Hindawi
Complexity
Volume 2022, Article ID 6927400, 8 pages
https://doi.org/10.1155/2022/6927400

WILEY | Hindawi

*Research Article*

# DLD: An Optimized Chinese Speech Recognition Model Based on Deep Learning

**Hong Lei**,[1] **Yue Xiao**,[2] **Yanchun Liang**,[2] **Dalin Li**,[2] **and Heow Pueh Lee**[3]

[1]*Faculty of Data Science, City University of Macau, Macau 999078, China*
[2]*Zhuhai Laboratory of Key Laboratory for Symbol Computation and Knowledge Engineering of Ministry of Education, Zhuhai College of Science and Technology, Zhuhai 519041, China*
[3]*Department of Mechanical Engineering, National University of Singapore, Singapore 117575, Singapore*

Correspondence should be addressed to Yanchun Liang; ycliang@jlu.edu.cn and Dalin Li; lidalin1982@foxmail.com

Received 31 August 2021; Revised 26 February 2022; Accepted 15 April 2022; Published 2 May 2022

Academic Editor: Daniele Salvati



Speech recognition technology has played an indispensable role in realizing human-computer intelligent interaction. However, most of the current Chinese speech recognition systems are provided online or offline models with low accuracy and poor performance. To improve the performance of offline Chinese speech recognition, we propose a hybrid acoustic model of deep convolutional neural network, long short-term memory, and deep neural network (DCNN-LSTM-DNN, DLD). This model utilizes DCNN to reduce frequency variation and adds a batch normalization (BN) layer after its convolutional layer to ensure the stability of data distribution, and then use LSTM to effectively solve the gradient vanishing problem. Finally, the fully connected structure of DNN is utilized to efficiently map the input features into a separable space, which is helpful for data classification. Therefore, leveraging the strengths of DCNN, LSTM, and DNN by combining them into a unified architecture can effectively improve speech recognition performance. Our model was tested on the open Chinese speech database THCHS-30 released by the Center for Speech and Language Technology (CSLT) of Tsinghua University, and it was concluded that the DLD model with 3 layers of LSTM and 3 layers of DNN had the best performance, reaching 13.49% of words error rate (WER).

## 1. Introduction

As we all know, artificial intelligence has been developing rapidly, and the intelligent interaction between human and machine has also become a key research area. Speech recognition is one of the important technologies for realizing intelligent interaction between human and machine. Currently, almost all of the commercial speech recognition technologies are provided online, such as WeChat, Baidu, and Xunfei. When the user recognition content needs to be made confidential, this kind of online speech recognition technology makes its process unable to ensure the privacy of user's information. Therefore, the online speech recognition technology is not suitable for all scenarios. To solve this problem, this study will focus on the analysis of model for offline Chinese speech recognition.

Speech recognition technology is mainly composed of signal preprocessing, feature extraction, acoustic model, language model, and decoder. At present, the most popular feature extraction methods are Mel frequency cepstral coefficients (MFCC), Fbank, and spectrogram. Spectrogram is a picture that converts speech signals into information represented by dots of different colours, which can retain most information among these methods. Therefore, based on the characteristics of spectrogram, this paper uses this method as feature extraction method. Convolutional neural network (CNN) has shown outstanding performance in the field of image recognition, so in the process of speech recognition, CNN is widely used as the main part of acoustic model to learn the acoustic information contained in spectrogram and has been shown to achieve good results. The N-gram model, which is a contiguous sequence of $n$ items from a given sample of text or speech, is currently the

most widely used in the language model, which is to calculate the probability of the arrangement relationship between words.

The acoustic model and the language model are the two most important parts in Chinese speech recognition. The relationship between them is that the acoustic model outputs, Pinyin sequence, a Romanization of the Chinese characters based on their pronunciation, after training and learning, will be used as the input of the language model and finally outputs the text sequence.

Among the traditional speech recognition models, the GMM-HMM (Gaussian mixture model-hidden Markov model) model has been widely used as a very effective acoustic model. Mohamed et al. applied deep belief networks (DBN) for the first time to construct an acoustic model [1]. The tests on the TIMIT, a standard data set used for evaluation of automatic speech recognition systems of the optimal DBN acoustic model, had achieved a phone error rate of 23%. Compared with the GMM-HMM model, it had been proven that the recognition performance of the DNN-HMM (deep neural network-hidden Markov model) model had significantly improved [2, 3], so that deep neural networks began to replace the Gaussian mixture model in traditional acoustic models.

