ISSN 1060-992X, Optical Memory and Neural Networks, 2018, Vol. 27, No. 4, pp. 272–282. © Allerton Press, Inc., 2018.

Performance Optimization of Speech Recognition System with Deep Neural Network Model

Wei Guan*

College of Modern Science and Technology, China Jiliang University, Hangzhou, Zhejiang, China

**e-mail: gwcjlu@163.com*

Received June 29, 2018; in final form, October 8, 2018

Abstract—With the development of internet, man-machine interaction has tended to be more important. Precise speech recognition has become an important means to achieve man-machine interaction. In this study, deep neural network model was used to enhance speech recognition performance. Feedforward fully connected deep neural network, time-delay neural network, convolutional neural network and feedforward sequence memory neural network were studied, and their speech recognition performance was studied by comparing their acoustic models. Moreover, the recognition performance of the model after adding different dimension human voice features was tested. The results showed that the performance of the speech recognition system could be improved effectively by using the deep neural network model, and the performance of feedforward sequence memory neural network was the best, followed by deep neural network, time-delay neural network and convolutional neural network. Different extraction features had different improvement effects on model performance. The performance of the model which was added with Fbank extraction features was superior to that added with Mel-frequency cepstrum coefficient (MFCC) extraction feature. The model performance improved after the addition of vocal characteristics. Different models had different vocal characteristic dimensions.

Keywords: deep neural network, acoustic model, speech recognition, discriminative training, performance optimization

DOI: 10.3103/S1060992X18040094

1. INTRODUCTION

Language is one of the means of communication between people. Direct verbal communication is more convenient and faster than text communication. The current man-machine communication is realized by inputting literal code, then can it be realized through speech like verbal communication between people? The answer is yes. Speech recognition technology is born for this reason. The main function of speech recognition technology is to identify effective information in speech signals and then convert them into text information which can be understood by machine [4]. There are a variety of speeches, and the classification standards include vocabulary size, pronunciation and speaker. Before 2012, the main acoustic model in speech recognition technology was Gaussian mixture model-hidden Markov model, but neural network structure began to be applied in speech recognition technology and gradually become the mainstream modeling method after 2000. Chan et al. [1] put forward a neural network for large vocabulary conversational speech recognition and obtained a word error rate of 14.1% in the test of Google voice search. Wang et al. [2] reduced the size of deep neural network model using split vector quantization algorithm and found that the algorithm could reduce 75 ~ 80% of model size on the premise of ensuring original performance. To address the problem of difficult parameter analysis in learning process of deep neural network, Wu et al. [3] put forward motivation training to improve node output communication of relevant regions in network and obtained consistent performance gains in two speech recognition tasks: a U.S. English broadcast news task and a Javanese conversational telephone speech task. With the deepening of research and to solve the problem that single structure cannot satisfy the modeling demand of the current speech recognition technology, different neural network structures have been applied into modeling, including time delay neural network, convolution neural network and feedforward sequence memory neural network [5]. In this study, the influence of deep neural network, time delay neural network, convolutional neural network and feedforward sequence memory neural network based acoustic model on the performance of English speech recognition system.

Figure captions

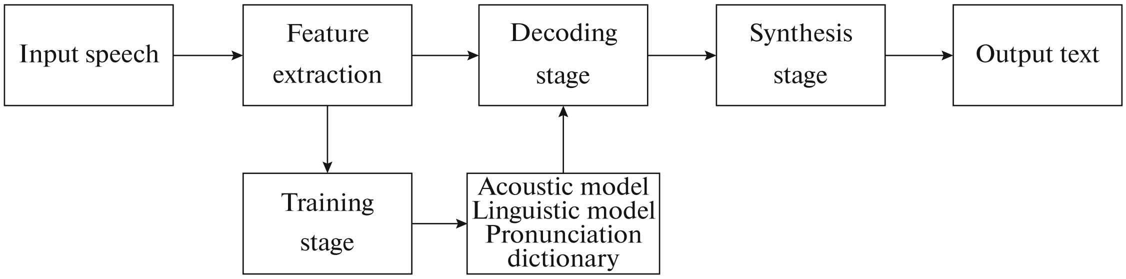

Fig. 1. The basic flow of the speech recognition system.

2. MAN-MACHINE INTERACTION AND SPEECH RECOGNITION

Computer cannot directly understand speech signals of human because of different operation principles, let alone multiple languages around the world. To make computer be able to understand speech information of human, speech recognition system emerged. The main function of speech recognition system is to transform speech signals to text information which can be understood by computer. The working process of speech recognition system is shown in Fig. 1.

There are two key stages in the whole speech recognition system, feature extraction stage and training stage. The training stage includes construction of acoustic model and language model. Pronunciation dictionary is taken as the comparison database.

