End-to-End Chinese Dialects Identification in Short Utterances using CNN-BiGRU

Qiuxian Zhang[1], Yong Ma[1], Mingliang Gu[1*], Yun Jin[1, 2], Zhaodi Qi[1], Xinxin Ma[1], Qing Zhou[3]
[1]School of Physics and Electronic Engineering, Jiangsu Normal University
[2]Kewen College, Jiangsu Normal University
[3]School of Languistic Science and Arts, Jiangsu Normal University
Xuzhou, Jiangsu, China [1, 3]
Nanjing, Jiangsu, China [2]
e-mail: zhang_qiuxian@126.com, mlgu@jsnu.edu.cn

***Abstract*—The performance of dialect identificatification in short utterances is obviously degraded. To address the issue, this paper proposes an end-to-end approach to reduce frequency variations and extract the global context feature information vector using CNN-BiGRU (Convolutional Neural Networks, Bidirectional Gated Recurrent Unit), which is beneficial to enrich the feature expression of short utterances. With the method, it makes higher-order abstraction of features and increases recognition accuracy. In addition, we perform frame-level feature splicing to obtain high dimensional eigenvectors. Experiments on 10 Chinese dialects showed the proposed method achieved 9.93% relative improvement in Acc than the mainstream i-vector system.**

Keywords—Chinese dialects identification; end-to-end; CNN-BiGRU; short utterances

I. INTRODUCTION

Dialect identification is an important research topic in speech recognition. Compared with speaker recognition and language identification (LID) [1], dialect identification has not been studied in depth. One of the main reasons is the lack of a common data set, and another reason is that dialect identification is generally considered to be a special case of LID. In this case, dialect identification becomes a challenging task. In particular, improving the performance of the dialect identification in short utterances is very important.

For LID and dialect identification, i-vector is regarded as the state-of-the-art for general tasks[2]. Based on the joint factor analysis technique (JFA), it was proposed that speaker and session differences can be characterized by a single subspace. With this subspace, digital vectors obtained from one speech data can be further converted into low-dimensional vectors (i-vectors). I-vector greatly facilitates the modeling and testing process of many systems for both speaker recognition and dialect identification.

Recently, deep neural networks (DNNs) replace the traditional speech recognition based on GMM-HMM model to model the observation probability of speech. In order to extract language discriminant features and representations, more and more end-to-end [3] NNs have been proposed to cross the framework level to the utterance level LID identity-avoiding the need for discriminative back-end algorithm [4]. For example, Gonzalez et.al proposed buiding Long Short Term Memory-Recurrent Neural Networks (LSTM-RNN) to identify languages [5]. It is suitable for modeling timing signals. Fernando et al. showed that bidirectional long short term memory network (BiLSTM) performs well for short durations (3 seconds) LID tasks by modelling temporal dependencies between past and future frame based features in short utterances [6]. Lozano-diez et al. proposed convolutional deep neural networks (CDNN) for short test durations (segments up to 3 seconds of speech) [7]. The experimental results showed that CDNN perform well in short utterances. The system is trained from the beginning to distinguish a given set of languages, so the previous speech recognition phase is not needed.

In this paper, our work forces on short utterances (less than or equal to 3s, average 2.5s) using the proposed CNN-BiGRU model to construct an end-to-end Chinese dialects identification system. GRU is a variant of LSTM. We use a fixed-length spelling frame as input, the CNN layer further captures the spatial information and passes the depth features to the BiGRU layer. BiGRU extracts the global context feature information vectors of short utterances. Then, softmax is performed after connecting the fully connected layer to discriminate target dialects.

The rest of this paper is organized as follows. In Section Ⅱ presents the proposed the model for short utterences of Chinese dialects identification in detail. Experimental results and analysis are presented in Section Ⅲ, and our whole work is summarized in Section Ⅳ.

II. METHODS CONSTRUCTION

In this section, we describe the network structure related with the proposed model and use it for Chinese dialects identification.

