2021 Ninth International Conference on Advanced Cloud and Big Data (CBD) | 978-1-6654-0745-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/CBD54617.2021.00043

Deep Neural Network Based Chinese Dialect Classification

Miao Wan[§], Jie Ren[§]*, Miao Ma[§], Zhiqiang Li[§], Rui Cao[¶], Quanli Gao[+]
[§]*Shaanxi Normal University, China,* [¶]*Northwest University, China,* [+]*Xian Polytechnic University, China*
renjie@snnu.edu.cn

***Abstract*—With the recent advance of neural networks in audio speech recognition (ASR), Deep Neural Network Based ASR has been widely used in multiple application scenarios such as smart homes, intelligent customer service, meeting minutes, and real-time translation. However, due to the ethnic variety of China, the pronunciation for the same word is different, which is a big challenge to the speech recognition system, especially for Short Utterances. This paper analyzes three kinds of commonly used speech feature parameters: spectrogram, MFCC (Mel-scale Frequency Cepstral Coefficients), and Fbank (Filter bank) and builds the dialect classification model based on a deep neural network for a dataset of 10 dialects in China. In detail, we study the geographical features of dialects and propose a multi-task model that uses the area of the dialect as an auxiliary task and builds a hard parameter sharing based multi-task learning model. The results show that the performance of this model can achieve up to 79.96%. Furthermore, as the hard parameter sharing model cannot effectively learn the correlation between sub-tasks, we then propose a sparse parameter sharing based multi-task learning model. The model uses joint training to automatically learn the correlation between sub-tasks, prune redundant networks, and share network parameters. The experiment results show that the sparse parameter sharing for the multi-task classification model achieves the best accuracy, with an average of 83.59%.**

Keywords—*Chinese dialect recognition; Hard-sharing parameter; Sparse-sharing parameter; Multi-task learning； neural network;*

I. INTRODUCTION

Speech recognition is an important part of human-computer interaction. Nowadays, deep neural learning-based speech recognition systems are becoming more mature and are widely used in many fields such as navigation, translation, smart home, in-vehicle systems, and teaching [1]. However, at present, due to the accent, dialect and other features of the user input voice, the intelligent speech recognition system often has the problem of inaccurate recognition [2], which requires the user to correct the accent and repeatedly input voice, which seriously affects the user experience. Therefore, automatically determining the input audio language in advance is a key step to improve the back-end performance of the speech recognition system.

Speech recognition is to analyze and determine the language genus of the input audio data, and further input the input audio data into the corresponding natural language processing model for reasoning based on the language classification result. This paper studies the classification of dialects in my country builds classification models of dialects based on neural networks, and conducts experiments to verify the classification of ten dialects in Northwest, North, Southwest, and Central China. Specifically, the main contributions of this paper are as follows:

(1) We consider the similarities between different dialect languages in the same geographic area and the correlation between different regional dialect languages, this paper is based on the three phonetic features of spectrogram, MFCC, and Fbank, the dialect area is used as an auxiliary task, and the construction is based on hard-sharing dialect language recognition model;

(2) We also found that the model based on parameter hard sharing only has good performance when the task relevance is strong, and the relevance of the dialect classification task cannot be clearly defined. Therefore, this paper proposes a multi-task model based on parameter sparse sharing. The model independently determines the correlation between tasks through joint training, and adaptively shares the overlapping parts of the subtask network.

II. RELATED WORK

A. Speech Language Recognition

Speech language recognition is an important branch of pattern recognition. Its first task is to extract features from the input sound as a multi-dimensional vector. Feature extraction is achieved by converting the speech waveform into a parameter representation form at a relatively minimum data rate for subsequent processing and analysis. Mel Frequency Cepstral Coefficients (MFCC) [3], Linear Prediction Cepstral Coefficients (LPCC) [4], Discrete Wavelet Transformation (DWT) [5] and Perceptual Linear Prediction (PLP) [6] is a common speech feature parameter. These methods have been tested in a wide range of applications and have high reliability and acceptability. Dialect research usually preprocesses speech into the above-mentioned characteristic parameters.

