Acta Technica *62* (2017), No. 6B, 321–334

Tour guide english speech recognition with dialect features and cultural background

Dejun Wang[2]

Abstract. Aimed at the non-ideal problem of recognition result in traditional automatic speech recognition application, a kind of automatic speech recognition with limited local weight-sharable convolutional neural network (CNN), based on the feature of mel-frequency spectral coefficient (MFSC), has been put forward. First, with the reference of treatment method on image information input in image treatment, the mapping input form of speech signal with two-dimensional array feature has been constructed and each mapping has been expressed as MFSC coefficient features including static data, first derivative and second derivative to make it convenient for applying image treatment method to recognize; second, convolutional neural network of image treatment has been introduced and aimed at the local speciality of speech signal feature, the learning algorithm of limited local weight-sharable convolutional neural network has been constructed to improve the speech signal identification and reduce the algorithm complexity; last, the algorithm put forward has been verified through experiment and the variation of algorithm parameter affecting the experiment has been given to provide basis for the specific application.

Key words. Convolutional neural network, Automatic speech recognition, Dialect features and feature acquisition, Training and learning, English pronunciation.

1. Introduction

Under the tour guide's dialect features and specific cultural background of the travel destination, the purpose of automatic speech recognition (ASR) is to realize the conversion from human speech to language. Due to the property of different loudspeakers and different speaking forms, uncertainty of environment noise and so on, human's speech signal is highly variable, which leads to make it quite a challenging task [1, 2]. Furthermore, ASR needs to convert the speech signal mapping with variable length to words or phonics with variable length sequence. As is known to

[1]Department of Tourism and Foreign Languages, Ma'anshan Teacher's College, Ma'anshan, 243041 China

all, hidden markov model[3] (HMM) has been able to successfully handle the problems like speech signal recognition with variable length sequence and time property modeling of state sequence of speech signal and so on and to be related to the specific observation probability distribution. However, Gaussian mixed model (GMM) has always been considered as the strongest estimation tool for probability distribution to process speech signal based on hidden markov model. Meanwhile, based on the common expectation-maximization, GMM-HMMs training method[4] has been widely used in GMM model building. Furthermore, literature [5, 6] and so on have put forward different types of discriminant training methods to further improve the accuracy of ASR recognition system.

Convolutional neural network is a kind of weight-shareable classification network[7-11], close to the biology neural network structure and able to effectively reduce the complexity of network model building and simplify the weight to set the numbers. Moreover, due to the scale invariance of deformation form like translation and so on possessed by it, it has been widely used in image treatment. The research thought of the Paper is to use the application method of convolutional neural network on image treatment for reference and combine the speciality of speech signal to build the mapping input form based on the two-dimensional array speciality of mel-frequency spectral coefficient (MFSC) and design the speech recognition algorithm of limited local weight-sharable convolutional neural network in order to improve the speech signal identification and reduce the algorithm complexity.

2. Deep neural network

Generally speaking, the deep neural network(DNN) means the feedforward neural network of multiple hidden layers. Each hidden layer has a certain amount of units (or neurons) and uses the low-layer output as the input to carry out neural network calculation[12] by weight vector and nonlinear activation function:

$$o_i^{(l)} = \sigma(\sum_j o_j^{(l-1)} w_{j,i}^{(l)} + w_{0,i}^{(l)}) . \tag{1}$$

