Tibetan Multi-Dialect Speech and Dialect Identity Recognition

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

Abstract: Tibetan language has very limited resource for conventional automatic speech recognition so far. It lacks of enough data, sub-word unit, lexicons and word inventories for some dialects. And speech content recognition and dialect classification have been treated as two independent tasks and modeled respectively in most prior works. But the two tasks are highly correlated. In this paper, we present a multi-task WaveNet model to perform simultaneous Tibetan multi-dialect speech recognition and dialect identification. It avoids processing the pronunciation dictionary and word segmentation for new dialects, while, in the meantime, allows training speech recognition and dialect identification in a single model. The experimental results show our method can simultaneously recognize speech content for different Tibetan dialects and identify the dialect with high accuracy using a unified model. The dialect information used in output for training can improve multi-dialect speech recognition accuracy, and the low-resource dialects got higher speech content recognition rate and dialect classification accuracy by multi-dialect and multi-task recognition model than task-specific models.

Keywords: Tibetan multi-dialect speech recognition, dialect identification, multi-task learning, wavenet model.

1 Introduction

Tibetan language is one of the most widely used minority languages in China. It is partly used in India, Bhutan and Nepal. The automatic speech recognition technology for Tibetan language has drawn more and more attention of researchers. It has shown that Tibetan speech recognition has a wide demand and immeasurable application prospects in many practical, real-life situations.

During the long-term development of Tibetan language, different dialects have been formed. Tibetan language is divided into three major dialects in China, including Ü-Tsang, Kham and Amdo dialect. Three dialects are divided into several local sub-dialects. Tibetan dialects pronounce very differently in different regions, such as Ü-Tsang and Kham dialects are tonal, but Amdo dialect is toneless. However, the written characters are unified for all Tibetan dialects. Since Lhasa of Ü-Tsang dialect is Tibetan standard

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

[2] Rensselaer Polytechnic Institute, JEC 7004, Troy NY 12180-3590, USA.

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

CMC. doi:10.32604/cmc.2019.05636

speech, there are much more research works than other dialects on linguistics, speech recognition and corpus [Zhang (2016); Yuan, Guo and Dai (2015); Pei (2009); Li and Meng (2012); Wang, Guo and Xie (2017); Cai and Zhao (2008); Cai (2009); Han and Yu (2010)].

Dialect identification has recently gained substantial interest in the field of language identification. It is more challenging than a general language identification task, since the similarities among dialects of a language are much more in terms of their phoneme set, word pronunciation, and prosodic traits [Shon, Ali and Glass (2018)]. Traditionally, speech content recognition and dialect classification are treated as two independent tasks and modeled respectively. The work in [Shon, Ali and Glass (2018)] explored end-to-end model only for dialect recognition using both acoustic and linguistic feature on Arabic dialect speech data. However, the way humans process speech signals always decipher speech content and other meta information together and simultaneously, including languages, speaker characteristics, emotions, etc. [Tang, Li and Wang (2016)]. The recent works in Li et al. [Li, Sainath, Sim et al. (2018); Toshniwal, Sainath, Weiss et al. (2018); Watanabe, Hori and Hershey (2017)] discussed how to learn a single end-to-end model for joint speech and language recognition. The work in Li et al. [Li, Sainath, Sim et al. (2018)] adopted listen, attend and spell (LAS) model for 7 English dialects and it has shown good performance compared to other LAS models for single dialect tasks. Similar work in Toshniwal et al. [Toshniwal, Sainath, Weiss 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. The work in Watanabe et al. [Watanabe, Hori and Hershey (2017)] was based on hybrid attention/connectionist temporal classification (CTC) architecture where the model used 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. These works suggested that end-to-end model can contribute to handling the variations between different languages or between different tasks by learning and optimizing a single neural network.

