows slight but consistent improvement versus F-CTC. All results shown in the table are obtained with CTC fine-tuning.

4.3. Experiments on unlabeled data

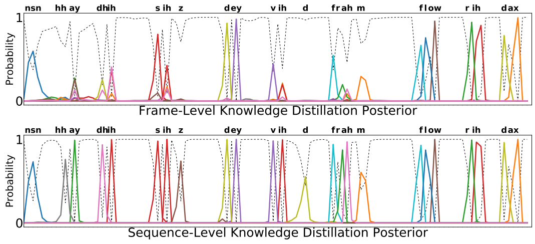
As discussed in Section 2.2, frame-level knowledge distillation can work in an unsupervised manner. Hence large amount of unlabeled data can be used to make the student model more similar to the teacher [13]. Another 2000 hours data extracted from the same source described in Section 4.2 is used. We utilize all data to do the frame-level knowledge distillation in an unsupervised manner, which obtains 3.5% WER reduction in line 2 of Table 3. After that, we continue the S-CTC procedure in line 3, which brings another 2.1% improvement. Both improvements are statistically significant.

Table 3: *Knowledge distillation utilizing both labeled and unlabeled data.*

Model	Method	CER (%)
	F-CTC	19.7
Student	+ unlabeled data	19.0
	+ S-CTC	**18.6**

4.4. Analysis

We selected an utterance from Switchboard corpus and forward propagated it on the teacher model described in Section 4.1. Figure 2 shows the frame-level phoneme probabilities emitted by the teacher CTC model (different color for each phoneme, dotted black line for *blank*), along with the sequence-level phoneme posterior from forward-backward calculation using the teacher outputs. As we can see, the sequence-level posterior is more sharp and evident at phoneme 'v', 'ih' and 'ah', and it can always focus on the right labels at each frame, hence gives student model more correct knowledge. Note that the frame-level posterior assigns high probability to wrong labels when the truth label is 'ay'. On the whole, a better method to distill the knowledge is the proposed S-CTC as shown in the result.



Frame-Level Knowledge Distillation Posterior

Sequence-Level Knowledge Distillation Posterior

Figure 2: *Posterior comparison.*

5. Conclusions

In this work, knowledge distillation for Connectionist Temporal Classification (CTC) is investigated for the first time. Moreover, the *sequence-level knowledge distillation* is proposed. Namely, the posterior probability of the student model is still optimized at sequence level. Nevertheless, the posterior probability of the label sequence given the feature sequence is obtained from the teacher model. Experiments are conducted on both Switchboard corpus and a larger Chinese corpus. The proposed method achieves significant and consistent improvement versus the student model and the traditional frame-level knowledge distillation. The proposed method benefits from utilizing the transcription and can be extended to other sequence discriminative training criteria. Future works include extending the sequence-level knowledge distillation to several newly-proposed sequence criteria, e.g. RNN transducer [27], recurrent neural aligner (RNA) [28], neural segmental model [29] and attention based encoder-decoder [30].

6. References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[3] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition." in *Interspeech*, 2013, pp. 2365–2369.

[4] T. He, Y. Fan, Y. Qian, T. Tan, and K. Yu, "Reshaping deep neural network for fast decoding by node-pruning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 245–249.

[5] L. Lu and S. Renals, "Small-footprint deep neural networks with highway connections for speech recognition," in *Interspeech*, 2016, pp. 12–16.

[6] X. Xiang, Y. Qian, and K. Yu, "Binary deep neural networks for speech recognition," in *Interspeech*, 2017, pp. 533–537.

[7] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[8] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[9] L. Lu, M. Guo, and S. Renals, "Knowledge distillation for small-footprint highway networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4820–4824.

[10] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[11] Z. Chen, J. Droppo, J. Li, and W. Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 184–196, 2018.

[12] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *Interspeech*, 2013, pp. 2345–2349.

[13] J. Li, M. L. Seltzer, X. Wang, R. Zhao, and Y. Gong, "Large-scale domain adaptation via teacher-student learning," in *Interspeech*, 2017.

[14] S. Watanabe, T. Hori, J. Le Roux, and J. R. Hershey, "Student-teacher network learning with enhanced features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5275–5279.

[15] A. Zeyer, E. Beck, R. Schlüter, and H. Ney, "CTC in the context of generalized full-sum HMM training," in *Interspeech*, 2017, pp. 944–948.

[16] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1317–1327.

[17] J. Wong and M. Gales, "Sequence student-teacher training of deep neural networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech*, vol. 8, 2016, pp. 2761–2765.

[18] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," in *Interspeech*, 2015.

[19] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[21] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.

[22] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 1992, pp. 517–520.

[23] I. McGraw, R. Prabhavalkar, R. Alvarez, M. G. Arenas, K. Rao, D. Rybach, O. Alsharif, H. Sak, A. Gruenstein, F. Beaufays *et al.*, "Personalized speech recognition on mobile devices," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5955–5959.

[24] Z. Chen, Y. Zhuang, Y. Qian, and K. Yu, "Phone synchronous speech recognition with CTC lattices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 86–97, Jan 2017.

[25] V. Vanhoucke, M. Devin, and G. Heigold, "Multiframe deep neural networks for acoustic modeling," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7582–7585.

[26] W. Deng, Y. Qian, Y. Fan, T. Fu, and K. Yu, "Stochastic data sweeping for fast DNN training," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 240–244.

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

[28] H. Sak, M. Shannon, K. Rao, and F. Beaufays, "Recurrent neural aligner: An encoder-decoder neural network model for sequence to sequence mapping," in *Interspeech*, 2017, pp. 1298–1302.

[29] H. Tang, L. Lu, L. Kong, K. Gimpel, K. Livescu, C. Dyer, N. A. Smith, and S. Renals, "End-to-end neural segmental models for speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, 2017.

[30] W. Chan, "End-to-end speech recognition models," Ph.D. dissertation, Carnegie Mellon University Pittsburgh, PA, 2016.