Design and Implementation of Cyrillic Mongolian Speech Input System for Thyroid Ultrasound Report

Galbadrakh Gantumur*
School of Computer Science and Engineering, Donghua University, Shanghai 201620, China

*Corresponding author e-mail: galbadrakhgantumur@yahoo.com.au

Abstract. This paper introduces a Cyrillic Mongolian speech input system for ultrasound examination report. The system uses the developing Mongolian speech recognition technology to complete the transformation from manual input to speech input system and speech input system for ultrasound examination is developed and designed. In view of the current situation that the report system of hospitals cannot be effectively and quickly input in human-computer interaction, the description standard of ultrasound examination is low, the operation of the system is relatively complex, and the retrieval process is inefficient. Experimental results on Mongolian speech corpus tasks show that the WER in Cyrillic Mongolian speech recognition is reduced by 18.97% and 3.55% respectively.

1. Introduction

If the computer is able to automatically recognize the doctor's language and generate a thyroid ultrasound report, it will greatly improve the doctor's work efficiency and reduce patient's waiting time. In recent years, with the advancement of computer technology and speech recognition technology, this idea has the potential to be transformed into reality. Foreign countries had been using speech recognition technology in the medical field for many years [1]. The application of Mongolian speech recognition technology in the Mongolian medical related industry is basically blank, and the corresponding independent intellectual property rights and patents are few, and there is no relevant technical standard. In case speech recognition technology is introduced during the diagnosis process, the doctor will perform speech recording and generate a thyroid ultrasound report, which will help to improve. The speech recognition based on Chahar Mongolian began in 2006 and was established in [2]. In literature review [3,4], the acoustic model based on neural network has been widely used in Chahar Mongolian language ASR and has achieved remarkable promotion. In recent years, Hybrid Frame Neural Networks (HFNs) have achieved significant improvement in [5], and the WER score of the optimal system is 6.99%. However, the research of Halkha Mongolian speech recognition system is still in its infancy. Zhang Hui, for the first time, applied the neural network to the acoustic model of Mongolian speech recognition system, which made a breakthrough in speech recognition performance and made the industrialization of Mongolian speech recognition system possible. Yonghe Wang and Fei Longbao introduced the first application of FSMN network in Mongolian large vocabulary continuous speech recognition task. The experimental results show that the network has better performance than DNN network [6]. Lin Yanshi, Fei Longbao, Wang Yonghe and Gao Guanglai used LF-MMI standard to establish the first Halkha Mongolian ASR based on TDNN chain [7], and applied weight transfer technology. It is found that TDNN can obtain more abundant context information, and
the chain model constructed can reduce WER performance. Researchers such as Bao Shien [9], Biligut [2], Feilong [10-12, 4] have carried out various studies on different aspects of Mongolian speech recognition. For example, Bao Shien and Biligtu constructed Mongolian phonetic corpus. Their experimental system is a Mongolian large vocabulary continuous speech recognition system based on HMM, which makes a useful attempt to the acoustic characteristics and language rules of Mongolian. When building HMMs, they use triphones as the basic sub-word speech unit and use statistical language model to carry out experiments. In addition to building models and programming, they have accumulated experience in building and optimizing corpus. The recognition result of Mongolian speech recognition system LVCSR can reach 85.16%. Reference [13] already described Mongolian Grammar, Word, Morpheme. DNN model used to directly model the posterior probability of the acoustic unit. The convolutional neural network (CNN), as a kind of deep neural network, has received increasing attention through the reduction of dependence on data through specific model structures. However, the DNN model has a large training parameter, although it has improved the word error rate, but it still cannot achieve the desired effect. Based on this, in this paper proposes to use convolutional neural network to realize Cyrillic Mongolian speech recognition, adopt different input features, and combine maxout and dropout algorithms respectively to reduce the training parameters and further improve the recognition accuracy.

2. System Structure And Design

The description of thyroid ultrasound and the conclusion of ultrasound are two important parts of the report of ultrasonography. Current ultrasound reports have a fixed frame format and specifications. Within the fixed format specifications, besides key medical terms, every doctor has his own descriptive habits on the phenomenon of ultrasound examination, and there is no relatively perfect standard to regulate it. If the speech input system can accurately and quickly identify the content of the doctor's dictation, it is necessary to build a vocabulary, and put the most important and commonly used medical vocabulary into the speech data corpus. The system generates digital speech by collecting the voice data during the doctor's examination through microphone, and then the Cyrillic Mongolian speech recognition system will convert the speech data into text. Thirdly, the formed text will be output in the form of a report. Cyrillic Mongolian speech input system for Thyroid Ultrasound report mainly consists of three part (Figure 1).