End-to-end model has more advantages for low-resource languages than conventional DNN/HMM systems because it avoids the need of linguistic resources such as dictionaries and phonetic knowledge [Li, Sainath, Sim et al. (2018)]. The work in [Sriram, Jun, Gaur et al. (2018)] proposed a general, scalable, end-to-end framework that uses the generative adversarial network (GAN) that was also used in many fields, including computer vision [Li, Jiang and Cheslyar (2018)], to enable robust speech recognition, which do not need domain expertise and simplifying assumptions. Considering limited linguistic resource for Kham dialect and Amdo dialect in Tibetan, our work tries to build an end-to-end model for Tibetan multi-task recognition. It can reduce the efforts of language-dependent processing including the use of pronunciation dictionary and word segmentation which are the big barriers when we build a conventional ASR for new Tibetan dialect. Meanwhile, we try to explore the capability of end-to-end model for capturing the variations between some small-data dialects and a big-data dialect.

In this work, we utilize WaveNet-CTC model to train multi-task recognition on three Tibetan dialects speech data. Since WaveNet is a deep generative model with very large receptive fields, it can capture the characteristics of many different speakers with equal fidelity and model the long-term dependency on speech data [Van Den Oord, Dieleman, Zen et al. (2016)]. It was efficiently applied for Multi-speaker speech generation and text-to-speech. Generative model can capture the underlying data distribution as well as the mechanisms used to generate data, we believe that such ability is crucial for shared representation across speech data from different dialects in a language. WaveNet can also give the predict distribution for speech data conditioned on all previous input, so we use the dialect information as an additional label output during training in order to perform the joint speech and dialect recognition. Experimental results show the advantage of WaveNet-CTC for multi-task Tibetan speech recognition, and multi-dialect model can improve the speech content recognition accuracy for limited-resource dialects.

2 Related works

In Tibetan speech recognition, most of research works is about Lhasa of Ü-Tsang dialect. The recent work in Wang et al. [Wang, Guo and Xie (2017)] applies the end-to-end model based on CTC technology to Lhasa-Ü-Tsang continuous speech recognition, achieving better performance than the state-of-the-art bidirectional long short-term memory network. The work in Huang et al. [Huang and Li (2018)] used end-to-end model training by applying the cyclical neural network and CTC algorithm to the acoustic modeling of Lhasa-Ü-Tsang speech recognition, and introduces time domain convolution operations on the output sequence of the hidden layer to reduce the time domain expansion of the network's hidden layer which improve the training and decoding efficiency of the model. The works in Li et al. [Li, Wang, Wang et al. (2018)] introduces the tone information into Lhasa-Ü-Tsang continuous speech recognition, and designs a set of phonemes with tones, which shows that the tones plays an important role in speech recognition of Lhasa-Ü-Tsang recognition.

In the speech recognition task on Tibetan-Chinese bilingual language, the work in Wang et al. [Wang, Guo, Chen et al. (2017)] solved the problem of sparsity caused by characters as a modeling unit through selecting Tibetan characters and Mandarin nontonal syllables as modeling units and adding noise algorithms. As for the speech recognition for other Tibetan dialects, due to the resources of Kham and Amdo dialect are relatively scarce, a few of related studies are about the endpoint detection, speech feature extraction, and isolated word recognition [Cai and Zhao (2008); Cai (2009); Han and Yu (2010); Li, Yu, Zheng et al. (2017)].

As the topic of Tibetan dialect identity recognition, to our knowledge, there is almost no relevant research. Therefore, the open corpus provided in this paper can make up for this gap for relevant researchers.

In the aspect of multi-task framework for speech recognition, many researchers have done some related works. The work in Ruder [Ruder (2017)] introduced the motivation, learning methods, working mechanism and important auxiliary task selection mechanism of multi-task framework, which provides guidance for applying multi-task framework to speech recognition. The work in Chen et al. [Chen and Mak (2015)] used multi-task framework to conduct joint training of multiple low-resource languages, exploring the universal phoneme set as a secondary task to improve the effect of the factor model of each language. The work in Siohan et al. [Siohan and Rybach (2015)] proposed two methods, namely, early system fusion and multi-task system fusion strategy to reduce the computational complexity of running multiple recognizers in parallel to recognize the speech of adults and children. The work in Tang et al. [Tang, Li and Wang (2016)] integrated speaker recognition and speech recognition into a multi-task learning framework using a recursive structure which attempts to use a unified model to simultaneously identify the two work. The work in Qian et al. [Qian, Yin, You et al. (2015)] combined two different DNNs (one for feature denoising and one for acoustic modeling) into a complete multi-task framework, in which all parameters can be used in real multi-task mode with two criteria training from scratch. The work in [Thanda and Venkatesan (2017)] combined the speaker's lip visual information with the audio input for speech recognition to learn the mapping of an audio-visual fusion feature and the frame label obtained from the GMM/HMM acoustic model, in which the secondary task is mapping visual features to frame labels derived from another GMM/HMM model. The work in Krishna et al. [Krishna, Toshniwal and Livescu (2018)] proposed a hierarchical multi-task model which step further on standard multi-task framework, and the performance in high resource and low resource language recognition were compared. The work in Yang et al. [Yang, Audhkhasi, Rosenberg et al. (2018)] conducted joint learning of accent recognizers and multi-task acoustic models to improve the performance of acoustic models. The above works have one thing in common, that is, the transfer of knowledge between tasks, which is a part of reason why the multi-task framework works. All these works demonstrate the effectiveness of multi-task mechanism.

