INVESTIGATION OF MODELING UNITS FOR MANDARIN SPEECH RECOGNITION USING DFSMN-CTC-SMBR

Shiliang Zhang, Ming Lei, Yuan Liu, Wei Li

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

ABSTRACT

The choice of acoustic modeling units is critical to acoustic modeling in large vocabulary continuous speech recognition (LVCSR) tasks. The recent connectionist temporal classification (CTC) based acoustic models have more options for the choice of modeling units. In this work, we propose a *DFSMN-CTC-sMBR* acoustic model and investigate various modeling units for Mandarin speech recognition. In addition to the commonly used context-independent Initial/Finals (CI-IF), context-dependent Initial/Finals (CD-IF) and Syllable, we also propose a *hybrid Character-Syllable* modeling units by mixing high frequency Chinese characters and syllables. Experimental results show that DFSMN-CTC-sMBR models with all these types of modeling units can significantly outperform the well-trained conventional hybrid models. Moreover, we find that the proposed hybrid Character-Syllable modeling units is the best choice for CTC based acoustic modeling for Mandarin speech recognition in our work since it can dramatically reduce substitution errors in recognition results. In a 20,000 hours Mandarin speech recognition task, the DFSMN-CTC-sMBR system with hybrid Character-Syllable achieves a character error rate (CER) of 7.45% while performance of the well-trained DFSMN-CE-sMBR system is 9.49%.

Index Terms— Connectionist temporal classification, Mandarin speech recognition, modeling units, DFSMN-CTC-SMBR, hybrid character-syllable

1. INTRODUCTION

Acoustic modeling with neural networks involves three key elements: neural network architecture, acoustic modeling units and optimization objective function. For the conventional neural networks hidden Markov model (NN/HMM) hybrid systems, neural networks are used to estimate the posterior probabilities of hidden Markov model states. The most popular neural networks architecture is the long short term memory recurrent neural networks (LSTM-RNN) [1, 2]. Neural networks based acoustic modeling is usually trained by using the frame-level cross-entropy (CE) criterion followed by some sequence discriminative training methods such as maximum mutual information (MMI) [3] and state-level minimum Bayes risk (sMBR) criterion [4]. One of the drawbacks of conventional hybrid systems is that they require a frame-level alignment for cross-entropy training, which not only makes the training pipeline cumbersome but also limits the choice of acoustic modeling units. As a result, conventional hybrid systems usually use the context-dependent states (CD-state) [5] or the context-dependent phones (CD-phone) [6] as modeling units.

Recently, acoustic modeling with connectionist temporal classification (CTC) [7, 8] has attracted more and more attention due to its faster decoding speed and better performance. 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 demand for frame-level training targets. Thereby, CTC based acoustic models have more options for the choice of modeling units. In [9], the LSTM-CTC acoustic models with either CD-state or CD-phone, or even CI-phone can all outperform the conventional hybrid systems using CD-state targets. Furthermore, the acoustic-to-word CTC models in [10] propose to use the sub-words (single-letter, double-letter, triple-letter) and words as modeling units for English speech recognition and results show that bigger modeling units always lead to better performance.

In previous works [8, 11, 12, 13, 14], neural networks used together with CTC are normally the LSTM-type networks. More recently, the proposed DFSMN-CTC acoustic models in [15] try to replace the LSTM with DFSMN [16] in CTC based acoustic modeling and show that DFSMN-CTC acoustic models with CI-phone and CD-phone both can significantly outperform the conventional hybrid DFSMN-CE model using CD-phone modeling units. In [9, 12], the LSTM-RNN acoustic models first trained with CTC loss can be further optimized with sequence-level discriminative training criteria such as state-level minimum Bayes risk (sMBR) criterion [4]. Along this line, we try to evaluate the performance of DFSMN acoustic models trained with CTC followed by the sMBR criterion in this work. As a result, we propose a *DFSMN-CTC-sMBR* acoustic model for Mandarin speech recognition.