On the basis of DNN, Recurrent Neural Network (RNN), LSTM, CNN etc. could also improve the modeling ability. Graves et al. proposed Deep Bidirectional Long Short-term Memory RNNs. After proper regularization methods for end-to-end training and noise weighting, phoneme recognition on the TIMIT reached a test set error rate of 17.7% [4]. Tian et al. proposed a training framework for hierarchical training of hierarchical LSTM models and exponential moving average method. The character error rate of this training framework model was 14% lower than that of the model with similar real-time functions [5]. C. Z. Liu and L. Liu proposed the use of fractional order theory to process the activation functions of nodes in convolutional neural networks. Through calculation, they found that the reciprocal of fractional order could speed up the convergence of the Sigmoid function and reduce training time [6]. Yang and Wang added Fisher criterion and L2 regularization to the convolutional neural network, making the trained network weight and bias closer to the optimal value, and effectively alleviating the overfitting problem caused by the small amount of data. And using a new type log activation function, compared with the sigmoid function, it could further improve the accuracy of speech recognition [7].

The study by Sainath et al. found that the DNN-LSTM-CNN model is suitable for speech recognition and performs well. However, since Chinese speech recognition has complex features such as voice intonation, synonyms, etc., we replace DNN with DCNN to capture these high-dimensional features. In Liu deep convolutional neural network speech recognition paper, it is also highlighted that the DCNN model has better robustness and accuracy in Chinese speech recognition. Therefore, based on the research of these two papers, we believe that the DCNN-LSTM-DNN model has a more prominent performance in Chinese speech recognition.

In the past few years, deep neural networks (DNNs) have achieved great success on the task of large-vocabulary continuous speech recognition (LVCSR) compared to the Gaussian mixture model/hidden Markov model (GMM/HMM) systems. Both Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) have shown improvements over Deep Neural Networks (DNNs) in various speech recognition tasks. CNN, LSTM and DNN are complementary in modeling capabilities, as CNN is good at reducing frequency variation, LSTM is good at temporal modeling, and DNN is suitable for mapping features into a more separable space. Therefore, exploiting their complementarity by combining CNN, LSTM and DNN into a unified architecture can improve speech recognition performance [8]. In addition, the traditional speech signal recognition results are susceptible to noise interference. This problem can be solved by a multinoise speech recognition method based on deep convolutional neural network (DCNN), which can improve the signal-to-noise ratio of speech signals and improve the accuracy of speech recognition rate. Based on these theoretical foundations, we propose a DCNN-LSTM-DNN architecture suitable for Chinese speech recognition.

According to the above work, involving deep learning to improve the accuracy of audio recognition has become an important research direction. The present work focuses on optimizing the Chinese speech recognition model by deep learning methods. Firstly, the DCNN-CTC acoustic model is designed, and the effectiveness of this model is proved by experiments. Then, we combine several neural networks, i.e., DCNN, LSTM, and DNN, propose the DCNN-LSTM-DNN (DLD) acoustic model. We employed the Deep Convolutional Neural Network (DCNN) acoustic model as the basic acoustic model and used the statistical language model based on hidden Markov as the language model. Our mainly work is to optimize the DCNN acoustic model and try to improve the performance of the Chinese speech recognition model. After adjusting the accuracy of the algorithm many times, we proposed the DCNN-LSTM-DNN acoustic model, which we call DLD has the lowest WER. This model achieves the goal of improving the performance of Chinese speech recognition. According to the experimental results on THCHS-30, the word error rate (WER) of the DLD is 13.49%, which is about 7% lower than that of the DCNN-CTC.

In this paper, a novel DCNN-LSTM-DNN model was proposed, which combines three different network structure models of DCNN, LSTM and DNN for Chinese speech recognition, and this new acoustic model significantly improves the accuracy. We addressed the two main issues in the field of Chinese speech recognition. The first issue is the poor performance of the existing Chinese speech recognition model, while the second issue is most Chinese speech recognition is online, and there are few local speech recognition systems. We designed the new DCNN-LSTM-DNN acoustic model that added the BN layer and incorporated the LSTM and DNN network structure which effectively improve the model accuracy and created local Chinese speech recognition.

The approach we proposed solved three critical problems in the current existing research. At present, the existing research uses convolutional neural network in the training process, each iteration of the network needs to relearn the data distribution, which causes the problems of network training difficulty and slow network convergence. We proposed to use the BN algorithm to solve this problem. In convolutional neural networks, the BN layer is usually added after the convolutional layer. It standardizes the input data of each layer to make it a standard normal distribution. This method ensures the stability of data distribution and achieves the purpose of accelerating network training. Another challenge in current research is the disappearance of gradients during training. The solution we proposed is to add a Long Short-term Memory Network. This network structure not only has a memory function, but also effectively alleviates the problem of disappearance of gradients. Finally, we designed a DCNN-LSTM-DNN Chinese speech recognition model with batch normalization processing to make the model performance better.

## 2. Related Work

Our approach is based on recent excellent work that proposed to use a novel end-to-end age and gender recognition convolutional neural network (CNN) with a specially designed multiattention module (MAM) from speech signals [9]. This model uses MAM to extract spatial and temporal salient features from the input data effectively.