A. Convolution Network Architecture

With the application of end-to-end technology in deep neural networks, we can perform end-to-end modeling directly from the filter bank (FBank) features to the dialect ID based on

CNN. The hidden layer of CNN usually consists of two parts: the convolutional layer and the pooling layer [8]. The convolutional layer is composed of several convolutional units, and the parameters of each convolutional unit are optimized by the backpropagation algorithm to extract different features of the input. According to the size of the given filter , each cell of the layer is connected to the local cell subset of the hidden layer below. Then, applying Rectified Linear Unit (ReLU) to convolve the input and a filter (weight, $\boldsymbol{W}$) and add the bias term ($\boldsymbol{b}$), which can be defined as:

$$\boldsymbol{h} = \max\left(0, \boldsymbol{W}^{T} * \boldsymbol{x} + \boldsymbol{b}\right) \tag{1}$$

where $\boldsymbol{h}$ is the nonlinear output of the input vector $\boldsymbol{x}$ after linear transformation.

The pooling layer is a feature with a larger dimension obtained after the convolution layer, by taking the maximum value or the average value to obtain a new feature with a small dimension.

B. Bidirectional GRU

For a given input sequence $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t)$, the standard recurrent neural network calculates the state vector sequence of the hidden layer $\boldsymbol{h} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_t)$ and the output vector $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_t)$ by iterating from 1 to T：

$$\boldsymbol{h}_t = \sigma_h\left(\boldsymbol{W}_{xh}\boldsymbol{x}_t + \boldsymbol{W}_{hh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_h\right) \tag{2}$$

$$\boldsymbol{y}_t = \sigma_y\left(\boldsymbol{W}_{hy}\boldsymbol{h}_t + \boldsymbol{b}_y\right) \tag{3}$$

where $\boldsymbol{W}$ represents the weight matrix between the layers; $\boldsymbol{b}_h$ and $\boldsymbol{b}_y$ are the offset vectors of the hidden layer and the output layer respectively; σ_h and σ_y are activation functions.

GRU is a variant of LSTM that has a simpler structure and better convergencee. GRU consists of an update gate and a reset gate [9]. The structure diagram of the GRU is shown in Fig. 1.

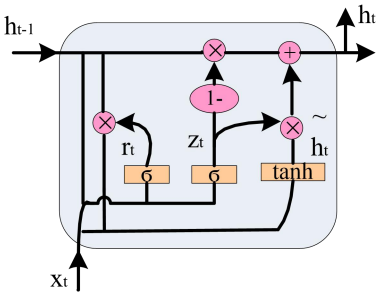

Fig. 1. The architecture of GRU cell

The update process of GRU is as follows:

$$\boldsymbol{r}_t = \sigma\left(\boldsymbol{W}_r\boldsymbol{x}_t + \boldsymbol{U}_r\boldsymbol{h}_{t-1} + \boldsymbol{b}_r\right) \tag{4}$$

$$\boldsymbol{z}_t = \sigma\left(\boldsymbol{W}_z\boldsymbol{x}_t + \boldsymbol{U}_z\boldsymbol{h}_{t-1} + \boldsymbol{b}_z\right) \tag{5}$$

$$\tilde{\boldsymbol{h}}_t = \tanh\left(\boldsymbol{W}_h\boldsymbol{x}_t + \boldsymbol{U}_h\left(\boldsymbol{r}_t \Theta \boldsymbol{h}_{t-1}\right) + \boldsymbol{b}_h\right) \tag{6}$$

$$\boldsymbol{h}_t = \left(1 - \boldsymbol{z}_t\right)\Theta\tilde{\boldsymbol{h}}_t + \boldsymbol{z}_t\Theta\boldsymbol{h}_{t-1} \tag{7}$$

where σ is the sigmoid function and tanh is the hyperbolic tangent function. $\boldsymbol{U}$ is the weight matrixes for the previous hidden state vector $\boldsymbol{h}_{t-1}$, $\tilde{\boldsymbol{h}}_t$ is a candidate activation and Θ is an element-wise multiplication. Vector $\boldsymbol{r}_t$, $\boldsymbol{z}_t$ denote the reset gate and the renew gate vector.

However, the speech itself has a certain context correlation. The language model in the traditional speech recognition system has insufficient memory ability for historical information, and cannot fully learn the relevance of the speech sequence. Bidirectional GRU neural networks (BiGRU) is proposed in this paper. The architecture is shown in Fig. 2.

