Interspeech 2018
2-6 September 2018, Hyderabad

Acoustic Modeling with DFSMN-CTC and Joint CTC-CE Learning

Shiliang Zhang, Ming Lei

Machine Intelligence Technology, Alibaba Group
{sly.zsl, lm86501}@alibaba-inc.com

Abstract

Recently, the connectionist temporal classification (CTC) based acoustic models have achieved comparable or even better performance, with much higher decoding efficiency, than the conventional hybrid systems in LVCSR tasks. For CTC-based models, it usually uses the LSTM-type networks as acoustic models. However, LSTMs are computationally expensive and sometimes difficult to train with CTC criterion. In this paper, inspired by the recent DFSMN works, we propose to replace the LSTMs with DFSMN in CTC-based acoustic modeling and explore how this type of non- recurrent models behave when trained with CTC loss. We have evaluated the performance of DFSMN-CTC using both context-independent (CI) and context-dependent (CD) phones as target labels in many LVCSR tasks with various amount of training data. Experimental results shown that DFSMN-CTC acoustic models using either CI-Phones or CD-Phones can significantly outperform the conventional hybrid models that trained with CD-Phones and cross-entropy (CE) criterion. Moreover, a novel joint CTC and CE training method is proposed, which enables to improve the stability of CTC training and performance. In a 20000 hours Mandarin recognition task, joint CTC-CE trained DFSMN can achieve a 11.0% and 30.1% relative performance improvement compared to DFSMN-CE models in a normal and fast speed test set respectively.

Index Terms: speech recognition, connectionist temporal classification, CTC, DFSMN-CTC, CTC-CE

1. Introduction

In the past few years, deep neural networks have become the state-of-the-art acoustic models in large vocabulary continuous speech recognition (LVCSR) systems. Depending on how the networks are connected, there exist various types of deep neural networks, such as feedforward fully-connected neural networks (FNN) [1, 2] , convolutional neural networks (CNN) [3, 4], recurrent neural networks (RNN) [5, 6] and long short-term memory networks (LSTM) [7]. In the conventional *hybrid* approach, deep neural networks are used to generate the individual frames of acoustic data, and their distributions are reformulated as emission probabilities for a hidden Markov model (HMM). Model training can then be carried out by using the frame-level cross-entropy (CE) criterion followed by some sequence discriminative training methods such as maximum mutual information (MMI) [8].

For conventional deep neural networks hidden Markov model hybrid systems, an additional problem is that the frame-level training targets must be inferred from an alignments determined by the HMM. More recently, researchers have paid more and more attention to the end-to-end speech recognition systems. Recent works on end-to-end speech recognition can be categorized into two main approaches: Connectionist Temporal Classification (CTC) [9, 10, 11, 12, 13, 14, 15] and attention-based encoder-decoder [16, 17, 18, 19, 20]. Both methods regard speech recognition as a sequence-to-sequence mapping problem and address the problem of variable-length input and output sequences.

The key idea of CTC is to use intermediate label representation allowing repetitions of labels and occurrences of blank label to identify less informative frames. CTC-based acoustic models can automatically learn the alignments between speech frames and target labels, which removes the need for frame-level training targets. In previous works [9, 12, 13, 15], the acoustic models used together with CTC are normally recurrent neural networks (RNNs), especially the Long Short-Term Memory (LSTM). Because of the memory mechanism of LSTM models, it means that the outputs no longer need to occur at the same time as the input features. Thereby, LSTM has become the most popular or somewhat default choice for end-to-end speech recognition systems with CTC. Experimental results in [11, 12, 13] shown that CTC-based acoustic models have achieved better performance than the conventional hybrid models. Moreover, experimental results also shown that CTC with bidirectional LSTM (BLSTM) can significantly outperform the unidirectional one. However, the output of the BLSTM is available after all of the frames in the input sequence are fed into the BLSTM because the future information is backward propagated from the end of the sequence. This latency problem prevents the application of CTC with BLSTM to low-latency online speech recognition. Another additional problem is that the CTC training of both unidirectional and bidirectional LSTM require to unroll the LSTM by the length of the input sequence, which consumes a huge amount of memory especially when the sequence is very long. In conventional hybrid approach, some variation architectures are proposed to handle these problems, such as the latency-controlled bidirectional LSTMs [21, 22]. However, these methods haven't been verified in CTC based models.

