ELEKTROTEHNIŠKI VESTNIK 88(1-2): 33-40, 2021
ORIGINAL SCIENTIFIC PAPER

A Parameter Transfer Method for an HMM-DNN Heterogeneous Model Using a Scarce Mongolian Language Data Set

Zhiqiang Ma[1, 2], Junpeng Zhang[1], Tuya Li[1], Rui Yang[1] and Hongbin Wang[1,2]

[1] College of Data Science and Application. Inner Mongolia University of Technology, China
[2] Inner Mongolia Autonomous Region Engineering & Technology Research Centre of Big Data Based Software Service, China
E-mail: mzq_bim@imut.edu.cn

Abstract. The Hidden Markov Model-Deep Neural Network (HMM-DNN) is one of the most successful architectures in the speech recognition. Although HMM-DNN achieves state-of-the-art results in the English and Mandarin language, and there are many un-updated parameters in the training of the HMM-DNN acoustic model on a small-scale Mongolian data set. This involves the model network training under-fitted to learn the data set features. In a speech-recognition scenario, the under-fitting of speech features leads to a problem of the system accuracy. The paper defines a concept of a homogeneous heterogeneous model and proposes a parameter learning method for the HMM-DNN heterogeneous model for a scarce Mongolian data set. KALDI is used as an experimental platform, the TIMIT English data set as a source data set, and the scarce Mongolian data set as a target data set. Using the proposed parameter transfer method, a considerably improved recognition accuracy for the Mongolian data set is achieved.

Keywords: HMM-DNN Model; Homogeneous Model; Heterogeneous Model; Parameter Transfer; Transfer Learning

Metoda prenosa parametrov za heterogeni model HMM-DNN z redkim mongolskim naborom podatkov

Globoka nevronska mreža s prikritimi Markovovi modeli (HMM-DNN) je ena od najuspešnejših arhitektur pri prepoznavanju govora. Čeprav je HMM-DNN dosegel najsodobnejše rezultate v angleščini in mandarinščini, ugotavljamo, da med učenjem akustičnega modela HMM-DNN na majhnem mongolskem naboru podatkov obstaja vrsta neposodobljenih parametrov. To je povzročilo pomanjkljivosti pri učenju mreže, ki ne more razpoznati pomena iz nabora podatkov. V prispevku smo predlagali metodo učenja parametrov za HMM-DNN pri redkem mongolskem naboru podatkov. Pri eksperimentalnem delu smo izbrali platformo KALDI z izvornimi podatki TIMIT English in redki mongolski nabor podatkov kot ciljni nabor podatkov. S predlagano metodo smo dosegli boljše rezultate pri razpoznavanju mongolskih podatkov.

1 INTRODUCTION

Speech recognition is a process of automatic translating the human speech into a text in re-al-time [1-3]. The field of speech recognition has been divided into a large-scale continuous vocabulary recognition (LSCVR) and small-scale continuous vocabulary recognition (SSCVR) according to the data-set size. Since it is relatively difficult to collect and annotate the Mongolian speech data, the Mongolian data-set open-sourced is pretty scarce. Therefore, the Mongolian speech-recognition research is mainly focused on SSCVR. In acoustic modelling of the Mongolian speech recognition, the Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) is one of the most successful acoustic models, for its high efficiency and low data-set demand. But GMM-HMM is also limited in out of vocabulary (OOV) and parameter overfitting [4-6]. The Hidden Markov Model-Deep Neural Network (HMM-DNN) approach beats the GMM-HMM accuracy in the English and Mandarin speech recognition. But, the HMM-DNN approach has a pretty strong demand of a massive training data compared to the GMM-HMM approach [7, 8]. How to resolve the HMM-DNN training underfitting with a small-scale data set like the Mongolian data-set has become a research hotspot [9-11].

To improve the performance of deep learning models with a small-scale data set, some researchers prompt transfer learning which can solve the basic problem of an insufficient training data [12-15]. Based on the above, the paper proposes a parameter transfer method for the HMM-DNN Mongolian speech recognition.

The main contributions of the paper are as follows. The concept of a homogeneous and heterogeneous

Received 17 December 2020
Accepted 5 February 2021

model is defined. A parameter transfer method for the heterogeneous Mongolian model is proposed. English is used as the source language and Mongolian as the target language to set-up on HMM-DNN homogeneous model. Based on experimented results, the proposed parameter transfer method achieves a reduction of 1.11% on the word error rate (WER) and 4.35% on the sentence error rate (SER).

2 Related Work

For the Mongolian language, the annotated speech data set to create high a quality training HMM-DNN model is scarce. To meet the demand for high-quality ASR systems for the Mongolian language, different approaches can be taken.