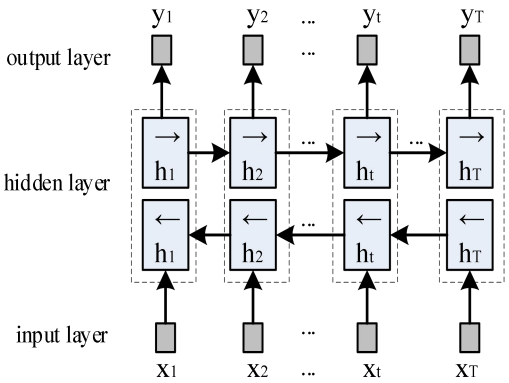

Fig. 2. The architecture of BiGRU for short utterances

BiGRU includes the forward GRU and the reverse GRU. The forward GRU is used to capture the above information of the speech feature vector, and GRU generates a forward implicit state sequence $\left\{\vec{\boldsymbol{h}}_1, \vec{\boldsymbol{h}}_2, \cdots, \vec{\boldsymbol{h}}_t, \cdots, \vec{\boldsymbol{h}}_T\right\}$ from left to right. While the reverse GRU captures the speech feature vector and the following information, and a reverse implicit state sequence $\left\{\overleftarrow{\boldsymbol{h}}_1, \overleftarrow{\boldsymbol{h}}_2, \cdots, \overleftarrow{\boldsymbol{h}}_t, \cdots, \overleftarrow{\boldsymbol{h}}_T\right\}$ is generatated from right to left . Finally, the global context information is obtained by combining the captured speech context feature information vectors. That is, the forward implicit state and the reverse implicit state are combined to form a hidden state of the speech segment. Through this model, the deep features of Chinese dialects will be obtained:

$$\boldsymbol{h}_t = \boldsymbol{g}\left(\vec{\boldsymbol{h}}_t, \overleftarrow{\boldsymbol{h}}_t\right) \tag{8}$$

C. Proposed system

In our work, we propose the model that combines CNN and BiGRU to identify short utterances based on end-to-end. An overall architecture is shown in Fig. 3.

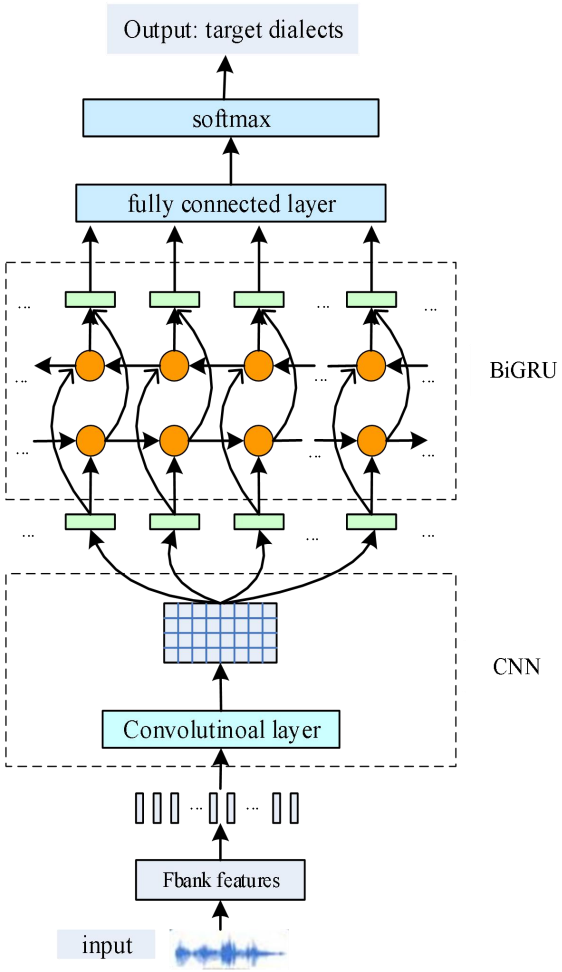

Fig. 3. Architecture of the proposed CNN-BiGRU for end-to-end Chinese dialects identification in short utterences

Firstly, we extract 40-dimensional FBank features from short utterances and perform the whole sentence mean processing on the feature. The regularized features are spliced with adjacent frames (5 frames left and right) to obtain high-dimensional feature vectors, and the spliced supervectors are reduced by principal component analysis (PCA) as input to the CNN-BiGRU model. Then, the CNN layer extracts the information implicit in the time dimension of short utterances and passes the depth features to the BiGRU layer. BiGRU models the long-term correlation of speech by modeling the time dependence between past and future frame-based features in short utterances. The extracted global context feature information vectors are fed to a fully connected DNN layer as the output vector of the last BiGRU layer, which transform the features into a space that makes that output easier to classify for short utterances. The output new feature vector is optimized by the fully connected layer, and the frame-level output values are average. Finally, the output vectors are mapped to the (0,1) interval by the softmax function, which is regarded as the probability of the output of various dialects, and is used to predict the category label of Chinese dialects. Among them, the model is optimized by cross-entropy (CE) criterion.