Where $o_i^{(l)}$ is the ith neuron output on the lth hidden layer of network; $w_{j,i}^{(l)}$ is the connection weight of neuron-i from the jth to the lth hidden layer on the $l-1$th hidden layer neuron; $w_{0,i}^{(l)}$ is the deviation term of neuron-i; σ is the nonlinear activation function. Sigmoid activation function has been selected in the paper:

$$\sigma(x) = 1/(1+\exp(-x)) . \tag{2}$$

The vector representation of formula (1) is:

$$o_i^{(l)} = \sigma(\mathbf{o}^{(l-1)} \cdot \mathbf{w}_i^{(l)}) . \tag{3}$$

In the formula, the deviation term can be included in the column vector-$\mathbf{w}_i^{(l)}$ by expanding the vector dimensionality of $\mathbf{o}^{(l-1)}$. Furthermore, the following matrix

form can be used to represent the calculation of neuron in each layer:

$$\mathbf{o}^{(l)} = \sigma(\mathbf{o}^{(l-1)}\mathbf{W}^{(l)}) \quad (l = 1, 2, \cdots, L-1) \tag{4}$$

In the formula, $\mathbf{W}^{(l)}$ represents the neural network weight in the lth hidden layer.

The first (bottom) network of DNN is the input layer and the uppermost layer is the output layer. For multi-class questions, the posterior probability of each class can be estimated with output softmax layer:

$$y_i = \frac{\exp(o_i^{(L)})}{\sum_j \exp(o_j^{(L)})}\,. \tag{5}$$

In the formula, the calculation form of $o_i^{(L)}$ is $o_i^{(L)} = \mathbf{o}^{(L-1)} \cdot \mathbf{w}_i^{(L)}$.

If the random gradient descent algorithm is used as the minimum objective function or small batch of each training sample, the updating process of weight matrix is:

$$\Delta\mathbf{W}^{(l)} = \varepsilon \cdot (\mathbf{o}^{(l-1)})'\mathbf{e}^{(l)} \quad (l = 1, 2, \cdots, L) \tag{6}$$

In the formula, ε is the learning rate and error signal vector on the lth layer and $\mathbf{e}^{(l)}$ can be of back propagation from sigmoid hidden unit. The calculation is as follows:

$$\begin{cases} \mathbf{e}^{(L)} = \mathbf{d} - \mathbf{y} \\ \mathbf{e}^{(l)} = \left(\mathbf{e}^{(l+1)}(\mathbf{W}^{(l+1)})'\right) \cdot \mathbf{o}^{(l)} \cdot (1 - \mathbf{o}^{(l)}) \end{cases} \tag{7}$$

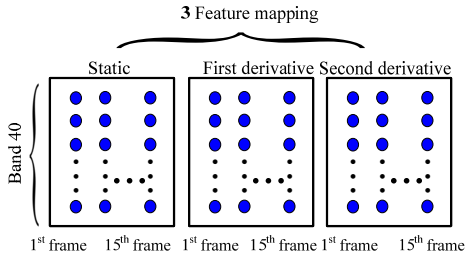
In the formula, "$\cdot$" represents two matrixes or multiply operation of vector with equal size.

3. Automatic speech recognition of convolutional neural network

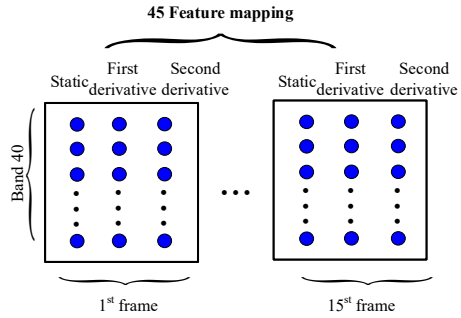
3.1. Building of network mapping

Based on the pattern recognition algorithm of CNN network, the data input needs to be expressed as feature mapping form, which uses the application program of image treatment for reference to express the data input as the input of two-dimensional array and be presented with the horizontal-x and vertical-y pixel form. CNN network runs in the little window where the images are input to make the network weight be able to observe speech feature from the data input by this window in the training and testing stage. The specific process is shown in Fig. 1.