So it is very significant to establish an accurate Tibetan multi-dialect recognition system using the existing Lhasa-Ü-Tsang speech recognition model and limited amount of other dialect data. It can not only relieve the burdensome data requirements, but also quickly expand the existing recognition model to other target language. It can accelerate the application of Tibetan speech recognition technology.

3 WaveNet-CTC for Tibetan multi-task recognition model

3.1 Wavenet

WaveNet, a deep neural network, is used for generating raw audio waveforms in Van Den Oord et al. [Van Den Oord, Dieleman, Zen et al. (2016)]. The model is fully probabilistic and autoregressive, with the predictive distribution for each audio sample conditioned on all previous ones. It yielded state-of-the-art performance for text-to-speech. A single WaveNet can capture the characteristics of many different speakers and model distributions over thousands of random variables. The work in Van et al. [Van Den Oord, Dieleman, Zen et al. (2016)] also shows that it can be used as a discriminative model, returning promising results for speech recognition. In our work, we employ it to model the distribution of speech data from different dialects and different speakers.

WaveNet model is composed of stacked dilated causal convolutional layers. The network models the joint probability of a waveform as a product of conditional probabilities as Eq. (1).

$$p(\mathrm{x}) = \prod_{t=1}^{T} p(x_t | x_1, \cdots, x_t) \tag{1}$$

The causal convolutions shown in Fig. 1. cannot make the prediction $p(x_{t+1} | x_1, \ldots, x_t)$ of model at timestep t depend on any of the future timesteps $x_{t+1}, x_{t+2}, \ldots x_T$. At training time, the conditional predictions for all timesteps can be made in parallel because all timesteps of ground truth x are known. When generating with the model, the predictions are sequential: after each sample is predicted, it is fed back into the network to predict the next sample. When modeling a long sequence, causal convolutions are faster to train than RNNs, since they do not have recurrent connections.

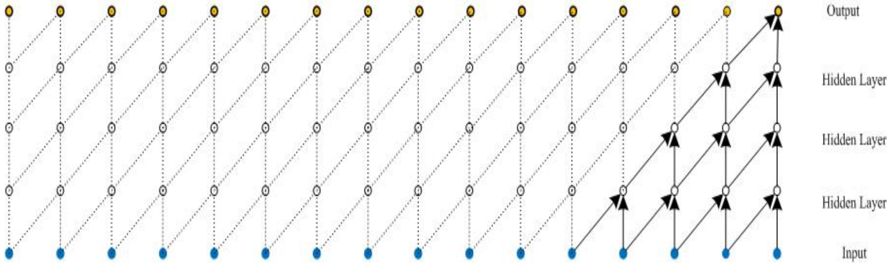

Figure 1: A stack of causal convolutional layers [Van Den Oord, Dieleman, Zen et al. (2016)]

A stack of dilated causal convolutional layers with dilation {1, 2, 4, 8} is shown in Fig. 2. It is more efficient than a causal convolution layers to increase the receptive field, since the filter is applied over an area larger than its length by skipping input values with a certain step.