One of the solutions is transfer learning, which is a machine learning technique to improve a model performance by pre-training on data from a related domain. Based on the techniques used in transfer learning, it can be classified into four categories: instances-based deep transfer learning [16, 17], mapping-based transfer learning [18, 19], network-based transfer learning [20, 21], and adversarial-based transfer learning [22, 23]. The instances-based transfer learning refers to the use a specific weight adjustment strategy, i.e., selection of partial instances from the source domain as supplements to the training set in the target domain by assigning appropriate weight values to the selected instances. The mapping-based transfer learning maps instances from two domains into a new data space with a better similarity. The network-based transfer learning refers to reusing a partial network pre-trained in the source domain, including its network structure and connection parameters, and transfer it to be a part of the neural network used in the target domain. Adversarial-based transfer learning uses on adversarial technology to find transferable features suitable for the two domains. The focus of the paper is on the network-based transfer learning.

The predominant one applied to ASR with scarce dataset are transfer learning [24] and training the model on multiple languages simultaneously [25]. To achieve a good performance, while a large amount of data is needed.

In terms of how much data is required and which architecture performs better, model training experiments are needed to transfer from one language to another language [24, 26]. The parameters learning from the source domain language to target domain language is similar to pre-training.

Unlike above, the presented approach trains the HMM-DNN heterogeneous model. The model architecture, too, differs from the traditional transfer learning models.

3 Definition of the Homogeneous Model

Definition 1: Model architecture. $T=(N,P,F,L)$, where N are the nodes of the network. $N=(N_1,N_2,...,N_i,...,N_l)$, N_i is the number of the nodes in the i-th layer of the deep neural network, and P is the parameter matrix of the network. $P=\{P_1^2,P_2^3,...,P_{i+1}^i,...,P_{l-1}^l\}$, where $P_{i+1}^i=\{w_i^{i+1},B_i\}$. w_i^{i+1}, B_i is the parameter weight matrix and the bias vector between the i-th layer and the (i+1)-th layer in the neural network, respectively. $F=\{g(\cdot),\ o(\cdot)\}$, where F is the network function and $g(\cdot)$ and $o(\cdot)$ are the activation function in the hidden layer of the neural network and the function of the output layer in the neural network, respectively. $L=(L_1,L_2,...,L_i,...,L_l)$ where L is the depth of the neural network and L_i are the hidden layers of the i-th layer.

Definition 2: Network model architecture. The architecture of a deep neural network model is described as M. If consists of a three-layer model structure and is expressed as $M=(T_I,T_H,T_O)$. T_I is the layer model structure of the model input layer. $T_I=(N,F,P,L)$, where $T_IN=\{N_1\}$, $T_IF=\varnothing$, $T_IP=\{P_1^2\}$. T_H is the layer model structure of the hidden layer. $T_H=(N,F,P,L)$, where $T_H.P=\{P_2^3,\ldots,P_i^{i+1},\ldots,P_{L-2}^{L-1}\}$, $T_H.F=\{g(g)\}$, $T_H.P=\{P_2^3,\ldots,P_i^{i+1},\ldots,P_{L-2}^{L-1}\}$, $T_H.L=(L_1,\ldots,L_i,\ldots,L_{l-1})$. T_O is the layer model structure of the model output layer. $T_O=(N,F,P,L)$, where $T_O.N=\{N_l\}$, $T_l.F=\{o(g)\}$, $T_l.P=\{P_{l-1}^l\}$,$T_l.L=\{L_L\}$.

Definition 3: Data set. The data set is used to represent the data set of the training model described as $D=\{X,Y\}$, where X is the feature data and Y is the label data. $D_S=\{X_S,Y_S\}$ and $D_T=\{X_T,Y_T\}$ are the source and target data, respectively.

Definition 4: Homogeneous model. When the source model $M_S=\{T_{SI},T_{SH},T_{SO}\}$ is the same as the layer structure corresponding to the target model $M_T=\{T_{TI},T_{TH},T_{TO}\}$. $T_{SI}=T_{TI},T_{SH}=T_{TH}$, source model M_S is homogeneous with target model M_T . It is expressed as $M_S=M_T$, and $T_{SO}=T_{TO}$.