The choice of acoustic modeling units is not only related to the objective function but also related to the language. Mandarin is a tonal language that is different from English. Accordingly, the choice of acoustic modeling units for Mandarin show some differences to that of English. There exist different kinds of modeling units for Mandarin such as phoneme, Initial/Finals (IF), syllable (tonal or toneless) and Chinese character, etc. In [17], it has compared various acoustic modeling units (CI/CD-IF, CI/CD-phone, CI/CD-syllable) in DNNs based large vocabulary continuous speech recognition systems for Chinese. In [18, 14], the Chinese character is adopted as modeling units for end-to-end Mandarin speech recognition.

In this work, we will investigate the performance of DFSMN-CTC-sMBR models with CI-IF, CD-IF and tonal syllable as modeling units, respectively. The ultimate goal of Mandarin speech recognition is to convert the speech signals into the Chinese character sequences. However, a common phenomenon in Mandarin is the homophone that many Chinese characters share the same syllable. As a result, acoustic models adopt syllable as modeling units can not distinguish these characters with identical syllable. Usually, we need a powerful language model to perform beam search decoding in order to achieve promising results. On the other hand, one can choose the Chinese character as modeling units, but it will suffer from the

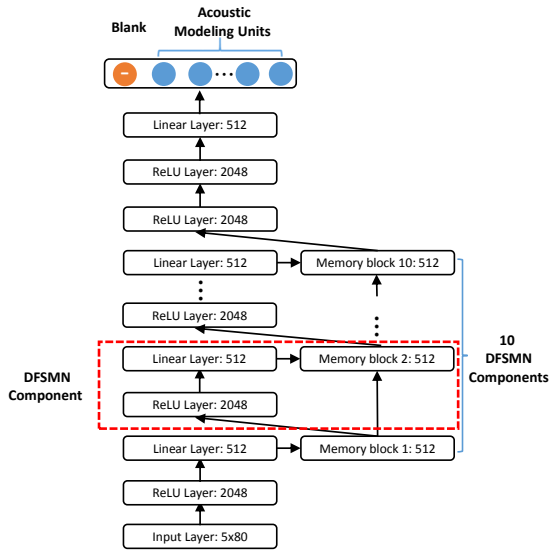

Fig. 1. Illustration of DFSMN-CTC-sMBR acoustic model.

OOV problem due to the huge number of the Chinese characters. In order to handle these homophone and OOV problems, we propose a novel *hybrid Character-Syllable* modeling units by mixing the high frequency Chinese characters (top 2000, 3000) and syllables. Experiments are conducted in a large Mandarin speech recognition task that consists of about 20,000 hours training data. We find that DFSMN-CTC-sMBR systems with all types of modeling units can significantly outperform the well-trained conventional hybrid systems. We also find that the proposed hybrid Character-Syllable modeling units is the best choice since it can significantly reduce substitution errors in recognition results. Finally, we can achieve a character error rate (CER) of 7.45% by using DFSMN-CTC-sMBR with hybrid Character-Syllable as modeling units while the performance of a well-trained DFSMN-CE-sMBR system is 9.49%.

2. DFSMN-CTC-SMBR

2.1. DFSMN

DFSMN [16] is an improved FSMN [19, 20] architecture by introducing the skip connections and the memory strides. As shown in Figure 1, it is a DFSMN with ten DFSMN-components followed by 2 fully-connected ReLU layers and a linear projection layer on the top. The output layer is the corresponding acoustic modeling units, which will be introduced in Section 3. The key element in DFSMN is the so-called DFSMN-component as shown in the red curve box in Figure 1, which enable DFSMN to model the long-term dependency in sequential signals. 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. The operations of the ℓ-th DFSMN component take the following form:

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

$$\mathbf{p}_t^\ell = \mathbf{V}_t^\ell \mathbf{h}_t^\ell + \mathbf{v}_t^\ell \tag{2}$$

$$\mathbf{m}_t^\ell = \mathbf{m}_t^{\ell-1} + \mathbf{p}_t^\ell + \sum_{i=0}^{N_1^\ell} \mathbf{a}_i^\ell \odot \mathbf{p}_{t-s_1*i}^\ell + \sum_{j=1}^{N_2^\ell} \mathbf{c}_j^\ell \odot \mathbf{p}_{t+s_2*j}^\ell \tag{3}$$

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. s_1 is the stride factor of look-back filter and s_2 is the stride of lookahead filter.