2.1. Feature Extraction

It is very important to remove speech signals that will affect the speech characteristics of text information during feature extraction. The better the feature extraction is, the better the speech recognition effect will be. Currently the methods of speech feature extraction include linear prediction coefficient method, perceptual linear prediction method and Mel cepstrum coefficient method, among which, Mel cepstrum coefficient method is the most widely used [6].

In speech recognition system, acoustic models play an important role in both training and decoding stages. Gaussian mixture model which mixes multiple Gauss models can simulate any distribution patterns theoretically [7]. Taking English phonemes as an example, a Gaussian mixture model corresponds to a phoneme classification, and then the probability of the phoneme classification corresponding to the current frame can be calculated. Hidden Markov model has variable state residence time and can effectively solve the time varying problem of speech signals. Therefore it has been a maintain acoustic modeling method.

2.2. Linguistic model

Phoneme sequences obtained by acoustic models can be converted into word sequences by means of pronunciation dictionaries, but normal word order cannot be guaranteed. In this case, a linguistic model is needed to filter out the most qualified sentences. Excellent language models can greatly improve the decoding speed and recognition accuracy [14]. The current widely applied statistical linguistic model include N-gram linguistic model and neural network based linguistic model [8].

The main research subject of this study is the deep neural network based acoustic model rather than linguistic model. Therefore N-gram linguistic model was used in the following experiment, and the model parameters remained unchanged in the experiment.

3. STRUCTURE OF DEEP NEURAL NETWORK

3.1. Feedforward Fully-Connected Deep Neural Network

Feedforward fully-connected deep neural network essentially is a multilayer perceptron of forward neural network structure which contains multiple hidden layers. Its structure is shown in Fig. 2. Feedforward fully-connected deep neural network has a hierarchical structure, including input layer, hidden layer and output layer. Nodes in the neighbourhood layers fully connect, and the nodes in the same layer are

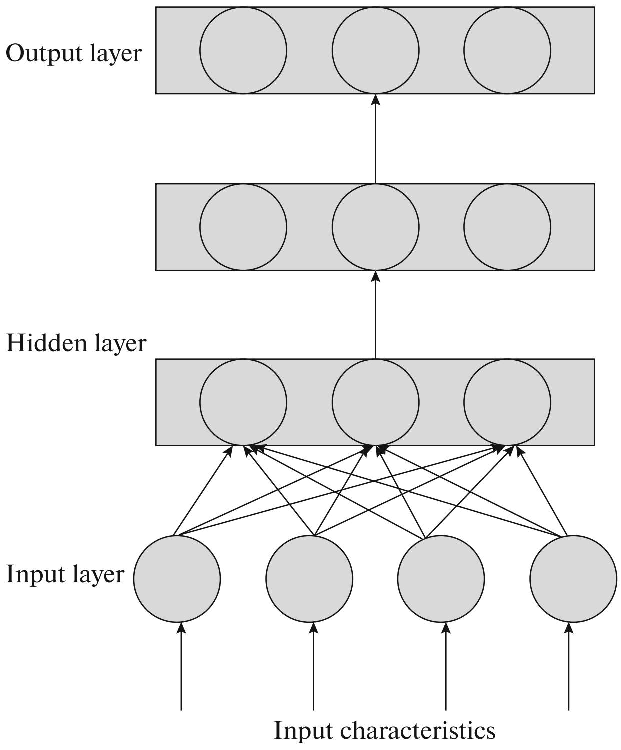

Fig. 2. The structure of deep neural network.

independent. If input characteristic $g^0 = N$, then the formula of activation value of nodes in the hidden layer of feedforward fully-connected deep neural network [9] is as follows.

$$b^m = W^m g^{m-1} + a^m \quad (1 \le m \le M + 1), \tag{1}$$

$$g^m = f(b^m) \text{ with } g_j^m = \frac{1}{1 + e^{-b_j^m}} \quad (1 \le m \le M), \tag{2}$$

where N stands for the number of the hidden layer, W^m and a^m stand for the weight and bias vector of the m-th hidden layer, and $f(\cdot)$ stands for the nonlinear activation function signoid of nodes in the hidden layer.

The output layer of feedforward fully-connected deep neural network usually models the posterior probability of input characteristics using softmax function, and its formula is:

$$y_s = g_s^{M+1} = \Pr(s|N) = \mathrm{softmax}_s(b^{M+1}), \tag{3}$$

where y_s stands for the s-th element in output vector y.

In Fig. 2, a result will be output from the output layer after extraction characteristics are input. Such a process is called forward propagation process. The output result needs to be compared with directive signal, in which corresponding optimization algorithm is needed. Stochastic gradient descent based error back propagation algorithm is the current commonly used optimization algorithm.