The mainstream of speech-language recognition was the first artificial neural network, followed by Gaussian mixture model, HMM, MVC and other models are also widely used. Then developed to the application based on neural network. In 2011, Ge et al. proposed a pre-trained DNN method based on context dependence [7], and Jiang Bing proposed a language recognition method based on deep neural network extraction of phoneme-related deep bottleneck features (DBF) in 2015 [8]. CNN has a wide range of applications in the field of speech recognition. O Koller et al. proposed a hybrid CNN-HMM model, focusing on the continuity of speech input [9]. In 2017, Microsoft added the CNN-BLSTM acoustic model to the new dialogue speech recognition system [10]. These have achieved good results. In recent years, recurrent neural network (RNN) has also been widely used in the field of speech recognition [11]. LSTM and GRU networks have also been proposed one after another. Because they improve the long-term dependence of

978-1-6654-0745-8/21/$31.00 ©2021 IEEE
DOI 10.1109/CBD54617.2021.00043

RNN, they have good performance in the recognition of serialized speech features. Related research has also improved the original LSTM and GRU models to achieve performance optimization, such as using a deep neural network that is a mixture of CNN and GRU to classify speech. [12].

In recent years, dialect recognition has mostly adopted end-to-end recognition models, usually using a network combining convolutional and recurrent layers [14]. Literature [2] uses convolutional neural networks to classify and recognize dialects in county-level cities in Jiangsu Province. Literature [19] applies multi-task learning to solve the problem of dialect language recognition, but only considers Fbank features and uses voice and audio comparisons long.

B. Research on speech recognition based on multi-task learning

Multitask Learning (MTL, Multitask Learning) assumes that there is a certain similarity between the data distributions of different tasks, and on this basis, the connection between tasks is established through joint training and optimization. Fully promote the exchange of information between tasks and achieve the goal of mutual learning. Each subtask can share some information by other subtasks, and indirectly use the shared information to achieve the purpose of improving the learning performance of each subtask [16].

S Ruder proposed the two most commonly used methods of MTL in deep learning [17]. Clarified the principles of MTL and provided guidelines for natural language processing and speech recognition.

Recently, multi-task learning and the LSTM recurrent network have provided research room for dialect language recognition. This paper uses the LSTM network to construct a single-task learning model and optimize its parameters. Further using the correlation of different regional dialects, a multi-task model based on parameter soft sharing is proposed to share hidden layer information to improve accuracy. This paper considers the similarities of dialects in the same dialect region, and uses the dialect region as an auxiliary task to propose a multi-task learning model based on hard sharing of parameters to improve the model's generalisation. The existing multi-task models for dialect classification research are limited to two sharing methods: hard parameter sharing and parameter soft sharing. These methods have great limitations on the choice of network parameters and network architecture. This article considers the flexibility of parameter selection, Proposed sparse sharing. In the selection of dialect languages, many studies mostly select a single regional dialect, such as the Hunan dialect. Some speech recognition related studies select minority languages such as Uyghur and Tibetan. This paper selects 10 Chinese dialect languages covering six major dialect regions. Furthermore, ten dialects have certain representative significance, covering a wide area, providing ideas for the classification of Chinese dialects.

III. VOICE FEATURES

Audio contains a wealth of voice feature information, and different feature vectors can often represent different acoustic meanings. From this, extracting voice features and mining the associated information of voice features are vital to language recognition. Speech feature extraction refers to selecting effective audio representations from a segment of speech. The original speech signal is a typical non-stationary signal, and the analysis method of speech feature extraction is for the stationary signal. Therefore, it is usually assumed that the short-term signal of 10ms-30ms is a stable signal, and the component analysis is performed based on this. In this paper, the long speech signal of the original audio is windowed and framed, and then multiple stable short-term signals are obtained. On this basis, the analysis and extraction of speech features are carried out. This paper extracts three kinds of speech features: spectrogram, Fbank and MFCC.

