Sequence-To-One Neural Networks for Japanese Dialect Speech Classification

1st Ryo Imaizumi
Tokyo Metropolitan University
Tokyo, Japan
imaizumi-ryo@ed.tmu.ac.jp

2nd Ryo Masumura
Nippon Telegraph and Telephone Corporation
Tokyo, Japan
ryou.masumura.ba@hco.ntt.co.jp

3rd Sayaka Shiota
Tokyo Metropolitan University
Tokyo, Japan
sayaka@tmu.ac.jp

4th Hitoshi Kiya
Tokyo Metropolitan University
Tokyo, Japan
kiya@tmu.ac.jp

***Abstract*—Automatic speech recognition (ASR) is usually constructed for recognizing standard language. Thus, when input speech includes dialect which is a variety of a language, performance of ASR is seriously degraded. To relax this problem, an approach is to use dialect-specific ASR recognizers by introducing a dialect speech classification module. In this situation, the performance of dialect-specific ASR depends on that of dialect speech classification. We propose a Japanese dialect speech classification method using sequence-to-one neural networks that are one of the successful methods in speech classification research fields. The experimental results showed that a classification system provided high classification accuracy.**

***Index Terms*—Japanese dialect speech classification, sequence-to-one neural network, LSTM, BLSTM**

I. Introduction

Dialect speech classification can be regarded as one of preprocessing modules for dialect-specific ASR. Recently, Arabic dialect speech classification trained from neural networks has been reported [1]. As this research has mentioned, the dialect speech classification becomes one of the important research topics for robust ASR [2], [3]. However, there are few Japanese dialect speech classification reports. Thus, this paper proposes Japanese dialect speech classification using sequence-to-one neural networks [4], [5]. To realize the framework, a bidirectional long short-term memory (BLSTM) [6] network with attention mechanism is adopted. In our experiments, we used Japanese dialect speech database including six regions; Aomori, Hiroshima, Kumamoto, Nagoya, Sapporo, and Sendai. From the dialect speech classification experiments, the BLSTM network with a preferable parameter combination achieved the highest accuracy.

II. Sequence-to-one neural network

One type of neural networks to handle variable length sequences is Recurrent Neural Network (RNN). RNN had been reported in order to treat continuous data such as sentences and speech [7]. Long short-term memory (LSTM) is an architecture that is an extension of RNN. An aim of LSTM is to represent long time series data which is difficult to model by RNN [8], [9]. Moreover, BLSTM had been proposed as one of the applications of LSTM. Figure 1 shows BLSTM network structure. Figure 2 shows the internal structure of the LSTM used in the BLSTM network. The BLSTM has a pair layer of a LSTM layer and a reverse LSTM layer as hidden layer for predicting past output from future input, and it is possible to utilize both past and future information. The BLSTM also contains an attention mechanism and importance weights of hidden expressions. Since BLSTM is one of the most successful methods in speech classification research fields, this paper adopts BLSTM as a sequence-to-one neural network to realize Japanese dialect speech classification.

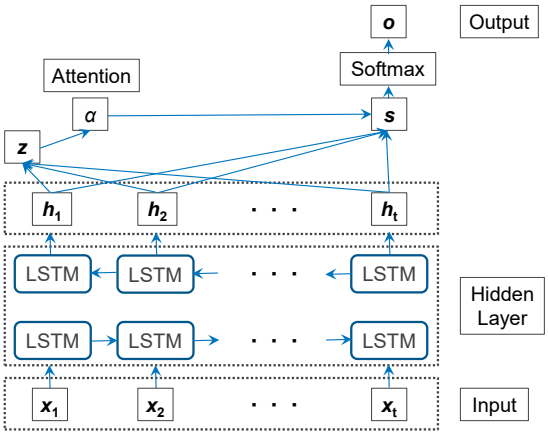

Fig. 1. BLSTM network structure

III. Japanese dialect speech database

A home-made speech database of Japanese dialects was used in all experiments. The dialect database consisted of six dialects; Aomori, Hiroshima, Kumamoto, Nagoya, Sapporo, and Sendai [10]. The number of utterances for each dialect is shown in Table I. All utterances are recorded by an iPhone 5 or a Xperia Z1, and the spoken content is about daily conversations. All utterances are sampled at 16 kHz with