**Figure 1.** Cyrillic Mongolian speech input system for Thyroid Ultrasound report

a) **Speech input system**

The speech input system of ultrasound examination report is connected with Cyrillic Mongolian speech recognition system. It can automatically import the information of the patients and doctor’s speech and save it.

b) **Cyrillic Mongolian Speech Recognition System**

The main task of speech system is to recognize a piece of speech information into corresponding text information. The recognition process is to extract the corresponding speech features of the original speech through a certain algorithm. The speech features are taken as network input composed of acoustic model, pronunciation dictionary and language model. The probability of each possible text information is calculated through the network, and the text information with the greatest probability is selected as speech recognition. The acoustic model used in this process needs to be trained with its corresponding speech database, and the language model needs to be trained with its corresponding text set.
c) Thyroid ultrasound reporting

Doctors can make ultrasound diagnosis more quickly and accurately according to the relevant information. They can also make reference to the ultrasound description of other similar patients and generate ultrasound examination more accurately.

3. CNN Model For Cyrillic Mongolian As

3.1 CNN model

The basic structure of convolution neural network (CNN) is shown in Figure 2, including input layer, convolution layer, full connection layer and output layer. The input layer represents the input characteristics; the convolution layer and the pooling layer can be one layer or multiple layers; the full connection layer can be one layer or multiple layers; and the output layer can be softmax layer. The particularity of convolutional neural networks is mainly embodied in three basic structures: convolutional layers, pooling layers and full-connected layers. The input of each node in the convolution layer is obtained by multiplying the node in a small rectangular area of the preceding layer by the weight matrix \( W \). Nodes sharing the same weight form a feature map. Each feature map is a local convolution method, which can also be considered as a local filter. A complete convolution layer contains multiple feature maps, which are derived from different weight matrices in order to obtain better local representation patterns for each region. For speech recognition, input space includes two dimensions: time domain and frequency domain.

![Figure 2. Convolutional Neural Network](image)

In fact, the frequency axis corresponds to a 40-dimensional log-mel filter bank coefficients. The convolutional neural network feature map only considers the convolution on the frequency axis, and the change in the time domain is handled by the Hidden Markov Model (HMM), mainly because the study of the references [14-15] shows the shift invariance in frequency. (Shift-in-variance) is more important than shift invariance in time. Assuming that the input feature map is one-dimensional, each feature map of the convolutional layer (1):

\[
q_{j,m} = (\sum_{i=1}^{l} \sum_{n=1}^{F} o_{i,n+m-1} w_{i,j,n} + w_{0,j})
\]

where \( q_{j,m} \) is the \( m \)-th unit of the \( i \)-th input feature map \( o_{i,m} \), \( q_{j,m} \) is the \( m \)-th unit of the \( j \)-th feature map \( Q_{j} \) in the convolution ply, \( w_{i,j,n} \) is the \( n \)-th element of the weight vector, \( w_{i,j} \) which connects the \( i \)-th input feature map to the \( j \)-th feature map of the convolution ply, is called the filter size, which determines the number of frequency bands in each input feature map that each unit in the convolution ply receives as input.

3.2 Dropout Algorithm

Dropout [16] is an effective method to overcome the over-fitting phenomenon of neural network training. In the feedforward phase of neural network training, each hidden layer node uses a hidden drop factor (HDF) as a probability to decide whether to participate in the updating of this parameter, which avoids the complex co-adaptation between hidden units and forces each hidden unit to be independent of other hidden units. When dropout is used, the output of layer \( l \) is (2):

\[
y^l = \frac{1}{1-p} r^* f(W^l y^{l-1} + b^l)
\]

Among them, \( y^{l-1} \) is the input of layer \( l \), \( W^l \) is the weight of layer \( l \), \( b \) is the bias, \( f \) is the non-linear activation function (such as sigmoid function, ReLU function, etc.), \( R \) is the binary mask (the binary matrix determines the real activation output of each layer), where \( r_j \)-Bernoulli(\( p \)) obeys the Bernoulli
probability distribution with probability \( p \). Because dropout is not used in the decoding phase, the factor \( \frac{1}{1-p} \) used in the training phase remains unchanged in the testing phase, and when dropout is not used, all inputs are fed into each layer. \(*\) is the point-by-point multiplication operator of elements. Figure 3 is a dropout diagram, in which the number of input nodes is 4, the output is 3, \( y^{l-1} \) is the input node, \( W^l \) is the weight matrix, \( f(*) \) is the activation function, and \( R(*) \) is the binary mask.