III. EXPERIMENT

A. Dataset Description

The database is provided by IFLYTEK for experiments. The data is stored in a PCM format with a sampling rate of 16000 Hz and 16 bits of quantization. We selecte 10 Chinese dialects: changsha, hebei, hefei, kejia, minnan, nanchang, ningxia, shan3xi, shanghai and sichuan.

The data is collected by various models of smartphones, the recording environment includes a quiet environment and a noisy environment. There were 75.42 hours of training data, and 12.38 hours of testing data, totally. The number of utterances for the training data was 60000, and for each dialect is 6000. We divide the testing data into long utterances, short utterances and mixed duration utterances, where long utterances are greater than 3s (average 6.3s) and short utterances are less than or equal to 3s (average 2.5s). For the different duration of testing data, it is 250 utterances for each dialect. Details of the number of utterances and the data size are shown in Table 1.

TABLE I. EXPERIMENTAL DATA SETS

Dataset	Training data	Testing data		
		long utterances	short utterances	mixed utterances
dialects	Utt/Size	Utt/Size	Utt/Size	Utt/Size
changsha	6000/8.37h	250/0.44h	250/0.23h	250/0.67h
hebei	6000/7.04h	250/0.43h	250/0.24h	250/0.67h
hefei	6000/8.40h	250/0.50h	250/0.25h	250/50.75h
kejia	6000/6.91h	250/0.35h	250/0.22h	250/0.57h
minnan	6000/7.27h	250/0.37h	250/0.19h	250/0.56h
nanchang	6000/8.38h	250/0.46h	250/0.26h	250/0.72h
ningxia	6000/6.92h	250/0.38h	250/0.17h	250/0.55h
shan3xi	6000/7.74h	250/0.57h	250/0.34h	250/0.91h
shanghai	6000/7.19h	250/0.57h	250/0.21h	250/0.78h
sichuang	6000/7.21h	250/0.42h	250/0.21h	250/0.63h
ALL	60000/75.42h	2500/3.89h	2500/2.30h	2500/6.19h

B. Baseline systems and experiments setup

The i-vector system was evaluated on the same testing data using Kaldi[10]. Firstly, an Universal Background Model (UBM) composed of 512 Gaussian components is trained from 40-dimensional FBank features. Then, Baum-Welch statistics calculates on this UBM, and a TV space of 400-dimension is derived from them by using LDA . Table 2 shows the results of experiments on data with long utterances, short utterances and mixed utterances.

In order to evaluate the effectiveness of the proposed method, we use LSTM for comparison. There are 2 LSTM in

series and 256 units in the hidden layer of each LSTM. The mini-batch size is set to 64, and stochastic gradient descent (SGD) with learning rate 0.1 is used in this experiment. Framing is performed with a frame length of 25 ms and a frame shift of 10 ms, and 40-dimensional FBank features are extracted as input, and the nonlinear deep features are extracted through the model. Then performing softmax after connecting the fully connected layer. Meanwhile, the average the frame-level output values are used to predict target dialects.

TABLE II. COMPARISION OF EXPERIMENTAL RESULTS ON SPEECH DATA WITH DIFFERENT DURATIONS OF ACC (%)

Method	i-vector	LSTM + DNN	GRU + DNN
long utterances	73.16	75.80	76.32
short utterances	70.20	71.95	72.17
mixed utterances	73.80	77.32	77.90

We also use GRU to instead of LSTM, which can reduce computation time and achieve better convergence. Table 2 shows that the performance of the GRU is the best in the three duration testing data. However, the recognition rate of short utterances are the lowest relative to long utterances and mixed utterances. For this, we propose an effective method for short utterances.

C. *The proposed system*

Considering the contextual relevance of speech, we use BiLSTM and BiGRU to apply the system with the same experimental configuration.

Because the convolutional layer of CNN can extract information implicit in the time dimension, it is used to capture the spatial information of the speech [11]. The features are modeled at the frame level by the connection of each hidden unit of the layer, and the pooling layer is used to convert the feature representation at the frame level to a fixed vector representation on the speech segment. Experiments prove that the performance has improved by applying this structure to the front end of BiGRU. Table 3 compares the results of six systems on short utterances.