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

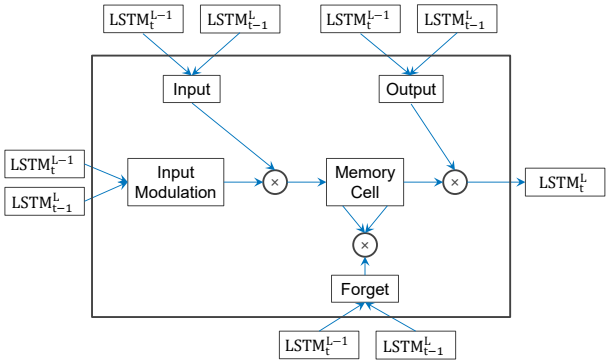

Fig. 2. The internal structure of the LSTM used in the BLSTM network

TABLE I
NUMBER OF UTTERANCES FOR EACH DIALECT IN THE JAPANESE DIALECT SPEECH DATABASE

	Training	Validation	Test
Aomori	10,203	538	1,352
Hiroshima	17,736	934	1,133
Kumamoto	8,861	467	1,438
Nagoya	17,680	931	1,102
Sapporo	15,157	798	1,356
Sendai	15,686	826	1,070
All	85,323	4,494	7,451

TABLE II
DIALECT CLASSIFICATION ACCURACY (%) (COMBINED ALL DIALECTS)

Number of LSTM units	Number of input dimensions	Number of hidden layers: 2	Number of hidden layers: 4
256	80	75.9	80.3
	240	76.0	78.5
	400	68.1	80.4
512	80	74.4	**81.3**
	240	74.7	77.5
	400	68.6	69.8

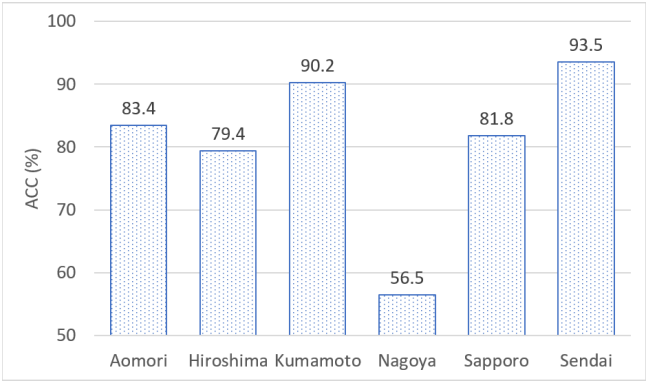

Fig. 3. ACCs for each dialect under the best parameter combination in Table II (512 LSTM units, 80 input dimensions, 4 layers)

16 bits. The number of speakers is about 80 for each dialect. The gender ratio is approximately equal. The distribution of ages is from 20 to 70 years old uniformly. The length of each dialect utterance was about 7 seconds, and the content of the dialect database was daily conversations. Each utterance has corresponding hand-labeled text and dialect.

IV. EXPERIMENT

A. Condition

In our Japanese dialect speech classification experiments, 80-dimensional log-mel filter bank was extracted from a frame as an input vector for the networks. The frame length and the frame shift were 25 ms and 10 ms, respectively. For the network structures, the BLSTM networks with attention mechanism were used. Some parameters of the BLSTM networks were set as follows; batch size was 16, dropout ratio was 0.2, and optimization method was adam. Early-stopping was carried out when the loss did not improve five times. The number of input dimensions was set to 80, 240, and 400 by connecting the preceding and following frames. The BLSTM networks had two or four hidden LSTM layers and each LSTM layer consisted of 256 or 512 units. For the evaluation, dialect speech classification accuracy (ACC (%)) was defined as follows;

$$\mathrm{ACC}(\%) = \frac{\#\text{ of Correct utterances}}{\#\text{ of Test utterances}} \times 100. \qquad (1)$$

B. Result

Table II shows the ACCs of each BLSTM network with the parameter combinations. It can be seen that all ACCs of using four hidden layers were higher than those of using two hidden layers. The results also had a tendency to increase the ACCs when the input dimension was smaller. Thus, the deeper hidden layers leaded to improve the representation ability of the network. And, the smaller dimensions of the input layer meant the network extracted the information with high time resolution. Consequently, the ACC of the BLSTM network with the preferable parameter combination achieved 81.3%.