A. Spectrogram

$$X_a(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\Pi k/N} \qquad (1)$$

Where x(n) represents the input voice signal, and N represents the number of Fourier transform points. In order to amplify the energy difference at low energies, it is necessary to perform logarithmic operations on the linear spectrum, and finally obtain the spectrogram by calculating as shown in Equation 2. The abscissa represents time, and the ordinate represents frequency. The gray value of each pixel reflects the energy signal density at the corresponding time and frequency.

$$X(k) = 10 * log(X_a(k)) \qquad (2)$$

B. Filter Bank

The Fbank(Filter Bank) algorithm refers to the working principle of the human ear to process audio. The algorithm performs pre-emphasis on the speech signal, windowing, framing, Fourier transform, Mel filtering and logarithmic operations, and finally obtains the Fbank feature value. The human ear's perception of vibration is arranged according to the frequency range. Fbank takes the square of the absolute value on the basis of formula (1) to obtain the energy spectrum [18] , the abscissa of the energy spectrum is frequency, and the ordinate is The energy density is in line with the perception of vibration by the human ear. Furthermore, Mel filtering is performed on the energy spectrum to convert the linear natural frequency spectrum into a Mel frequency spectrum that conforms to the characteristics of human hearing. Set L filters (usually forty filters), use formula (3) to calculate the Mel filter bank, where m(l) is the value of the l-th Mel filter, W(k;l) Is the triangular window function, |S(k)| represents the energy spectrum, and k is the frequency. Finally, take the logarithm to get the Fbank feature.

$$m(l) = \sum_k W(k;l) \cdot |S(k)| \ (l = 1, \dots, L) \qquad (3)$$

C. Mel Frequency Cepstrum Coefficient

MFCC(Mel Frequency Cepstrum Conefficient) is a cepstrum parameter extracted in the frequency domain of the Mel scale. On the basis of formula 3, the discrete cosine transform (DCT) is used to separate the frequency components to obtain the cepstrum. As shown in Equation 4, the speech signal is composed of spectral details and a filter characteristic envelope equivalent to pronunciation. In order to better analyze the language signal, the cepstrum analysis separates the

spectrum details and the envelope. MFCC is the low-frequency part extracted from the cepstrum.

$$C(i) = \sqrt{\frac{2}{L}} \sum_{l=1}^{L} logm(l) \cdot cos\{(l - \frac{1}{2}) \frac{i\Pi}{L}\} \quad (4)$$

D. Comparative Analysis of Three Voice Features

Fig.1 shows the three speech feature extraction methods and their relationships among spectrogram, MFCC, and Fbank. The spectrogram is obtained by taking the logarithm after Fourier transform of the speech signal. Compared with MFCC and Fbank, the spectrogram saves more original speech signal information.

The Fbank feature is to further use the Mel filter bank for processing after the Fourier transform of the speech signal. And get the logarithm of the result. The MFCC feature is the cepstrum parameter obtained after the DCT off-chord transformation is performed on the basis of Fbank.

MFCC and Fbank refer to the sensitivity of humans to the level of sound, and use the Mel scale to quantify this sensitivity. The process of obtaining MFCC and Fbank is similar to the process of human ear processing sound. But the conversion from Fbank to MFCC is a lossy conversion, and only part of the information of the voice signal can be obtained. The transformation relationship between the three is shown in Fig.1

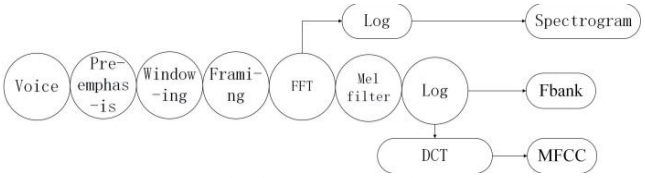


Fig. 1 Spectrogram, Fbank, MFCC Transformation Relationship Diagram

IV. DIALECT CLASSIFICATION MODEL BASED ON SINGLE TASK LEARNING

This section builds a single-task neural network model based on the three features of spectrogram, MFCC, and Fbank. Model construction includes three processes: data collection, feature extraction, and model training, as shown in Fig.2. The training and testing data set used in this paper are provided by the iFLYTEK AI Challenge Contest [28].