Stacking a few blocks of dilated causal convolutional layers has very large receptive fields size. For example, 3 blocks of dilated convolution with the dilation {1, 2, 4, 8} are stacked, where each {1, 2, 4, 8} block has receptive field of size 16, and then the dilation repeats as {1, 2, 4, 8, 1, 2, 4, 8, 1, 2, 4, 8}. So the stacked dilated convolutions have receptive field of size 4096.

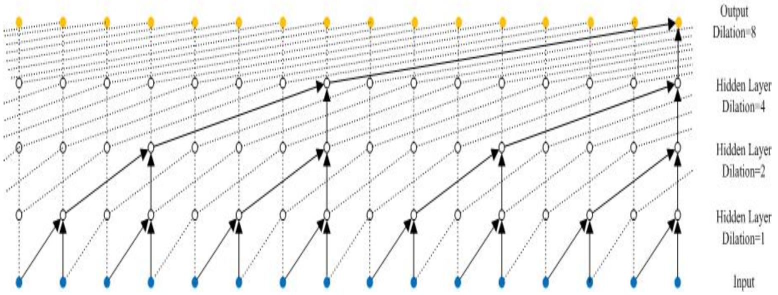

Figure 2: A stack of dilated causal convolutional layers [Van Den Oord, Dieleman, Zen et al. (2016)]

WaveNet uses the gated activation unit as same as the one used in the gated PixelCNN [Oord, Kalchbrenner and Kavukcuoglu (2016)]. Its activation function is as Eq. (2).

$$h_i = \tanh(W_{f,i} * x_i) \odot \sigma \left(W_{g,i} * x_i \right) \tag{2}$$

where * denotes a convolution operator, Θ denotes an element-wise multiplication operator, $\sigma(\cdot)$ is a sigmoid function. i is the layer index. f and g denote filter and gate, respectively, and W is learnable weight.

WaveNet uses residual and parameterised skip connections to speed up convergence and enable training of much deeper models. The more details on WaveNet can be found in [Van Den Oord, Dieleman, Zen et al. (2016)].

3.2 End-to-end Tibetan multi-task model

We adopt the architecture of Speech-to-Text-WaveNet [Namju (2017)] for Tibetan multi-task speech recognition. It uses a single CTC to sit on top of WaveNet and trains WaveNet with CTC loss. The forward-backward algorithm of CTC can map speech to text sequence. The architecture is shown as Fig. 3.

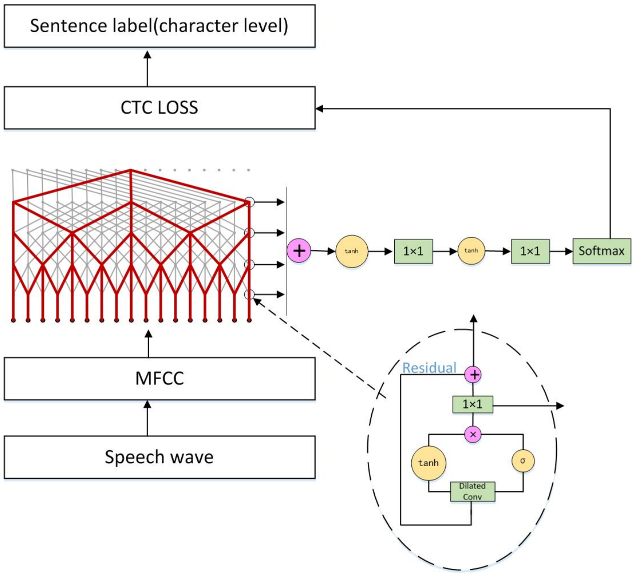

Figure 3: The architecture of WaveNet-CTC [Huang and Li (2018)]

The difference among Tibetan dialects is 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 with written language. Even if there is a small amount of differences in vocabulary, 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 (syllables are separated by delimiter “་”.), which is a two-dimensional planar character shown as Fig. 4. A Tibetan sentence is shown in Fig. 5, where the sign “|” is used as the end sign of a Tibetan sentence. Tibetan letters are not suitable for the output symbols of end-to-end model, because the output is not a recognized Tibetan characters sequence. So a syllable of Tibetan characters is used as the CTC output unit.