At present, there are many types of DNN matrix input forms, the frequently used two of which have been given in Fig. 1. First, as is shown in Fig. 1 (a), the input speech matrix can be set as three two-dimensional feature mappings and each mapping is expressed as the feature (static, first derivative and second derivative) of mel-frequency spectral coefficient (MFSC) to be expressed along the distribution form of frequency (frequency band index) and time (data frame). Under this circumstance, two-dimensional convolution operation can be carried out and mean-



(a) Matrix method 1 for feature mapping



(b) Matrix method 1 for feature mapping

Fig. 1. Matrix form input of different speeches

while normalization operation can be carried out for frequency and time. Or, only frequency normalization can be considered here.

Under this circumstance, the same MFSC feature can be combined as 1-dimensional (1-D) feature mapping (along the band index), as shown in Fig. 1 (b). For example, if the speech window includes 15 frames and each frame includes 40 filters, the 1-dimensional matrix of 45 feature mappings can be built, the result of which is that the 1-dimensional convolution can be exerted along the frequency coordinate axis. It mainly considers using the latter 1-dimensional convolution along the frequency in the Paper, as shown in Fig. 1(b). Once the input feature vector has been built, the activation operation shall be carried out respectively in the convolution and convergence layer. Similar to the input layer, the convolution and convergence layer of this unit can also be expressed as mapping matrix form and the convolution and convergence layer can be expressed as 1 CNN layer in CNN terms.

3.2. CNN convolution layer

As shown in Fig. 2, for each input feature mapping, it shall assume that I is the total number of mappings, $O_i(i = 1, \cdots, I)$, which is connected with multiple feature mappings (assuming that the total number is J), $Q_j(j = 1, \cdots, J)$ and based on the convolution layer ($I \times J$ and $\mathbf{w}_{i,j}$ $(1, \cdots, I; j = 1, \cdots, J)$) of local-weight matrix. The mapping can be expressed as the convolution operation in signal treatment. Assuming that the input feature mapping is 1-dimensional, the feature mapping of

each neuron convolution can be calculated as:

$$q_{j,rm} = \sigma\left(\sum_{i=l}^{I}\sum_{n=1}^{F} o_{i,n+m-1}w_{i,j,n} + w_{o,j})\right),\ (j = 1, \cdots, J)\ . \tag{8}$$

In the formula, $o_{i,m}$ is the ith input of feature mapping-O_i in neuron-m; $q_{j,m}$ is the jth input of neuron-m in feature mapping- $O_j$of convolution layer; $\mathbf{w}_{i,j,n}$ the nth of weight vector-$w_{i,j}$; F is the size of filter and its value decides the number of feature mapping bands input in each convolution.

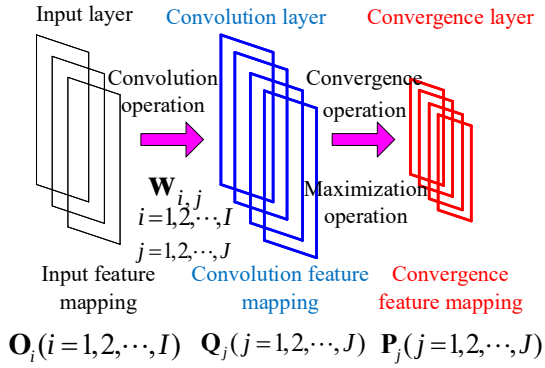


Fig. 2. CNN network operation structure

Due to the locality of MFSC feature mapping selected, these feature mappings are limited within the scope of frequency of speech signal to be defined. Based on the convolution operator, the matrix form can be simplified in formula (8):

$$Q_j = \sigma(\sum_{i=1}^{I} O_i * \mathbf{w}_{i,j}),\ (j = 1, \cdots, J)\,. \tag{9}$$

In the formula, O_i is expressed as the ith input feature mapping and $\mathbf{w}_{i,j}$ is expressed as the local-weight matrix. When it is based on 1-dimensional feature mapping, both of O_i and $\mathbf{w}_{i,j}$ are vectors and when it is based on two-dimensional feature mapping, both of O_i and $\mathbf{w}_{i,j}$ are matrixes.