In the homogeneous model, $T_{SI}.P=\{P_1^2\}$, $T_{SH}.P=\{P_2^3,...,P_i^{i+1},...,P_{l\text{-}2}^{l\text{-}1}\}$, . $T_{SO}.P=\{P_{l-1}^l\}$. In model M_S and $T_{TI}.P=\{P_1^2\}$, $T_{TH}.P=\{P_2^3,...,P_i^{i+1},...,P_{l\text{-}2}^{l\text{-}1}\}$, $T_{TO}.P=\{P_{l-1}^l\}$ in model M_T belongs to the same matrix shape. When the parameter transfer of the model is performed, the $T_{SI}.P$, $T_{SH}.P$ and $T_{SO}.P$ pairs in source model M_S are directly transferred to the positions corresponding to $T_{TI}.P$, $T_{TH}.P$ and $T_{TO}.P$ in the M_T to obtain transfer model M'_T . See algorithm 1 for the specific parameter transfer process.

Algorithm 1. Parameter transfer algorithm of the homogeneous model.

Input: X_S, Y_S, X_T, Y_T

Output: M'_T

1: $initialize(M_S)$;

2: $M_S \leftarrow train(X_S, Y_S, M_S)$;

3: $M_T \leftarrow M_S$;

4: $M'_T \leftarrow train(X_T, Y_T, M_T)$;

5: **return** M'_T ;

4 Definition of the Heterogeneous Model

Definition 5: Heterogeneous model. T_{SI} and T_{SH} of the source model and T_{SI} and T_{SH} of the target model $M_T = \{T_{TI}, T_{TH}, T_{TO}\}$ are the same. T_{SO} and T_{TO} are different and are denoted as $M_S.T_{SO} \neq M_T.T_{TO}$. It indicates that the M_S model and the M_T model are heterogeneous models, denoted as $M_S <> M_T$.

The heterogeneous model transfer training process refers to the parameters transfer of $T_{TH}.P$ in source model M_S constructed by source data D_S to target model M_T constructed by target data D_T to obtain transferred model M'_T .

In the heterogeneous model, since $T_{SO}.N = \{N_l\}$ of the layer structure in model M_S is different from $T_{TO}.N = \{N_l\}$ of the layer structure in model M_T , the parameter matrix of model M_S is not of the same shape are model M_T . Therefore, during the parameter transfer, the parameters of model M_S can't be directly transferred into the parameter matrix of model M_T , which increases the difficulty of the parameter transfer. The heterogeneous model parameter transfer process is shown in Fig. 1.

In Fig. 1, $T_{SO}.P_{l-1}^{l}$ in the M_S model is different from $T_{TO}.P_{l-1}^{l}$ in the M_T model, i.e. $M_S.T_{SO}.P_{l-1}^{l} \neq M_T.T_{TO}.P_{l-1}^{l}$. $T_{SI}.P$, $T_{SH}.P$ in the M_S model and $T_{TI}.P$, $T_{TH}.P$ in the M_T model belong to the same shape matrix, i.e., $M_S.T_{SI}.P = M_T.T_{TI}.P$,

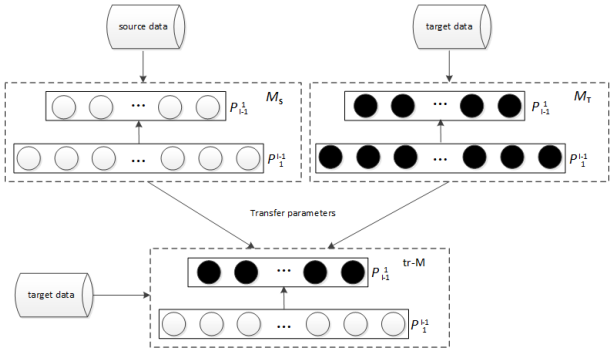


Figure 1. Parameter transfer process of the heterogeneous model.

Algorithm 2. Parameter transfer algorithm of the heterogeneous model.

Input: X_S, Y_S, X_T, Y_T

Output: M'_T

1: $initialize(M_S)$;

2: $M_S \leftarrow train(X_S, Y_S, M_S)$;

3: $M_T \leftarrow initialize(M_T)$;

4: $M_T \cdot T_{TI} \cdot P \leftarrow M_S \cdot T_{SI} \cdot P$;

5: $M_T \cdot T_{TH} \cdot P \leftarrow M_S \cdot T_{SH} \cdot P$;

6: $M'_T \leftarrow train(X_T, Y_T, M_T)$;

7: **return** M'_T ;

$M_S.T_{SH}.P = M_T.T_{TH}.P$. The heterogeneous model parameter transfer algorithm is shown in Algorithm 2.

5 Parameter Transfer Method for the Heterogeneous Model

The HMM-DNN model provides more labelled speech data to train better the model and creates transfer learning of more features [18]. Also, the constructed speech-recognition system ensures a lower word error rate (WER) and a lower sentence error rate (SER). However, for the languages like Mongolian, which are used for a relatively small area, it is difficult to label a large-scale speech data. Thus, improving the prediction accuracy and reducing the word error rate and the sentence error rate with a scarce speech data are a problem of the HMM-DNN acoustic model training [27].