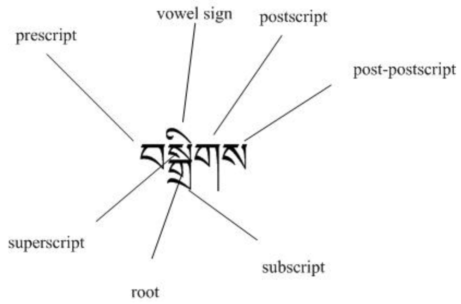

Figure 4: The structure of a Tibetan syllable

ང་ལ་སྒོར་བརྒྱད་ཡོད།

Figure 5: A Tibetan sentence (It means I have eight bucks)

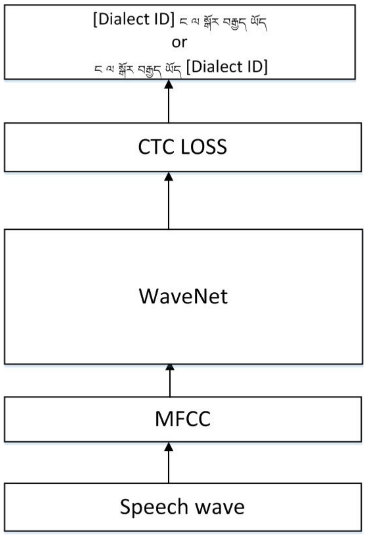
[Dialect ID] ང ལ སྒོར བརྒྱད ཡོད
or
ང ལ སྒོར བརྒྱད ཡོད [Dialect ID]
CTC LOSS
WaveNet
MFCC
Speech wave

Figure 6: Our end-to-end model for Tibetan multi-task recognition

In this paper, we explore to expand the Tibetan characters sequence with dialect symbols as output targets. For example, when including the Yushu-Kham dialect, we add the symbol 'Y' into the label inventory. We evaluate two ways to add the dialect information into label sequence. One is to add the symbol to the beginning of the target label sequence, like "Y ཐུགས རྗེ ཆེ" ("ཐུགས་རྗེ་ཆེ།" means "Thanks" in English). The other is to add the symbol at the end of the label sequence, like "ཐུགས རྗེ ཆེ Y".

Meanwhile, we remove the sign "|" in Tibetan sentence and replace the delimiter "་" with the space. In this work, we do not combine a language model. Our end-to-end model for Tibetan multi-task speech and dialect recognition is shown as Fig. 6.

4 Experiments

4.1 Data

Our experimental data is from an open and free Tibetan multi-lingual speech data set TIBMD@MUC, which can be downloaded from https://pan.baidu.com/s/14CihgqjA4AFFH1QpSTjzZw. The text corpus consists of two parts. One is 1396 spoken language sentences selected from the book "Tibetan Spoken Language" [La (2005)] written by La Bazelen, and the other part is collected to 8,000 sentences from online news, electronic novels and poetry of Tibetan on internet. All text corpuses include a total of 3497 Tibetan syllables.

There are 114 recorders who were from Lhasa City in Tibet, Yushu City in Qinghai Province, Changdu City in Tibet and Tibetan Qiang Autonomous Prefecture of Ngawa. They used different dialects to speak out the same text for 1396 spoken sentences, and other 8000 sentences are read loudly in Lhasa dialect. Speech data files are converted to 16K Hz sampling frequency, 16bit quantization accuracy, and wav format.

Our experimental data for multi-task speech recognition is shown in Tab. 1, which 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.

39 MFCC features of each observation frame are extracted from speech data using a 25ms window with 10ms overlaps.

Table 1: Tibetan multi-dialect dataset statistics

Dialect	Training data (hours)	Training utterances (#)	Test data (hours)	Test utterances (#)
Lhasa-Ü-Tsang	20.73	15870	0.3	264
Yushu-Kham	2.82	2203	0.2	137
Amdo pastoral dialect	2.15	2671	0.2	111
Total	25.7	20744	0.7	512

4.2 Model details

For multi-task speech recognition, the CTC output layer contains 3502 nodes (3497+1 blank+1 space+3 dialect ID labels). The WaveNet network consists of 15 layers, grouped into 3 dilated residual block stacks of 5 layers. In every stack, the dilation rate increases by a factor of 2 in every layer, starting with rate 1 (no dilation) and reaching the maximum dilation of 16 in the last layer. The filter size of causal dilated convolutions is 7. The number of hidden units in the gating layers is 128. The number of hidden units in the residual connection is 128. The model was trained for 100 epochs with the ADAM optimizer with batch size of 10. The learning rate was held constant at. The models were trained on one Nvidia GTX1070Ti GPU.