3.3. CNN convergence layer

As shown in Fig. 2, exert convergence operation in CNN convolution layer to generate the corresponding convergence lager. The convergence function shall be independently used in each convolution feature mapping. When the maximum function is used, CNN convergence layer can be defined as[14]:

$$p_{i,m} = \max_{n=1}^{G} q_{i,(m-1)\times s+n}\,. \tag{10}$$

Where G is the scale of convergence layer and s is the size of displacement, which

confirms the overlapping ratio of adjacent convergence windows. Similarly, if the mean function is used, the input can be calculated as:

$$p_{i,m} = r \sum_{n=1}^{G} q_{i,(rm-1)\times s+n} \cdot \tag{11}$$

In the formula, r is the scale factor which can be learnt. Generally, in the application of image recognition, under the constraint of $G = s$ and under the circumstance when the convergence window is not overlapped, there is no gap among them so under this circumstance, the representation of maximum convergence feature mapping is superior to the representation of mean convergence feature mapping. In the work of the Paper, the method to independently adjust G and s shall be adopted. Furthermore, the nonlinear activation function shall be adopted to generate the final output. The convergence process sketch with the size of convergence layer being 3 has been given in Fig. 3. Each convergence layer unit receives input from neuron in three convolution layers in the same feature mapping. If G=s, the size of convergence layer will be one third of convolution layer.

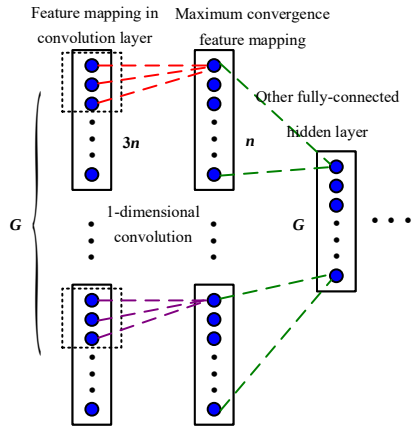


Fig. 3. Regularization CNN network of weight share

4. Learning algorithm for limited weight share of CNN network

4.1. Limited weigh share

The weight share scheme shown in Fig. 3 is the full weight share (FWS) method which is the standard application form in image treatment of CNN network, because this mode can appear in any local place in image. However, when it is specific to the speech signal feature, its speech feature will be different in different bands and the result by directly using full weight share method in speech recognition is not ideal,

while using different weight convergence will be more suitable for speech signal which has different speech signal feature under different band feature. Limited weight share example has been given in Fig. 4 for CNN network and only the convolutional neuron of neuron connected to the same convergence layer can share the same convergence weight.

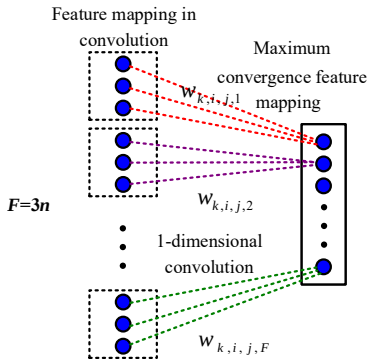


Fig. 4. CNN network of local weight share

These convolution units need to share their weight to make it convenient for calculating comparable speech features and converge these features to convergence layer. It can be explained that each band can be regarded as an independent subnet, which has independent sharable convolution weight. Each subnet part all includes some feature mappings in convolution layer and these feature mappings can scan all dimensionality input through weight vector to confirm whether this feature exists or not. The size of convergence layer has decided the number of usage of this weight vector in the adjacent position of input space, that is to say, in the convolution layer, the size of each feature mapping equals to the size of convergence layer. The complete convolution feature of each convergence neuron has been collected as a feature with convergence function in the Section. The convolution activation function can be calculated as:

$$q_{k,j,m} = \sigma\left(\sum_{i}\sum_{n=1}^{F} o_{i,(k-1)\times s+n+rm-1} \cdot w_{k,i,j,n} + w_{k,0,j}\right). \tag{12}$$