2.2. Connectionist Temporal Classification

Connectionist temporal classification (CTC) [7] 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 frames 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 during training, and then remove the blank labels and merging repeating labels to obtain the unique corresponding sequence during decoding.

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:

$$\left.\begin{array}{l} \mathcal{F}(a,-,b,c,-,-) \\ \mathcal{F}(-,-,a,-,b,c) \\ \mathcal{F}(a,b,b,b,c,c) \\ \mathcal{F}(a,-,b,-,c,c) \end{array}\right\} => (a,b,c) \tag{4}$$

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:

$$\mathbf{P}(\mathbf{z}|\mathbf{x}) = \sum_{\pi \in \mathbf{\Phi}(\mathbf{z})} \mathbf{P}(\pi|\mathbf{x}) \tag{5}$$

For CTC based acoustic modeling, the CTC is usually applied on the top of LSTM-type networks. Modeling training is to minimize the following CTC objective function :

$$\mathcal{L}_{ctc}(\mathbf{x}) = -\log \mathbf{P}(\mathbf{z}|\mathbf{x}) \tag{6}$$

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) [12, 13].

2.3. Sequence Discriminative Training

Connectionist temporal classification (CTC) is still a frame-wise discriminative training criterion, which is suboptimal for word error rate (WER) minimization objective in ASR. Previous works in [9, 12] demonstrate that CTC trained LSTM-RNN acoustic models can be further optimized with sequence-level discriminative training criteria such as state-level minimum Bayes risk (sMBR) criterion [4], which can achieve an additional about 10% relative performance improvement. Along this line, DFSMN-CTC can also be further optimized with sequence-level discriminative training criteria. In this work, we will evaluate the performance of DFSMN-CTC-sMBR acoustic models for Mandarin speech recognition.

3. ACOUSTIC MODELING UNITS FOR MANDARIN

The choice of acoustic modeling units is essential to the performance of Mandarin speech recognition systems. Mandarin is naturally a syllabic language that each Chinese character can be phonetically represented by a syllable. Furthermore, each Chinese syllable also has Initial/Final (IF) structure. In previous works[14, 17, 18, 21, 22], many different acoustic modeling units have been evaluated in many frameworks (such as GMM-HMM, DNNs, End-to-End ASR). In this work, we will try to investigate the performance of DFSMN-CTC-sMBR acoustic models with CI-IF, CD-IF, syllable and a novel *hybrid Character-Syllable* modeling units. The detailed composition of modeling units in this study is as shown in Table 1.

3.1. Initial-Finals

Initial/Finals (IF) is the commonly used basic modeling unit for Mandarin speech recognition. According to the official released scheme for Chinese phonetic alphabet, there are 23 Initials and 35 toneless Finals. Usually, each Final has five tones that results in 185 tonal Finals. For CI-IF modeling units, it includes 23 Initials and 185 tonal Finals. As to the number of tied CD-IF, it is determined by the data-driven decision tree, which is 7951 in our work.

3.2. Syllable

Mandarin is a syllabic language that each Chinese character can be represented by a syllable. A common syllable set (with tones) usually contains thousands of tonal syllables according to the pronunciation lexicon. In this study, the tonal syllable set consists of 1319 tonal syllables with tone.

3.3. Hybrid Character-Syllable

Homophone is a common phenomenon in Mandarin that many Chinese characters share the same syllable. The total number of syllables is thousands while the number of Chinese characters is tens of thousand. Thereby, acoustic models adopt syllable as modeling units itself can not distinguish characters with the same syllable. Usually, we need a powerful language model to perform beam search decoding in order to achieve promising result. On the other hand, one can choose the Chinese character as modeling units, but it will suffer from the OOV problem due to the huge number of the Chinese characters. In order to handle these homophone and OOV problems, we propose a novel hybrid Character-Syllable modeling units by mixing the high frequency Chinese characters and syllables. We choose the highly frequency 2000 and 3000 Chinese characters and combine them with 1319 tonal syllables, denoted as *Char(2k)-Syllable* and *Char(3k)-Syllable*, respectively. The coverages of the 2000 and 3000 frequency Chinese characters is 95.58% and 98.65% on our dataset, respectively. The raw transcripts are converted into the mixing characters and syllables sequences for model training.

