Metadata of the chapter that will be visualized in SpringerLink

Book Title	Intelligent Robotics and Applications	
Series Title		
Chapter Title	Towards End-to-End Speech Recognition with Deep Multipath Convolutional Neural Networks	
Copyright Year	2019	
Copyright HolderName	Springer Nature Switzerland AG	
Author	Family Name	**Zhang**
	Particle	
	Given Name	**Wei**
	Prefix	
	Suffix	
	Role	
	Division	School of Mechanical Engineering
	Organization	Jiangnan University
	Address	Wuxi, 214122, Jiangsu, China
	Division	
	Organization	Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology
	Address	Wuxi, 214122, Jiangsu, China
	Email	
Author	Family Name	**Zhai**
	Particle	
	Given Name	**Minghao**
	Prefix	
	Suffix	
	Role	
	Division	School of Mechanical Engineering
	Organization	Jiangnan University
	Address	Wuxi, 214122, Jiangsu, China
	Division	
	Organization	Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology
	Address	Wuxi, 214122, Jiangsu, China
	Email	
Author	Family Name	**Huang**
	Particle	
	Given Name	**Zilong**
	Prefix	
	Suffix	
	Role	
	Division	School of Mechanical Engineering
	Organization	Jiangnan University

	Address	Wuxi, 214122, Jiangsu, China
	Division	
	Organization	Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology
	Address	Wuxi, 214122, Jiangsu, China
	Email	
Author	Family Name	**Liu**
	Particle	
	Given Name	**Chen**
	Prefix	
	Suffix	
	Role	
	Division	School of Mechanical Engineering
	Organization	Jiangnan University
	Address	Wuxi, 214122, Jiangsu, China
	Division	
	Organization	Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology
	Address	Wuxi, 214122, Jiangsu, China
	Email	
Author	Family Name	**Li**
	Particle	
	Given Name	**Wei**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Suzhou Vocational Institute of Industrial Technology
	Address	Suzhou, 215104, Jiangsu, China
	Email	
Corresponding Author	Family Name	**Cao**
	Particle	
	Given Name	**Yi**
	Prefix	
	Suffix	
	Role	
	Division	School of Mechanical Engineering
	Organization	Jiangnan University
	Address	Wuxi, 214122, Jiangsu, China
	Division	
	Organization	Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology
	Address	Wuxi, 214122, Jiangsu, China
	Email	caoyi@jiangnan.edu.cn

Abstract

Approaches to deep learning have been used all over in connection to Automatic Speech Recognition (ASR), where they have achieved a high level of accuracy. This has mostly been seen in Convolutional

Neural Network (CNN) which has recently been investigated in ASR. Due to the fact that CNN has an increased network's depth on one branch, and may not be wide enough to work on capturing adequate features on signals of human speech. We focus on a proposal for an architecture that is deep and wide in CNN referred to as Multipath Convolutional Neural Network (MCNN). MCNN-CTC combines three additional paths with Connectionist Temporal Classification (CTC) objective function, and can be defined as an end-to-end system that has the ability to fully exploit spectral and temporal structures related to speech signals simultaneously. Results from the experiments show that the newly proposed MCNN-CTC structure enables a reduction in the error rate arising from the construction of end-to-end acoustic model. In the absence of a Language Model (LM), our proposed MCNN-CTC acoustic model has a relative reduction of 1.10%–12.08% comparing to the traditional HMM-based or DCNN-CTC-based models with strong generalization performance.