Our work is also inspired by the recent success of a stacked network with dilated CNN feature [10, 11], which is a one-dimensional dilated convolutional neural network (1D-DCNN) for speech emotion recognition (SER) that utilizes the hierarchical features learning blocks (HFLBs) with a bidirectional gated recurrent unit (BiGRU).

LSTM is a recurrent neural network architecture, which is mainly used to solve the problems of gradient disappearance and gradient explosion of traditional recurrent neural networks. Haim et al. [12] proposed a recurrent neural network structure based on a long short-term memory network (LSTM), which can use the parameters of the model to train the acoustic model. By comparing the experimental results of LSTM, RNN, and DNN models, it is found that the LSTM model has better recognition performance.

Convolutional neural network (CNN) is a well-known and widely used deep learning network structure. Convolutional neural network is gradually applied to the field of speech recognition due to its translation invariance. Abdel-Hamid et al. [3] proposed a limited-weight-sharing scheme that can better model speech features, which can better process speech features.

## 3. DCNN-LSTM-DNN Hybrid Acoustic Model

In this section, we propose our DCNN-LSTM-DNN hybrid acoustic model. First of all, we provide a brief review of the DCNN-CTC acoustic model that combines Connectionist Temporal Classification (CTC) and DCNN. Although DCNN-CTC achieves better performance than DCNN, its overall acoustic performance is still not good enough. In order to improve the offline Chinese speech recognition performance, we propose the DCNN-LSTM-DNN (DLD) acoustic model, DLD optimizes the previous network, adds batch normalization (BN) layers into DCNN, and insert DNN between LSTM and CTC. As a result, DLD reaches lower WER, and the speech recognition performance is improved.

*3.1. DCNN-CTC Acoustic Model.* The training process of the acoustic model belongs to supervised learning, which is necessary to know the label corresponding to each frame of the input data. Therefore, in the preparation stage of the training data, the audio data must be aligned with the corresponding pinyin sequence. An alignment rule for characters and audio is needed to ensure that it is effective even if different people's speaking rate are different. If all the audio manually are aligned manually, it will be unrealistic due to excessive workload and time-consuming when facing large data sets apparently. The CTC algorithm, proposed by Graves et al. [13], can solve the above problem very well. Its advantage is that when the model is introduced into the CTC algorithm, the training of the model only requires an input sequence and its corresponding output sequence, without the need for alignment and label.

Many scholars have proved that the acoustic model with CTC algorithm can effectively improve the recognition performance, at the same time, it can also present an end-to-end model training structure, reduces the difficulty of speech recognition [14–17]. Based on the DCNN acoustic model, the DCNN-CTC acoustic model is proposed, which consists of 10 convolutional layers and 5 maximum pooling layers, takes ReLU as the activation function of the convolutional layer, CTC algorithm as the loss function of the model. At the end of the DCNN-CTC, the pinyin sequence classified by the Softmax layer is the output. The specific structure and parameter settings are shown in Figure 1. The filter_size is set to 3, the pool_size is set to 2, and the number of feature maps is set to 32, 64, and 128, respectively.

Assuming that there are $m$ neurons in the $l-1$th layer and $n$ neurons in the $l$th layer, the linear coefficients $w$ of the $l$th layer form an $n \times m$ matrix $w^l$. The bias $b$ of the $l$th layer forms an $n \times 1$ vector $b^l$, the output $a$ of the $l-1$ layer forms a $m \times l$ vector $a^{l-1}$, the inactive linear output $z$ of the $l$th layer forms an $n \times l$ vector $z^l$, and the output $a$ of the $l$th layer forms an $n \times l$ vector $a^l$. Represented by matrix method, the output of the $l$ layer is

$$a^l = \sigma\left(z^l\right) = \sigma\left(w^L a^{l-1} + b^1\right). \tag{1}$$

*3.2. DCNN-LSTM-DNN Acoustic Model.* Based on DCNN-CTC acoustic model, we propose an optimization method that adds a batch normalization layer after each convolution operation and integrates the LSTM and DNN network structure behind the DCNN. The specific network structure is shown in Figure 2.

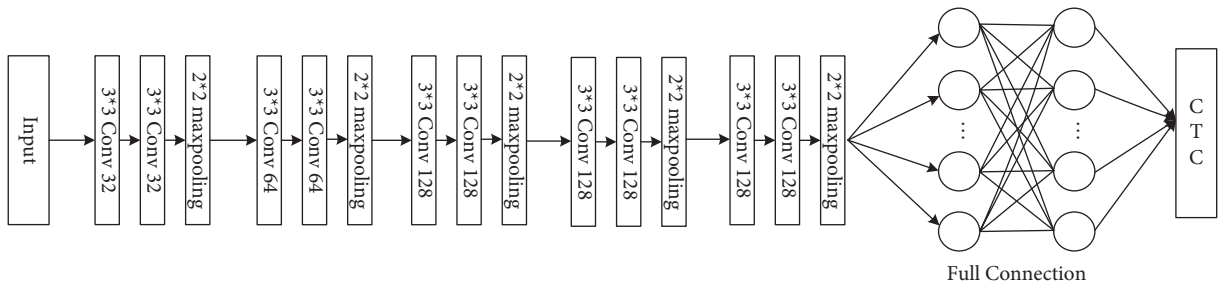


Figure 1: The structure of the DCNN-CTC acoustic model.