4. EXPERIMENTS

In this paper , we have evaluated the proposed DFSMN-CTC-sMBR acoustic models with various modeling units on a large vocabulary Mandarin speech recognition task. The training set consists of about 20000 hours data that collected from 20 domains, including *news*, *sport*, *tourism*, *game*, *literature*, *education* et al. A test set contains about 30 hours data is used to evaluated the performance of all models. During decoding, a pruned trigram language model trained with the training transcripts is used. Evaluations are performed in term of character error rate (CER in %).

Table 1. Detailed composition of acoustic modeling units for Mandarin in this study.

Modeling units	Detailed composition
CI-IF	23 Initials + 185 tonal Finals
CD-IF	7951 context-dependent Initial/Finals
Syllable	1319 tonal syllables
Char(2k)-Syllable	2000 high frequency Chinese characters + 1319 tonal syllables
Char(3k)-Syllable	3000 high frequency Chinese characters + 1319 tonal syllables

4.1. Experimental Setup

The 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 context window of 5 (2+1+2) to produce the 400-dimensional features and then down-sample the inputs frame rate. For most experiments, we down-sample the input features by 3 to produce the 30ms duration frames as suggestion in [9]. We will also investigate CTC based models with different frame durations in section 4.4. DFSMN are randomly initialized using the Glorot-Bengio strategy described in [23]. All models are trained in a distributed manner using BMUF [24] optimization on 16 GPUs. The initial learning rate is set to being 0.00001, which is halved if the frame accuracy on the validation set improves less than 0.2% in two consecutive epochs. Early stopping is adopted when the learning rate is halved for six times.

4.2. Baseline Systems

For the baseline conventional hybrid systems, we have trained the latency-controlled BLSTM (LCBLSTM) [25] and DFSMN with the lower frame rate (LFR) technology [6]. An existing CE trained hybrid DNN-HMM system with CD-states is used to realign and generate the 10ms frame-level target labels. We map the 14359 CD-states to 7951 CD-IF and subsample by averaging 3 one-hot target labels, producing the soft targets. The baseline LCBLSTM consists of 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. The model topology of DFSMN is as shown in Figure 1. The baseline DFSMN consists of 10 DFSMN-components, denoted as *DFSMN(10)*. The look-back and lookahead orders of the memory block are 5 and 2, respectively. And the stride factors of look-back and lookahead filters are 2 and 1, respectively. These baseline models are first trained with CE loss and then followed by 2 epochs of sMBR based sequence discriminative training. Experimental results in Table 2 show that DFSMN-CE-sMBR model can achieve about 10% relative improvement compared to LCBLSTM-CE-sMBR model.

4.3. DFSMN-CTC-sMBR with Various Modeling Units

In this experiment, we have evaluated the performance of DFSMN-CTC-sMBR acoustic models with various modeling units (as introduced in Sec.3), including CI-IF, CD-IF, Syllable and Hybrid

Table 2. CER(%) for models trained with different modeling units and objective functions.

Exp	Model	Modeling Units	CER(%)	
			CE	+sMBR
1	LCBLSTM	CD-IF	11.32	10.59
2	DFSMN(10)	CD-IF	10.53	9.49
			CTC	+sMBR
3	DFSMN(10)	CI-IF	10.38	9.37
4	DFSMN(10)	CD-IF	9.70	-
5	DFSMN(10)	Syllable	9.03	7.94
6	DFSMN(10)	Char(2k)+Syllable	8.87	7.61
7	DFSMN(10)	Char(3k)+Syllable	8.81	7.45
8	DFSMN(12)	Char(3k)+Syllable	8.46	7.28

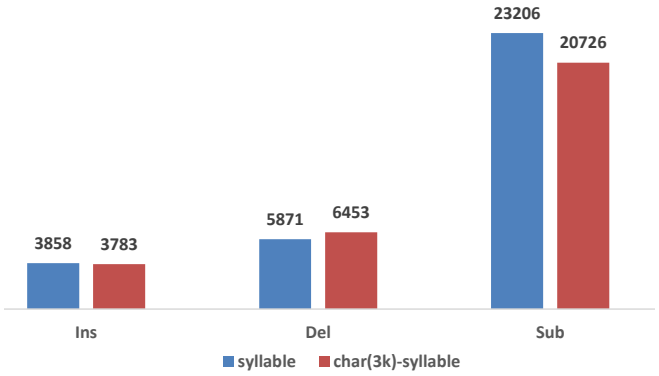

Fig. 2. Error analysis of DFSMN-CTC-sMBR models with *syllable* and *char(3k)-syllable* as modeling units.