Keywords (separated by '-')	Automatic Speech Recognition (ASR) - Acoustic Model (AM) - MCNN-CTC - Connectionist Temporal Classification (CTC)

Towards End-to-End Speech Recognition with Deep Multipath Convolutional Neural Networks

Wei Zhang[1,3], Minghao Zhai[1,3], Zilong Huang[1,3], Chen Liu[1,3], Wei Li[2], and Yi Cao[1,3](✉)

[1] School of Mechanical Engineering, Jiangnan University, Wuxi 214122, Jiangsu, China
caoyi@jiangnan.edu.cn
[2] Suzhou Vocational Institute of Industrial Technology, Suzhou 215104, Jiangsu, China
[3] Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology, Wuxi 214122, Jiangsu, China

Abstract. Approaches to deep learning have been used all over in connection to Automatic Speech Recognition (ASR), where they have achieved a high level of accuracy. This has mostly been seen in Convolutional Neural Network (CNN) which has recently been investigated in ASR. Due to the fact that CNN has an increased network's depth on one branch, and may not be wide enough to work on capturing adequate features on signals of human speech. We focus on a proposal for an architecture that is deep and wide in CNN referred to as Multipath Convolutional Neural Network (MCNN). MCNN-CTC combines three additional paths with Connectionist Temporal Classification (CTC) objective function, and can be defined as an end-to-end system that has the ability to fully exploit spectral and temporal structures related to speech signals simultaneously. Results from the experiments show that the newly proposed MCNN-CTC structure enables a reduction in the error rate arising from the construction of end-to-end acoustic model. In the absence of a Language Model (LM), our proposed MCNN-CTC acoustic model has a relative reduction of 1.10%–12.08% comparing to the traditional HMM-based or DCNN-CTC-based models with strong generalization performance.

Keywords: Automatic Speech Recognition (ASR) · Acoustic Model (AM) · MCNN-CTC · Connectionist Temporal Classification (CTC)

1 Introduction

Automatic Speech Recognition (ASR) is an automatic method designed to translate human form speech content into textual form [1]. Deep learning has in the past been applied in ASR to increase correctness [2–4], a process that has been successful. As of late, CNN has been successful in acoustic model [5, 6]. Which is applied in ASR combining with HMMs [5], in a way identical to the regular Deep Neural Networks (DNNs) [7, 8], which in turn lead to a hybrid system. DNN-HMM uses a discriminant

H. Yu et al. (Eds.): ICIRA 2019, LNAI 11745, pp. 1–10, 2019.
https://doi.org/10.1007/978-3-030-27529-7_29

model to replace the GMM-HMM generation model, which takes advantage of DNN's powerful fitting ability to model the posterior probability of each frame. The HMM still handles the operations in temporal modelling and decoding whereas the neural network generates posterior probability of the corresponding state [4].

A large amount of problems arise as a result of this hybrid system, where the modules' training which is done separately for different modules and with a different criteria that may certainly not be optimal in the solution of the final task. Consequently, additional hyperparameters turning throughout all training stages are required and can be not only time consuming but also highly laborious [9]. Contrary to the above system, end-to-end model is proposed recently because of its simplicity of modeling process, and also the recognition accuracy is gradually approaching the hybrid system [10–12]. CTC is a objective function introduced by Graves as a means to simplify this process [13, 14], which infers alignments in speech label automatically leading to an end-to-end system. This has generated promising results that can discovery in Deep Speech [15, 16] and EESEN [10].

We propose the MCNN model and construct the MCNN-CTC acoustic model in combination with the CTC objective function, which obtains a significant recognition results. Based on the CTC loss function, this paper studies the speech recognition of small and medium datasets in detail. The merits of the MCNN-CTC include: (a) The above acoustic model can extract more useful features, both in time dimension and frequency axis; (b) MCNN has wider network structure, which can extract sufficient features of speech, and has stronger nonlinear capability; (c) Thanks to the CTC loss, MCNN-CTC can take an end-to-end training manner [17].

The rest of this paper is organized as follows. Section 2 describes the network architecture of MCNN-CTC. A concise introduction to CTC objective function and decoding algorithm are given in Sect. 3. We represent the experimental results in Sect. 4 and conclude our future work in Sect. 5.

2 Multipath Convolutional Neural Networks

As we can see clearly from Fig. 1, MCNN is an augmentation of the CNN's width, and has the ability to extract additional detailed features from speech in terms of width as compared to the basic extraction of high-dimensional speech features in term of depth. Therefore, MCNN is able to increase the performance of the recognition.