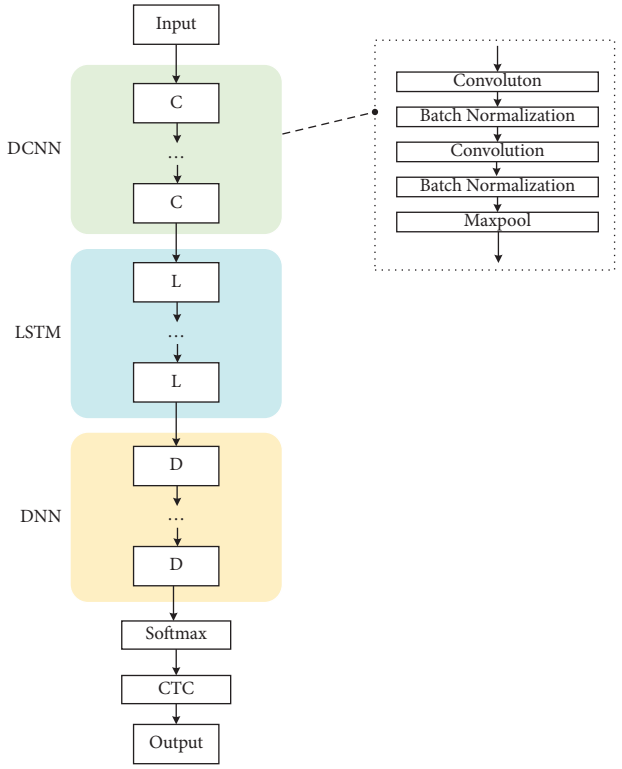


Figure 2: The structure of the DLD acoustic model.

Batch normalization layer can normalize the input data of each layer, make it easier and faster for the model to learn the laws in the data, the part of DCNN can reduce the number of parameters and get the key information during learn the input features of spectrogram. Then, the LSTM layer predicts the current time step information according to the previous time step information, and finally inputs the predicted results into DNN for classification to obtain the Pinyin sequence.

*3.2.1. DCNN Component.* Due to the characteristics of the translation invariance, CNNs are gradually applied to the field of speech recognition [18]. It plays an indispensable role in solving machine learning problems in fields involving high-dimensional data such as speech recognition and computer vision [19]. The main function of CNN is to collect key features in the data through convolution operation and at the same time eliminate features with less information through pooling operation. Therefore, convolution and pooling operations enable the model to learn data features more accurately and faster with fewer parameters.

Visual geometry group network (VGGNet) is very effective in image recognition, which is based on CNN [20, 21]. In order to realize speech feature extraction, our work first converts the audio data into the corresponding spectrogram, and then takes it as the input of acoustic model. We design a DCNN structure with 10 convolutional layers and 5 pooling layers by referring to the structure of VGGNet.

There are two commonly used activation functions in CNN, Sigmoid, and ReLU. Different activation functions are suitable for different data [21, 22]. In order to verify which of Sigmoid and ReLU would be more effective for voice information, we established Sigmoid-based network DCNN and ReLU-based network DCNN (ReLU), respectively.

The DCNN consists of 10 convolutional layers and 5 maximum pooling layers. The specific process of spectrogram feature extraction is convolution-convolution-max-pooling. Cross entropy loss function is used as the loss function of the model. At the end of the model, the results are classified through the Softmax layer.

The structure of DCNN (ReLU) acoustic model is the same as DCNN, and the only difference is that the activation function of the DCNN (ReLU)'s convolutional layers is ReLU, while the DCNN is Sigmoid.

*3.2.2. Batch Normalization (BN) Layers.* As one of the most widely used optimization methods in deep learning, batch normalization (BN) [23] can improve the performance and stability of neural networks. By adding BN layers, the impact of data shift and increase in the training process can be effectively resolved, and the training speed of the model can be accelerated while avoiding the disappearance of the gradient.