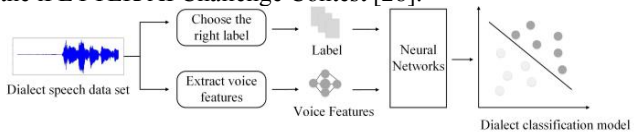


Fig. 2 Speech classification Model construction Process Diagram

This paper uses the Python wave library [23] to extract speech signal spectrograms, MFCC and HTK to extract 40-dimensional Fbank features. The data of 10 dialects (Minman, Kejia,Shanghai,Hefei,Shannxi,Ningxia,Changsha,Hebei, Nanchang, Sichuan) are processed into [data, label].

V DIALECT CLASSIFICATION MODEL BASED ON MULTI-TASK LEARNING

TABLE 1 DATASET DESCRIPTION

Dialect	Train-set		Test-set	
	The number of participants	Dialect audio number	The number of participants	Dialect audio number
Ningxia	30	6000	5	500
Hefei	30	6000	5	500
Sichuan	30	6000	5	500
Shaanxi	30	6000	5	500
Changsha	30	6000	5	500
Hebei	30	6000	5	500
Nanchang	30	6000	5	500
Shanghai	30	6000	5	500
Kejia	30	6000	5	500
Minnan	30	6000	5	500

Multi-task learning (MTL) has achieved certain success in the fields of natural language processing [20] and computer graphics [21]. The core is to train multiple related tasks at the same time and share parameters between different tasks to improve the overall generalization ability of the model. This section constructs a dialect classification model based on hard sharing of parameters, which uses dialect region classification as an auxiliary task, and a dialect classification model based on sparse parameter sharing.

A. Dialect classification model with dialect region recognition as an auxiliary task

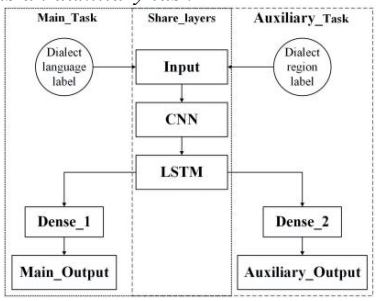


Fig.3 Model architecture diagram based on parameter hard-sharing

This section introduces the main task of dialect language classification, and the construction of a multi-task learning model based on parameter hard sharing with dialect region recognition as an auxiliary task. Chinese dialects can be divided into seven dialect regions: Mandarin dialects, Hakka dialects, Xiang dialects, Wu dialects, Cantonese dialects, Fujian dialects, and Gan dialects. Different dialect regions have their own characteristics. At the same time, the regional characteristics of different dialects also have a certain degree of intersectionality. For example, Gan dialects has long been influenced by the Hakka dialects close to the geographical position, resulting in its characteristics similar to Hakka dialects. This section uses the regional classification of dialects as an auxiliary task to provide the main task with different regional feature information of different dialects and the cross information of dialect regions, thus constructing a multi-task learning model based on parameter hard sharing. The parameter hard sharing mechanism shares the hidden layers between tasks and retains their respective output layers, which is suitable for situations with high task relevance. A dialect language classification

model that uses dialect region recognition as an auxiliary task. The main task dialect language classification and auxiliary task region recognition tasks share the same network architecture, share the hidden information of each dialect region, retain the output of the main task and auxiliary tasks, and the network will jointly train all the mission loss value, through the output evaluation result. The corresponding relationship between dialect regions and labels is shown in Table 2, and the experimental results are shown in Section VI.

The model structure is shown in Fig.3. The model implements hard parameter sharing in the sharing layer, calculates loss1 and loss2 of the main task and auxiliary tasks, and performs a weighted summation. The weights are all 0.5. Backward and update loss_sum, train and save the model.