In the formula, $w_{k,i,j,n}$ is the nth convolution weight from the ith input feature mapping to the jth convolution mapping in subnet k; the value range of m is $1 \sim G$. Under this circumstance, the form of activation function in convergence layer is:

$$p_{k,j} = \max_{m=1}^{G} q_{k,j,m}\,. \tag{13}$$

In a similar way, LWS convolution layer above can also be expressed as multiplication form by using large-scale sparse matrix, but the building method of $\hat{\mathbf{o}}$ and $\hat{\mathbf{W}}$ is different from FWS convolution layer's. First, the sparse matrix $\hat{\mathbf{W}}$ can be built in accordance with Fig. 5, where each $\mathbf{W}_k$ can be built in accordance with local

weight $w_{k,i,j,n}$ and the form is:

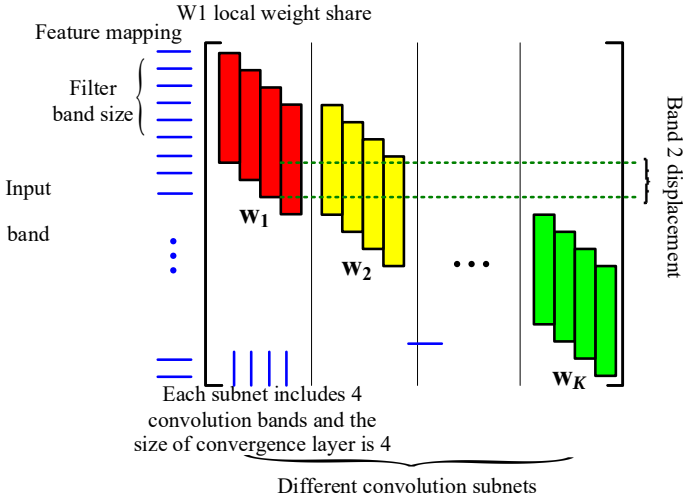


Fig. 5. CNN network calculation process of local weight share

$$\mathbf{W}_k = \begin{bmatrix} w_{k,1,1,1} & w_{k,1,2,1} & \cdots & w_{k,1,J,1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,I,1,1} & w_{k,I,2,1} & \cdots & w_{k,I,J,1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,I,1,2} & w_{k,I,2,2} & \cdots & w_{k,I,J,2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,I,1,F} & w_{k,I,2,F} & \cdots & w_{k,I,J,F} \end{bmatrix} . \tag{14}$$

In the formula, $k = (1, 2, \cdots, K)$, so when the same weight of each subnet is repeated to use G, but the weight of different subnets is different. Secondly, the calculation form of feature mapping can be expressed as a lager-scale vector form:

$$\hat{\mathbf{q}} = [|\mathbf{v}_{1,1}| \cdots |\mathbf{v}_{1,G}| \cdots |\mathbf{v}_{K,1}| \cdots |\mathbf{v}_{K,G}|] . \tag{15}$$

In the formula, K is the total number of subnets, G is the size of convergence layer and $\mathbf{v}_{k,m}$ is the row vector input by subnet neuron k of band m included in the feature mapping.

$$\hat{\mathbf{v}}_{k,rm} = [q_{k,1,m}, q_{k,2,m}, \cdots, q_{k,I,m}] . \tag{16}$$

In the formula, I is the total number of input feature mapping in each subnet. Hence, based on the weight of LWS, the learning method is as follows:

$$\begin{cases} \hat{\mathbf{q}} = \sigma\left(\hat{\mathbf{o}}\hat{\mathbf{W}}\right), \\ \Delta\hat{\mathbf{W}} = \varepsilon\hat{\mathbf{o}}'\mathbf{e}. \end{cases} \tag{17}$$