The HMM-DNN structure consists of two parts: the HMM and the DNN model. The main function of the former is to use phonemes to predict different states and of the latter is to classify the input speech features by phoneme states. Therefore, the parameter transfer training for the HMM-DNN acoustic model is the parameter transfer training for the DNN model. The DNN model consists of the input layer, hidden layer, and softmax layer. The input layer receives the voice and audio feature data. The hidden layer is responsible for a deep extraction of the audio features. The softmax layer classifies and predicts of the speech features. The number of the neuron nodes in the softmax layer corresponds to the number of the phonemes in the language. So, in different languages, P_{l-1}^{l} of the softmax layer of the DNN model will be different. The DNN models constructed for diverse languages are usually of different shapes, so they are heterogeneous models. In the paper, the parameter transfer approach is tested with four different scales of the source data set using the DNN model as the original model. The parameter transfer process of the DNN model is shown in Fig. 2.

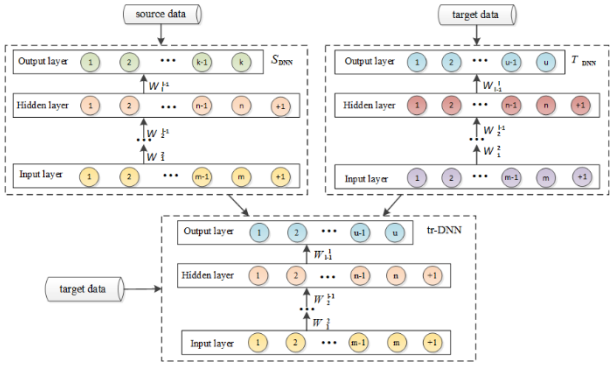


Figure 2. Parameter transfer process of the HMM-DNN acoustic model.

In Fig. 2, the model trained by data set D_s is denoted as M_S. The model trained by the data set D_T is denoted as M_T. The number of the nodes in $M_S.T_{Sl}$ is the same as the number of the nodes in $M_T.T_{Tl}$, both are m. The number of the nodes in each layer of $M_S.T_{SH}$ is the same as the number of the nodes in each layer of $M_T.T_{TH}$. If is n. The number of the nodes of $M_S.T_{SO}$ is k and the number of the nodes of the $M_T.T_{TO}$ is u. This indicates that $M_S.T_{Sl}.P.W$ and $M_T.T_{Tl}.P.W$, $M_S.T_{SH}.P.W$ and $M_T.T_{TH}.P.W$, $M_S.T_{Sl}.P.B$ and $M_T.T_{Tl}.P.B$, $M_S.T_{SH}.P.B$ and $M_T.T_{TH}.P.B$ are of the same matrix shape. While $M_S.T_{SO}.P.W$ and $M_T.T_{TO}.P.W$ are of a different shape represented as $M_S.T_{Sl}.P.W = M_T.T_{Tl}.P.W$, $M_S.T_{SH}.P.W = M_T.T_{TH}.P.W$, $M_S.T_{Sl}.P.B = M_T.T_{Tl}.P.B$, $M_S.T_{SH}.P.B = M_T.T_{TH}.P.B$, $M_S.T_{SO}.P.W \neq M_T.T_{TO}.P.W$, so $M_S.T_{Sl} = M_T.T_{Tl}$, $M_S.T_{SH} = M_T.T_{TH}$, $M_S.T_{SO} \neq M_T.T_{TO}$. This shows that the M_S and the M_T model are heterogeneous models denoted as $M_S <> M_T$.

According to the parameter transfer algorithm obtained with the heterogeneous model, the parameter transfer algorithm obtained with the DNN heterogeneous model is obtained algorithm 3.

Algorithm 3. Parameter transfer algorithm of the HMM-DNN heterogeneous model.

Input: X_S, Y_S, X_T, Y_T.

Output: M'_T

1: $initialize(M_S)$;
2: $M_S \leftarrow train(X_S, Y_S, M_S)$;
3: $M_T \leftarrow initialize(M_T)$;
4: $M_S.T_{Sl}.P.W = M_T.T_{Tl}.P.W$;
5: $M_S.T_{SH}.P.W = M_T.T_{TH}.P.W$;
6: $M_S.T_{Sl}.P.B = M_T.T_{Tl}.P.B$;
7: $M_S.T_{SH}.P.B = M_T.T_{TH}.P.B$;
8: $M'_T \leftarrow M_T$;
9: $M'_T \leftarrow train(X_T, Y_T, M_T)$;
10: **return** M'_T

6 Experimental Environment and Design

6.1 Experimental Environment

We use the TIMIT English data set is used as source data D_S to train source model M_S and use the Mongolian data set as target data D_T to train target model M_T. Due to the scarce population using the Mongolian language, it is not easy to collect a large size data set. The used Mongolian data set consists of 310 audios denoted as IMUT310. They are divided into a training and a test set. The training set has 287 Mongolian words, and the test set has 23 Mongolian words.