![dropout diagram](image)

**Figure 3. dropout Diagram**

### 3.3 Maxout Algorithm

Maxout [17] is a simple feedforward structure, which uses a new activation function: the maxout unit. Given the input \( x \in \mathbb{R}^d \), \( x \) can be either visible or hidden, the maxout function is defined as follows (3):

\[
h_i(x) = \max_{j \in [1,k]} z_{ij}
\]

(3)

Among them, \( z_{ij} \) can be learned according to (4) form:

\[
z_{ij} = x^T W_{ij} + b_{ij}
\]

(4)

Among them, \( W \in \mathbb{R}^{d \times m \times k} \) and \( b \in \mathbb{R}^{m \times k} \) are the parameter matrices and bias vectors that need to be studied respectively. There are \( m \) units in the current hidden layer. \( h_i(x) \) represents the output of the current hidden layer's unit \( i \), \( i = 1,..., m \), \( j \) represents the label of affine map, \( j = 1,..., k \). In convolutional networks, maxout feature maps are represented by the largest feature maps obtained from \( k \) affine feature maps. A single maxout element can be regarded as a piecewise linear approximation of any convex function. Maxout network can learn not only the correlation between hidden units, but also the activation function of each hidden unit. Although the gradient is highly sparse and dropout is sparse and effective in the training stage, the representation generated by maxout is not sparse. Although many common activation functions have curvature, maxout is almost linear in the local region. Given the deviations obtained according to standard conventions, maxout not only works well, but also has excellent robustness and is easy to train with dropout and produce good performance.

### 4. Experiment Preparation

#### 4.1 Dataset

There were no Mongolian speech database and text corpus for use in study. First, we created Mongolian speech corpus and it contains over 20000 utterances in total recorded from 996 different sentences spoken by 188 male and female speakers. The database contains about 5 hours of speech data, about 3M words of text data, including 15000 words of vocabulary table. The training set is used for acoustic model training, the dev set is used to select model parameters, and the test set is used for performance testing. The corpus mainly comes from Mongolian novels, newspapers, medical reports...
and books. The reading environment is the silent environment. The sampling frequency of voice file is 16 kHz and the sampling bit is 16 bit.

4.2 Evaluation

The results of continuous speech recognition are generally word sequences. Dynamic programming algorithm is used to compare the recognition results with the correct annotation sequence. The types of errors are divided into three categories: insertion error, deletion error and substitution error. Among them, insertion errors are caused by inserting other words between two adjacent tags, deletion errors are caused by the fact that the words corresponding to a tag cannot be found in the recognition results, and substitution errors are caused by the inconsistency between the identified words and the corresponding tags. Assuming that the total number of labels in a test set is N, the number of insertion errors is I, the number of deletion errors is D, and the number of errors is R. The word error rate was used as the evaluation index (5).

\[
WER = \frac{I + D + R}{N} \times 100\% 
\]

5. Experiment And Analysis

5.1 Contrast experiment of CNN baseline system

This experiment is using open source Keras, Tensorflow based on deep convolutional neural network (CNN). The CNN model consists of two convolution layers, one maximum pooling layer, four full connection layers and the last soft max output layer. The input layer contains 11 frames, 40-dimensional fbank features, and first-order and second-order difference. The number of nodes in the input layer is 3x11x40. The number of nodes in the output layer is 1415. The convolution size of the first convolution layer feature map is 9x9, the convolution size of the second convolution layer feature map is 4x3, the size of the pooling layer is 3x1, and the full connection layer is four hidden layers based on sigmoid activation function. The initial learning rate is 0.08, keeping four iterations and halving the subsequent iterations. Table 1 the performance of CNN, we can see the recognition performance had a great improvement. CNN reduced the error rate by 18.46% respectively.

| Model       | WER (%) |
|-------------|---------|
| CNN         | 18.46   |

5.1.1 Maxout and Dropout selection experiments

In convolutional neural network, the improvement of word error rate is not obvious, dropout is slightly improved, while maxout makes performance decline. Dropout decides which parameters to participate in this update according to certain rules. The application of dropout technology in this paper makes the system performance not significantly improved. Maxout may result in performance degradation due to the loss of some important weights, resulting in the loss of part of the system information and system performance degradation.