Meanwhile, the error vector is propagated through the largest convergence function, as follows:

$$e_{k,i,n}^{1ow} = e_{k,i} \cdot \delta(u_{k,i} - n). \tag{18}$$

Where,

$$u_{k,i} = \arg\max_{m=1}^{G} \; q_{k,i,m}. \tag{19}$$

4.2. LWS-CNN train

Because different weights are used in each band and it only needs to consider the band scope where the speech mode appears, LWS method contributes to the decreasing of the total number of neurons in convergence layer. However, for the smaller number of feature mappings, 1 band should be enough. On the other hand, it is not allowed to exert convolution layer again on the convergence layer in LWS scheme, because the features in different convergence layers of LWS are irrelevant.

In the Section, we have put forward LWS-CNN training form that has been modified on the base of convolution restriction Boltzman machine (CRBM). To carry out study for model parameter of CRBM, we need to define the conditional probability of state for hidden layer neuron. The conditional probability $h_{k,j,m}$ of activation condition for hidden neuron can be defined as:

$$P(h_{k,j,rm} = 1|\mathbf{v}) = \frac{\exp(I(h_{k,j,rm}))}{\sum_{n=1}^{p}\exp(I(h_{k,j,n}))}. \tag{20}$$

In the formula, $I(h_{k,j,rm})$ is the weighting signal sum from the signal of input layer, which can reach the neuron, and can be defined as:

$$I(h_{k,j,rm}) = \sum_{i}\sum_{n=1}^{f} v_{i,(k-1)\times s+n+rm-1} w_{k,i,j,n} + w_{k,i,j,0}. \tag{21}$$

Hence, the calculation form showing the conditional probability distribution $v_{i,n}$ for feature mapping i of band i in neuron for hidden neuron state meets Gaussian feature distribution:

$$P(v_{i,n}|\mathbf{h}) = N\left(v_{i,n}; \sum_{j,(k,m)\in C(i,n)} h_{k,j,m} w_{k,i,j,f(n,k,m)}, \sigma^2\right). \tag{22}$$

In the formula, $C(i,n)$ means the connection index receiving convolution band and subnet input from neuron on the apparent layer, $w_{k,i,j,f(n,k,m)}$ is the connection weight from band m in input feature mapping i to band m in feature mapping j on the k th layer of convolution subnet, $f(n,k,m)$ is the mapping function from

the index of connection node to the index of corresponding filter's element and σ^2 means the Gaussian distribution variance of fixed model parameter.

Based on the two conditional probabilities above, the connection weight value of all CRBM models above can use the regular ramification to carry out iteration estimation. Weight training value of CRBM model can be deemed as the initial value of the scheme in LWS convolution layer. After the weight in the convolution layer has been learnt and trained, based on formula (12-13), calculate the output in convolution later and convergence layer. The output of convergence layer continues to serve as the training output of the deep network in the next layer till the convergence of algorithm.

5. Experimental analysis

5.1. Experiment setting

Matlab 2013a. In the experiment of the Section, the effective evaluation of CNNs algorithm has been carried out in two tour guides' speech recognition tasks: carry out mobile phone speech recognition and large vocabulary speech search task in TIMIT test library. Use a 25ms window and fixed 10ms frame rate to carry out speech recognition. Speech feature vector can be generated through the conversion of Fourier filter, where 40 logarithmic energy coefficients together with its first and second time derivative are included. Normalization operation for all speech data shall be carried out to make each vector dimensionality have zero mean and unit variance feature. Hardware configuration of experiment: CPU i3-2440k and 6G RAM. Simulation platform of software: Matlab 2013a.

In the speech library of TIMIT[14, 15], the standard 462 speech training dataset shall be used and one 50 speech dataset separately designed shall be used to adjust the neuron parameter, including learning plan and learning rate. Meanwhile, based on 24 speech test dataset, the test shall be carried out and this test set shall not be overlapped with the training set. Besides to record MFSC features, a logarithmic energy feature shall be recorded for each frame. Full weight share CNN algorithm and comparative test for limited-local weight share scheme shall be carried out here first. In the Section, first, evaluate ASR recognition performance of CNN algorithm under different parameter settings. The method is to set all parameter values and show the situation of speech recognition performance with the change of some parameter.