3.2. Time Delay Neural Network

Time delay neural network is used for speech recognition, and it evolves from error back propagation algorithm mentioned above. The basic structure of time delay neural network is shown in Fig. 3. The training and learning of feedforward fully-connected deep neural network starts from context, while time delay neural network starts from short context and then extends to a larger range according to the number of hidden layer. Therefore time delay neural network can reflect the relationship of context and time. Because of the relationship between structure transformation and input time, it is usually equivalent to the initial model of convolutional neural network. The general algorithm of time delay neural network [10] is:

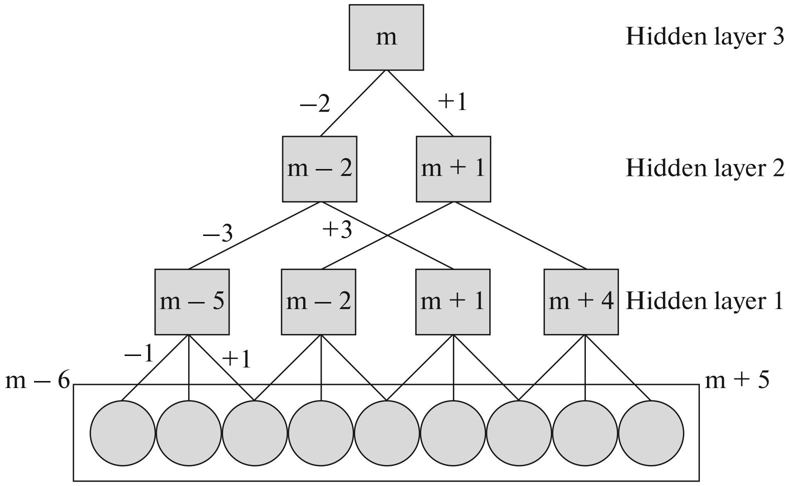

Fig. 3. The structure of time delay neural network.

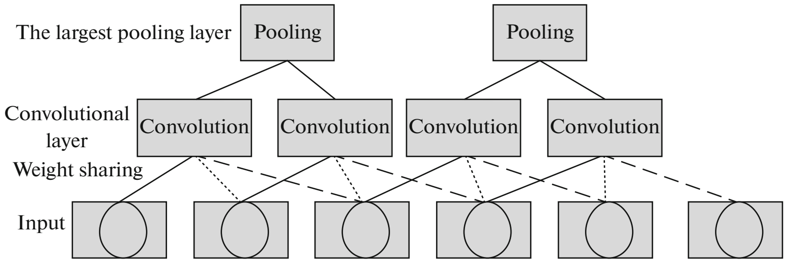

Fig. 4. The structure of convolutional neural network.

The output layer is:

$$\sigma_j = (m_j - p_j)p_j(1 - p_j), \tag{4}$$

$$\Delta M_{ju}(m+1) = \eta\sigma_j p_u + \alpha\Delta M_{ju}(m). \tag{5}$$

The hidden layer is:

$$\sigma_u = p_u(1 - p_u)\sum_j M_{ju}\sigma_j, \tag{6}$$

$$\Delta M_{uk}(m+1) = \eta\left(\sum_m \sigma_u x_{k(1+m)}\right) + \alpha\Delta M_{uk}(m), \tag{7}$$

where p_j stands for the actual output of the j-th layer, δ_j stands for the correction of the j-th layer, m_j stands for the input time of the j-th layer, ΔM_{ju} stands for the difference between the j-th and u-th layers, and η and α are coefficients.

3.3. Convolutional Neural Network

Convolutional neural network is an algorithmic model designed under the inspiration of receptive field in biology, which is essentially a mathematical model with supervised learning module. As shown in Fig. 4, convolutional neural network can extract features of the input layer alternately using multiple convolutional layers and then integrate and transform the extracted local features with the fully connected layer. Convolutional neural network can effectively obtain generic obvious features from a large amount of learning data. A typical convolutional neural network has great invariance in input features because of its special structure.

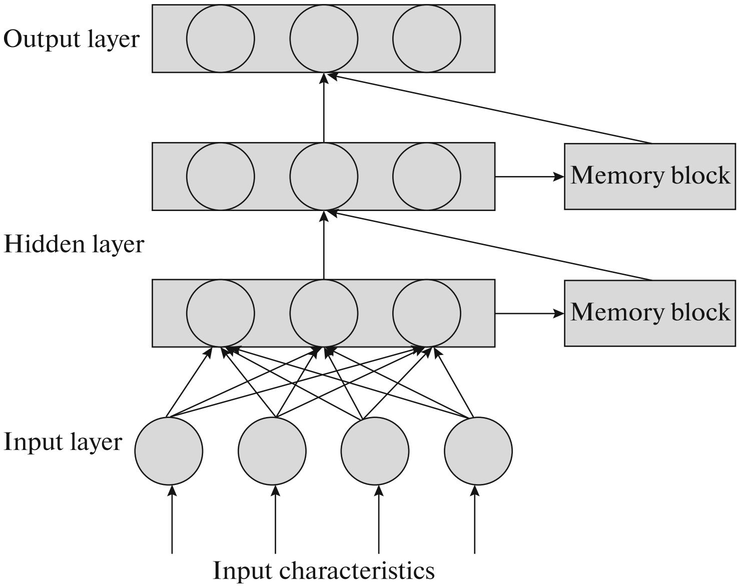

Fig. 5. The structure of feedforward sequence memory neural network.