In the experiment, a 39-dimensional MFCC is used to extract the audio features. The employed Kaldi platform uses 300 hidden nodes, the initial-learning-rate is 0.015, the final learning rate is 0.002 and the activation function is a tanh function.

6.2 Experimental Design

The goal is to transfer the parameters from model M_S to model M_T. The obtained model is model M'_T. To verify the effectiveness of the proposed parameter transfer method, the following three comparative experiments are designed.

(1) Experiment of the Optimal data set D_S and the number of the layers of model M_S. Data set D_S is obtained from four different scales, which are 1000 sentences denoted as D_{S1000}, 2000 sentences denoted as D_{S2000}, 3000 sentences denoted as D_{S3000}, and 3696 sentences denoted as D_{S3696}. A three-tone scale is used as the acoustic model unit. The number of the hidden layers is 4, 5, 6 and 7, respectively. The number of the nodes in the hidden layer is 300. The epoch values are set to 5, 10, 15, 20, 25 and 30, respectively.

(2) Experimental comparison of the transfer and original model. The target data set is used to train original model M_T. Source data set D_S is selected to train source model M_S, the parameters of source model M_S are transferred to obtain the target model M'_T.

(3) Experimental comparison of different layers. Layers 4, 5, 6 and 7 of the original and transfer model our compared. The word error rate and the sentence error rate are compared.

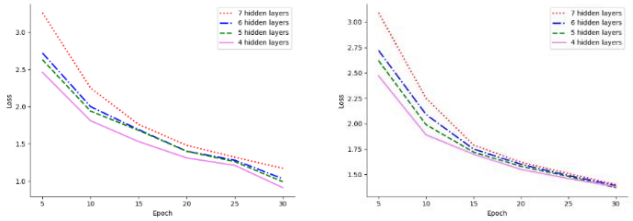
7 Analysis of Experimental Results

7.1 The Optimal Experimental Results of the Size of Data Set D_S and the Number of Layers of Model M_S

The experiments to select optimal model are carried out according to the scheme shown in Section 6.

7.1.1 The Optimal Experimental Results of the Number of Layers of Model M_S

Source model M_S with four different layers is trained with four different scales of data set D_S. The loss value is used as the evaluation index. Model M_S is trained with data sets D_{S1000}, D_{S2000}, D_{S3000} and D_{S3696}, respectively. The training result of model M_S is shown in Fig. 3.



(a) Loss of model M_S trained with data set D_{S1000}

(b) Loss of model M_S trained with data set D_{S2000}

(c) Loss of model M_S trained with data set D_{S3000}

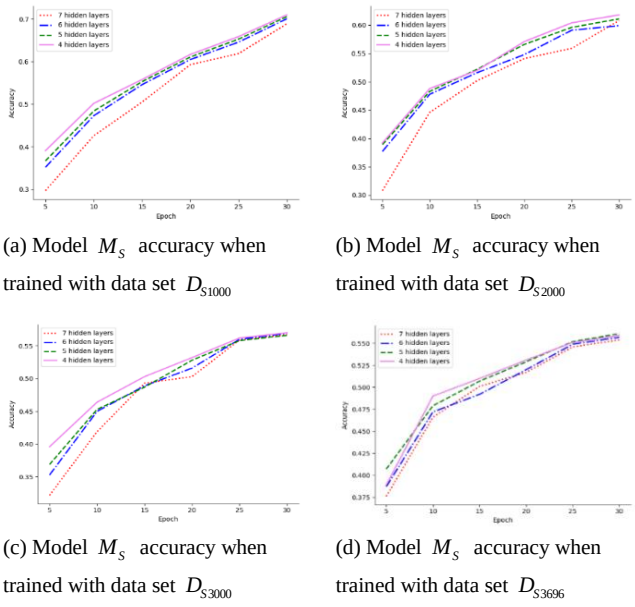
(d) Loss of model M_S trained with data set D_{S3696}

Figure 3. Loss of the HMM-DNN model trained with different data set scales.

Model M_S trained with four different data sets shows a gradual loss descending and finally reaching the bottom at around 1.0 to 1.5. The model M_S convergence is considered stable.