*3.2.3. Long Short-Term Memory Network Component.* Long short-term memory (LSTM) network is a neural network architecture with memory function and gating mechanism, which has the advantage of being better for the processing of time-related data. For the large-vocabulary speech recognition, the LSTM model had the best experimental results and the fastest convergence speed compared to the RNN model and the DNN model [12]. In essence,

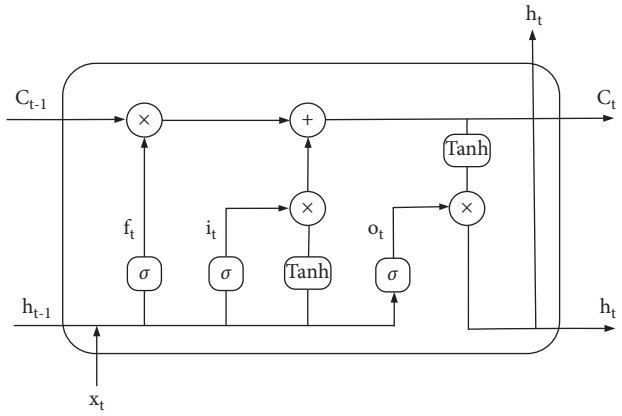


Figure 3: The structure of long short-term memory network.

speech recognition is a process of recognition based on the relevance of context front and back.

LSTM has a recurrent neural network architecture, which solves the problems of gradient disappearance and gradient explosion in traditional recurrent neural networks by adding a gate mechanism [24], including a forget gate, an input gate, and an output gate. The role of the three gates is to decide which information to discard, to keep, and to update.

As shown in Figure 3 [25], $x_t$ is the input information of the cell and $f_t$, $i_t$, and $o_t$ represent the output of the forget gate, input gate, and output gate, respectively. All of the gates use Sigmoid as the activation function, denoted by $\sigma$, while the cell unit uses Tanh. $h_{t-1}$ and $C_{t-1}$, $h_t$ and $C_t$ are the hidden states and cell states of the previous time step and the current time step, respectively. $w_f$, $w_i$, $w_o$, $w_c$ and $b_f$, $b_i$, $b_o$, $b_c$ represent the weight matrix and bias coefficient of the forget gate, input gate, output gate, and cell unit, respectively, then the specific calculation process is as follows:

$$\begin{aligned} f_t &= \sigma\left(w_f\left[x_t, h_{t-1}\right] + b_f\right), \\ i_t &= \sigma\left(w_i\left[x_t, h_{t-1}\right] + b_i\right), \\ o_t &= \sigma\left(w_o\left[x_t, h_{t-1}\right] + b_o\right), \\ C_t &= f_t \times C_{t-1} + i_t \times \tanh\left(w_c\left[x_t, h_{t-1}\right] + b_c\right), \\ h_t &= o_t \times \tanh\left(C_t\right). \end{aligned} \tag{2}$$

One problem with LSTM is that temporal modeling is done on the input features. However, advanced modeling of features can help clarify the underlying factors for changes in the input, which will make it easier to learn the temporal structure between successive time steps. Studies have shown that DCNNs can learn speaker-adaptive discriminatively trained features that eliminate variations in the input. Therefore, it would be beneficial to have several fully connected DCNN layers for the LSTM.

The specific implementation of the combined use of DCNN-LSTM is divided into three steps. First, we reduce the frequency variation of the input signal by passing the input through several convolutional layers. The size of the last layer of the DCNN is large due to the number of feature map temporal-frequency contexts. Therefore, we pass it to the LSTM layer before adding a linear layer to reduce the feature size and find that adding this linear layer after the DCNN layer reduces parameters without the loss of accuracy. Next, after frequency modeling, the DCNN output is passed to the LSTM layer, which is suitable for modeling the signal in time. Finally, after performing frequency and temporal modeling, we pass the output of the LSTM to several fully connected DNN layers. These higher layers are suitable for generating higher order feature representations of different classes that are easier to distinguish. Therefore, we believe that combining CNN-LSTM is helpful to improve the performance of Chinese speech recognition.

Table 1: Statistics of THCHS-30 database.

| Data set | Speaker | Male | Female | Age | Utterance | Duration (hour) |
|---|---|---|---|---|---|---|
| Training | 30 | 8 | 22 | 20–55 | 10893 | 27.23 h |
| Test | 10 | 1 | 9 | 19–50 | 2496 | 6.24 h |

Table 2: Specific environment configuration information for experiments.

| Environment | Version | Package | Version |
|---|---|---|---|
| Operating system | Ubuntu | python_speech_features | 0.6 |
| Python | 3.7.9 | SciPy | 1.5.2 |
| TensorFlow-GPU | 1.15.0 | NumPy | 1.19.2 |
| CUDA | 10.0.130 | Matplotlib | 3.3.2 |
| cuDNN | 7.6.5 | H5py | 2.10.0 |
| Keras | 2.3.1 | Wave | 0.0.2 |

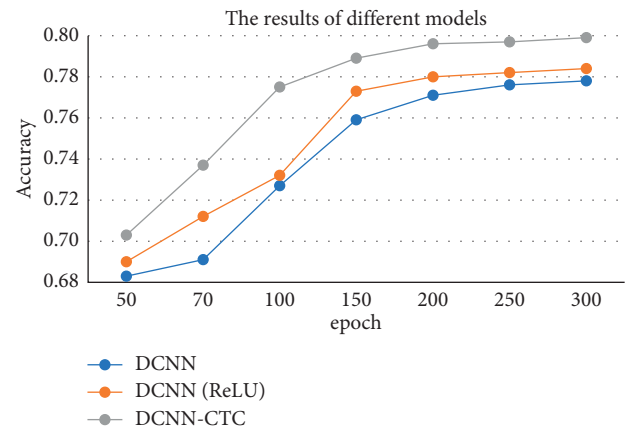


Figure 4: Relationship between the model accuracy and the iteration times.