The MCNN's structure is shown in Fig. 1. The full structure of MCNN comprises of a total of three sub-networks, extracting features of speech and concatenating them. The calculation formulas are shown in Eq. (1)–(3):

$$h^{(l)} = \sigma\left(W^{(l)} * h^{(l-1)} + b^{(l)}\right) \tag{1}$$

In formula (1), where $h^{(l-1)}$ and $h^{(l)}$ represent two adjacent feature layers, * represents convolution calculation, and $W^{(l)}$ and $b^{(l)}$ represent weights and bias matrices obtained from network training, respectively; $W^{(l)}$ is convoluted with $h^{(l-1)}$, and $\sigma(\bullet)$ represents the activation function. In formula (2), t_{nl}^{out} represents the output value of the

l'th neuron in the n'th feature map; t_{nq}^{in} represents the input value of the q'th neuron in the n'th feature map; $f_{\text{pool}}(\bullet)$ is the pooling function.

$$t_{nl}^{out} = f_{pool}\left(t_{nq}^{in}, t_{n(q+1)}^{in}\right) \tag{2}$$

$$H^{l} = Concat\left(h_i^l, h_j^l, h_k^l\right) \tag{3}$$

In formula (3), where h_i^l, h_j^l, h_k^l represent the i, j, and k feature maps of three different branches, respectively, and the $Concat(\bullet)$ function represents the spliced feature map to obtain the total feature map H^l of the current layer.

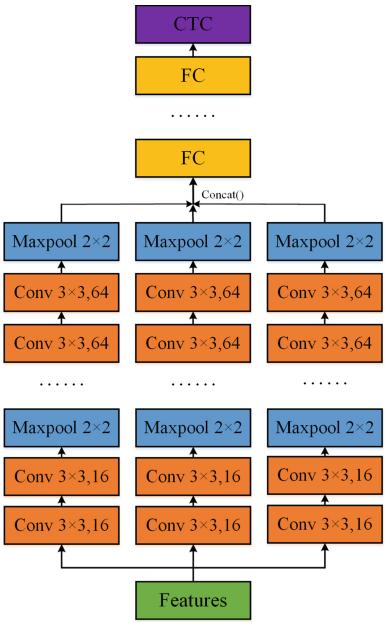


Fig. 1. The structure of multipath convolutional neural network

3 Connectionist Temporal Classification

Connectionist Temporal Classification (CTC) [12, 13] can basically be defined as a target function that maximizes the possibilities of any output sequence [18], which enables this by employing a softmax output layer and summing over all likely input sequences efficiently. It characterizes a separate output circulation $P(k|t)$ throughout every progression t in the input succession and an extra "blank" symbol which is the representation for non-output. The network makes decision on whether to remove any label at each step or not. The probability of removing or emitting the blank or label

given T as length, x as input sequence and y_t as output vectors, is given as follow with k and t as index and time respectively:

$$P(k \mid t, x) = \frac{\exp\left(y_t^k\right)}{\sum_{k'} \left(y_t^{k'}\right)} \tag{4}$$

y_t^k is component k of y_t. Then, π is a length T representing blank indices as well as label indices for a CTC path. $P(\pi|x)$ is the probability representing the emission probabilities' product present at each time interval:

$$p(\pi \mid x) = \prod_{t=1}^{T} p(\pi_t \mid t, x) \tag{5}$$

There are tons of paths and ways of separating labels using blanks for any given transcription sequence. In order to map all paths to the given transcription, one can apply methods such as a many-to-one map ψ, which can be outlined as a means that removes first the repeated labels. Then, y which is the output interpretation can be determined by including the probabilities of all the paths mapped onto it by ψ:

$$P(y \mid x) = \sum_{\pi \in \psi^{-1}(y)} P(\pi \mid x) \tag{6}$$

$$\left\{ \begin{array}{c} \psi(a,b,c,-,-) \\ \psi(a,b,-,-,c) \\ \psi(a,b,b,-,c) \\ \cdots \\ \psi(a,-,b,b,c) \end{array} \right\} = (a,b,c) \tag{7}$$