Figure 3 shows the ACCs for each dialect under the network using the preferable parameter combination in Table II. Comparing Figure 3 with Table I, the ACCs tended to be high when the amount of the training data was small, and the ACCs become worse when the amount of the training data was large. To investigate the trend of the misrecgnitions, a confusion matrix is depicted in Figure 4. The BLSTM conditions were the same as Figure 3. In the misrecognition case of Aomori, Sapporo, and Sendai, Kumamoto dialect was often selected. The misrecognized utterances of Nagoya dialect were likely to classify into Hiroshima dialect, and the misrecognized utterances of Hiroshima dialects were usually classified into Aomori. These trends indicated that there was little relationship between the places of the regions and the misrecognition pattern. Comparing the ACCs with the amount of the training data, it can be seen that the networks of Nagoya and Hiroshima dialects were suffered from the overfitting problem, and that of Kumamoto was able to keep its robustness.

V. CONCLUSION

This paper proposed Japanese dialect speech classification using sequence-to-one neural network. The BLSTM network

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

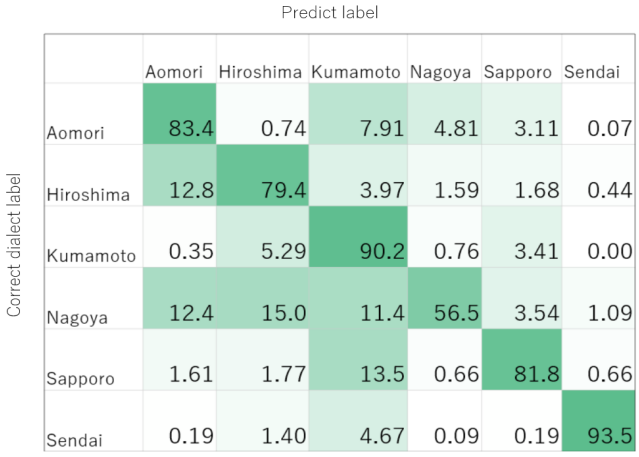

Fig. 4. Confusion matrix of the dialect classification result

was adopted to the classification network. In our experiments, six Japanese dialects were used to construct the classification networks. From the results, the preferable parameter combination contributed to obtain the higher ACC.

As future work, we will investigate relationship between the performance accuracy and the data amount. Additionally, many other network architectures using the other dialects or databases will be performed.

REFERENCES

[1] Xiaoxiao Miao and Ian McLoughlin. Lstm-tdnn with convolutional front-end for dialect identification in the 2019 multi-genre broadcast challenge. *arXiv preprint arXiv:1912.09003*, 2019.

[2] Thomas Purnell, William Idsardi, and John Baugh. Perceptual and phonetic experiments on american english dialect identification. *Journal of language and social psychology*, 18(1):10–30, 1999.

[3] Omar F Zaidan and Chris Callison-Burch. Arabic dialect identification. *Computational Linguistics*, vol.40,(no.1):pages 171–202, 2014.

[4] Wang Geng, Wenfu Wang, Yuanyuan Zhao, Xinyuan Cai, Bo Xu, Cai Xinyuan, et al. End-to-end language identification using attention-based recurrent neural networks. In *proc. INTERSPEECH*, pages 2944–2948, 2016.

[5] Weicheng Cai, Danwei Cai, Shen Huang, and Ming Li. Utterance-level end-to-end language identification using attention-based cnn-blstm. In *proc. ICASSP*, pages 5991–5995. IEEE, 2019.

[6] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *proc. NAACL HLT*, pages 1480–1489, 2016.

[7] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *proc. INTERSPEECH*, pages 2877–2880, 2011.

[8] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *proc. INTERSPEECH*, pages 338–342, 2014.

[9] Suman Ravuri and Andreas Stolcke. A comparative study of recurrent neural network models for lexical domain classification. In *proc. ICASSP*, pages 6075–6079. IEEE, 2016.

[10] Ryo Imaiuzmi, Ryo Masumura, Sayaka Shiota, and Hitoshi Kiya. Japanese dialect speech classification using sequence-to-one neural networks. in japanese. In *proc. SP2019-57*, volume 119, pages 41–46, 2020.

 978-1-7281-9802-6/20/$31.00 ©2020 IEEE