 JIOT, vol.1, no.1, pp.17-23, 2019

Tibetan Multi-Dialect Speech Recognition Using Latent Regression Bayesian Network and End-To-End Mode

Yue Zhao[1], Jianjian Yue[1], Wei Song[1,*], Xiaona Xu[1], Xiali Li[1], Licheng Wu[1] and Qiang Ji[2]

Abstract: We proposed a method using latent regression Bayesian network (LRBN) to extract the shared speech feature for the input of end-to-end speech recognition model. The structure of LRBN is compact and its parameter learning is fast. Compared with Convolutional Neural Network, it has a simpler and understood structure and less parameters to learn. Experimental results show that the advantage of hybrid LRBN/Bidirectional Long Short-Term Memory-Connectionist Temporal Classification architecture for Tibetan multi-dialect speech recognition, and demonstrate the LRBN is helpful to differentiate among multiple language speech sets.

Keywords: Multi-dialect speech recognition, Tibetan language, latent regression bayesian network, end-to-end model.

1 Introduction

Tibetan language is one of the most widely used minority languages in China and the cross-boundary languages. The application of automatic speech recognition technology to Tibetan language has been paid more and more attention by researchers. Tibetan speech recognition has shown wide demand and immeasurable application prospects.

During the long term development of Tibetan language, different dialects have been formed. It is divided into three major dialects of Ü-Tsang, Kham and Amdo. Three dialects are divided into several local sub-dialects. The pronunciation of Tibetan dialect is very different in different regions, but the written characters based on Tibetan language is unified. Since Lhasa-Ü-Tsang dialect is a Tibetan standard speech, there are much more research works than other dialects on its linguistics, speech recognition and corpus establishment [Zhang (2016); Yuan, Guo and Dai (2015); Li and Meng (2012); Cai and Zhao (2008); Cai (2009)].

End-to-end automatic speech recognition has more advantages for low-resource languages than conventional DNN/HMM systems because it avoids the need for linguistic resources such as dictionaries and phonetic knowledge [Wang, Guo and Xie (2017)]. Some recent works for multilingual speech recognition explored end-to-end

[1] School of Information and Engineering, Minzu University of China, Beijing, 100081, China.

[2] Rensselaer Polytechnic Institute, 110 Eighth Street, Troy NY 12180-3590, USA.

[*] Corresponding Author: Wei Song. Email: songwei@muc.edu.cn.

model to build a single language independent system. The work in Li et al. [Li, Tara and Khe (2018)] adopted the attention-based sequence-to-sequence model for 7 English dialects and it has shown good performance compared to other sequence-to-sequence language-specific models. Similar work in Toshniwal et al. [Toshniwal, Tara, Ron et al. (2018)] with multi-task end-to-end learning for 9 Indian languages obtained the largest improvement by conditioning the encoder on the speech language identity. These works showed that end-to-end model can contribute to handling the variations between different languages by learning and optimizing a single neural network. The work in Watanabe et al. [Watanabe, Hori and Hershey (2017)] was based on hybrid attention/connectionist temporal classification (CTC) architecture where the model uses a deep convolutional neural networks (CNNs) followed by bidirectional long short-term memory (BLSTM) in encoder networks, and showed that it achieved the state-of-the-art performance in several ASR benchmarks including English, Japanese, Chinese mandarin, German etc.

In our works, we utilize BLSTM-CTC model trained on 3 Tibetan dialects speech data. In addition, to gain a higher accuracy, inspired by Watanabe et al. [Watanabe, Hori and Hershey (2017)] we introduce a latent regression Bayesian network (LRBN) before BLSTM-CTC to produce a generative model for multilingual speech data representation as end-to-end input. Since generative model can capture the underlying data distribution as well as the mechanisms used to generate data, we believe that such characteristics are crucial for shared representation across speech data from different languages. However, CNNs are typically discriminative models and they are built mainly for discriminative tasks, such as classification [Fang, Zhang, Sheng et al. (2018); Meng, Rice, Wang et al. (2018)]. Although the restricted Boltzmann machine (RBM) is one of the most successful generative models, it typically assumes latent variables are independently given data. Such an assumption weakens their data modeling and representation power. In contrast, the latent variables are dependent on each other given the observations through the so-called "explain-away" principle in LRBN. Through their inter dependency, latent variables coordinate with each other to better explain the patterns in the visible layer. Experimental results show that the advantage of hybrid LRBN/BLSTM-CTC architecture for Tibetan multi-dialect speech recognition, and demonstrate the LRBN is helpful to differentiate among multiple language speech set.