The "crumbling together" as seen throughout different paths apparently in the similar translation allows the utilization of unsegmented information by CTC. This is as a result of the removal of all requirements needed to know the location in which the input sequence occurs. When given a certain transcription y^*, CTC objective function can be minimized by training the network:

$$CTC(x) = -logP(y^* \mid x) \tag{8}$$

As a means to generate predictions, the best path decoding algorithm is applied from to a trained model using CTC which in turns generates predictions. The highest probability latent sequence are obtained by removing the most likely at each interval since the model assumes that there is independence between the latent symbols given a frame-wise case in the network. By applying $\sigma(\cdot)$ to the prediction of latent sequence, the predicted sequence can be identified as follows:

$$L \approx \sigma(\pi^*) \tag{9}$$

In this case, π^* becomes the most probable concatenation formalized using $\pi^* = Argmax_\pi P(\pi|x)$. In this situation you have to consider the outcome as it is not really the highest probable output arrangement. This sequence requires search procedures that are approximate such as beam search, and the search for this sequence is not tractable.

4 Experiments

This section focuses on the proposed model, where we evaluate it based on the phonetic recognition in relation to the Thchs30 and ST-CMDS datasets. Figure 1 shows MCNN-CTC architecture.

4.1 Data and Experimental Equipment

In order to verify the superiority of the proposed model, we test on two standard Chinese Mandarin speech datasets, Thchs30 and ST-CMDS. For the sake of ensuring the reliability of the experimental results, we have adopted different methods for the two datasets. In the Thchs30 dataset, the number of training set, validation set and test set are 10000, 893 and 2495 sentences respectively. However, in the ST-CMDS dataset, since the original corpus did not divide the dataset, we referred to the division method of Thchs30 dataset and randomly selected 100000 sentences as the training set, 600 sentences as the validation set, and the remaining 2000 sentences as the test set. For the two previous datasets, there are no overlapping between the corpus. GTX-1080Ti graphics card is used for training to ensure the smooth operation of the experiment.

4.2 Modeling Unit and Feature Extraction

The speech recognition of Chinese speech uses the traditional method of modelling which comprises of characters, state, phoneme, and a few phonetic methods of modeling [19]. As a means of making up for the lack of phonetic modeling research, this paper utilizes experiments with phonetic as the only method of modeling. Two major advantages of phonetic modelling are: (a) With the Chinese dictionary having about 200 phonemes, the phonetics are about 1400; (b) Direct modelling leads to inaccurate classification of networks due to many parameter as words in the dictionary as about 16,000 [20, 21].

We use two different data preprocessing methods for two different datasets. For the Thchs30 dataset and the ST-CMDS dataset, we use a frame length of 25 ms and a frame shift of 10 ms to frame the speech signal. However, it is worth noting that for the Thchs30 dataset, we extract the 200-dimensional spectrogram as the speech feature. Nevertheless, in the ST-CMDS dataset, 120-dimensional FBank are applied to the speech feature with splicing one frame before and after, and the total feature dimension is 360 dimension.

4.3 Training and Evaluation

In order to fully advance the model, we apply Adam [22] with learning rate at 1e−2 in training stage. The stochastic gradient descent is used for fine-tuning and has a learning rate 1e−4. During training, batch size 32 are also used. With a 0.3 probability, dropout [23] is applied to all layers with an exception for the layers of input and output. Applied at the fine-tuning stage is L2 norm with coefficient 1e−5 [24]. The pool size and kernel size are 2*2 and 3*3 respectively [25], where at the same time the predicted sequences are acquired using the best path decoding [26].

4.4 Experimental Results of Thchs30

Table 1 shows the test results that help in determining the influence as a result of the layers of the fully connected layer.

