ISSN 2616-5775 Vol. 6, Issue 2: 23-27, DOI: 10.25236/AJCIS.2023.060204

Research on dialect speech recognition based on DenseNet-CTC

Yijie You[a,*], Xiangguo Sun

Mechanical Engineering College, Sichuan University of Science and Engineering, Yibin, China
[a]Amon_you@163.com
**Corresponding author*

Abstract: *At present, most of the speech recognition research is based on a wide range of regions, and there are few studies on speech recognition of urban dialects. The mainstream speech recognition methods are mostly based on ResNet network, using ResNet network as acoustic model and N-gram as language model. In this study, DenseNet is used as the basic network, and the data set of Zigong dialect subdivided by Sichuan dialect is taken as the research object of speech recognition. DenseNet-BiGRU + CTC is constructed as the acoustic model of speech recognition, and RNN is used as the speech recognition model of language model. Experiments show that the speech recognition model using DenseNet network as the basic network has higher accuracy than the model based on ResNet. Compared with the GRU-CTC network word error rate (WER) decreased by 3 %, compared with the DPCNN-Attention-CTC speech recognition method error rate decreased by 5 %.*

Keywords: *Dialect, Speech Recognition, DenseNet, CTC, Acoustic Model*

1. Introduction

Auto Speech Recognition (ASR) converts the input digital speech signal into text form output through machine learning or deep learning algorithm model. The output text is the content that is considered by the model to be close to the maximum probability of the original speech signal. Speech recognition belongs to the field of natural language processing. It originated from the 10 English-based digital speech recognition systems studied by Davis et al.in 1952. After the related algorithms of machine learning were widely used, the speech recognition model based on support vector machine (SVM) and hidden Markov model (HMM) appeared [1]. Until 2010, deep learning began to be widely used in the field of speech recognition. At the same time, the performance of computers and other hardware has reached a certain height, which can support a large number of calculations required for deep learning. Therefore, a large number of researchers began to try various deep learning models in order to achieve higher recognition accuracy in speech recognition. For example, the acoustic model uses CNN-HMM [2], DNN [3,4] etc., and the RNN model is used as the language model for speech recognition research.

However, most of the current research on speech recognition is based on the more classic deep learning network. In order to obtain higher accuracy, the number of layers of the network will be deeper, and the hardware equipment required for training will have higher requirements. The training process will be more difficult, and there are relatively few studies on speech recognition of more subdivided dialects. Based on the rapid development of the current society, the population flow is greater, and the dialect will bring a lot of inconvenience in the process of communication. Therefore, in this study, a speech recognition model based on DenseNet, BiGRU network and CTC is proposed. Compared with the model used in other studies, it has smaller network parameters and training difficulty.

2. Data set preparation

The research object is Zigong dialect, a branch of Sichuan dialect system. Zigong dialect has a clear understanding of Blade-alveolars and rolling tongues. Zigong dialect classifies all the entering tone characters of ancient Chinese into falling tone, and retains the first tone of Mandarin. Two tones change into three tones, three tones change into four tones, and four tones change into two tones contains most of the pronunciation of Zigong dialect.

In the research, XiaoMi 10 Android mobile device is used as the audio acquisition device. The

ISSN 2616-5775 Vol. 6, Issue 2: 23-27, DOI: 10.25236/AJCIS.2023.060204

collected speech is a dual-channel signal, and then Audition is used to preprocess the speech signal. In the preprocessing stage, it is necessary to remove the blank audio, noise, noise and other useless audio in the speech signal. The collected speech signal is divided into short-term audio of 3-5s, and the integrity of the speech is guaranteed as much as possible when the speech is segmented. After segmentation, 1650 voices are obtained. After the calculation process shown in Figure 1, the feature map of the speech signal can be obtained. At present, the widely used speech features are spectrogram, FBank and MFCC [5,6]. FBank and MFCC have more Mel filter banks than the spectrogram, and filter out the part of the speech signal that the human ear cannot perceive [7]. In this study, the spectrum map is used as the input feature of the neural network. After extracting the features, it is necessary to mark the feature map of the dialect. The mark document is txt. The document needs to mark the Chinese character content in the speech feature map, the pinyin corresponding to the character, and the phoneme corresponding to the pinyin. It is also necessary to make a dialect corresponding vocabulary for speech recognition decoder to translate the recognition results of the acoustic model into Chinese character.