Convolutional layer is a core part of the whole network, and its output is called characteristic pattern. The convolution is like linear weighed operation. The formula [11] is:

$$R(i,j) = (O * H)(i,j) = \sum_{c}\sum_{d} O(i+c, j+d)H(c,d). \tag{8}$$

The generation formula of characteristic pattern is:

$$\alpha_j^m = f(\gamma^m) = f\left(\sum_{i \in N_i} \alpha_i^{m-1} * H_j^m + \beta_j^m\right), \tag{9}$$

where O stands for characteristic pattern, H stands for convolutional kernel, α_j^m stands for the output characteristic pattern of the j-th convolutional kernel of the m-th layer, N_j stands for the set of the output characteristic pattern of the m—1-th layer, H_j^m stands for the j-th convolutional kernel of the m-th layer, β_j^m stands for the bias term of the characteristic pattern of corresponding convolutional kernel, and * stands for convolution operation. Different from the activation function in time delay neural network, the activation function in convolutional neural network is rectified linear unit rather than the traditional signoid function.

Pooling layer, also called subsampling layer, is mainly responsible for compressing the characteristic pattern obtained by the convolutional layer.

$$\alpha_j^m = f(\mathrm{down}(\alpha_j^{m-1}) + \beta_j^m), \tag{10}$$

where down(·) stands for subsampling function. The pooling layer can reduce training parameters through compressing data size and moreover make network obtain certain invariance, i.e., increase error-tolerant rate for speech recognition system.

3.4. Feedforward Sequence Memory Neural Network

Feedforward sequence memory neural network essentially is an improved version of the traditional deep neural network. It can be found from the comparison between Figs. 2 and 5 that the improvement of feedforward sequence memory neural network is the addition of memory block in the hidden layer. Memory block is a block structure used for storing historical information of the hidden layer, and it will be output to the next layer along with normal hidden layer. Memory block is the key of feedforward sequence memory neural network, which makes the neural network be capable of learning the long-term dependence relationship between sequence data in case of no feedback layer. Then feedforward sequence memory neural network can be divided into scalar coefficient based feedforward sequence memory neural network and vector coefficient based feedforward sequence memory neural network. The memory module in scalar coefficient based feedforward sequence memory neural network [12] can be expressed as:

$$\mathbf{s}_t^m = \sum_{i=0}^{R} \alpha_i^m \mathbf{s}_{t-i}^m, \tag{11}$$

where $\mathbf{s}_t^m$ stands for the output of memory module at time t, and α_i^m stands for the scalar coefficient of memory module at time i.

The memory module in vector coefficient based feedforward sequence memory neural network can be expressed as:

$$\mathbf{s}_t^m = \sum_{i=0}^{R} \boldsymbol{\alpha}_i^m \bullet \mathbf{s}_{t-i}^m, \tag{12}$$

where $\boldsymbol{\alpha}_i^m$ stands for the vector coefficient of memory module at time i, and • stands for the multiplication of two vector elements.

It is found from the comparison between equation (11) and (12) that the difference of memory module between the two types of feedforward sequence memory neural network is the selection of coding coefficient. Scalar coefficient based feedforward sequence memory neural network shares one scalar coefficient, while vector coefficient based feedforward sequence memory neural network uses different vector coefficients.

4. VOCAL CHARACTERISTICS AND DISCRIMINATIVE TRAINING

4.1. Vocal Characteristics

The ultimate goal of speech recognition system is to convert text information in speech signals to texts. But actually speech signals contain the special characteristics of voice, and undesired impurities cannot be thoroughly removed. Especially when speech signals are continuous, long paragraph, the influence of impurities on speech recognition is larger. The current common solutions to the problem include establishing an acoustic model which normalizes speaker and is irrelevant to speaker and establishing a speak related acoustic model. Speaker related acoustic model *i*-vector was selected in this study. The expression of the model [13] is:

$$N = n + Uv, \tag{13}$$

where n stands for mean value supervector of universal background model, U stands for the total variability subspace, and v stands for the vocal characteristic *i*-vector obtained by calculation.