Table 1. The influence of different fully connected layers on phonetic error rate

Fully connected layer	Modeling unit	Number of parameters	Phonetic error rate
512-1422	Phonetic	1.95 M	26.65%
1024-512-1422	Phonetic	2.22 M	26.49%

Table 1: 512-1422 represents the neurons in the fully connected layer as 512 and 1422, and the DCNN model utilizes two layers from the fully connected layers with the quantity of neurons as 512-1422; MCNN makes use of the three layers of the fully connected layer with the quantity of neurons as 512-1024-1422.

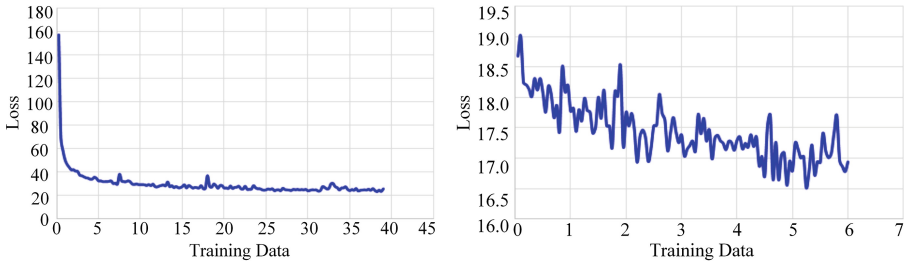
The error rate is lowest when the fully connected layer has three layers as shown in Table 1, however, the model's performance does not improve a lot than that of two layers. The network parameters are also improve when the connected layers are three as compared to when they are two. Therefore, DCNN-CTC uses two layers of fully connected layers, but MCNN-CTC uses a three layers fully connected layer to further classify features.

Table 2. Experimental results of different acoustic models

Modeling structure	Modeling unit	Number of parameters	Word error rate	Phonetic error rate
GMM-HMM [27]	Phone	-	30.53%	-
DNN-HMM [27]	Phone	-	25.16%	-
BLSTM-CTC [28]	Phone	-	25.35%	-
DCNN(7)-CTC	Phonetic	1.95 M	-	26.65%
DCNN(8)-CTC	Phonetic	2.20 M	-	25.66%
DCNN(9)-CTC	Phonetic	2.25 M	-	25.42%
MCNN(6)-CTC	Phonetic	4.66 M	-	25.37%
MCNN(7)-CTC	Phonetic	4.77 M	-	**23.43%**
MCNN(8)-CTC	Phonetic	3.61 M	-	24.85%
MCNN(9)-CTC	Phonetic	3.72 M	-	25.18%

Table 2: MCNN(7) represents that we use three paths CNN and convolution layers are seven. Phone and phonetic represent the modeling units of acoustic models, and we use phonetic modeling in this paper. The training and fine-tune loss of MCNN(7)-CTC are shown in Fig. 2(a) and (b) respectively.

The error rate is highly reduced in the testing process with a 23.43% phonetic error rate as compared to the GMM-HMM, DNN-HMM and BLSTM-HMM, and the error rate reduces by a 7.10, 1.73 and 1.92% respectively. It can be clearly seen from Table 2 that depth is very import for CNN, and we can come up with it development has the following main trends: as the number of DCNN layers increases, the error rate decreases gradually.



(a) (b)

Fig. 2. Curve of Thchs30 dataset's loss function in MCNN(7)-CTC

4.5 Experimental Results of ST-CMDS

In the experiment of the ST-CMDS dataset, in order to verify the generalization performance of the models proposed in this paper, we refer to the experimental results of Thchs30 and use DCNN(7)-CTC and MCNN(7)-CTC acoustic model for the experiment. Finally, the experimental results of DCNN-CTC and MCNN-CTC for the acoustic model are shown in Table 3.