The model accuracy trained with four different data sets in different hidden layers is shown in Fig. 4. As seem from Figure 4, the accuracy of the model using different data sets and different hidden layers is around 0.4 to 0.5 in the first 10 epoch, then if gradually ascends reaching the peak value at approximately 0.55 to 0.7.

According to the experimental results shown in Fig. 3 and Fig. 4, data set D_{S1000} is used to train model M_S with four hidden layers, data set D_{S2000} to train model M_S with four hidden layers, data set D_{S3000} to train model M_S with four hidden layers and data set D_{S3696} to train the model M_S with five hidden layers.



(a) Model M_S accuracy when trained with data set D_{S1000}

(b) Model M_S accuracy when trained with data set D_{S2000}

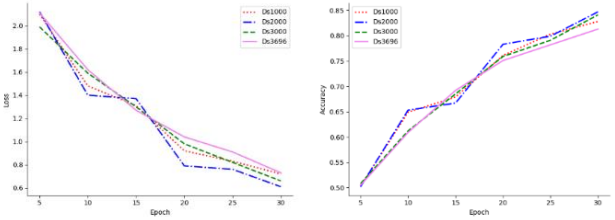
(c) Model M_S accuracy when trained with data set D_{S3000}

(d) Model M_S accuracy when trained with data set D_{S3696}

Figure 4. Model accuracy for different data sets D_S and different hidden layers

7.1.2 The Optimal Experimental Results of the Data Set D_S Size

The optimal M_S model obtained in (1) and M_T model are used to train transfer learning. Transfer target model M_T' is measured with the loss and accuracy. The experimental results are shown in Fig. 5.



(a) Loss of model M_T' when trained with different data sets (D_S)

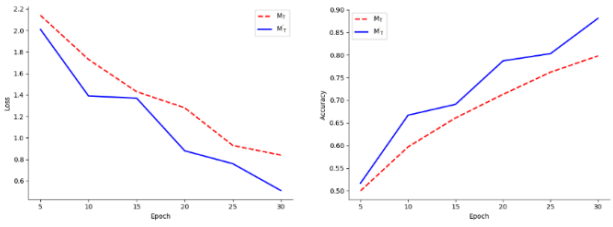
(b) Accuracy of model M_T' when trained with different data sets (D_S)

Figure 5. Loss and accuracy of model M_T' when trained with different scale data sets (D_S).

As seem from Fig. 5 (a), the model M_T' loss values for different data sets descend gradually, and the model M_T' loss value for data set D_{S2000} is 0.607. The line chart in Fig. 5 (b) shows that model M_T' accuracy for the data set D_{S2000} outperforms the others. It is shown that model M_T' transferred from model M_S and model M_T can exhibit the best performance. Therefore, an exhibit further experiments, source model M_S is trained using data set D_{S2000}.

7.2 Comparative Experimental Results of the Transfer and Original Model

To set up the parameter transfer method, the best source model obtained above is used. Original model M_T and target model $M_T^{'}$ are evaluated in terms of the loss and accuracy. Experimental results are shown in Fig. 6.



(a) Loss of model M_T and model $M_T^{'}$.

(b) Accuracy of model M_T and model $M_T^{'}$.

Figure 6. Loss and accuracy of model M_T and model $M_T^{'}$.

In Fig. 6, the loss of transfer model $M_T^{'}$ is lower than the loss of original model M_T and the accuracy of transfer model $M_T^{'}$ is higher than the accuracy of original model M_T. These results indicate that the Mongolian acoustic model is better after adding the parameter transfer method.

7.3 Comparative Experimental Results of the Different Layers

To verify the impact of the proposed parameter transfer method on the performance of the Mongolian speech-recognition model, the word error rate, and the sentence error rate of the Mongolian original model and transfer model with different hidden layers are determined. The experimental results are shown in Table 1.

Table 1. WER and SER of the Mongolian speech-recognition based on original model M_T and transfer model $M_T^{'}$

Acoustic model	Number of hidden layers	Word error rate (%)		Sentence error rate (%)	
		Training set	Validation set	Training set	Validation set
Original Model M_T	4	11.47	23.89	32.06	52.17
	5	10.19	21.67	27.18	43.48
	6	10.10	21.11	27.87	39.13
	7	9.30	22.22	25.44	47.83
Transferred Model $M_T^{'}$	4	9.53	23.33	23.54	47.83
	5	8.21	20.56	21.95	39.13
	6	7.93	20.00	20.91	34.78
	7	7.27	21.67	19.16	43.48