*3.2.4. DNN Component.* DNN is composed of an input layer, hidden layers, and an output layer. The difference between DNN, CNN, and LSTM is that DNN is fully connection structure and does not include convolutional units or temporal associations. So, the fully connection properties can effectively map the input features into a separable space, which is helpful for data classification.

Table 3: Results of the network models on THCHS-30.

| Acoustic model | Convolutional layer activation function | Fully connected layer activation function | WER (%) |
|---|---|---|---|
| DCNN | Sigmoid | ReLU | 22.23 |
| DCNN (ReLU) | ReLU | ReLU | 21.76 |
| DCNN-CTC | ReLU | ReLU | 20.18 |

## 4. Experiments

*4.1. Training Data and Environment Setup.* In the experiments, we use the THCHS-30, an open source Chinese acoustic data. The THCHS-30 consists of audio news longer than 30 hours, as shown in Table 1, recorded by college students who can speak Mandarin fluently in quiet environments. The sampling rate of the recording is 16,000 Hz, and the sample size is 16 bits. The 1000 sentences (text prompts in recording) of THCHS-30 were selected from a large volume of news. In the whole data, 80% is divided into training data and the rest 20% is test set data.

All our experiments are implemented on the server with powerful computational capabilities, and Python is used as the programming language. We use TensorFlow to build an acoustic model and choose the version of TensorFlow-GPU to speed up the training speed. Among the packages that use during the experiment, python_speech_features and SciPy are used to complete the audio data processing, and matplotlib is to draw results diagrams. The specific configuration information is given in Table 2.

*4.2. Results of DCNN, DCNN(ReLU), and DCNN-CTC Acoustic Model.* In this part, we train DCNN, DCNN (ReLU), and DCNN-CTC, respectively, by using the training data set of THCHS-30. In the experiment, epoch is set to 50, 100, 150, 200, 250, and 300 to explore the change of accuracy of each model with the increase of the number of training rounds, as shown in Figure 4. It can be seen from the figure that with the increase of epoch, the accuracy of the model also increases. However, when epoch increases from 200 to 300, the improvement of the accuracy of the model begins to flatten. Therefore, we will not increase the value of epoch for training and take the training model when epoch is 300 as the final model for performance comparison.

After completing the training of the model, we use the test set in THCHS-30 to test DCNN, DCNN(ReLU), and DCNN-CTC acoustic model, and the test results are shown in Table 3. The WER of the DCNN on the test set is 22.23%. When the Sigmoid of convolutional layer of DCNN is replaced with ReLU, the WER of the DCNN(ReLU) is 21.76%, which is 0.47% lower than that of DCNN. Due to the import of CTC, DCNN-CTC has a better result among the three models that is 20.18%, which has improved by 2.05% compared with DCNN. According to the results in Table 3, we know that the import of CTC improves the performance of the acoustic model to a certain extent, and the recognition effect taking ReLU as the activation function is better.

*4.3. Results of DLD Acoustic Model.* We have trained the data on the network with different LSTM layers and DNN layers to find the best architecture based on DCNN-CTC. First, we test the DCNN-CTC acoustic model combined with LSTM to determine the optimal number of layers. Then, we test the DCNN-LSTM acoustic model combined with DNN and determine the best architecture of DLD acoustic model. The number of nodes of LSTM and DNN in each layer is different. Table 4 illustrates the number of nodes corresponding to each layer of the LSTM and DNN.

Table 4: Number of nodes for each layer of LSTM and DNN.

| Net | Layer no. | Number of nodes |
|---|---|---|
| LSTM | Layer 1 | 128 |
| | Layer 2 | 256 |
| | Layer 3 | 512 |
| | Layer 4 | 512 |
| DNN | Layer 1 | 512 |
| | Layer 2 | 1024 |
| | Layer 3 | 1048 |

Table 5: Results with LSTM of different layers on THCHS-30.

| Acoustic mode | WER (%) |
|---|---|
| DCNN + 2 LSTM | 14.53 |
| DCNN + 3 LSTM | 14.01 |
| DCNN + 4 LSTM | 14.03 |

Due to the memory function of LSTM, it can effectively process the text content related to time series, which also indicates that LSTM can play a good role in the acoustic model. Therefore, we test LSTM with different layers and get the best layer number of the model. Table 5 shows that the WER drops by 0.52% from DCNN + 2 LSTM to DCNN + 3 LSTM. However, when the number of LSTM layers increases to 4, the WER does not change significantly, and the speed of model training will be reduced. Thus, integrating LSTM can effectively improve the performance of audio recognition, and the layer number has an optimal value.