Universal background model was established using Gaussian mixture model in this study. English speeches were used as training data. Total variability subspace was constructed using expectation maximization algorithm. As a large amount of data were needed in the construction of complete *i*-vector characteristics, *i*-vector characteristics of short sentences were used.

5. PERFORMANCE ANALYSIS OF ACOUSTIC MODEL IN SPEECH RECOGNITION

5.1. Model Comparison

5.1.1. Experimental environment. The test was carried out on laboratory servers (Windows 7, I7 processor and 16 G memory). Kaldi (stable version) was used for writing speech recognition system [15].

5.1.2. Experimental data. English speech data set [15] in UCI machine learning library was used. Speakers in the dataset nearly covered people who aged 12 ~ 70 years. 10000 speeches with clear and standard pronunciation were selected; 9000 speeches were randomly selected as training sets, and the remaining 1000 speeches were selected as testing sets. Experiment personnel read sentences, and the speech feature parameters of the sentences were collected at the sampling frequency of 16 KHz and using 16-bit code.

5.1.3. Performance evaluation criteria. Word error rate was regarded as the main evaluation criterion [13]. It refers to the ratio of the number of words wrongly recognized to the total number of words; the smaller the rate was, the better the recognition performance was. The calculation methods are shown below.

Table 1. The recognition performance of different models

Extraction features	Word error rate of Gaussian mixture model/%	Word error rate of deep neural network/%	Word error rate of time delay neural network/%	Word error rate of convolutional neural network/%	Word error rate of feedforward sequence memory network/%
MFCC	24.51	12.44	13.04	14.56	11.49
Fbank	24.11	11.94	12.44	14.06	11.19

The formula of word error rate is:

$$WER = \frac{X + Y + Z}{P} \times 100\%, \tag{14}$$

where X stands for the number of the replaced words, Y stands for the number of the deleted words, Z stands for the number of the inserted wrong words, and P stands for the number of all the words in the data set.

5.1.4. Setting of experimental model. Traditional Gaussian mixture model-hidden Markov model acoustic model. It was used as the basic model for neural network acoustic model. Firstly 39-dimensional Mel-frequency cepstral coefficient (MFCC) and Fbank characteristics were extracted from the training set for training Gaussian mixture model-hidden Markov model model. Dimension refers to the unit of characteristic vector number. During training, optimization was firstly performed using maximum likehood estimate criterion and criterion of minimum phoneme error rate. N-gram linguistic model was used in decoding. The dataset of the training linguistic model was the same with the acoustic model. Models were regarded as replacement basis for the neural network model in the following after training. Moreover the performance of the traditional model was tested using the testing set.

(1) Feedforward fully-connected deep neural network based acoustic model;

(2) The structure of feedforward fully-connected deep neural network included one input layer, one output layer and five hidden layers. Each hidden layer included 1024 nodes. The activation function of all hidden layers was Relu function. The MFCC and Fbank characteristics were extracted from the training set for training. The initial working number was set as 3, and the initial learning rate was set as 0.3. At the beginning, the learning rate remained unchanged; after 4 times of iteration, the learning rate reduced by half in each iteration until the 10th iteration ended. The performance of the model was tested using testing set after training;

(3) Time delay neural network based acoustic model. The structure of time delay neural network included five hidden layers. Each hidden layer included 1024 nodes. The activation function of the nodes was ReLU function. In every layer, 1500 was input, and 300 was output. The five hidden layers were downsampled as $\{-3, -2, -1, 0, 1, 2, 3\}$, $\{0\}$, $\{-1, 1\}$, $\{0\}$ and $\{-3, 3\}$ respectively. The performance of the model was tested using the testing set after training;

(4) Convolutional neural network based acoustic model.

The structure of convolutional neural network based acoustic model included five hidden layers, 2 convolutional layers and 3 ordinary hidden layers. Firstly 2 convolutional layers were trained, and one of them included 10 convolutional kernels whose size was 5 × 5. After the first convolutional operation, 10 feature vectors were obtained. Next the maximum value was downsampled to 3 × 3. The second convolutional layer included 15 convolutional kernels whose size was also 5 × 5. 150 feature vectors were obtained after the first convolutional operation. Then the maximum value was downsampled to 3 × 3. The output was processed by one-dimensional expansion and regarded as the input of the perception layer. Then a multi-layer perceptron which included 3 hidden layers and 2048 nodes and whose activation function of nodes was ReLU function was trained. The convolutional layer was merged with the hidden layer. The convolutional neural network obtained after merging was retrained following frame cross entropy criterion. The performance of the model was tested using the testing set after training.