Character-Syllable. The model topology of DFSMN is denoted as DFSMN(N_f): 400-N_f $\times$ [2048-512(N_1; N_2; s_1; s_2)]-2 $\times$ 2048-512-N_{out}. Here, $N_1 = 5$, $N_2 = 2$, $S_1 = 2$, $S_2 = 1$, which are the same to the configurations in baseline systems. N_f denotes the number of DFSMN-component. We have evaluated two DFSMN architectures: DFSMN(10) and DFSMN(12) that consist of 10 and 12 DFSMN-components, respectively. N_{out} denotes the output layer size which is equal to the number of acoustic modeling units plus 1 (denote the *blank*). These models are first trained with the CTC loss and then further optimized by the sMBR based sequence discriminative training. Detailed experimental results are as shown in Table 2.

The objective function, modeling units and neural network architecture all play a crucial role in acoustic modeling. Experimental results in Table 2 show that sMBR training can further bring more than 10% relative performance improvement compared to the base CE or CTC training. Moreover, DFSMN-CTC-sMBR models with all types of modeling units (CI-IF, CD-IF, Syllable, Char(2k)+Syllable, Char(3k)+Syllable) can significantly outperform the baseline DFSMN-CE-sMBR model with CD-IF as acoustic modeling units. Comparison of *Exp2* and *Exp7* demonstrates that DFSMN-CTC-sMBR model achieves +21.5% relative CER reduction compared to the DFSMN-CE-sMBR model when using the same neural network architecture. This improvement comes from the CTC loss on the one hand and from the modeling units on the other hand.

Mandarin speech recognition is to convert speech waveform into the corresponding Chinese character sequence. Thereby, CTC based models with modeling units closer to Chinese character will result in better performance. From *Exp3* to *Exp7*, we can see that the CTC based system with Syllable modeling units performs much better than the CI-IF and CD-IF based systems. Moreover, the proposed hybrid Character-Syllable based models (Char(2k)+Syllable, Char(3k)+Syllable) can further outperform the syllable based model. For acoustic modeling with hybrid Character-Syllable units, it can effectively handle the *homophone* and OOV problems in Mandarin. As shown in Figure 2, DFSMN-CTC-sMBR model with *char(3k)-syllable* modeling units significantly reduce the substitution (sub) errors compared to the *syllable* based model. Thereby, hybrid Character-Syllable modeling units is the best choice for Mandarin speech recognition in our study. From *Exp7* and *Exp8*, we can see that deeper DFSMN can further improve the performance when using the same objective functions and modeling units.

Table 3. Comparison of DFSMN-CTC-sMBR models with different frame duration. (Modeling units: *Char(3k)+Syllable*)

Model	Frame Duration	RTF	CER(%)	
			CTC	+sMBR
DFSMN(12)	30ms	0.164	8.46	7.28
	40ms	0.128	8.63	7.37
	50ms	0.098	8.76	7.45

4.4. DFSMN-CTC-SMBR with Various Frame Durations

In previous work [9], the CTC based acoustic models with CI-phone or CD-phone as modeling units usually adopt the input features with frame duration being 30ms. However, the duration of Chinese syllables and characters is about 300ms, which is much longer than the CI-phone or CD-phone in English. So we continue to evaluate the DFSMN-CTC-sMBR model with longer frame durations (40ms, 50ms) for Mandarin speech recognition when using the hybrid character-syllable as modeling units. Experimental results in Table 3 show that it only suffers from about 2% relative degradation in performance when increasing the frame duration from 30ms to 50ms. The benefits of longer frame duration is that it can not only speedup the modeling training but also reduce the computation cost during decoding. For example, the real-time factor (RTF) of model with frame duration being 50ms is 0.098 which is much smaller than the RTF of model with frame duration being 30ms.