TABLE 2 DIALECT REGION LABEL CORRESPONDENCE TABLE

Dialect area	Dialect	Lable
Min-Dialect	Minnan	0
Kejia-Dialect	Kejia	1
Wu-Dialect	Shanghai,Hefei	2
Guanhua-Dialect	Shaanxi,Ningxia,Hebei,Sichuan	3
Gan-Dialect	Nanchang	4
Xiang-Dialect	Changsha	5

B. Dialect classification model based on sparse parameter sharing

This section introduces the dialect classification model based on sparse parameter sharing [22]. The sparse sharing automatically discovers the commonalities between subtasks, trims unimportant network parameters, and completes the sharing of some parameters between subtasks, which improves the flexibility of sharing mechanisms. At the same time, the amount of model parameters is reduced, and the efficiency of model execution is improved.

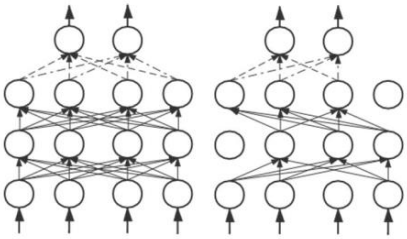


Fig. 4 Hard sharing model and Sparse sharing model

The framework of the sparse sharing model is shown in Fig.4(b). The training process consists of 3 steps. First, build a hyperparameter model called the base network. Then, the subtask starts training based on the base network, and during the training process, the subtask dynamically tailors the redundant parameters according to the demand. Finally, combine the sub-networks of each task to form a sparse shared network. Due to automatic training, sub-tasks networks with high task relevance often share part of the network weights, while sub-tasks networks with low task relevance are alone. Unlike hard sharing, sparse sharing achieves the purpose of parameter sharing because of the overlap of some networks between subtasks. Fig.5 shows the sparse sharing mechanism used in this paper. First, choose the LSTM-based model as the basic network for the sparse sharing mechanism, and define two subtasks: dialect language classification and dialect region

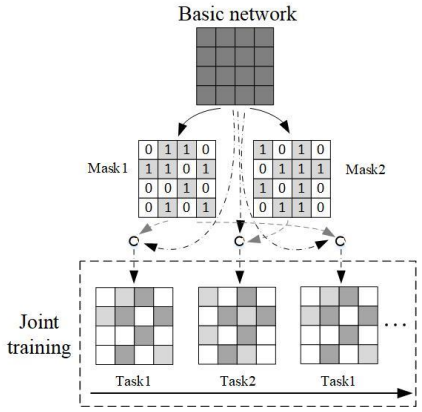


Fig.5 Dialect classification model based on sparse sharing

classification. Task1 is the task of dialect language classification, and Task2 is the task of dialect region classification. Furthermore, a binary mask matrix Mε{0,1} is used to select subnets for each task. As shown in Fig.5, Mask1 and Mask2 are sub-task networks, and all sub-task networks are merged to form a sparse shared structure. Then, two sub-task data are used for joint training, and the data of each task is extracted during joint training to train the corresponding sub-network. In the basic network, the overlapping parameters of the sub-tasks are trained multiple times. The overlapping parameters are shown in dark color in Fig.5 Square representation. The sparse parameter sharing mechanism not only pays attention to the difference between the two classification tasks, but also pays attention to the intersection and correlation between the characteristics of the dialect and the dialect region to which it belongs

1) Subtask network

The subtask network is obtained by iterative magnitude pruning (IMP) [23] on the base network, and the training process is shown in Algorithm 2. The network structure and parameters are determined by the binary mask matrix and the basic network. Suppose the parameter of the basic network is θ, the Mask matrix of subtask t is M_t, and the network parameter of subtask t is M_t·θ. α is the percentage of the remaining weight in each round of training, that is, the α parameters in the subnet are retained, the weights are arranged from small to large, and the 1-α parameter weights before cropping, if the parameters are retained, the corresponding M_t value It is 1; otherwise, it is 0.