(5) Feedforward sequence memory neural network based acoustic model

Feedforward sequence memory neural network cannot be directly generated by Kaldi, but can be created using the functional components in the software. The model included 5 hidden layers, each layer included 1024 nodes, and the activation function of the nodes was ReLU function. There were 3 memory blocks, which were placed at the 1st, 3rd and 5th layer. The order of look ahead and back of all the memory blocks was set as 30, and MFCC and Fbank characteristics were input. The working number was 3 at first

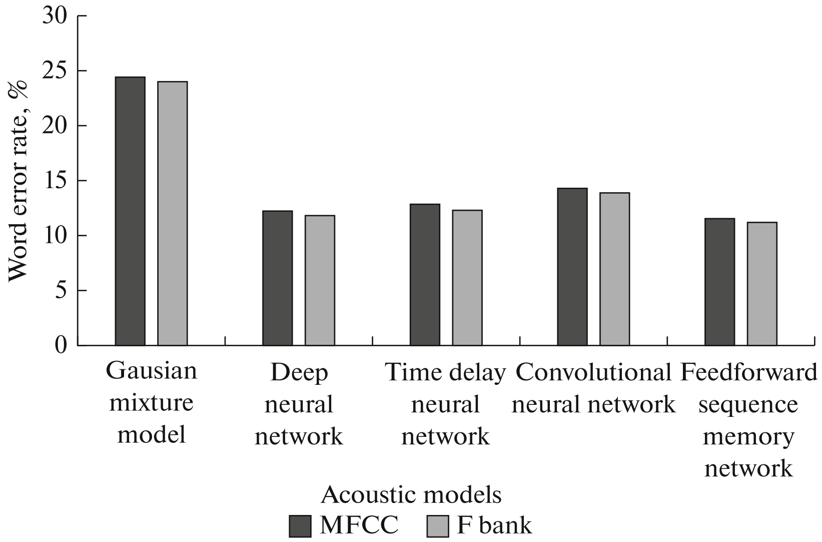

Fig. 6. The performance of speech recognition systems based on different models under different extraction features.

and increased to 5 with the increase with training times. The performance of the model was tested using the testing set after training.

(6) Optimization of neural network based acoustic model added with vocal characteristics

Firstly Gaussian mixture model-universal background model was trained, and the parameters used in the training were the mean value of some characteristics which were randomly selected and the variance of all characteristics. The total variability subspace U was randomly initialized. U parameters were adjusted using 10 times of expectation maximization algorithm iterations. When the *i*-vector characteristics were extracted from the training set and testing set, Fbank characteristics were used. Then the original Fbank characteristics and *i*-vector characteristics were input into the aforementioned neural network based acoustic model. The extraction dimension of *i*-vector characteristics was 40, 80, 120 and 160 respectively. Finally the performance of the model was tested.

5.2. Results

5.2.1. Performance of different models. Extraction feature MFCC in Table 1 makes equidistant partition of voice on Mel scale, which is more approximate to human auditory system than the linear interval band used in normal logarithmic cepstrum. Extraction feature Fbank can effectively filter the frequency point at a specific frequency in an audio to get audio signal containing the specific frequency.

As shown in Fig. 6, the word error rate of the traditional Gaussian mixture model based on MFCC, deep neural network model, time delay neural network model, convolutional neural network and feedforward sequence memory network model were 24.51, 12.44, 13.04, 14.56 and 11.49% respectively. The traditional Gaussian mixture model based on Fbank, deep neural network, time delay neural network model, convolutional neural network model and feedforward sequence memory network model were 24.11, 11.94, 12.44, 14.06 and 11.19% respectively. The lateral comparison of performance of speech recognition sys-

Table 2. The influence of vocal characteristics on the performance of deep neural network model

Vocal feature dimensions	0 dimension	40 dimension	80 dimension	120 dimension	160 dimension
Gaussian mixture model	24.11%	24.01%	23.88%	23.67%	23.98%
Deep neural network model	12.74%	12.42%	12.21%	12.51%	12.63%
Time delay neural network model	12.94%	12.64%	12.25%	12.11%	12.33%
Convolutional neural network model	14.56%	14.15%	13.99%	13.88%	14.05%
Feedback sequence memory network model	11.49%	10.88%	10.98%	11.28%	11.45%