5. CONCLUSIONS

In this work, we propose a *DFSMN-CTC-sMBR* acoustic model and investigate various modeling units for Mandarin speech recognition. Except for the commonly used CI-IF, CD-IF and tonal Syllable, we also propose a *hybrid Character-Syllable* units by mixing the high frequency Chinese character and syllable. Experimental results in a 20,000 hours Mandarin recognition task show that the proposed DFSMN-CTC-sMBR models with all types of modeling units can significantly outperform the well-trained conventional hybrid models. Particularly, the hybrid Character-Syllable modeling units is the best choice for Mandarin recognition in our work since it is helpful to handle the homophone and OOV problems. Compared to the well-trained DFSMN-CE-sMBR model, the proposed DFSMN-CTC-sMBR with hybrid Character-Syllable as modeling units can achieve +21.5% relative character error rate (CER) reduction.

6. REFERENCES

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

[2] Haşim Sak, Andrew Senior, and Françoise Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.

[3] Karel Veselý, Arnab Ghoshal, Lukás Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks," in *Interspeech*, 2013, pp. 2345–2349.

[4] Brian Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3761–3764.

[5] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2012.

[6] Golan Pundak and Tara N Sainath, "Lower frame rate neural network acoustic models.," in *Interspeech*, 2016, pp. 22–26.

[7] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen 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.

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

[9] Haşim Sak, Andrew Senior, Kanishka Rao, and Françoise Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] Jinyu Li, Guoli Ye, Amit Das, Rui Zhao, and Yifan Gong, "Advancing acoustic-to-word CTC model," *arXiv preprint arXiv:1803.05566*, 2018.

[11] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al., "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[12] Haşim Sak, Andrew Senior, Kanishka Rao, Ozan Irsoy, Alex Graves, Françoise Beaufays, and Johan 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.

[13] Yajie Miao, Mohammad Gowayyed, and Florian 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] 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 *International Conference on Machine Learning*, 2016, pp. 173–182.

[15] ShiLiang Zhang and Ming Lei, "Acoustic modeling with DFSMN-CTC and joint CTC-CE learning," *Proc. Interspeech 2018*, pp. 771–775, 2018.

[16] Shiliang Zhang, Ming Lei, Zhijie Yan, and Lirong Dai, "Deep-FSMN for large vocabulary continuous speech recognition," *arXiv preprint arXiv:1803.05030*, 2018.

[17] Xiangang Li, Yuning Yang, Zaihu Pang, and Xihong Wu, "A comparative study on selecting acoustic modeling units in deep neural networks based large vocabulary chinese speech recognition," *Neurocomputing*, vol. 170, pp. 251–256, 2015.

[18] Jie Li, Heng Zhang, Xinyuan Cai, and Bo Xu, "Towards end-to-end speech recognition for chinese mandarin using long short-term memory recurrent neural networks," in *Sixteenth annual conference of the international speech communication association*, 2015.

[19] Shiliang Zhang, Cong Liu, Hui Jiang, Si Wei, Lirong Dai, and Yu Hu, "Feedforward sequential memory networks: A new structure to learn long-term dependency," *arXiv preprint arXiv:1512.08301*, 2015.

[20] Shiliang Zhang, Cong Liu, Hui Jiang, Si Wei, Lirong Dai, and Yu Hu, "Nonrecurrent neural structure for long-term dependence," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 4, pp. 871–884, 2017.

[21] Hao Wu and Xihong Wu, "Context dependent syllable acoustic model for continuous chinese speech recognition," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[22] Sheng Gao, Tan Lee, Yiu Wing Wong, Bo Xu, Pak-Chung Ching, and Taiyi Huang, "Acoustic modeling for chinese speech recognition: A comparative study of mandarin and cantonese," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. IEEE, 2000, vol. 3, pp. 1261–1264.

[23] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[24] Kai Chen and Qiang 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.

[25] Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yaco, Sanjeev Khudanpur, and James 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.