Table 3. Comparison of experimental results between DCNN-CTC and MCNN-CTC

Modeling structure	Modeling unit	Number of parameters	Error rate of validation	Error rate of test
DCNN(7)-CTC	Phonetic	7.80 M	23.86%	23.80%
MCNN(7)-CTC	Phonetic	6.74 M	22.92%	22.97%

From Table 3, Compared with the DCNN-CTC acoustic model, the MCNN-CTC has a relative error reduction of 3.94% and 3.45% in the validation set and the test set respectively. Moreover, thanks to the reduction in the number of convolution kernels, the parameter amount of the MCNN-CTC acoustic model is greatly reduced, and the parameter amount are relatively reduced by 14.10%. In view of this, the structure of

MCNN proposed in this paper is applied to acoustic model with remarkable effect and strong generalization performance.

4.6 Experimental Summary

In summary, this paper conducts a detailed study on the end-to-end acoustic model built by combining DCNN with CTC objective function, and proposes MCNN-CTC acoustic model which has a superior performance in the Chinese standard corpus Thchs30 and ST-CMDS datasets. It is worth noting that, based on the experimental results of Tables 2 and 3, we can conclude that the MCNN-CTC proposed in this paper have greatly reduced the error rate of the model compared with the traditional DCNN-CTC acoustic model. The best results for the Thchs30 and ST-CMDS datasets were reduced by 12.08% and 3.45%, respectively, with reasonable experimental parameters. Moreover, it can be vividly seen from the experimental results that the generalization performance of the acoustic model constructed by MCNN-CTC is excellent, and compared with the acoustic model constructed by the traditional GMM-HMM, the error rate is greatly reduced in a exceedingly simple manner.

5 Conclusion and Future Works

In this paper, an end-to-end system for Chinese Mandarin is established, which is based on CNN and CTC objective function. We deeply analyze the influence of different convolution layers, pooling layers and fully connected layers on DCNN-CTC. Based on the above acoustic model, we propose MCNN-CTC, which is combined MCNN with CTC objective function. We can also find that data is very significant and with the increase of data, MCNN-CTC perform much better than DCNN-CTC or hybrid system. Further, as shown by the promising results from Thchs30 and ST-CMDS, the generalization performance of MCNN-CTC is strong.

In the future, an enormous amount of research will be done on MCNN-CTC with the purpose of building a better acoustic model. In the decoding phase, we will incorporate Language Model to further reduce the error rate.

Acknowledgements. This work reported here was supported by the National Natural Science Foundation of China (Grant No. 51375209), 111 Project (Grant No. B18027), the Six Talent Peaks Project in Jiangsu Province (Grant No. ZBZZ-012), the Research and the Innovation Project for College Graduates of Jiangsu Province (Grant No. SJCX18-0630 and KYCX18-1846). Finally, the authors would like to thanks for the support of Thchs30 and ST-CMDS datasets.

References

1. Lecun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. In: The Handbook of Brain Theory and Neural Networks. MIT Press, USA (1995)
2. Abdel, H.O., Mohamed, A.R., Jiang, H.: Convolutional neural networks for speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(10), 1533–1545 (2014)