2)Parallel training sub-task network

Train the sub-task network for each task, and the overlapping part of each sub-task network will share network parameters, so these parameters will be updated by the data of multiple tasks.

VI EXPERIMENTAL ANALYSIS

A.Experimental platform

Hardware platform: This paper uses a high-performance server for model training. The platform is configured with Intel(R) Core(TM) CPU, 64GB memory, and 2 GeForce RTX 2080 Ti GPUs.

Software platform: The server runs Ubuntu system, equipped with Pytorch (1.6.1), CUDNN (1.6.0), CUDA (10.0). The models in this paper are all implemented based on the Pytorch framework.

B. Data set

The experimental data set in this paper is the dialect data provided by the iFLYTEK AI Developer Competition [28]. The data set includes ten dialects, and each dialect contains speech data of 40 local people. The data is stored in a PCM format with a sampling rate of 16000 Hz and 16-bit quantization. The data set contains two parts: a training set and a test set. Training set Each dialect has 5000 voices, including 15 males and 15 females, with a total of 30 speakers, each with 200 voices; each dialect in the test set contains 5 speakers, including 3 females and 2 males The speakers in the training set and test set are not repeated, as shown in Table 1. At the same time, this paper has carried on the unified duration and the de-drying pretreatment to the used data set.

Uniform duration: Due to the inconsistent duration of data and audio, and there are a large number of blank periods in some audio, this paper uses audio cropping or filling methods to process all audio data into 2 seconds and extract three features to construct a single-task model.

C.Analysis of the results of the multi-task language recognition model

1) Hard shared dialect classification model with dialect region recognition as an auxiliary task

First, perform regional labeling for each piece of audio information in the data set. The format is [data, label1, label2], where label1 represents the data dialect language label, label2 represents the data dialect area label, the correspondence between label1 and the dialect language, and the correspondence between label2 and the area is shown in Table 1. Then, the original data of the three speech features are input into the model, and the experimental results are shown in Fig.6(a). Compared with the single-task model, the accuracy of the parameter hard-sharing model for joint training of the main task and the auxiliary task is improved by an average of 1%. Among them, the hard shared multi-task model based on MFCC features has the best performance, reaching 79.96%. The model recall rate, accuracy rate, and optimal value of F1_score based on the three characteristics are shown in Fig.6(b)

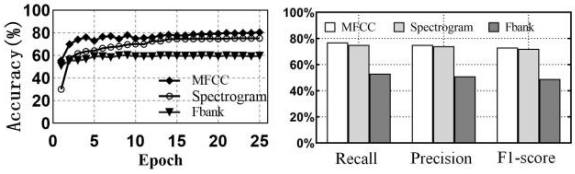


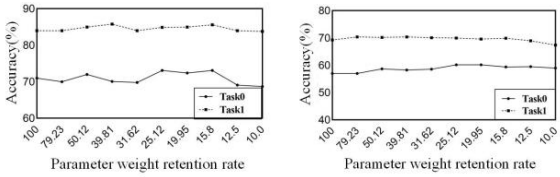
Fig. 6 Accuracy (a) , recall, precision, and F1-score(b) of multi-task models based on parameter hard-sharing

2) Dialect classification model based on sparse sharing

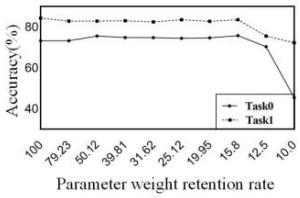
In this section,choose a model with the same architecture as the hard shared subtask model as the base network. The optimal sub-task models are selected for spectrogram,MFCC,Fbank, Task0 is the task of dialect language classification, and Task1 is the task of regional dialect classification. Fig.7 compares the accuracy of subtasks under different parameter weight retention rates. The experiment selects the subtask model with the highest accuracy. The selection of subtask model is shown in Table 3.

