End-to-end Tibetan Ando dialect speech recognition based on hybrid CTC/attention architecture

Jingwen Sun[†] Gang Zhou[†] Hongwu Yang[*†#] and Man Wang[†]
[*]School of Educational Technology, Northwest Normal University, Lanzhou, China
[†]College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, China
[#]National and Provincial Joint Engineering Laboratory of Learning Analysis Technology in Online Education, Lanzhou, China
E-mail: yanghw@nwnu.edu.cn Tel: +86-18193125028

***Abstract*— End-to-end automatic speech recognition reduces the difficulty of building a speech recognition system through single network architecture. The tokenization, pronunciation dictionary and phonetic context-dependency trees required in the traditional deep learning-based speech recognition are omitted in this system to simplify the complex modeling process. This paper proposes a method to realize Tibetan Ando dialect speech recognition with end-to-end speech recognition model based on hybrid connectionist temporal classification (CTC)/attention. A bidirectional long short-term memory network (BLSTM) is used for the encoder network through 80 mel-scale filter-bank coefficients alone with pitch features form total 83-dimensionals acoustic features to train the network. We compared proposed method with the methods only based on CTC architecture and the structure only based on attention architecture by adjusting CTC weight of the system. The result shows that the hybrid model can obtain optimal weight to achieves the highest recognition rate of 64.5% when the CTC weight is 0.2.**

***Index Terms*: Tibetan Ando speech recognition; attention mechanism; CTC; hybrid CTC/attention; end-to-end system.**

I. INTRODUCTION

The aim of automatic speech recognition (ASR) technology is to make computers understand human languages. With the development of information technology, deep learning network is widely used in the field of automatic speech recognition. Traditional speech recognition mainly uses ASR system based on a hidden Markov model (HMM)/ Gaussian mixture model (GMM)[1]. These systems use HMM to model the speech unit and GMM to generate the observation probability of HMM. However, the independence assumption in these methods ignore the inter-frame correlation.

In recent years, deep learning has been widely used in the field of speech recognition. The framework based on Hidden Markov Model/Deep Neural Networks (HMM/DNN) has become a typical speech recognition model. DNN is a neural network structure with multiple hidden layers[2]. The system built by this model has made considerable progress in training and recognition. However, the speech recognition system based on HMM/DNN consists of acoustic model, pronunciation dictionary and language model. It is very difficult for minority languages to build such a system because of lacking of linguistic knowledge.

To solve these problems, Graves et al. proposed the connectionist temporal classification (CTC)[3], which regards the speech recognition as a sequence conversion (speech sequence to annotation sequence) and abandons a series of hypotheses of the traditional HMM based speech recognition system. An end-to-end speech recognition framework was then proposed that does not require linguistic correlation and simplifies the deep network model[4]. At present, there are two main end-to-end speech recognition systems. One is the Attention-based method, which is used to align the acoustic frame and the recognition symbols. The other is based on the CTC method, which uses Markov assumption to solve the sequence problem effectively through dynamic programming[5]. The end-to-end speech recognition system adopts the method of constructing an acoustic model based on BLSTM network[6]. BLSTM can automatically learn the correspondence between acoustic features and annotation sequences during the training process, and ban the complex state alignment in the traditional speech recognition process[6].

Tibetan language belongs to the Tibetan-Burmese branch of the Sino-Tibetan language family[7]. It is a phonetic language. The Ando dialect is one of the three major dialects of Tibetan language and is mainly used in the Ando Tibetan area[8]. The pronunciation of the three major dialects of Tibetan language is quite different, but the same Tibetan text system is used. The biggest feature of the Ando dialect compared to other Tibetan dialects is the indistinguishable tone. The study of Tibetan Ando speech recognition technology can effectively solve the language barrier between Ando Tibetan people and people of other nationalities in China and promote intercultural exchanges[9].

Due to the difficulty of collecting Tibetan Ando dialect data, the lack of corpus will cause serious data sparsity in the training process, which makes the research progress of Tibetan Ando dialect recognition very slow. Furthermore, the study of Tibetan linguistics is not perfect[10]. To sort out Ando dialect pronunciation dictionary and define phoneme set for Ando dialect requires professional knowledge of Tibetan linguistics. Because of the constraints of these factors, it is very difficult to recognize Ando dialect speech using traditional methods such as HMM/DNN[11]. The end-to-end speech recognition technology does not rely on linguistic

knowledge, only the speech and its corresponding pronunciation text[12]. We propose to apply the end-to-end speech recognition technology to the Tibetan Ando dialect speech recognition. The hybrid CTC/attention architecture is used with continuously adjusting the CTC weights to find the weight value that maximizes the recognition rate.

II. END-TO-END ASR

This section introduces the formulation of CTC and attention-based end-to-end ASR system separately. Their respective disadvantages are also illustrated. Meanwhile, combined with their advantages, the working principle of hybrid CTC/attention architecture and how to change the weight of CTC is explained.

A. Connectionist Temporal Classification

Traditional speech recognition methods need to know which pronunciation corresponding to each frame. The end-to-end approach is concerned with whether the output sequence is the same as the input sequence, rather than whether the predicted sequence and the input sequence are aligned at a certain point in time.

Speech recognition mainly solves the matching from a speech feature sequence, X, to an word sequence, W[13]. It is formulated as follows:

$$\hat{W} = \arg\max_{W \in V^*} p(W \mid X). \tag{1}$$

ASR is estimated in all possible word sequences and finds the most probable word sequence[14]. Therefore, the problem is transformed into the calculation of the posterior probability.

