Chinese Speech Recognition System based on Deep Learning

Pengyuan Shao*
Shanghai Pinghe School, Shanghai, China

*Corresponding author e-mail: hzhe_wjdi@163.com

Abstract. This paper builds a complete Chinese speech recognition system, including acoustic model and linguistic model, which can recognize the input audio signal into Chinese characters. The system realizes the modeling of acoustic model and linguistic model in speech recognition based on deep framework, of which the acoustic model is CNN-CTC and linguistic model is transformer. The data set uses THCHS-30, which refers to 30-hour Chinese speech database of Tsinghua University. The experimental results show that the Chinese speech recognition system based on deep learning achieves 90% accuracy on the test set and has an excellent effect on Mandarin speech recognition in quiet environment.

1. Research Status and Basic Constitution

1.1. Research Status of Speech Recognition
After decades of development and exploration, people have finally broken through the three major obstacles of large vocabulary, continuous speech and speaker independent in the laboratory. Integrated these three characteristics into one system for the first time, and determined the mainstream position of statistical method and model in speech recognition and processing accordingly. From the perspective of acoustic recognition, the speech sequence modelling approach HMM(Hidden Markov Model) based on large-scale speech data of multiple speakers and Markov chain, effectively solves the characteristics of short-term stability and long-term time variation of speech signal, and constructs the sentence model of continuous speech according to some basic modeling units, which achieves high modeling accuracy and flexibility[1] [2].

1.2. Basic Constitution of Speech Recognition System
The typical implementation scheme of speech recognition system is that the input analog speech signal should be preprocessed firstly, including pre-filtering, sampling and quantification, windowing, endpoint detection, pre-emphasis and so on. After the speech signal is preprocessed, the next important link is the characteristic parameter extraction. The requirements of characteristic parameters are as follows: 1. The extracted characteristic parameters can effectively represent phonetic features and have excellent discrimination. 2. There is excellent independence between the parameters of each order. 3. The characteristic parameters should be calculated conveniently, and it is better to have an efficient calculation method to ensure the real-time realization of speech recognition [3].

In the training stage, after processing the characteristic parameters, each vocabulary obtains a model, which is saved as the template. In the recognition stage, the speech preference obtains the
speech parameters through the same channel, generates the test template, matches with the reference template, and takes the reference template with the highest matching score as the recognition result. At the same time, with the help of a lot of prior knowledge, the accuracy of recognition can be improved.

Based on this, this paper attempts to use deep learning method to solve the accuracy of standard Mandarin speech recognition in quiet environment, including establishing the acoustic model of CNN+CTC and the linguistic model based on Transformer.

2. Method and Principle Introduction

2.1. Composition and Structure of Convolutional Neural Network (CNN)

The neural network is a hierarchical structure, which is usually composed of input layer, hidden layer, and output layer, as shown in the following figure:

![Artificial neural network structure](image)

**Figure 1.** Artificial neural network structure

2.2. Connectionist Temporal Classification (CTC)

The full name of CTC algorithm is: Connectionist temporal classification. CTC introduces blank (the frame does not have a predicted value). One spike in an entire segment of speech corresponds to each predicted classification, and the other positions without spike is considered as blank. For a segment of speech, the final output of CTC is the sequence of spike and does not care how long each phoneme lasts.

![Classification output of ‘Nihao’ speech signal after introducing CTC](image)

**Figure 2.** Classification output of ‘Nihao’ speech signal after introducing CTC

2.3. Transformer [4]

The model structure is as follows:
Figure 3. The transformer architecture

3. Data Set Introduction
THCHS-30 Speech Database. This database is called TCMSD (Tsinghua University Continuous Mandarin Speech Database). THCHS-30 is recorded through a single carbon microphone in a quiet office environment for more than 30 hours. Most of the people that participate in the recording are college students with fluent Mandarin.