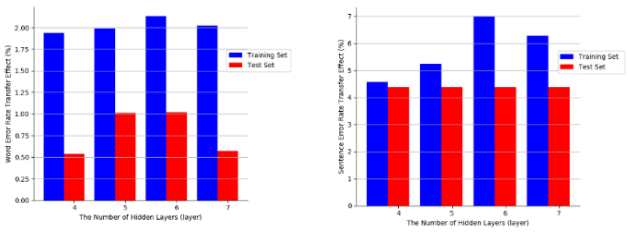
Table 1 shows that in the transferred Mongolian speech-recognition model. The word error rate and the sentence error rate are smaller than the original Mongolian speech-recognition model. This indicates that the proposed parameter transfer method performance well. Also, with the increase in the number of the hidden layers, the word error rate and the sentence error rate of the transfer model gradually decreased on the training set and the word error rate and the sentence error rate first decrease and then increase on the validate set. This indicates that the Mongolian speech-recognition model is over-fitting when the number of the hidden layers is over six. When the number of the hidden layers is six, the word error rate and the sentence error rate on the training set decrease by 2.17% and 6.96%, respectively. The word error rate and the sentence error rate on the validated set are decrease by 1.11% and 4.35%, respectively. Following the alone, the Mongolian speech-recognition model achieves the best performance using the proposed parameter transfer method.

The paper demonstrates the effectiveness of the proposed parameter transfer method by comparing it with the original Mongolian speech-recognition model with different hidden layers. The parameter transfer effect (PTE) is evaluated. It defines the difference between the error rate of the original model and the error rate of the transfer model (Equation 1). PTE is further divided into the word error rate of the transfer effect (WERTE) and the sentence error rate of the transfer effect (SERTE). WERTE defines the difference between the word error rate of the original model and the word error rate of the transfer model (Equation 2). SERTE defines the difference between the sentence error rate of the original model and the sentence error rate of the transfer model Equation (3).

$$PTE = ER_b - ER_p \tag{1}$$

$$WERTE = WER_b - WER_p \tag{2}$$

$$SERTE = SER_b - SER_p \tag{3}$$



(a) Word error rate

(b) Sentence error rate

Figure 7. Parameter transfer effect with different hidden layers.

In the current experimental environment, while the number of the hidden layers increases from four layers to six layers, the effect of the speech-recognition system ascends. When the number of the hidden layers of the model increases to seven layers, the experimental results show that the speech-recognition effect begins

descending and the model parameter transfer ability is limited.

8 CONCLUSION

The paper defines the concept of a homogeneous and a heterogeneous speech-recognition model. A parameter transfer method is proposed for the HMM-DNN heterogeneous model for the scarce Mongolian data set. The method assesses a 1.11%/4.35% WER/SER reduction compared to the original HMM-DNN model. Experiments show that the performance of the parameter transfer heterogeneous model is affected by the scale of the source data set and the number of the hidden layers. Therefore, the method is reliable on the scarce Mongolian data set. In the future work, an effort will be taken to improve the training methods for scarce data sets.

ACKNOWLEDGEMENT

Funding project: National Natural Science Foundation of China (61762070, 61862048), Natural Science Foundation of Inner Mongolia (2019MS06004), Inner Mongolia Autonomous Region Special Program for Engineering Application of Scientific and Technical Payoffs (2020CG0073), Inner Mongolia Key Technological Development Program (2019ZD015), Key Scientific and Technological Research Program of Inner Mongolia Autonomous Region (2019GG273).

REFERENCES

[1] Yu, Dong, and Li Deng. (2016) *Automatic Speech Recognition*. Springer london limited.

[2] Bijl, David, and Henry Hyde-Thomson. (2001) "Speech to text conversion." U.S. Patent No. 6,173,259. 9 Jan.

[3] Schroeder, Manfred R. (2013) *Computer speech: recognition, compression, synthesis.* Vol. 35. Springer Science & Business Media.

[4] Kanda, Naoyuki, Ryu Takeda, and Yasunari Obuchi. (2013) "Elastic spectral distortion for low resource speech recognition with deep neural networks." 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE.

[5] Dawa, I., Y. Sagisaka, and S. Nakamura. (2008) "Investigation of asr systems for resource-deficient languages." ACTA Automatica Sinica 1: 1-8.

[6] Long, F. E. I., G. A. O. Guanglai, and Yan Xueliang. (2013) "Research on Mongolian spoken term detection method based on segmentation recognition." Computer science 40.9: 208-211.

[7] Dahl, George E., et al. (2011) "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." IEEE Transactions on audio, speech, and language processing 20.1: 30-42.

[8] Li, Longfei, et al. (2013) "Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition." 2013 Humaine association conference on affective computing and intelligent interaction. IEEE.

[9] Singh, Atma Prakash, Ravindra Nath, and Santosh Kumar. (2018) "A Survey: Speech Recognition Approaches and Techniques." 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). IEEE.