2 Hybrid LRBN/BLSTM-CTC architecture

This section introduces the hybrid architecture of latent regression Bayesian network proposed by Nie et al. [Nie, Zheng and Ji (2018)] and BLSTM-CTC model for multiple dialect speech recognition. Fig. 1 shows the hybrid architecture.

2.1 Latent regression bayesian network

The LRBN is a directed generative model consisting of one latent layer and one visible layer as shown in Fig. 2. Due to the "explaining away" effect in Bayesian networks, LRBN is able to capture both the dependencies among the latent variables given the observation and the dependencies among visible variables.

The LRBN is a special kind of Bayesian Network, and every latent variable connects to every visible variable with a directed edge as shown in Fig. 2. According to the chain rule

in Bayesian Networks, the joint probability of all variables is factorized into the product of prior probabilities and conditional probabilities as shown in Eq. (1),

$$P(x,h)=\prod_{j=1}^{n_h}P(h_j)\prod_{i=1}^{n_d}(x_i\mid h) \tag{1}$$

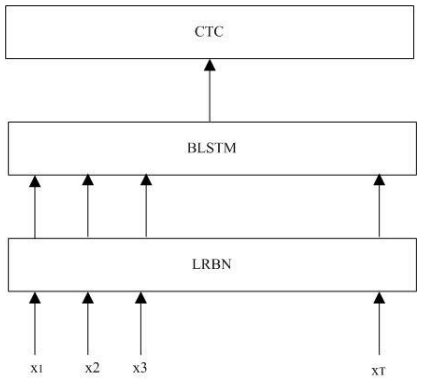

Figure 1: Hybrid LRBN/BLSTM-CTC multi-dialect speech recognition: the LRBN extracts the shared feature from multiple dialect speech data; BLSTM encodes the shared feature trained by CTC loss; CTC predicts an output label sequence

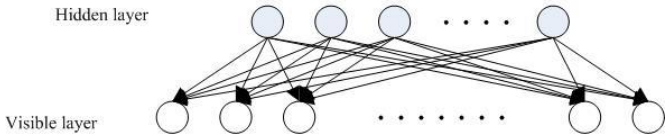

Figure 2: Latent regression Bayesian network

The prior probability for a latent variable is defined as Eq. (2),

$$P(h_j=1)=sigm(d_j) \tag{2}$$

where $sigm(d_j)=1/(1+\exp(-d_j))$ is the sigmoid function, and d_j is the parameter. This formulation is essentially a Bernoulli distribution. The conditional probability of a visible variable given all the latent variables is defined as a linear Gaussian, as shown in Eq. (3),

$$P(x_i\mid h)\sim N(\omega_i^T h+b_i,\sigma_i) \tag{3}$$

where the mean is a linear combination of the values of the latent variables. ω_{ij} is the weight for node h_j and x_i; b_i is a constant term; and σ_i is the standard deviation. Thus, the LRBN can be viewed as a mixture of Gaussian with the number of components exponential in the number of latent variables.

Plugging in the prior distributions and conditional distributions, the joint distribution of visible variables and hidden variables has the following formulation,

$$P_{\Theta}(x,h)=\prod_j \frac{\exp(d_j h_j)}{1+\exp(d_j)} \prod_i \mathcal{N}(x_i:\omega_i^T h+b_i,\sigma_i)$$

$$=\frac{\exp(-\psi_{\Theta}(x,h))}{(2\pi)^{n_d/2}\prod_j(1+\exp(d_j))} \tag{4}$$

where $\Theta=\{W,\sigma,b,d\}$, and

$$\psi_{\Theta}(x,h)=\sum_i \frac{(x_i-b_i)^2}{2\sigma_i^2}-\sum_i \frac{x_i-b_i}{\sigma_i^2}\omega_i^T h+\sum_i \frac{1}{2\sigma_i^2}(\omega_i^T h)^2-d^T h, \tag{5}$$