| Model          | WER% |
|----------------|------|
| CNN            | 18.17|
| CNN+dropout    | 18.15|
| CNN+maxout     | 19.15|

6. Conclusion

In conclusion a method of Cyrillic Mongolian speech recognition based on convolutional neural network is proposed, and a recognition system is constructed by combining maxout and dropout
algorithms. By optimizing the input features, full-connection layer activation functions and convolution layer feature maps of convolution neural network, and then introducing maxout and dropout algorithms, the proposed method is more effective. The speech recognition system based on convolutional neural network reduces the error rate of Cyrillic Mongolian speech recognition by 18.97% and 3.55% respectively. The scale of training parameters for each convolution layer of convolution neural network is also effectively improved. In speech recognition, we need to consider the complex working environment, the standard level of doctor's language dialect, improve the technology of speech enhancement, and speech recognition, and develop a wider range of speech recognition system.

7. Acknowledgments
The authors would like to thank the Donghua university.

References
[1] Chiu, C., Tripathi, A., Chou, K., Co, C., Jaitly, N., Jaunzeikare, D., Kannan, A., Nguyen, P., Sak, H., Sankar, A., Tansuwan, J., Wan, N., Wu, Y., Zhang, X. (2018) Speech Recognition for Medical Conversations. Proc. Interspeech 2018, 2972-2976, DOI: 10.21437/Interspeech.2018-40.
[2] G. Gao, Biligetu, Nabuqing, and S. Zhang, “A mongolian speech recognition system based on HMM,” in Computational Intelligence. Springer, 2006, pp. 667–676.
[3] H. Zhang, et al., “Comparison on Neural Network based acoustic model in Mongolian speech recognition,” International Conference on Asian Language Processing IEEE, 2017.
[4] Hui Zhang, Feilong Bao and Guanglai Gao,”Mongolian Speech Recognition Based on Deep Neural Networks” in Researchget,2015, pp.180-188
[5] Y. Wang, F. Bao, G. Gao, Research on Mongolian speech recognition based on TDNN-LSTM, Natural Language Processing and Chinese Computing, 2018.
[6] Yonghe Wang, Feilong Bao, Hongwei Zhang, Guanglai Gao “Research on Mongolian Speech Recognition Based on FSMN.” NLPCC 2017: 243-254
[7] Linyan Shi, Feilong Bao, Yonghe Wang, Guanglai Gao “Research on Transfer Learning for Khalkha Mongolian Speech Recognition Based on TDNN.” IALP 2018: 85-89
[8] A. Graves, S. Fernandez, 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. ACM, 2006, pp. 369–376.
[9] Bao Shien. Research and implementation of continuous speech recognition system for Mongolian and specific human vocabulary [D]. Hohhot: Inner Mongolia University, 2005.
[10] Fei Long Bao, Gao G L. Improving of acoustic model for the Mongolian speech recognition system[C]. Chinese Conference on Pattern Recognition (CCPR). 2009: 1-5
[11] Fei Long Bao, Gao G L, Yan X L, etc. Segmentation-Based Mongolian LVCSR approach[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2013: 8136-8139
[12] Fei Long Bao, Gao Guanglai, Wang Hongwei. Research on Mongolian speech keyword detection method based on stems[J]. Journal of Chinese Information Processing. 2016, 30(1): 124-128
[13] Seiichi Nakagawa, Erdenebat Turmunkh, Hiroshi Kibishi, Kengo Ohta, Yasuhiisa Fujii, Masatoshi Tsuchiya and Kazumasa Yamamoto “Development of large vocabulary continuous speech recognition system for mongolian language
[14] Abdel-Hamid O, Mohamed A, Jiang H, et al Convolutional neural network speech recognition IEEE Transaction on Audio, Speech, And Language Processing, 2014, 22 10 1533-1545
[15] Sainath T N, Mohamed A, Kingsbury B, et al Deep convolutional neural networks for LVCSR Proceedings of International Conferene on Acoustics, Speech and Signal Processing ICASSP 2013 8614-8618
[16] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors. 2013
[17] Goodfellow I J, Warde Farley D, Mirza M, et al. Maxout networks Proceedings of International Conference on Machine Learning (ICPR) 2013 1319-1327