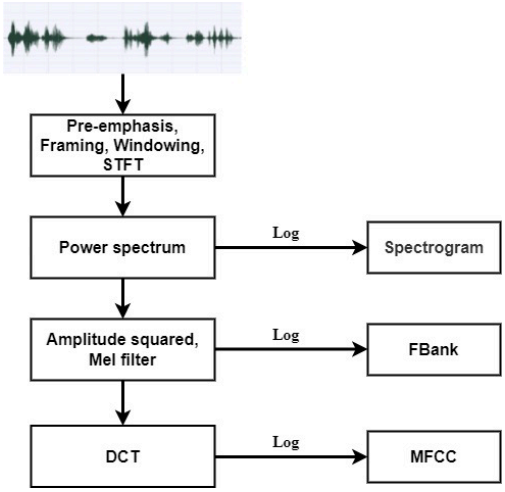

Figure 1: Process of feature extraction

The sex ratio of speakers in the data set is 2:1, of which male speakers account for 20%. Considering that the final neural network model may have poor generalization performance for gender, the tone modulation function of Audition software is used to increase and decrease the male and female voices respectively. The male voice can simulate the female voice after the rising tone, and the female voice can simulate the male voice after the falling tone, so the final data set have 3300 Zigong dialect voices.

3. Deep learning model

3.1. DenseNet-BiGRU model

The Dense Connection Network (DenseNet) used in the research is a new convolutional neural network structure proposed by Huang et al.in 2017. The dense connection network is named after the dense block in its structure. Each convolutional layer is connected with a dense block, and the dense connection is shown in Figure 2. The residual network directly adds the features before and after the calculation, while the dense connection network concatenates the features before and after the calculation. Dense connection is applied to all layers of the dense connection network, and the connection between layers is dense connection.

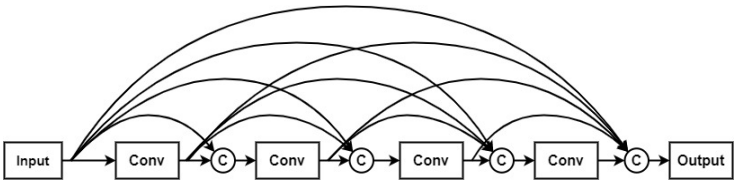

Figure 2: Dense block

The Gated Recurrent Unit and the Long Short-Term Memory network (LSTM) belong to the recurrent neural network (RNN). The difference between the two is that the overall structure of the gated recurrent unit is more concise than the latter, so the former can converge faster during training. On smaller data

ISSN 2616-5775 Vol. 6, Issue 2: 23-27, DOI: 10.25236/AJCIS.2023.060204

sets, the Gated Recurrent Unit (GRU) can provide the same accuracy as the latter. Suppose the input time series is$[t_1, t_2, t_3 \ldots\ldots t_{n-2}, t_{n-1}, t_n]$, the hidden states of the input of the forward gated recurrent unit in the first layer correspond to the input sequence, while the hidden states of the reverse gated recurrent unit in the second layer are opposite to those in the first layer. The input corresponding to the first hidden state t_1 on the right of the second layer is t_n. The second layer is transferred and calculated from right to left. According to the reset gate and update gate, the state of the previous hidden unit is judged, and whether it is transferred to the next unit is judged. The bidirectional gated recurrent unit structure is shown in Figure 3.

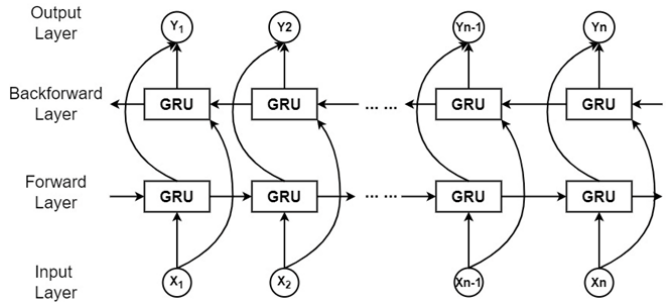

Figure 3: Architecture of BiGRU

3.2. Language Model

The RNN model is used as the language model in the study. The traditional N-gram model is divided into 1-gram grammar, 2-gram grammar and 3-gram grammar model. A statistical model is established according to the n-gram grammar in the label. In the final decoding process, the final speech recognition result is decoded according to the vocabulary with higher frequency. However, the traditional N-gram model has certain limitations. In the process of speaking, all words have a correlation. If the RNN model is used to train a language model, a relevant model can be established based on the text information before and after. The information in the original audio signal can be decoded more accurately, and the use of the RNN model as a language model can reduce the size of the entire speech recognition model.

3.3. CTC Decode

Connection temporal classification (CTC) is a loss function calculation method, which can reduce the complicated work of label division and alignment in speech recognition [8,9]. When the speech recognition model is trained, only the audio sequence is input into the model, and then the character label corresponding to the input audio is given. The model can start to learn by itself and finally get one. CTC outputs a probability space, which is the probability of each character at a certain time, so it is necessary to add a decoding search algorithm to participate in the training process. The search algorithm is used to find a route in the probability space according to the specified requirements and constraints. The characters on the route are the final output recognition results. There are three methods for CTC decoding. Beam Search is used as the search decoding method in the study. CTC decoding is applied to the acoustic model of DenseNet-BiGRU.