3. Mohamed, A., Dahl, G.E., Hinton, G.E.: Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. **20**(1), 14–22 (2012)
4. Hinton, G.E., Deng, L., Yu, D.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. **29**(6), 82–97 (2012)
5. Abdel, H.O., Mohamed, A.R., Jiang, H.: Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: International Conference on Acoustics, Speech and Signal Processing, pp. 4277–4280. IEEE, Kyoto, May 2012
6. Sainath, T.N., Mohamed, A.R., Kingsbury, B.: Deep convolutional neural networks for LVCSR. In: International Conference on Acoustics, Speech and Signal Processing, pp. 8614–8618. IEEE, Vancouver, May 2013
7. Zhang, Y., Pezeshki, M., Brakel, P.: Towards end-to-end speech recognition with deep convolutional neural networks. arXiv preprint arXiv:1701.02720, January 2017
8. Qian, Y.M., Woodland, P.C.: Very deep convolutional neural networks for robust speech recognition. In: Spoken Language Technology Workshop, pp. 481–488. IEEE, Berkeley, June 2017
9. Bahdanau, D., Chorowski, J., Serdyuk, D.: End-to-End attention-based large vocabulary speech recognition. In: International Conference on Acoustics, Speech and Signal Processing, pp. 4945–4949. IEEE, Shanghai, March 2016
10. Miao, Y.J., Gowayyed, M., Metze, F.: EESEN: end-to-end speech recognition using deep RNN models and WFST-based decoding. arXiv preprint arXiv:1507.08240, October 2015
11. Zhang, H., Bao, F., Gao, G.: Mongolian speech recognition based on deep neural networks. In: Sun, M., Liu, Z., Zhang, M., Liu, Y. (eds.) CCL 2015. LNCS (LNAI), vol. 9427, pp. 180–188. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25816-4_15
12. Tan, T., Qian, Y.M., Hu, H.: Adaptive very deep convolutional residual network for noise robust speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(8), 1393–1405 (2018)
13. Graves, A., Santiago, F., Gomez, F.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: International Conference on Machine Learning, pp. 369–376. IEEE, Pittsburgh, June 2006
14. Graves, A., Mohamed, A., Hinton, G.E.: Speech recognition with deep recurrent neural networks. In: International Conference on Acoustics, Speech and Signal Processing, pp. 6645–6649. IEEE, Hong Kong, April 2003
15. Hannun, A., Case, C., Casper, J.: Deep speech: scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014)
16. Amodei, D., Anubhai, R., Battenberg, F.: Deep speech 2: end-to-end speech recognition in English and Mandarin. arXiv preprint arXiv:1512.02595 (2015)
17. Wang, Y., Deng, X., Pu, S.: Residual convolutional CTC networks for automatic speech recognition. arXiv preprint arXiv:1702.07793, February 2017
18. Li, J., Zhang, H., Cai, X.Y.: Towards end-to-end speech recognition for Chinese Mandarin using long short-term memory recurrent neural networks. In: Interspeech 2015, pp. 3615–3619. IEEE, Berlin, September 2015
19. Zhou, S.Y., Dong, L.H., Xu, S., Xu, B.: Syllable-based sequence-to-sequence speech recognition with the transformer in Mandarin Chinese. arXiv preprint arXiv:1804.10752, June 2018
20. Zou, W., Jiang, D.W., Zhao, S.J., Li, X.G.: A comparable study of modeling units for end-to-end Mandarin speech recognition. arXiv preprint arXiv:1805.03832, May 2018
21. Dong, L.H., Xu, S., Xu, B.: Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In: International Conference on Acoustics, Speech and Signal Processing, pp. 4437–4441. IEEE, Calgary, April 2018

22. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980, July 2015
23. Srivastava, N., Hinton, G.E., Krizhevsky, A.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
24. Zhou, Z.H.: Machine Learning. Tsinghua University Press, Beijing (2016)
25. Simonyan, K., Andrew, Z.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2015)
26. Awni, Y.H., Andrew, L.M., Daniel, J.: First-pass large vocabulary continuous speech recognition using bi-directional recurrent DNNs. arXiv preprint arXiv:1408.2873, December 2014
27. Wang, D., Zhang, X.: THCHS-30: a free chinese speech corpus. arXiv preprint arXiv:1512.01882, December 2015
28. Zhang, L.M., Wang, Y.Z., Zhang, B.Q.: Chinese Mandarin recognition and improvement based on CTC criterion. Comput. Eng. (2019)

AQ1

Author Query Form

Book ID : **488092_1_En**

Chapter No : **29**

Please ensure you fill out your response to the queries raised below and return this form along with your corrections.

Dear Author,
During the process of typesetting your chapter, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query Refs.	Details Required	Author's Response
AQ1	Kindly provide the volume id and page range for Ref. [28], if possible.	