Compared with the Gaussian-Bernoulli Restricted Boltzmann Machine (GRBM), the LRBN adopts directed links between visible nodes and hidden nodes instead of undirected links. This leads to an extra term $\sum_i \frac{1}{2\sigma_i^2}(\omega_i^T h)^2$ in Eq. (5) compared with the energy function of GRBM. This extra term explicitly captures the relationship among latent variables. The dependencies among the latent layer given the visible layer can help better explain the patterns in the input data. Furthermore, unlike the GRBM, the LRBN has no intractable partition function issue, because the joint distribution is obtained by multiplying all the prior probabilities and conditional probabilities.

2.2 BLSTM-CTC

In end-to-end model, we apply the BLSTM trained with the CTC objective function. This model has been proved to be comparable/superior to deep neural network (DNN)/hidden Markov model in large vocabulary speech recognition.

LSTM contains memory cells with self-connections to store the temporal states of the network. Also, multiplicative gates are added to control the flow of information. We use a single layer bidirectional LSTM with the same block of LSTM as Watanabe et al. [Watanabe, Hori and Hershey (2017)]. CTC objective is adopted to automatically learn the alignments between speech frames X and their label sequences C . By using conditional independence assumptions, the posterior distribution $p(C \mid X)$ is factorized as Eq. (6),

$$p(C \mid X) \approx \sum_Z \prod_t p(z_t \mid z_{t-1}, C)\, p(z_t \mid X)\, p(C) \tag{6}$$

where z_t is a label in time frame t . Since we use the LRBN before BLSTM, the posterior probability $p(z_t \mid X)$ can be computed by Eq. (7),

$$p(z_t \mid X) = Softmax(Lin(\mathbf{h_t})) \tag{7}$$

where *Softmax(.)* is a softmax activation function; *Lin(.)* is a linear layer to convert hidden vector $\boldsymbol{h_t}$ to a vector with the dimension of the label set size; $\boldsymbol{h_t}$ = *BLSTM(LRBN(X))*, *LRBN(.)* is a LRBN layer followed by a BLSTM layer *BLSTM(.)*.

3 Experiments

Our experimental data consists of 20.73 hours Lhasa-Ü-Tsang, 2.82 hours Yushu-Kham, and 2.15 hours Amdo pastoral dialect, and their corresponding texts contain 3497 syllables for training. We collect 0.3 hours Lhasa-Ü-Tsang, 0.2 hours Yushu-Kham, and 0.2 hours Amdo pastoral dialect respectively to test.

All speech data is sampled at 16 KHz, and 39 MFCC features of each observation frame were extracted from speech data using a 25 ms window with 10ms overlaps. We construct 11 frames of MFCC as long-term dependent speech input to LRBN, and set 39 hidden nodes in hidden layer of LRBN.

The differences in Tibetan dialects are mainly expressed in phonetics, but minor in vocabulary and grammar. In the period of Tubo Dynasty, many works had been done for the determination in Tibetan language writing, which still kept the basic unity of Tibetan written language. So far, Tibetan people have no major obstacles in communication of written language. Even if there are a small amount of differences in vocabulary, and it will tend to be unified. The rules of grammar have changed slightly. Tibetan characters are written in Tibetan letters from left to right, but there is a vertical superposition in syllables, which is a two-dimensional planar character. Therefore, Tibetan letters are not suitable for the output symbols of end-to-end model, since the output is not a recognized Tibetan characters sequence. So, monosyllable of Tibetan characters is used as the CTC output unit.

We evaluate our method by comparing with the multi-dialect CNN/BLSTM-CTC, and dialect-specific BLSTM-CTC. The BLSTM was three layers with 258 cells in each layer and direction, and linear projection layer is followed by each BLSTM. The model was trained for 100 epochs with the ADAM optimizer with batch size of 10. The learning rate was held constant at 0.001. We used 4-layer CNN architecture: convolution1(*f=5, s=1*), maxpool (*f=2,s=2*), convolution2(*f=5,s=1*), maxpool(*f=2,s=2*). The initial three input channels are composed of spectral feature, delta, and delta features. Input speech feature images are down sampled to (1/4*1/4) images along with the time-frequency axises through the two max-pooling layers.