On the other hand, some non-recurrent neural architectures have been proposed to model the long-term dependency, such as the time delay neural network (TDNN) [23, 24, 25], very deep CNN [26], feedforward sequential memory networks (FSMN) [27, 28, 29]. In [29], the proposed Deep-FSMN (DFSMN) can significantly outperform the BLSTM while faster in training speed and less in model parameters when trained with the frame-level targets using cross-entropy. Moreover, DFSMN can easily control the latency by designing the lookahead filters order and the stride. In this work, we firstly try to replace the LSTM with DFSMN in CTC-based acoustic models. We have evaluated the performance of DFSMN-CTC acoustic models in various LVCSR tasks that consist of about 1000, 4000 and 20000 hours of training data. Experimental result shown that DFSMN-CTC with either CI-Phone or CD-Phone targets can significantly outperform the conventional hybrid DFSMN-CE model using CD-Phone targets. We also found that CTC-based acoustic models are more robust to the speed rate than CE-based models. Unfortunately, CTC-based models can somehow suf-

10.21437/Interspeech.2018-1049

fer from the latency problem that at which an output target is detected can be arbitrarily delayed after its corresponding input event [30]. In this work, we have proposed a novel joint CTC-CE learning framework by using CTC-blank posterior as regularization term to handle this problem. More importantly, the joint CTC-CE learning method helps to improve the stability of CTC training and the performance of DFSMN-CTC based acoustic models. Finally, in a 20000 hours Mandarin recognition task, joint CTC-CE trained FSMN can achieve a 11.0% and 30.1% relative performance improvement compared to DFSMN-CE models in a normal and fast speed test set respectively.

2. Connectionist Temporal Classification

Connectionist temporal classification (CTC) [14] is a loss function for sequence labeling problems that converts the sequence of labeling with timing information into the shorter sequence of labels by removing timing and alignment information. When applied to acoustic modeling, CTC can automatically learn the alignments between input speech frame sequences and their label sequences (e.g., phonemes or characters) without employing the frame-level alignment information. The main idea is to introduce the additional CTC blank ($-$) label, and remove the blank labels and merging repeating labels to obtain the unique corresponding sequence.

For a set of target labels, Ω, and its extended CTC target set is defined as $\bar{\Omega} = \Omega \cup \{-\}$. Given an input sequence $\mathbf{x}$ and its corresponding output label sequence $\mathbf{z}$. The CTC path, π, is defined as a sequence over $\bar{\Omega}$, $\pi \in \bar{\Omega}^T$, where T is the length of the input sequence $\mathbf{x}$. The label sequence $\mathbf{z}$ can be represented by a set of all possible CTC paths, $\Phi(\mathbf{z})$, that are mapped to $\mathbf{z}$ with a sequence to sequence mapping function $\mathcal{F}$, $\mathbf{z} = \mathcal{F}(\Phi(\mathbf{z}))$. The mapping function $\mathcal{F}$ maps the CTC path to the label sequence by first merging the consecutive same labels into one and then discard the blank labels, such as:

$$\left.\begin{array}{l}\mathcal{F}(a,-,b,c,-,-)\\ \mathcal{F}(-,-,a,-,b,c)\\ \mathcal{F}(a,b,b,b,c,c)\\ \mathcal{F}(a,-,b,-,c,c)\end{array}\right\} => (a,b,c) \tag{1}$$

Thereby, the log-likelihood of the reference label sequence $\mathbf{z}$ given the input $\mathbf{x}$ can be calculated as an aggregation of the probabilities of all possible CTC paths:

$$\mathbf{P}(\mathbf{z}|\mathbf{x}) = \sum_{\pi \in \mathbf{\Phi}(\mathbf{z})} \mathbf{P}(\pi|\mathbf{x}) \tag{2}$$

For CTC based acoustic modeling, the CTC is usually applied on the top of deep recurrent neural networks (RNNs). During training, the RNNs can then be trained to minimize the following CTC objective function :

$$\mathcal{L}_{ctc}(\mathbf{x}) = -\log \mathbf{P}(\mathbf{z}|\mathbf{x}) \tag{3}$$

The forward-backward algorithm can be used to compute the gradient of $\mathcal{L}_{ctc}$ with respect to the RNNs outputs. Decoding a CTC network can be performed with a beam search algorithm by using the weighted finite-state transducers (WFSTs) [13].