TABLE 3. SUBTASK MODEL SELECTION RESULT

Feature	Task	Parameter pruning rate
MFCC	Task0	25.12%
	Task1	39.81%
Spectrogram	Task0	15.8%
	Task1	15.8%
Fbank	Task0	25.12%
	Task1	31.62%



(a)MFCC (b)Fbank



(c) Spectrogram

Fig. 7 Performance of subtask network based on MFCC, Fbank and Spectrogram

Joint training is performed on tasks with different characteristics, and the accuracy is shown in Fig.8(a). The recall rate, precision rate, and F1_score performance are shown in Fig.8(b). The accuracy of the model based on MFCC features can reach 83.59%, which is 3% higher than the accuracy of the dialect recognition model based on parameter hard sharing. The model joint training based on the features of the spectrogram has an accuracy rate of 81.5% for denoising data. The joint training effect of the model based on the Fbank feature is not significant, the highest accuracy rate is 71.88%, but the accuracy rate of the corresponding parameter hard sharing model is improved.

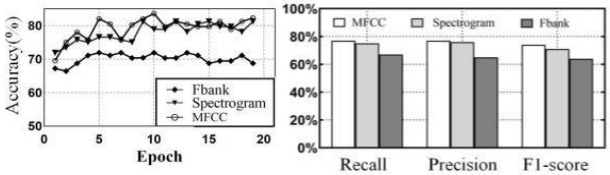


Fig.8 Accuracy (a) , recall, precision, and F1-score(b) of joint training based on three features

3) Experimental comparison

This paper compares the dialect classification model based on sparse sharing (here referred to as SSNet, Sparse-sharing network) with the ATLNet dialect classification model proposed in [19]. In order to ensure the unity of the data, this paper uses the same 2-second data set Train the ATLNet network. The experimental results are shown in Figure 9. As can be seen from the figure, because the ATLNet (accuracy rate is only 78.3%)

model can not fully explore the correlation between dialect tasks, so in the 2-second data set, the performance of the dialect classification model based on sparse sharing proposed in this paper is better.

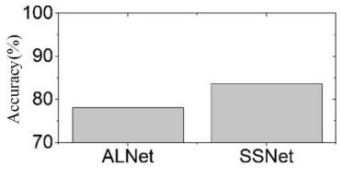


Fig.9 Comparison of SSNet and ALNet experimental results

VII CONCLUSION

China has a vast territory and a wide variety of dialects. The recognition of Chinese dialects has always been a difficult point in speech recognition. Improving the accuracy of dialect classification in my country is crucial for the next step of dialect recognition. This paper first extracts typical speech features MFCC, Fbank, and spectrogram for comparative study, and builds LSTM dialect classification models based on the three speech features, with an accuracy rate of up to 79.04%. Furthermore, this paper excavates the regional characteristics of ten Chinese dialects, and establishes a hard-sharing multi-task dialect classification model with dialect regions as auxiliary tasks. Due to the correlation between multiple sub-tasks, compared with the single-task model, joint training multi-task The model can dig into more data features, and thus obtain better performance, with an accuracy rate of 79.96%. Finally, in order to further increase the scale of the multi-task model and remove the influence of redundant network nodes on the results, this paper proposes a multi-task dialect classification model based on parameter sparse sharing. During the joint training of multiple sub-tasks, the base network is automatically tailored. And share the highly relevant network weights, and then reduce the complexity of the network while improving the recognition performance, the accuracy rate can reach 83.59%.

ACKNOWLEDGMENT

This work was Supported by the National Natural Science Foundation of China (No. 61902229) , Fundamental Research Funds for the Central Universities (No. GK202103084) .

REFERENCES

[1] Otter, Daniel W., Julian R. Medina,et al. A survey of the usages of deep learning for natural language processing [J]. IEEE Transactions on Neural Networks and Learning Systems (2020)

[2] LiMIng Wei, Research on dialect classification method based on convolutional neural network[D]. Southeast University.(in Chinese)

[3] Murty, K. Sri Rama, and Bayya Yegnanarayana. Combining evidence from residual phase and MFCC features for speaker recognition[J]. IEEE signal processing letters 13.1 (2005): 52-55.