The fully connected DNN layer is used for classification, if the number of layers is set to be too many, the model parameters will increase, resulting in the decrease of training speed. Therefore, we only test the model combining two layers and three layers. From Table 6, on the basis of DCNN-3LSTM model, the WER of the model incorporating 2-layer DNN is 13.87%, while 3-layer DNN is 13.49%. In comparison, the recognition effect is improved by 0.38%. Therefore, we can determine that the performance of the DCNN-CTC acoustic model after integrating 3 layers of LSTM and 3 layers of DNN in turn has the best performance. In addition, we did a 10-fold cross-validation. First, we divide the entire training set $S$ into $k$ disjoint subsets. Assuming that the number of training examples in $S$ is $m$, each subset has m/10 training examples, and the corresponding

Table 6: Results of DNN with different number of layers on THCHS-30.

| Acoustic mode | WER (%) |
|---|---|
| DCNN + 3 LSTM + 2 DNN | 13.87 |
| DCNN + 3 LSTM + 3 DNN | 13.49 |

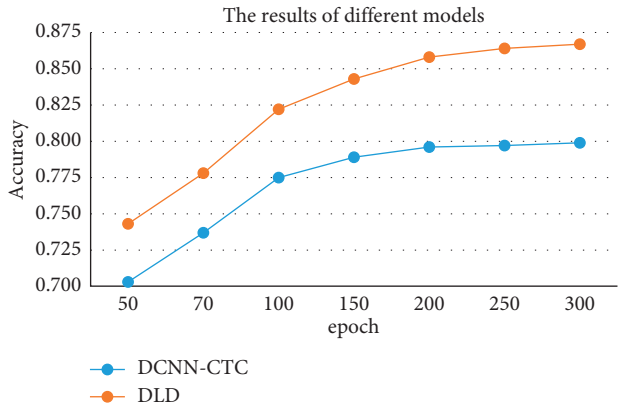


Figure 5: The accuracies of DCNN-CTC and DLD.

subsets are called {s1, s2, . . ., s10}. Then, we take 2 of the divided subsets as the test set each time and the other 8 as the training set to train the model. In deep learning, we need a validation set that is independent of training and testing sets for parameter selection. In the divided two test sets, we select one as the validation set to detect whether the model has overfitting, underfitting problems, and the best training rounds of the model during the training process. Then put this model on the test set to get the classification rate. Finally, calculate the average of the classification rates obtained 10 times as the true classification rate of the model. This method makes full use of all samples, but the computation is cumbersome, requiring 10 training and 10 testing. We found that the DLD acoustic model has the best performance with 3 layers of LSTM and 3 layers of DNN.

From the above experimental results, the final WER of the DLD acoustic model is 13.49%, while that of the DCNN-CTC acoustic model is 20.18%. In contrast, the WER of the acoustic model has dropped by about 7%. It also can be clearly seen from Figure 5 the change trend of the accuracy of the DLD acoustic model and the DCNN-CTC acoustic model.

*4.4. Comparison with Other Acoustic Models.* The SE-MCNN-CTC acoustic model [26] was proposed by Zhang et al. that consisted of MCNN and SENet (squeeze-and-excitation Nnetworks) on the basis of the DCNN-CTC acoustic model. MCNN is a multipath convolutional neural network, which combines multiple branch networks in parallel.

The DCNN-BGRU-CTC acoustic model [27] was proposed by Lv, which integrated bidirectional gate recurrent unit (GRU) into DCNN-CTC. The bidirectional GRU is used for capturing the context information, and the CTC is used as the loss function for an end-to-end training.

Table 7 shows the comparison results among DCNN-CTC designed at the beginning, the optimized DLD, and another two models. According to the table, our DLD has a 14.28% WER, which is the lowest among all the results. At the same time, it also proves that our DLD has better recognition performance on THCHS-30 than the other models.

Table 7: WER of different acoustic model on THCHS-30.

| Acoustic model | WER (%) |
|---|---|
| DCNN-CTC | 20.18 |
| SE-MCNN-CTC | 23.05 |
| DCNN-BGRU-CTC | 18.13 |
| DLD | 14.28 |

## 5. Conclusions

In this paper, we mainly optimize the acoustic model in Chinese speech recognition. Based on the DCNN-CTC acoustic model, we propose the deep convolutional neural network-long short term memory-deep neural network (DCNN-LSTM-DNN, DLD) acoustic model. This model adds a batch normalization (BN) layer, a long short-term memory network (LSTM), and a deep neural network (DNN) structure. As a result, the WER of the acoustic model is reduced by 7%. As shown in our work, batch normalization and integration of LSTM and DNN is found to improve the training speed and accuracy of the acoustic model. However, since the THCHS-30 dataset we used is recorded in a quiet environment by college students who can speak Mandarin fluently, its accuracy will decrease when it recognizes accented speech or when it comes to noisy data in the actual application. In the future work, we will try to introduce a self-attention mechanism into the language model to strengthen the language model's ability on learning homophones, so as to further improve the Chinese speech recognition accuracy.