For dialect-specific model for small-data dialects, we first took the multi-dialect model without dialect ID, i.e., “Model”, as the starting point and retraining the same architecture for each dialect using a small amount of training data. We refer to this type of models as “Model-R” in Tab. 2. These models got acceptable recognition rates. We also build dialect-specific models on each dialect data, as “Dialect-specific model” in Tab. 2. The Model-R by retaining achieved better performance than Dialect-specific model for small-data dialects for speech content recognition.

Table 2: Syllable error rate (%) of dialect-specific models

	Yushu-Kham	Amdo pastoral dialect
Model-R	45.91	49.74
Dialect-specific model	52.46	53.47

For dialect identity recognition model, we used a two-layer LSTM (300 hidden units in each layer) network followed by a softmax layer to classify the dialect identities, in which cross entropy was adopted as loss function. The model was trained for 500 epochs with the ADAM optimizer with batch size of 50. The learning rate was held constant at 0.001. The weight parameters of the softmax layer were initialized with random uniform distribution of range [0, 1]. We also crop the gradient to within [-1, 1] to alleviate the gradient vanishing.

4.3 Results

The experimental results are shown in Tab. 3 and Tab. 4. We refer to the model integrated with dialect ID at the beginning of the output as “ID-Model”, the model with dialect ID at the end of the output as “Model-ID”, the model without dialect information in output as “Model” respectively, and compared them with the end-to-end Dialect-specific model.

From Tab. 3, 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. WaveNet-CTC model can capture the shared speech features and linguistic features among different dialects of a language. The underlying shared knowledge in one language can transfer from one dialect to other dialects. For Lhasa-Ü-Tsang, a big-data dialect, all multi-dialect speech recognition models performed worse than dialect-specific model. It shows that the added two small-data dialect does harm to big-data dialect for multi-dialect speech recognition. In spite of that, the ID-Model trained with dialect information at the beginning of label sequence has closer recognition rate to dialect-specific model for Lhasa.

Table 3: A comparison on syllable error rate (%) of multi-task models and task-specific models

	Lhasa-Ü-Tsang	Yushu-Kham	Amdo pastoral dialect
Model	50.45	46.08	49.84
ID-Model	39.89	**41.18**	**43.14**
Model-ID	48.26	45.65	49.01
Dialect-specific model	**35.46**	52.46	53.47

Inserting the dialect symbol into label sequence performed better than the model without dialect information for multi-dialect speech recognition. It shows that dialect information helps to improve the speech content recognition for multi-task models.

From Tab. 4, we can observe that multi-task learning models have the very high accuracy for dialect identity recognition. ID-Model and Model-ID outperformed dialect ID recognition model. It presents that multi-task speech recognition models can decipher speech content and dialect information together and simultaneously, and perform both well. This is the same way that human process speech signals.

Table 4: Dialect ID recognition accuracy (%) of multi-task models and task-specific model

	Lhasa-Ü-Tsang	Yushu-Kham	Amdo pastoral dialect
ID-Model	100	99.27	100
Model-ID	100	97.81	100
Dialect-specific model	99.1	73.3	85.2

Besides, the ID-Model has higher accuracy than Model-ID for both speech recognition and dialect identification. Based on this observation, it shows that the speech content recognition depends upon the accuracy of dialect classification in this multi-dialect and multi-task recognition model.

5 Conclusion

In this paper, we proposed to use the WaveNet-CTC model for Tibetan multi-dialect and multi-task recognition. It provides a simple and effective solution for building new Tibetan dialect model without the use of dialect-specific linguistic resource. It is optimized to predict the Tibetan character sequence appended with the dialect symbol as the output target, which effectively forces the model to learn shared hidden representation that are suitable for both character prediction and dialect prediction for different dialect of a language. In future work, we will improve the speech content recognition accuracy using a Tibetan language model.