3. Our Approach

3.1. DFSMN-CTC

Deep-FSMN (DFSMN) [29] is an improved FSMN architecture by introducing the skip connections and the memory strides. As shown in Figure 1, it is a DFSMN with 10 DFSMN components followed by 2 fully-connected ReLU layers and a linear projection layer on the top. The DFSMN component consists of four parts: a ReLU layer, a linear projection layer, a memory block and a skip connection from the bottom memory block, except for the first one that without the skip connection from the bottom layer. Thereby, the formulations of the ℓ-th DFSMN component take the following form:

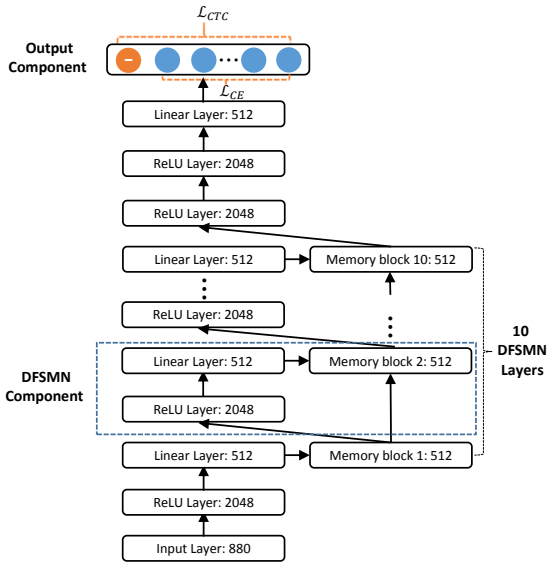

Figure 1: *Joint CTC and CE learning framework for DFSMN based acoustic modeling.*

$$\mathbf{h}_t^\ell = \max(\mathbf{W}^\ell \mathbf{m}_t^{\ell-1} + \mathbf{b}_t^\ell, 0) \tag{4}$$

$$\mathbf{p}_t^\ell = \mathbf{V}_t^\ell \mathbf{h}_t^\ell + \mathbf{v}_t^\ell \tag{5}$$

$$\mathbf{m}_t^\ell = \mathbf{m}_t^{\ell-1} + \mathbf{p}_t^\ell + \sum_{i=0}^{N_1^\ell} \mathbf{a}_i^\ell \odot \mathbf{p}_{t-s_1*i}^\ell + \sum_{j=1}^{N_2^\ell} \mathbf{c}_j^\ell \odot \mathbf{p}_{t+s_2*j}^\ell \tag{6}$$

Here, $\mathbf{h}_t^\ell$ and $\mathbf{p}_t^\ell$ denote the outputs of the ReLU layer and linear projection layer respectively. $\mathbf{m}_t^\ell$ denotes the output of the ℓ-th memory block. N_1^ℓ and N_2^ℓ denotes the look-back order and lookahead order of the ℓ-th memory block, respectively. As shown in eq.(6), by adding the skip connections between the memory blocks of DFSMN components, the output of the bottom layer memory block can be directed flow to the upper layer. During back-propagation, the gradients of higher layer can also be assigned directly to lower layer that help to overcome the gradient vanishing problem. s_1 and s_2 are the strides for look-back and lookahead filters respectively, which help to remove the redundancy in adjacent acoustic frames.

In previous works, DFSMN is evaluated in the conventional hybrid approach that using the frame-level target labels and the cross-entropy loss function. In this work, we will try to evaluate the performance of CTC-based DFSMN trained both with CD-Phones and CI-Phones. We will explore how this type of non-recurrent models behave when trained with CTC loss.

3.2. Joint CTC-CE Learning Framework

In previous work [11], it observed that training with CTC is unstable that sometimes training will fail to converge. In [11], it suggests to handle this problem by using two output layers with

CTC and the conventional CE loss during the training, or initializing from a CE loss pre-trained model. In our experiments, we found that even with CE pre-trained networks as initialization, CTC training can sometime still fail to converge. Moreover, we found that CTC training with CI-Phones is more stable than CD-Phones. This is because the searching space of CD-Phones alignments is more huge than that of CI-Phones. As to solve this unstable problem, we have proposed a novel joint CTC and CE learning framework. Comparison of the DFSMN-CTC and DFSMN-CE acoustic models, except for the training loss function, the only difference is the additional CTC blank label. Thereby, instead of using two softmax output layers for CTC and CE loss, we only use a single softmax output layer as shown in Figure 1, and define a novel optimization framework by joint the CTC loss and a regularized CE loss as followings:

$$\mathcal{L}_{ctcce}(\mathbf{x}) = \mathcal{L}_{ctc}(\mathbf{x}) + \alpha \cdot \mathcal{L}_{ce}(\mathbf{x}) \tag{7}$$

$$\mathcal{L}_{ce}(\mathbf{x}) = -\sum_{i=2}^{K} (1 - p(y_1|\mathbf{x}))t_i \log p(y_i|\mathbf{x}) \tag{8}$$