Table 1. Specific information of data set in database (THCHS-30) (2)

| Data set   | Audio length (h) | The number of sentences | The number of words |
|------------|------------------|-------------------------|---------------------|
| Train      | 25               | 10000                   | 198252              |
| Development| 2:14             | 893                     | 17743               |
| Test       | 6:15             | 2495                    | 49085               |

4. Experimental Design and Process

4.1. Acoustic Model Building
The training input is the time-frequency graph, and the label is the corresponding pinyin label as follows:

Figure 4. The speech recognition model adopts the structure of CNN+CTC
The components of the model include: the convolutional layer with the size of 3*3. The activation function adopts 'relu'. The convolutional filling method is ‘same’. Normalized layer; Maximum pooling layer; Fully connected layer and CTC loss function.

4.2. Linguistic Model Building
The model uses self-attention, which includes:
- Layer norm layer. The neuron input in this layer has the same mean and variance, and different input samples have various means and variances. The embedding layer transforms positive integers (subscript) into vectors with fixed size, so as to explore which words have mutual similarity in higher space, and can excavate the semantic relations between words [5]. Multihead layer is the core idea of self-attention. It carries on attention mechanism with itself, but the original input will be mapped to eight different spaces for calculation through linear transformation before this, and finally they are connected together. Feedforward layer with two fully connected layers, uses the convolution simulation to accelerate operation. The dense layer can also be used.

5. Experimental Result Analysis

5.1. Data Analysis

5.1.1. Data Preprocessing. Input is the input audio data, which requires to be transformed to spectrogram data and carry on recognition through the ability of cnn to process the picture. The process is as follows: read audio file, construct Hamming window, data subframe length: 25ms, frame shift: 10ms, subframe windowing, and finally Fourier transform.

![Figure 5](image)

**Figure 5.** The left figure is the randomly read audio file, and the right figure is the corresponding spectrogram

5.2. Model Training
In the training process, we set batch_size as 4, that is, the sample size of each training is 4. The number of iterations is 2500, and the epoch used is 20. The training data is prepared and shuffle is designed to disrupt the training data sequence.

5.3. Model Test
Our model is generally realized by python. By using Thchs-30 corpus, the total training time is more than 5 hours and 20 epoch trainings have been carried out. The average error falls from 4.66% in the first time to 1.00% in the 20th. The training accuracy has been obviously improved. In the subsequent test set, the test results of the model show that the word error rate is 1.74%, which is very small. The training result of this model is very successful.
Figure 6. Average error after data training

At the same time, we use any audio data to test model output in the data set. By comparing the output result with the original result, it is found that our model based on CNN+CTC can transform speech signal into pinyin label very well.

Figure 7. Comparison between the output result of the model and the original result

6. Conclusion and Prospect

This paper focuses on the Chinese speech recognition system in python based on convolutional neural network and transformer linguistic model. The model is built by using the transformer, which is the most advanced linguistic model self-attention system. The training time is reduced by the parameter sharing mechanism in cnn convolutional neural network and the blank in ctc model (The frame has no predicted value). The model can recognize statements in speech environment with certain complex syntactic structure and quiet background, so as to achieve the basic accuracy of Chinese recognition.

Based on the limitations of hardware devices, we only use a relatively small free and open source data set, that is, THCHS-30. The characteristic of this data set is the standard Mandarin recording in a quiet environment with a single person. Therefore, the dialogue of multi-people in the real life scene based on the noisy environment will be the development direction of our speech recognition model in the future. Once the technology of the model is mature, major breakthroughs will be made in the fields of intelligent sound control, security and protection, man-machine conversation and so on, which will help human life environment become more convenient and safe.

References

[1] He Xiangzhi. Research and Development of Speech Recognition [J]. Computer and Modernization, 2002 (3): 3-6.
[2] Zhang Jianhua. Study on Speech Recognition Application based on Deep Learning [D]. Beijing
University of Posts and Telecommunications, 2015.

[3] Zhang Deliang. Deep Neural Networks for Chinese Speech Recognition [D]. Beijing Jiaotong University, 2015.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in neural information processing systems. 2017: 5998-6008.

[5] Bian J, Gao B, Liu T Y. Knowledge-powered deep learning for word embedding [C]//Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, 2014: 132-148.