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

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.

Chen, D.; Mak, B. K. W. (2015): Multitask learning of deep neural networks for low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 7, pp.1172-1183.

Han, Q. H.; Yu, H. (2010): Research on speech recognition for Ando Tibetan based on HMM. *Software Guide*, vol. 9, no. 7, pp.173-175.

Huang, X. H.; Li, J. (2018): The acoustic model for Tibetan speech recognition based on recurrent neural network. *Journal of Chinese Information Processing*, vol. 32, no. 5, pp. 49-55.

Krishna, K.; Toshniwal, S.; Livescu, K. (2018): Hierarchical multitask learning for CTC-based speech recognition. arXiv preprint arXiv:1807.06234.

La, B. (2005): *Tibetan Spoken Language*. Publishing House of Minority Nationalities.

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. 2163-2168.

Li, C.; Jiang, Y. M.; Cheslyar, M. (2018): Embedding image through generated intermediate medium using deep convolutional generative adversarial network. *Computers, Materials & Continua*, vol. 56, no. 2, pp. 313-324.

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.

Li, G.; Yu, H.; Zheng, T. F.; Yan, J.; Xu, S. (2017): Free linguistic and speech resources for Tibetan. *Proceedings of APSIPA Annual Summit and Conference*, vol. 2017, pp. 12-15.

Li, J.; Wang, H.; Wang, L.; Dang, J.; Khuru, K. et al. (2016): Exploring tonal information for Lhasa dialect acoustic modeling. *10th International Symposium on Chinese Spoken Language Processing*, pp. 1-5.

Li, Y. J.; Zhao, X. M.; Xu, W. Q.; Yan Y. H. (2018): Cross-lingual multi-task neural architecture for spoken language understanding. *Proceeding of Interspeech*, pp. 566-570.

Namju, K. B. (2017): Speech-to-text-wavenet: end-to-end sentence level English speech recognition using deepmind's wavenet. https://github.com/buriburisuri/speech-to-text-wavenet .

Oord, A. V. D.; Kalchbrenner, N.; Kavukcuoglu, K. (2016): Pixel recurrent neural networks. arXiv:1601.06759.

Pei, C. B. (2009): *Research on Tibetan Speech Recognition Technology Based on Standard Lhasa (Ph.D. Thesis).* Tibet University.

Qian, Y.; Yin, M.; You, Y.; Yu, K. (2015): Multi-task joint-learning of deep neural networks for robust speech recognition. *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 310-316.

Ruder, S. (2017): An overview of multi-task learning in deep neural networks. arXiv:1706.05098.

Shon, S.; Ali, A.; Glass, J. (2018): Convolutional neural networks and language embeddings for end-to-end dialect recognition. arXiv:1803.04567.

Siohan, O.; Rybach, D. (2015): Multitask learning and system combination for automatic speech recognition. *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 589-595.

Sriram, A.; Jun, H.; Gaur, Y.; Satheesh, S. (2018): Robust speech recognition using generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 5639-5643.

Tang, Z.; Li, L.; Wang, D. (2016): Multi-task recurrent model for speech and speaker recognition. *Signal and Information Processing Association Annual Summit and Conference*, pp. 1-4.

Thanda, A.; Venkatesan, S. M. (2017): Multi-task learning of deep neural networks for audio visual automatic speech recognition. arXiv:1701.02477.

Toshniwal, S.; Sainath, T. N.; Weiss, R. J.; Li, B.; Moreno, P. et al. (2018): Multilingual speech recognition with a single end-to-end model. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4904-4908.

Van Den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O. et al. (2016): Wavenet: A generative model for raw audio. arXiv:1609.03499.

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.

Wang, Q.; Guo, W.; Chen, P.; Song, Y. (2017): Tibetan-Mandarin bilingual speech recognition based on end-to-end framework. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1214-1217.

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

Recognition and Understanding Workshop, pp. 265-271.

Yang, X.; Audhkhasi, K.; Rosenberg, A.; Thomas, S.; Ramabhadran, B. et al. (2018): Joint modeling of accents and acoustics for multi-accent speech recognition. arXiv:1802.02656.

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.