In the Experiment, a convolution layer, a convergence layer and 2 fully-connected hidden layers located on the top have been used. There are 1000 neurons on the fully-connected layer. Parameter setting of convolution and convergence: the convergence dimension is 6, the displacement dimension is 2, the filter size is 8, FWS feature mapping is 150 and the feature mapping on each band of LWS is 80. Comparing the algorithms, select FWS-CNN and LWS-CNN (in the Paper) algorithm.

5.2. Algorithm parameter influence

In the Section, it mainly verifies the influence of different set CNN parameter on the performance of algorithm. The experiment result by comparing the algorithm in test and dev. has been given in Fig. 6-9. The data shows that the size of convergence layer and the number of feature mappings have the most significant influence on the recognition performance of final ASR.

It has been shown in Fig. 6 that the better tour guide's speech recognition performance will be generated after the parameters in convergence layer are increased to 6 for all algorithms. The bigger the parameters in convergence layer are set, the better the LWS algorithm performance is. It has been shown in Fig. 6 and 7 that the overlapping convergence window will not generate obvious performance gain and the similar performance is generated by using the parameters with the same values and the parameter setting with same displacement size in the convergence layer and meanwhile the complexity of model can be reduced. It has been shown in Fig. 8 that the larger number of feature mappings will usually result in better speech recognition performance, especially in FWS algorithm. It has been shown in Fig. 9 that the size dimension of filter has a smaller influence on algorithm performance. Meanwhile the experiment data shows that compared to FWS algorithm, because LWS algorithm has the learning capacity aimed at the features of different frequency bands, it can get better tour guide's speech recognition performance when the feature mapping parameter is smaller. This shows that LWS scheme is more efficient on the aspect of the number setting of hidden layer neurons.

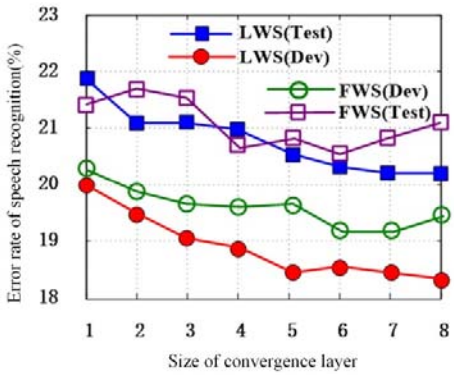


Fig. 6. Parameter influence of convergence layer

6. Conclusion

To further improve the tour guide's speech recognition result, use the image treatment method for reference to build the two-dimensional array feature mapping input of speech signal and combine the convolutional neural network to carry out speech recognition; meanwhile to improve the speech recognition result and combine the