[10] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. (2011) "Survey on speech emotion recognition: Features, classification schemes, and databases." Pattern Recognition 44.3: 572-587.

[11] Besacier, Laurent, et al. (2014) "Automatic speech recognition for under-resourced languages: A survey." Speech Communication 56: 85-100.

[12] Weiss, Karl, Taghi M. Khoshgoftaar, and DingDing Wang. (2016) "A survey of transfer learning." Journal of Big data 3.1: 9.

[13] Tan, Chuanqi, et al. (2018) "A survey on deep transfer learning." International conference on artificial neural networks. Springer, Cham.

[14] Yang, Qiang, et al. (2020) *Transfer learning*. Cambridge University Press.

[15] Ruder, Sebastian. (2019) *Neural transfer learning for natural language processing.* Diss. NUI Galway.

[16] Li N, Hao H, Gu Q, et al. A transfer learning method for automatic identification of sandstone microscopic images[J]. Computers & Geosciences, 2017, 103: 111-121.

[17] Liu X, Liu Z, Wang G, et al. Ensemble transfer learning algorithm[J]. IEEE Access, 2017, 6: 2389-2396.

[18] Arjovsky M, Chintala S, Bottou L. Wasserstein gan[J]. arXiv preprint arXiv:1701.07875, 2017.

[19] Long M, Zhu H, Wang J, et al. Deep transfer learning with joint adaptation networks[C]//International conference on machine learning. PMLR, 2017: 2208-2217.

[20] Chang H, Han J, Zhong C, et al. Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(5): 1182-1194.

[21] George D, Shen H, Huerta E A. Deep Transfer Learning: A new deep learning glitch classification method for advanced LIGO[J]. arXiv preprint arXiv:1706.07446, 2017.

[22] Long M, Cao Z, Wang J, et al. Domain adaptation with randomized multilinear adversarial networks[J]. arXiv preprint arXiv:1705.10667, 2017.

[23] Luo Z, Zou Y, Hoffman J, et al. Label efficient learning of transferable representations acrosss domains and tasks[J]. Advances in neural information processing systems, 2017, 30: 165-177.

[24] Wang, Dong, and Thomas Fang Zheng. (2015) "Transfer learning for speech and language processing." 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE.

[25] Chen, Dongpeng, and Brian Kan-Wing Mak. (2015) "Multitask learning of deep neural networks for low-resource speech recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing 23.7: 1172-1183.

[26] MA, Zhiqiang, et al. (2018) "Mongolian acoustic modeling based on deep neural network." CAAI Transactions on Intelligent Systems 3: 27.

[27] Ma, Zhiqiang, et al. (2017) "A pipelined Pre-training algorithm for DBNs." Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer, Cham. 48-59.

Zhiqiang Ma received his B.E. degree in computer application technology from the Hohai University, Nanjing, China, in 1995 and his M.E. degree in computer science and technology from the Beijing Information Science and Technology University, Beijing, China, in 2007. In 1995, he joined the Inner Mongolia University of Technology. Currently he is a Professor at the College of Data Science and Application. He is a reviewer of the Journal of Chinese Information Processing, IEEE ACCESS, and the International

Conference of Pioneering Computer Scientists, Engineers and Educators. His research interests include deep learning, speech recognition and natural language processing.

Junpeng Zhang received his B.E. degree in software engineer from the North University of China, Taiyuan, China, in 2018. In the same year, he joined the Inner Mongolia University of Technology. Currently he is a graduate student at the College of Data Science and Application. His research interests include deep learning and natural language processing with a special focus on speech recognition.

Tuya Li received her B.E. degree from the Inner Mongolia University of Technology, Inner Mongolia, China, in 2017. In the same year, she joined the Inner Mongolia University of Technology. Currently, she is a graduate student at the College of Information Engineering. Her research interests include deep learning and natural language processing with a special focus on emotional dialogue generation.

Rui Yang received her B.E. degree in computer science and technology from the Shanxi Agricultural University, Jinzhong, China, in 2017. In the same year, she joined the Inner Mongolia University of Technology. Currently she is a graduate student at the College of Information Engineering. Her research interests include deep learning and natural language processing with a special focus on emotional dialogue generation.

Hongbin Wang received his B.Sc. degree in Information and Computing Sciences from Shandong Agricultural University, Tai'an China, in 2012, and his M.Sc. degree in software engineering from the Inner Mongolia University, Hohhot, China, in 2018. He joined the Inner Mongolian University of Technology in 2018, where he is currently a teaching assistant with the College of Data Science and Application. His research interests include machine learning, machine translation, and automatic speech recognition.