TABLE III. COMPARION OF SIX SYSTEMS ON SHORT UTTERANCES FOR CHINESE DIALECTS IDENTIFICATION

Method	Acc(%)
i-vector	70.20
LSTM+DNN	71.95
GRU+DNN	72.17
BiLSTM+DNN	71.63
BiGRU+DNN	**75.46**
CNN-BiGRU+DNN	**79.25**

From the results, we observe that BiGRU and CNN-BiGRU methods achieve 5.26% and 9.05% relative improvements than the baseline system (i-vector) on short utterances.

In order to obtain more frame-level features to express speech information, we splice adjacent frames (left and right 5 frames) to obtain high-dimensional eigenvectors, and reduced the splicing supervector by principal component analysis (PCA) as input. The experimental results are shown in Table 4.

TABLE IV. COMPARION WITH EXPERIMENTAL RESULTS OF BASELINE SYSTEMS

Method	Acc(%)
i-vecrot	70.20
LSTM	71.95
CNN-BiGRU+DNN	**80.13**

The method significantly improves the performance. The experiment obtains 9.93% relative improvement with Acc 80.13%.

IV. CONCLUSIONS

The paper proposes an end-to-end approach of short utterances based on CNN-BiGRU that makes up for the lack of information on short utterances. The method achieves 9.93% relative improvement than i-vector, and obtains 8.18% relative improvement over LSTM. Experimental results show that the proposed method significantly improve the performance. For future work, we will study feature representation of short utterances based on knowledge distillation for Chinese dialects identification.

ACKNOWLEDGMENT

Thanks to the Chinese dialects data set provided by IFLYTEK. This work is supported by the National Natural Science Foundation of China (Grant No. 61708061 and 61673108), Xuzhou Science and technology project (KC18015), Industry-University-Research collaboration project of Jiangsu Province (BY2018077), Research Fund for the Doctoral Program of New Teachers of Jiangsu Normal University (Grant No. 17XLR029), Jiangsu University Natural Science Research Project (Grant No. 17KJB510016, 17KJB510018 and 18KJB510013), Jiangsu Normal University School Funding Project (2018YXJ591).

REFERENCES

[1] Peng Shen, Xugang Lu, Sheng Li, and Hisashi Kawai, "Feature Representation of Short Utterances based on Knowledge Distillation for Spoken Language Identification," in Proc. of Interspeech, pp. 1813–1817, 2018.

[2] Suwon Shon, Ahmed Ali, and James Glass, "Convolutional Neural Networks and Language Embeddings for End-to-End Dialect Recognition," arXiv:1803.04567v2, 2018.

[3] Rozental Alon and Fleischer Daniel, "Amobee at SemEval-2018 Task 1: GRU neural network with a CNN attention mechanism for sentiment classification," in North American Chapter of the Association for Computational Linguistics, pp. 218-225, 2018.

[4] Ma Jin, Yan Song, Ian McLoughlin, Wu Guo, and Li-Rong Dai, "End-to-End Language Identification Using High-Order Utterance Representation with Bilinear Pooling," in Proc. of Interspeech, pp. 2571–2575, 2017.

[5] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, and P. J. Moreno, “Automatic language identification using long short-term memory recurrent neural networks,” in Proc. of Interspeech, pp. 2155–2159, 2014.

[6] S. Fernando, V. Sethu, E. Ambikairajah, and J. Epps, “Bidirectional Modelling for Short Duration Language Identification,” in Proc. Of Interspeech, pp. 2809–2813, 2017.

[7] A. Lozano-Diez, R. Zazo Candil, J. G. Dominguez, D. T.Toledano, and J. G. Rodriguez, “An End-to-End Approach to Language Identification in Short Utterances using Convolutional Neural Networks,” in Proc. of Interspeech, pp. 403–407, 2015.

[8] Trung Ngo Trong, Ville Hautamaki, and Kong Aik Lee, “Deep Language : a comprehensive deep learning approach to end-to-end language recognition,” in Proceedings of Odyssey - The Speaker and Language Recognition Workshop, pp. 109–116, 2016.

[9] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” arXiv:1412.3555v1, 2014.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. StemmerN and K. Vesely, “The kaldi speech recognition toolkit,” in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB, 2011.

[11] Maryam Najafian, Sameer Khurana, Suwon Shon, Ahmed Ali, and James Glass, “Exploiting convolutional neural networks for phonotactic based dialect identification,” in ICASSP, pp. 5174-5178, 2018.