Where, α is a pre-set constant and the CTC loss, $\mathcal{L}_{ctc}(\mathbf{x})$, is the same as eq.(3). $p(y_1|\mathbf{x})$ in eq.(8) denotes the probability of the CTC blank label in the softmax output layer. $\mathbf{T} = \{t_2, t_3, \cdots, t_K\}$ denotes the frame-level target labels. $(1 - p(y_1|\mathbf{x}))$ can be regarded as the unassigned credit of the output without the CTC blank, which is then used to regularized the CE loss. This regularization term is helpful and important. At the beginning of training, the prediction of the acoustic model is just like a random guessing, then both the CTC and CE loss play a big role in guiding the training. During training, the CTC loss tend to generate the shape spike distribution that only a few spikes for each output target while predicting blank label with high probability the rest of time. Thereby, the regularized CE loss will help to produce the accurate alignment for the output target while won't effect the distribution of blank label. As the result, the proposed joint CTC-CE training will be more stable and help to relieve the delay problem. Decoding procedure of the Joint CTC-CE loss trained model is the same to the plain CTC model.

4. Experiments

4.1. Experimental Setup

In this work, we have evaluated the performance of the proposed DFSMN-CTC and the joint CTC-CE learning frame work on several large vocabulary Mandarin speech recognition tasks, with the total amount of training data being 1000 hours (1k), 4000 hours (4k) and 20000 hours (20k). We have also constructed two test sets, a normal test set and a fast speed test set, to evaluate the performance. Acoustic features used for all experiments are 80-dimensional log-mel filterbank (FBK) energies computed on 25ms window with 10ms shift. We stack the consecutive frames within a long context window of 11 (5+1+5) to produce the 880-dimensional features and then subsample the input frames with 3. These features are used as inputs for all the following experiments. All models are trained in a distributed manner using BMUF [31] optimization on 16 GPUs.

4.2. Baseline Systems

For the baseline CE-based models, we have trained the hybrid Latency-Controlled BLSTM (LCBLSTM) [21, 22] and DFSMN with the lower frame rate (LFR) [29]. An existing CE trained hybrid DNN-HMM system using CD-States is used to realign and generate the 10ms frame-level target labels. We firstly map the 14359 CD-states to 7951 CD-Phones and subsample by averaging 3 one-hot target labels (LFR is 30ms), producing the soft LFR targets. For the baseline LFR trained BLSTM system, we have trained a hybrid LCBLSTM model by stacking 3 BLSTM layers (500 memory cells for each direction), 2 ReLU DNN layers (2048 hidden nodes for each layer) and a softmax output layer. The center-context frames and right-context frames of LCBLSTM are $N_c = 27$ and $N_r = 13$ respectively. LCBLSTM is trained using the BPTT with a mini-batch of 30 sequences. For LFR DFSMN model, the model topology is the same as Figure 1, except the softmax output layer that without the *blank* unit. The DFSMN consists of 10 DFSMN components followed by 2 fully-connected ReLU layers, a linear projection layer and a softmax output layer with 7951 CD-Phone targets. The look-back order and lookahead order of the memory block is 5 and 2 respectively, and the strides are 2 and 1 respectively.

Table 1: *Performance of the CE and CTC based models.*

Method	Label	Data (Hours)	Test set (WER %) Normal	Fast
BLSTM-CE	CD-Phone	1k	19.77	47.56
		4k	16.53	37.17
		20k	13.97	31.71
DFSMN-CE	CD-Phone	1k	18.19	44.25
		4k	14.24	33.92
		20k	12.10	29.79
DFSMN-CTC	CI-Phone	1k	17.82	43.22
		4k	13.82	32.15
		20k	11.46	26.84
DFSMN-CTC	CD-Phone	1k	16.95	40.27
		4k	13.13	26.70
		20k	11.71	24.04

Table 2: *Comparison (model size in MB, training time per epoch in hour) of various acoustic models in the 1000 hours training dataset.*

Method	Label	Model Size (MB)	Time/Epoch (Hours)
BLSTM-CE	CD-Phone	155	3.67
DFSMN-CE	CD-Phone	114	0.50
DFSMN-CTC	CD-Phone	114	0.58
DFSMN-CTC	CI-Phone	97	0.43

The performances of the baseline CE trained LCBLSTM and DFSMN, denoted as *BLSTM-CE* and *DFSMN-CE*, are as shown in Table 1. Experimental results show that DFSMN can consistently outperform the LCBLSTM with different amount of training data. For example, in the normal test set, 20000 hours (20k) training data trained DFSMN can achieve a WER of 12.10% while the performance of the LCBLSTM is 13.97%, which is about 13.4% relative performance improvement. In the fast speed test set, DFSMN can still outperform the LCBLSTM, but the performance gain is less than the normal test set.