## Data Availability

The data called THCHS-30 used to support this research are completely free to academic users. THCHS-30 is a classic Chinese speech data set, which provides a toy database for new researchers in the field of speech recognition (https://www.openslr.org/18/).

## Conflicts of Interest

The authors of this paper declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (61972174), the Science and Technology Planning Project of Guangdong Province (2020A0505100018), the Guangdong Universities' Innovation Team Project (2021KCXTD015), and Guangdong Key Disciplines Project (2021ZDJS138).

## References


[1] A. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Proceedings of the Nips workshop on deep learning for speech recognition and related applications*, vol. 1, no. 9, p. 39, Whistler, BC, Canada, December 2009.

[2] A. 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, 2011.

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

[4] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of the 2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649, Vancouver, Canada, May 2013.

[5] X. Tian, J. Zhang, Z. Ma et al., "Deep LSTM for large vocabulary continuous speech recognition," 2017, https://arxiv.org/abs/1703.07090.

[6] C. Z. Liu and L. Liu, "Convolutional neural network optimization algorithm in speech recognition[J]," *Journal of Harbin University of Science and Technology*, vol. 21, no. 3, pp. 34–38, 2006.

[7] Y. Yang and Y. Wang, "Speech recognition based on improved convolutional neural network algorithm[J]," *Journal of Applied Acoustics*, vol. 37, no. 06, pp. 1–7, 2018.

[8] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proceedings of the 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4580–4584, IEEE, South Brisbane, QLD, Australia, April 2015.

[9] A. Tursunov, J. Y. Mustaqeem, J. Y. Choeh, and S. Kwon, "Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms," *Sensors*, vol. 21, no. 17, p. 5892, 2021.

[10] K. Soonil, "1D-Cnn: Speech emotion recognition system using a stacked network with dilated CNN features[J]," *CMC-COMPUTERS MATERIALS & CONTINUA*, vol. 67, no. 3, pp. 4039–4059, 2021.

[11] S. Kwon, "Optimal feature selection based speech emotion recognition using two-stream deep convo-lutional neural network[J]," *International Journal of Intelligent Systems*, vol. 36, no. 9, pp. 5116–5135, 2021.

[12] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," 2014, https://arxiv.org/abs/1402.1128.

[13] 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*, pp. 369–376, Pittsburgh, Pennsylvania, June 2006.

[14] S. H. Wang, "Research on Speech Recognition Based on Deep Learning Neural network," Guilin University of Electronic Technology, Guilin, China, Master, 2015.

[15] S. Wang, "Tibetan Lhasa Acoustic Model Based on LSTM-CTC Speech Recognition system," Northwest University for Nationalities, Lanzhou, China, Master, 2019.

[16] L. M. Zhang, Y. Z. Wang, and N. B. Zhu, "Mandarin recognition and improvement based on CTC criteria[J]," *Computer Engineering*, vol. 45, no. 6, pp. 249–253, 2019.

[17] Y. P. Jiang, "ASR research based on CTC," Xinjiang University, Ürümqi, China, Master, 2019.

[18] L. C. Yang and B. Yu, "Convolutional networks for images, speech, and time series[J]," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[19] J. Ephrath, M. Eliasof, L. Ruthotto, E. Haber, and E. Treister, "LeanConvNets: low-cost yet effective convolutional neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 4, pp. 894–904, 2020.

[20] J. Gu, Z. Wang, J. Kuen et al., "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[22] J. H. Liu, "Chinese speech recongition based on deep convolution nerual networks," TaiYuan University of Technology, Taiyuan, China, Msater, 2019.

[23] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International conference on machine learning. PMLR*, pp. 448–456, Paris, France, April 2015.

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

[25] S. Wang, "Tibetan Lhasa acoustic model based on LSTM-CTC speech recognition system," Northwest Minzu University, Lanzhou, China, Master, 2019.

[26] W. Zhang, M. H. Zhai, Z. L. Huang, W. Li, and Y. Cao, "Acoustic model of Chinese speech recognition based on SE-MCNN-CTC[J]," *Journal of Applied Acoustics*, vol. 39, no. 2, pp. 223–230, 2020.

[27] K. R. Lv, "Research on end-to-end speech recognition methods based on Language model," Jilin University, Changchun, China, Master, 2020.