[4] Yujin, Yuan, Zhao Peihua, and Zhou Qun. Research of speaker recognition based on combination of LPCC and MFCC [C]// 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems. Vol. 3. IEEE, 2010.

[5] Shensa, Mark J. The discrete wavelet transform: wedding the a trous and Mallat algorithms [J]. IEEE Transactions on signal processing 40.10 (1992): 2464-2482.

[6] learning for natural language processing [J]. IEEE Transactions on Neural Networks and Learning Systems (2020)

[7] Murty, K. Sri Rama, and Bayya Yegnanarayana. Combining evidence from residual phase and MFCC features for speaker recognition[J]. IEEE signal processing letters 13.1 (2005): 52-55.

[8] Yujin, Yuan, Zhao Peihua, and Zhou Qun. Research of speaker recognition based on combination of LPCC and MFCC [C]// 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems. Vol. 3. IEEE, 2010.

[9] Shensa, Mark J. The discrete wavelet transform: wedding the a trous and Mallat algorithms [J]. IEEE Transactions on signal processing 40.10 (1992): 2464-2482.

[10] Hermansky, Hynek. Perceptual linear predictive (PLP) analysis of speech [J]. the Journal of the Acoustical Society of America 87.4 (1990): 1738-1752.

[11] Dahl G E , Acero A.Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition[J].IEEE Transactions on Audio Speech & Language Processing , 2012 , 20(1):30-42.

[12] Jiang Bin, Research on deep learning methods of language recognition[D]. University of Science and Technology of China,2015(in Chinese)

[13] Koller, Oscar, et al. Deep sign: Hybrid CNN-HMM for continuous sign language recognition [C]//Proceedings of the British Machine Vision Conference 2016. 2016.

[14] Koller, Oscar, et al. "Deep sign: Hybrid CNN-HMM for continuous sign language recognition[C]// Proceedings of the British Machine Vision Conference 2016. 2016.

[15] Miao, Yajie, Mohammad Gowayyed, and Florian Metze. EESEN: End-to-end speech recognition using deep RNN models and WFST- based decoding[C]//2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015.

[16] Shewalkar, Apeksha, Deepika Nyavanandi, and Simone A. Ludwig. Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU [J]. Journal of Artificial Intelligence and Soft Computing Research 9.4 (2019): 235-245.

[17] Xiong, Wayne, et al. The Microsoft 2017 conversational speech recognition system [C]// 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018.

[18] Ali A R . Multi-Dialect Arabic Speech Recognition[C]// 2020 International Joint Conference on Neural Networks (IJCNN). 2020.

[19] Qin Chenguang, Wang Hai, Ren Jie, et al. Dialect language recognition based on multi-task learning[J]. Computer Research and Development(in Chinese)

[20] Zhang Yu, Liu Jianwei, Zuo Xin. Multi-task learning[J]. Chinese Journal of Computers, 2020(7):1340-1378. (in Chinese)

[21] Ruder, Sebastian. An overview of multi-task learning in deep neural networks[J]. arXiv preprint arXiv:1706.05098 (2017).

[22] Stevens K N. Toward a model of lexical access based on acoustic landmarks and distinctive features[J]. Journal of the Acoustical Society of America, 2002, 111(4):1372-1891.

[23] WAVE, https://docs.python.org/3/library/wave.html

[24] Kim, Suyoun, Takaaki Hori, and Shinji Watanabe. Joint CTC-attention based end-to-end speech recognition using multi-task learning [C]// 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017.

[25] Kendall, Alex, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[26] Sun, Tianxiang, et al. Learning sparse sharing architectures for multiple tasks [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 05. 2020.

[27] Elesedy, Bryn, Varun Kanade, and Yee Whye Teh. Lottery tickets in linear models: An analysis of iterative magnitude pruning [J]. arXiv preprint arXiv:2007.08243 (2020).

[28] Xunfei Open Platform, http://challenge.xfyun.cn/