4. Results and discussion

4.1. Evaluation criteria

The study uses the sub-error rate as the evaluation standard of the speech recognition model. The word error rate is calculated as shown in Equation (1). By calculating the proportion of the words that need to be inserted, replaced, and deleted in the final recognition result in the original sentence of the tag sentence.

$$\mathrm{WER} = 100\% * \frac{S+D+I}{N} \quad (1)$$

4.2. Results

The training feature used in the experiment is the spectrum map. After feature extraction and

ISSN 2616-5775 Vol. 6, Issue 2: 23-27, DOI: 10.25236/AJCIS.2023.060204

calculation, the collected speech signal is divided into short-term audio in the range of 3-5s, in which the interval of the number of characters in the speech content is 60-98 characters. The speech data set introduced above is divided according to the ratio of 7:1:1:1, of which 7 is the proportion of the training set, and the remaining data is used as the training result of the test model. When the acoustic model is trained, the initial learning rate is set to 0.003, the training batch is 16, the training period is 140, and the weight is attenuated to 1e-4. The early termination strategy is adopted to prevent the model from overfitting.

The RNN language model is trained with labels, and then combined with the trained acoustic model to output the character results of speech recognition. The whole speech recognition model is also tested with the test set divided in the previous text, and the word error rate is obtained as shown in Table 1. The average word error rate is 17.91 %.

Table 1: ASR model test results

Number	WER(%)
1	17.62
2	18.83
3	17.47

Compared with the speech recognition methods in article [11] and article [10], the method used in this study has obtained higher recognition accuracy. From Table 1, it can be seen that one of the word error rates of the three test sets has a higher word error rate than the other two groups. According to the distribution of the data set, there is no limit to the acquisition environment noise of the speech during data collection. Therefore, when the data in the data set is randomly selected as the test set, more data with large background noise will be extracted, resulting in a higher word error rate in the second set of test sets.

5. Conclusion

In the research, based on the dense connection network, the RNN network and CTC decoding are combined to form the acoustic model part of speech recognition. The RNN model is used as the language model modeling of speech recognition, which is most applied to the city-level dialect speech recognition. The city-level dialect data set was created. The final trained speech recognition network based on DenseNet-CTC obtained a word error rate of 17.91 %, compared with the DPCNN-Attention-CTC in literature [11] 22.9 % and the GRU-CTC speech recognition method in literature [10]. The word error rate decreased by 5 % and 3 %.

References

[1] Hartmann W, Hsiao R, Tsakalidis S. Alternative networks for monolingual bottl e-neck features[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017: 5290-5294.

[2] Mukhedimham Yiminjiang, Aiskar Aimudura, Mijiti Abrimiti. Uyghur speech recognition based on CNN-HMM and RNN [J]. Modern electronic technology, 2021,44 (11):172-176. DOI:10.16652/j.issn.1004-373x.2021.11.036.

[3] Krishna G, Tran C, Carnahan M, et al. Advancing speech recognition with no speech or with noisy speech[C]//2019 27th European Signal Processing Conference (E U-SIPCO). 2019: 1-5.

[4] Kang J, Zhang W Q, Liu J. Gated convolutional networks based hybrid acoustic models for low resource speech recognition[C]//2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 2017: 157-164.

[5] Nancuoji, Zhuoma, Dugecao. Tibetan speech recognition based on BLSTM and CTC [J]. Journal of Qinghai Normal University (Natural Science Edition), 2019,35 (04):26-33.DOI:10.16229/j.cnki.issn1001-7542. 2019. 04.005.

[6] Sadhu S, Li R, Hermansky H. M-vectors: Sub-band Based Energy Modulation Fea-tures for Multi-stream Automatic Speech Recognition[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019: 6545-6549.

[7] Yuliani A R, Sustika R, Yuwana R S, et al. Feature transformations for robust speech recognition in reverberant conditions[C]//2017 International Conference on Co m-puter, Control, Informatics and its Applications (IC3INA). 2017: 57-62.

[8] Pan Yuecheng, Liu Zhuo, Pan Wenhao, Cai Dianlun, Wei Zhengsong. An end-to-end Mandarin speech

ISSN 2616-5775 Vol. 6, Issue 2: 23-27, DOI: 10.25236/AJCIS.2023.060204

recognition method based on CNN/CTC [J]. Modern information technology,2020,4(05):65-68.DOI:10.19850/j.cnki. 2096-4706.2020.05.019.

[9] Yang Deju, Ma Liangli, Tan Linshan, Pei Jingjing. End-to-end speech recognition based on gated convolutional network and CTC [J]. Computer engineering and design, 2020, 41(09): 2650-2654. DOI:10.16208/j.issn1000-7024.2020.09.037.

[10] Dong Jiaren, Liu Guangcong. Research on speech recognition method based on GRU-CTC hybrid model [J].Modern computer. 2019(26):13-16.

[11] Hu Li, Huang Hongquan, Liang Chao, Song Yueyang, Chen Yanming. End-to-end speech recognition based on dual-channel CNN [J]. Sensors and microsystemms.2021, 40(11):69-72+83.DOI:10.13873/J.1000-9787 (2021)11-0069-04.