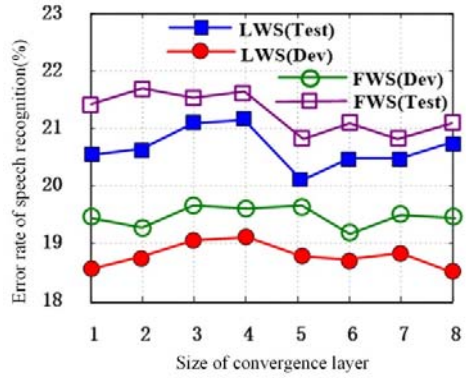


Fig. 7. Parameter influence of displacement size

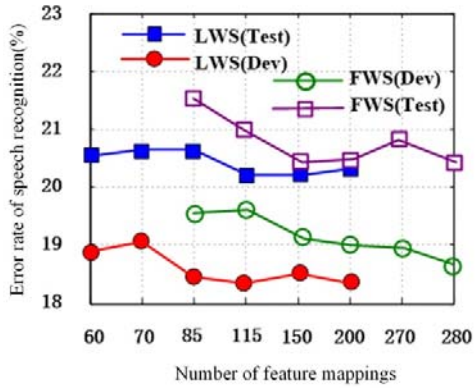


Fig. 8. Influence of the number of feature mappings

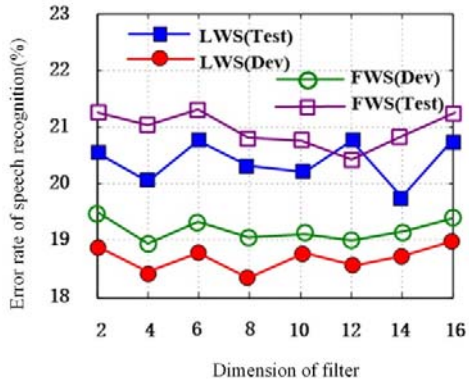


Fig. 9. Size dimension influence of filter

speech signal feature to build the learning method for local limited weight-shareable convolutional neural network, the experiment result shows that the algorithm put

forward has higher calculation performance. The convolutional neural network itself is a rather complex calculation method. How to realize the further optimization of algorithm, combine the calculation hardware to carry out programming and build real time recognition system is the research direction in future.

Acknowledgement

The higher occupation education innovation and development action plan (2015-2018) of major construction (XM-01): Maanshan Teacher's College tourism management professional group; Maanshan Teacher's College outstanding talent education and training: excellent English tour guide (2016xjzjjh02).

References

[1] Ververidis D, Kotropoulos C: (2006) *Emotional speech recognition: Resources, features, and methods*[J]. Speech Communication, 48(9):1162-1181.

[2] Taylor G W: (2005) *Two-way speech recognition and dialect system*[J]. Acoustical Society of America Journal, 113(5):2392-2392.

[3] Precoda K, Podesva R J: (2003) *What will people say? Speech system design and language/cultural differences [speech recognition]*[C]// Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on. IEEE Xplore, 2003:624-629.

[4] Soto V, Siohan O, Elfeky M, et al.: (2016) *Selection and combination of hypotheses for dialectal speech recognition*[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2016:5845-5849.

[5] Pedersen C, Diederich J: (2008) *Accent in Speech Samples: Support Vector Machines for Classification and Rule Extraction*[J]. Rule Extraction from Support Vector Machines, 80:205-226.

[6] Sinha S, Jain A, Agrawal S S :(2015) *Fusion of multi-stream speech features for dialect classification*[J]. CSI Transactions on ICT, 2(4):243-252.

[7] Wahlberg F, Dahllöf M, Mårtensson L, et al.: (2014) *Spotting Words in Medieval Manuscripts*[J]. Studia Neophilologica, 86(Supp 1):171-186.

[8] Agarwalla S, Sarma K K: (2016) *Machine learning based sample extraction for automatic speech recognition using dialectal Assamese speech*[J]. Neural Networks the Official Journal of the International Neural Network Society, 78:97-111.

[9] Aikhenvald A Y: (2014) *Language Contact and Language Blend: Kumandene Tariana of Northwest Amazonia1*[J]. International Journal of American Linguistics, 80(Volume 80, Number 3):323-370.

[10] Takada A: (2013) *Generating morality in directive sequences: Distinctive strategies for developing communicative competence in Japanese caregiver–child interactions*[J]. Language & Communication, 33(4):420-438.

[11] Ibrahim H S, Abdou S M, Gheith M: (2015) *Idioms-Proverbs Lexicon for Modern Standard Arabic and Colloquial Sentiment Analysis*[J]. International Journal of Computer Applications, 118(11):26-31.

[12] Agrawal S S, Jain A, Sinha S: (2016) *Analysis and modeling of acoustic information for automatic dialect classification*[J]. International Journal of Speech Technology, 2016(3):1-17.

Received May 7, 2017