4.3. DFSMN-CTC

The architecture of the DFSMN-CTC model is as shown in Figure 1. The output targets can be either CI-Phone or CD-Phone

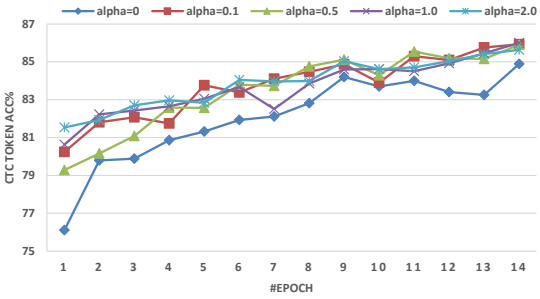

Figure 2: *Comparison of various learning curves of joint CTC-CE trained DFSMN models.*

Table 3: *Performance of the Joint CTC-CE trained DFSMN models on the 20000 hours training set.*

Method	Alpha	Test set (WER %)			
		Normal	Gain	Fast	Gain
CE	-	12.10	-	29.79	-
CTC	-	11.71	3.2%	24.04	19.3%
Joint CTC CE	0.1	10.92	9.8%	21.68	27.2%
	0.5	10.67	11.8%	21.98	26.2%
	1.0	10.77	11.0%	20.80	30.1%
	2.0	11.03	8.8%	22.86	23.3%

set. Here, the CD-Phone set is the same to that used in CE experiments. We firstly map the word level training data transcripts into the CI-Phone sequences by using a Mandarin lexicon. And then map these CI-Phone sequences into the CD-Phone sequences by using a context-dependent tree. These CI-Phone sequences and CD-Phone sequences are used as the targets to train the CI-Phone and CD-Phone based DFSMN-CTC models respectively. The performance of various DFSMN-CTC models are as shown in Table 1.

Compared to the baseline DFSMN-CE models, both CI-Phone and CD-Phone based DFSMN-CTC models can achieve much better performance whether in the normal test set or in the fast speed test set. Experimental results in the fast speed test set indicate that CTC model is more robust to the speed rate compared to the hybrid CE models. CD-Phone DFSMN-CTC models always perform much better than the CI-Phone models that is consistent with previous LSTM-type CTC works [11, 30]. Moreover, with the increasing of training data, the performances gap between the CTC-based models and CE-based models are more obvious. The performance of the 20k training data trained CD-Phone DFSMN-CTC seems not play as well as when the training data is 4k. This is because the training always failed with the general parameter configurations. Finally, we trained the model with a much smaller learning rate, being 0.000001, which is 10 times smaller than we used in other experiments. Maybe it can achieve better performance by more carefully tuning the learning rate schedule or introducing some normalization methods. Instead, we proposed a joint CTC-CE learning framework to solve this unstable problem. The experimental results will be presented in next section. In Table 2, we have compared the model size and training time of CE-based and CTC-based models. CTC training is only a slightly slower than the corresponded CE training.

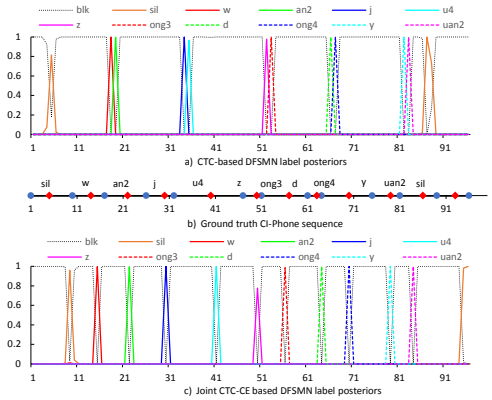

Figure 3: *Label posteriors estimated by CTC and Joint CTC-CE trained DFSMN. (The CD-Phone label posteriors are mapped into the CI-Phone label posteriors.)*