Table 1: Syllable error rate (%) of multi-dialect models and dialect-specific models

	Lhasa-Ü-Tsang	Yushu-Kham	Amdo pastoral dialect
LRBN/BLSTM-CTC	46.71	**52.92**	**49.53**
CNN/BLSTM-CTC	47.15	53.60	50.21
Dialect-specific BLSTM-CTC	**39.80**	55.34	54.35

From Tab. 1, we can see that all multi-dialect speech recognition models outperform dialect-specific models for low-resource dialects, including Yushu-kham dialect and

Amdo pastoral dialect. LRBN/BLSTM-CTC model can represent the shared speech features among different dialects of a language better than CNN/BLSTM-CTC. A generative model of LRBN is more suitable for multi-language speech feature representation.

4 Conclusion

In this paper, we proposed to combine the latent regression Bayesian network with BLSTM-CTC for Tibetan multi-dialect end-to-end speech recognition. The latent regression Bayesian network can extract the shared speech feature for the input of end-to-end speech recognition model. It has a simpler and understood structure and less parameters to learn than CNN. Experimental results show that the LRBN/BLSTM-CTC outperforms CNN/BLSTM-CTC for Tibetan multi-dialect speech recognition.

Acknowledgement: This work is supported by the ministry of education research in the humanities and social sciences planning fund (15YJAZH120), National Natural Science Foundation (61602539, 61873291), and MUC 111 Project.

References

Cai, L.; Zhao, C. X. (2008): Method and implementation of endpoint detection in Ando Tibetan Language. *Gansu Science and Technology*, vol. 24, no. 5, pp. 46-48.

Cai, L. (2009): *Study of Methods of Speech Features Extraction of Ando Tibetan (Ph.D. Thesis).* Qinghai Normal University.

Fang, W.; Zhang, F.; Sheng, V. S.; Ding, Y. (2018): A method for improving CNN-based image recognition using DCGAN. *Computers, Materials & Continua*, vol. 57, no. 1, pp. 167-178.

Li, B.; Sainath, T. N.; Sim, K. C.; Bacchiani, M.; Weinstein, E. et al. (2018): Multi-dialect speech recognition with a single sequence-to-sequence model. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4749-4753.

Li, G. Y.; Meng, M. (2012): Research on acoustic model of large-vocabulary continuous speech recognition for Lhasa Tibetan. *Computer Engineering*, vol. 38, no. 5, pp. 189-191.

Meng, R.; Rice, S. G.; Wang, J.; Sun, X. (2018): A fusion steganographic algorithm based on faster R-CNN. *Computers, Materials & Continua*, vol. 55, no. 1, pp. 1-1.

Nie, S. Q.; Zheng, M.; Ji, Q. (2018): The deep regression bayesian network and its applications. *IEEE Signal Processing Magazine*, pp. 101-111.

Toshniwal, S.; Tara, N. S.; Ron, J. W.; Zhang, Y. T. (2018): Multilingual speech recognition with a single end-to-end model. *ICASSP*, pp. 2040-2045.

Wang, Q. N.; Guo, W.; Xie, C. D. (2017): Towards end to end speech recognition system for Tibetan. *Pattern Recognition and Artificial Intelligence*, vol. 30, no. 4, pp. 359-363.

Watanabe, S.; Hori, T.; Hershey, J. R. (2017): Language independent end-to-end architecture for joint language identification and speech recognition. *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 265-271.

Yuan, S. L.; Guo, W.; Dai, L. R. (2015): Speech recognition based on deep neural networks on Tibetan Corpus. *Pattern Recognition and Artificial Intelligence*, vol. 28, no. 3, pp. 210-213.

Zhang, Y. (2016); *Research on Tibetan Lhasa Dialect Speech Recognition Based on Deep Learning (Ph.D. Thesis).* Northwest Normal University.