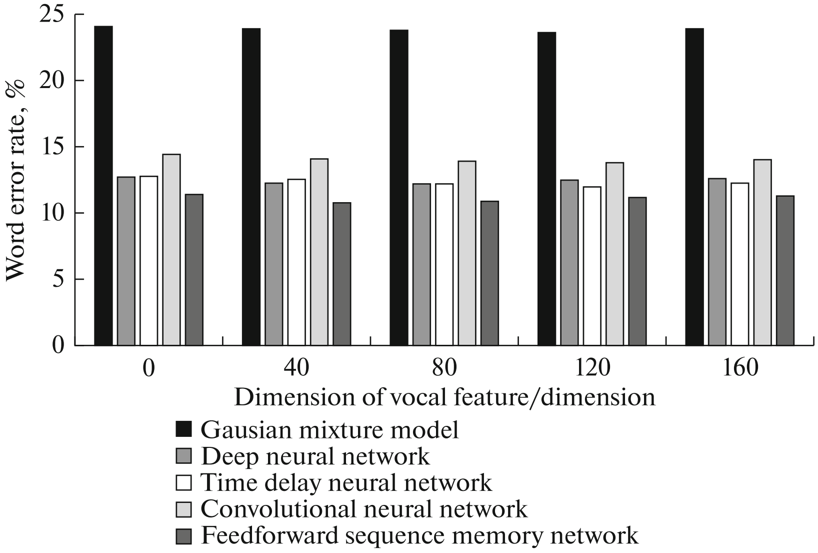

Fig. 7. The identification results of models after the addition of vocal features.

tems based on different acoustic models under the same extraction feature suggested that the word error rate of the four deep neural models was much lower than that of the traditional acoustic model, and the word error rate of feedforward sequence memory network was the smallest, convoltuional neural network the largest, and time delay neural network the second largest and deep neural network the third.

5.2.2. Influence of vocal characteristics on the performance of deep neural network model. As shown in Table 2 and Fig. 7, 0 dimension represents the absence of human voice characteristics. The vertical comparison of the word error rates of different models under the same dimension of human voice feature indicated that the performance of Gaussian mixture model was similar to the results in the last section. T he word error rate of Gaussian mixture model was the largest, about half higher than other models, the word error rate of feedforward sequence memory model was the smallest, and the word error rate of convolutional neural network model is the second largest. Before 80 dimensions, the word error rate of time delay neural network model was higher than that of deep neural network model, which was in the third place. At 120 dimensions and 160 dimensions, the word error rate of deep neural network model was higher than that of time delay neural network model, which was in the third place. The lateral comparison of the recognition performance of the same model with different dimensions of human voice features suggested that the word error rate of each model decreased after human voice features were added, that is, the recognition performance was improved; with the increase of the dimension of human voice characteristics, the word error rate of different models showed a trend of decreasing first and then increasing, and changes of word rates of different models were diffferent. As shown in Table 2 and Fig. 7, Gaussian mixture model had the smallest error rate, 23.67%, at 120 dimensions of human voice characteristics; deep neural network had the smallest error rate, 12.21%, at 80 dimensions of human voice characteristics; time delay neural network had the smallest error rate, 12.11%, at 120 dimensions of human voice feature; convolutional neural network had the smallest error rate, 13.88%, at 120 dimensions of human voice characteristics; feedforward sequence memory network had the lowest word error rate, 10.88%, at 40 dimensions of human voice characteristics.

6. CONCLUSION

The four neural network models had difference in performance under the same extraction feature. It was because that feedforward sequence memory network had a memory module which could store the information of previous input characteristics and context of the corresponding input layer and efficiently find out output information according to the stored information and current input characteristics compared to the other models; deep neural network connected context after characteristics were input; time delay neural network conected short context firstly, which was more advantegous than deep neural network in accuracy and than time delay neural network in output speed.

The vertical comparison of the performance of speech recognition systems based on the same model but under different extraction characteristics showed that the word error rate of the acoustic model using Fbank extraction feature was lower than that of the acoustic model using MFCC extraction feature, but the improvement was relatively small; the word error rate of the traditional Gaussian mixture model based on Fbank extration feature was 0.4% lower than that based on MFCC extraction feature; the word error rate of the deep neural network model based on Fbank extraction feature was 0.5% lower than that based on MFCC extration feature; the word error rate of time delay neural network model based on Fbank extraction feature was 0.6% lower than that based on MFCC extraction feature; the word error rate of convolutional neural network model based on Fbank extraction feature was 0.5% lower than that based on MFCC extraction feature; the word error rate of feedforward sequence memory network model based on Fbank extraction feature was 0.3% lower than that based on MFCC extraction feature. It shows that the performance of the model based on Fbank extraction feature was better than the same kind of model based on MFCC extraction feature. MFCC only divided the cepstrum of audio equally on the Mel scale. Although the frequency band was closer to the human ear, it does not remove the “impurity” from the audio, which caused information loss. Fbank filtered a specific frequency point in the audio, which effectively removed the “impurity” from the audio and retain the effective information as possible.

After adding human vocal features, the recognition performance of the five speech recognition models was improved. The reason was that the models could more accurately judge the unique frequency point of human voice in the audio when extracting features and retain the human voice frequency which conbtained effective information to the greatest extent. With the increase of dimension of vocal characteristics, the contained information increased, and the referable information increased during feature extraction; as a result, the accuracy improved. When the dimension of human vocal features exceeded a certain range, the redundant information became impurity in the process of feature extraction, which interfered the extraction of effective information. The reason why different models had different boundaries of human vocal feature dimension was that different models had different structures of judgment and learning.