4.4. Joint CTC-CE

We have trained DFSMN using the joint CTC-CE learning method with various α, being 0.1, 0.5, 1.0 and 2.0. If α is equal to 0, then joint CTC-CE is turn out to be the plain CTC. The training set consists of 20000 hours data. The CD-Phone sequences and the CD-Phone alignments are used as targets to train these models. Learning curves in Figure 2 show that the joint CTC-CE models converge much faster the plain CTC model, especially at the beginning of training. Experimental results in the normal and fast speed test sets are listed in Table 3. Joint CTC-CE trained DFSMN can significantly outperform the CE or CTC individual models. With α being 1, the joint CTC-CE model can achieve 11.0% and 30.1% relative performance improvement compared to the CE-based model in the normal and fast speed test sets respectively. The label posteriors estimated by CTC and joint CTC-CE trained DFSMN for a sentence (*w an2 j u4 z ong3 d ong4 y uan2*) in the training set are as shown in Figure 3(a) and Figure 3(c). Figure 3(b) is the ground truth CI-Phone sequence that the central location of each phone (marked with red dots). Consistent with previous work [12], the CTC-based model has learned an arbitrary alignment. For joint CTC-CE trained DFSMN, the constrained CE loss (in eq.(8)) helps to produce the accurate alignment for the output target while won't effect the distribution of the blank label that the spikes of label posteriors usually match the central locations of each phone (as shown in Figure 3(b)). Thereby, joint CTC-CE learning help to overcome the spike delay problem [12], which is essential to the real-time speech recognition.

5. Conclusions

In this work, we present a CTC-based acoustic model using DFSMN instead of the popular LSTM. Experimental results shown that DFSMN-CTC can significantly outperform the conventional CE-based model. Thereby, DFSMN-CTC can take advantage of both DFSMN and CTC that faster in training and decoding and better in performance. Moreover, we also propose a novel joint CTC-CE learning framework to handle the unstable and spike delay problems of CTC. In a 20000 hours Mandarin speech recognition task, the proposed method can achieve 11.0% and 30.1% relative performance improvement compared to the CE-based model in the normal and fast speed test sets respectively.

6. References

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

[2] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[3] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4277–4280.

[4] O. Abdel Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 10, pp. 1533–1545, 2014.

[5] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[6] A. Graves, A. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6645–6649.

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

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

[9] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[10] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[11] H. Sak, A. Senior, K. Rao, O. Irsoy, A. Graves, F. Beaufays, and J. Schalkwyk, "Learning acoustic frame labeling for speech recognition with recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4280–4284.

[12] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," *arXiv preprint arXiv:1507.06947*, 2015.

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

[14] 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.

[15] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[16] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," *arXiv preprint arXiv:1412.1602*, 2014.

[17] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[18] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4945–4949.

[19] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4960–4964.

[20] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," *arXiv preprint arXiv:1712.01769*, 2017.

[21] Y. Zhang, G. Chen, D. Yu, K. Yaco, S. Khudanpur, and J. Glass, "Highway long short-term memory rnns for distant speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5755–5759.

[22] S. Xue and Z. Yan, "Improving latency-controlled blstm acoustic models for online speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5340–5344.

[23] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proceedings of Interspeech*, 2015.

[24] V. Peddinti, G. Chen, D. Povey, and S. Khudanpur, "Reverberation robust acoustic modeling using i-vectors with time delay neural networks," *Proceedings of Interspeech*, 2015.

[25] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 3, pp. 328–339, 1989.

[26] T. Sercu, C. Puhrsch, B. Kingsbury, and Y. LeCun, "Very deep multilingual convolutional neural networks for lvcsr," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4955–4959.

[27] S. Zhang, H. Jiang, S. Xiong, S. Wei, and L. Dai, "Compact feed-forward sequential memory networks for large vocabulary continuous speech recognition." in *INTERSPEECH*, 2016, pp. 3389–3393.

[28] S. Zhang, C. Liu, H. Jiang, S. Wei, L. Dai, and Y. Hu, "Nonrecurrent neural structure for long-term dependence," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 871–884, 2017.

[29] S. Zhang, M. Lei, and a. D. L. Yan, Zhijie, "Deep-fsmn for large vocabulary continuous speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018, pp. 58 639–587.

[30] H. Sak, F. de Chaumont Quitry, T. Sainath, K. Rao *et al.*, "Acoustic modelling with cd-ctc-smbr lstm rnns," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 604–609.

[31] K. Chen and Q. Huo, "Scalable training of deep learning machines by incremental block training with intra-block parallel optimization and blockwise model-update filtering," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5880–5884.