Study on speech recognition system has lasted for a long time. The optimization of acoustic models and language models used in speech recognition system has gained some achievements. The influence of deep neural network based acoustic model on the performance of English speech recognition system was explored in this study, and moreover the performance of acoustic models was enhanced using several training methods. Finally the following results were obtained.

Gaussian mixture model based acoustic model and deep neural network, time delay neural network, convolutional neural network and feedforward sequence memory network based models were trained and tested using the same testing set. The results demonstrated that the performance of all the neural network based acoustic models was far better than that of Gaussian mixture model based acoustic model when MFCC and Fbank characteristics were selected; feedforward sequence memory neural network based model had the best performance, with word error rate of 11.49 and 11.09%; convolutional neural network had the poorest performance, with word error ate of 14.56 and 14.16%; time delay neural network based model and feedforward fully-connected deep neural network based model had similar performance as the word error rates of the former were 13.04 and 12.64% respectively and the word error rates of the latter were 12.44 and 12.04% respectively. The performance of the model based on Fbank extraction feature was slihghtly superior to that of the same kind of model based on MFCC extraction feature.

The models had the improved recognition performance after the addition of vocal characteristics, and different models had different improvement after different dimensions of vocal features were added. Feedforward sequence memory network model which was added with 40 dimension of vocal feature had a word error rate of 13.88%, indicating the best improvement; deep neural network model which was added with 80 dimension of vocal feature had a word error rate of 12.21%, indicating the best improvement; convolutional neural network model and time delay neural network model which was added with 120 dimension of vocal feature had a word error rate of 13.88 and 12.11%, indicating the best improvement.

REFERENCES

1. Chan, W., Jaitly, N., Le, Q., and Vinyals, O., Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, *IEEE International Conference on Acoustics, Speech and Signal Processing,* Shanghai, 2016, pp. 4960–4964.
2. Wang, Y., Li, J. and Gong, Y., Small-footprint high-performance deep neural network-based speech recognition using split-VQ, *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4984–4988.

3. Wu, C., Karanasou, P., Gales, M.J.F., and Sim K.C., *Stimulated deep neural network for speech recognition*, in *Interspeech*, San Francisco, 2016, pp. 400–404.
4. Graves, A., Mohamed, A.R. and Hinton, G., Speech recognition with deep recurrent neural networks, *IEEE International Conference on Acoustics, Speech and Signal Processing,* Vancouver, BC, 2013, pp. 6645–6649.
5. Salvador, S.W. and Weber, F.V., US Patent 9153231, 2015.
6. Cai, J., Li, F., Zhang, Y., and Liu, Y., Research on multi-base depth neural network speech recognition, *Advanced Information Technology, Electronic and Automation Control Conference,* Chongqing, 2017, pp. 1540–1544.
7. Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y., Attention-based models for speech recognition, *Comput. Sci.,* 2015, vol. 10, no. 4, pp. 429–439.
8. Miao, Y., Gowayyed, M., and Metze, F., *EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding*, *Automatic Speech Recognition & Understanding, Scottsdale,* 2015, pp. 167–174.
9. Schwarz, A., Huemmer, C., Maas, R. and Kellermann, W., Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments, *IEEE International Conference on Acoustics, Speech and Signal Processing,* 2015, pp. 4380–4384.
10. Kipyatkova, I., Experimenting with hybrid TDNN/HMM acoustic models for Russian speech recognition, *Speech and Computer: 19th International Conference,* 2017, pp. 362–369.
11. Yoshioka, T., Karita, S. and Nakatani, T., Far-field speech recognition using CNN-DNN-HMM with convolution in time', *IEEE International Conference on Acoustics, Speech and Signal Processing,* Brisbane, 2015, pp. 4360–4364.
12. Wang, Y., Bao, F., Zhang, H. and Gao, G.L., *Research on Mongolian speech recognition based on FSMN*, *Natural Language Processing and Chinese Computing,* 2017, pp. 243–254.
13. Alam, M.J., Gupta, V., Kenny, P., and Dumouchel, P., Speech recognition in reverberant and noisy environments employing multiple feature extractors and i-vector speaker adaptation', *EURASIP J. Adv. Signal Process.,* 2015, vol. 2015, no. 1, p. 50.
14. Brayda, L., Wellekens, C., and Omologo, M., N-best parallel maximum likelihood beamformers for robust speech recognition, *Signal Processing Conference,* Florence, 2015, pp. 1–4.
15. Ali, A., Zhang, Y., Cardinal, P., Dahak, N., Vogel, S., and Glass, J.R., A complete KALDI recipe for building Arabic speech recognition systems, *2014 Spoken Language Technology Workshop*, South Lake Tahoe, NV, 2015, pp. 525–529.