Especially, CTC uses the blank symbol "<b>" to represent the letter boundary of the letter sequence, C, which aims to solve the problem of overlapping words[14]. The state sequence is denoted by Z with "<b>". Formulas can be rewritten as follows:

$$p_{ctc}(C \mid X) = \sum_{Z} p(C \mid Z, X) p(Z \mid X) \tag{2}$$

$$\approx \sum_{Z} p(C \mid Z)\, p(Z \mid X). \tag{3}$$

In (3), $p(Z \mid X)$ is acoustic model and $p(C \mid Z)$ is letter mode.

In acoustic model, the posterior distribution is modeled by BLSTM. The input to BLSTM is speech feature sequence, X, and the output is a hidden vector.

$$p(Z \mid X) = \prod_{t=1}^{T} p(z_t \mid z_1, \cdots, z_{t-1}, X) \tag{4}$$

$$\approx \prod_{t=1}^{T} p(z_t \mid X). \tag{5}$$

In CTC letter model, $p(C \mid Z)$ can be rewritten by conditional independence hypothesis as:

$$p(C \mid Z) = \frac{p(Z \mid C) p(C)}{p(Z)} \tag{6}$$

$$= \prod_{t=1}^{T} p(z_t \mid z_1, \cdots, z_{t-1}, C) \frac{p(C)}{p(Z)} \tag{7}$$

$$\approx \prod_{t=1}^{T} p(z_t \mid z_{t-1}, C) \frac{p(C)}{p(Z)}. \tag{8}$$

They are a combination of the state transition probability. $p(C)$ is the language model of CTC. Put (5) and (8) into (3), we can obtain:

$$p_{ctc}(C \mid X) \approx \sum_{Z} \prod_{t=1}^{T} p(z \mid z_{t-1}, C)\, p(z_t \mid X) \frac{p(C)}{p(Z)}. \tag{9}$$

B. Attention

Attention adopts encoder-decoder structure. Attention mechanism is concerned with alignment between acoustic frames and recognition symbols. This structure usually encodes the input sequence into a fixed length vector.

Unlike the CTC method, the attention-based method does not need to make any conditional independence hypothesis[15], and directly estimates posterior probability, $p(C \mid X)$:

$$p_{att}(C \mid X) = \prod_{l=1}^{L} p(c_l \mid c_1, \cdots, c_{l-1}, X). \tag{10}$$

$p_{att}(C \mid X)$ is an attention-based posterior distribution function, $p = (c_l \mid c_1, \cdots, c_{l-1}, X)$ can be obtained by:

$$h_t = Encoder(X), \tag{11}$$

$$r_l = \sum_{t=1}^{T} a_{lt} h_t, \tag{12}$$

$$p(c_l \mid c_1, \cdots, c_{l-1}, X) = Decoder(r_l, q_{l-1}, c_{l-1}). \tag{13}$$

Eq. (11) and (13) are the networks of encoder and decoder. BLSTM is used in the encoder network. Letter-wise hidden vector is expressed as r. The decoding network is composed of their previous state.

C. Hybrid CTC/Attention

Since attention is generated by the decoder network, when it predicts the end label of sequence, it may not pay attention to all encoded frames and ends prediction label prematurely. Instead, it can predict the next label with high probability by focusing on the same part as the previous label. In this case, the same label sequence is predicted repeatedly.

CTC probability forces monotonic alignment, and does not permit large jumps or loops in the same frame[16]. It avoids the premature prediction of end-of-sequence label by attention mechanism. Frame-synchronously is done by CTC while output-label-synchronously is done by attention[16]. So we combined two methods.

In this paper, the end to end speech recognition based on hybrid CTC/attention architecture is applied to the Tibetan Ando dialect speech recognition, which effectively utilizes the advantages of both architectures in training and decoding. The input sequence, X, is transformed into the high level features, H, and the attention decoder generates the letter sequence, C. Fig. 1 illustrates the overall architecture of the framework. For the attention model, uses the labels <sos> as the starting symbol and <eos> represents the end of

a sequence. The architecture of hybrid CTC/attention-based end-to-end Mandarin speech recognition is shown in Fig. 1.

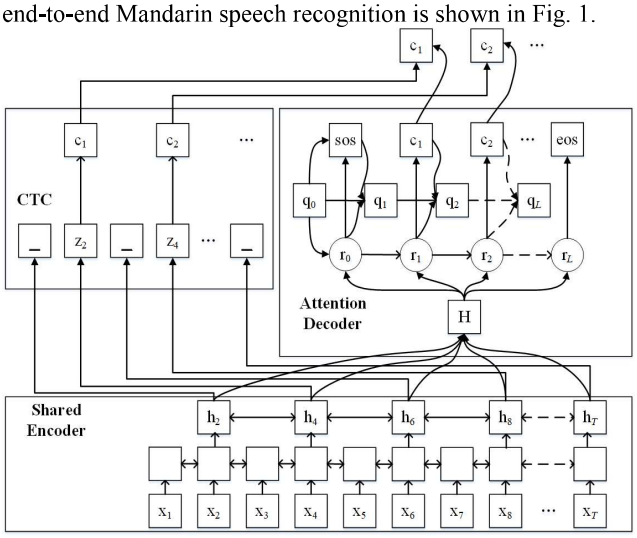

Fig.1 Architecture of Hybrid CTC/attention-based end-to-end Tibetan Ando dialect speech recognition

In the training process, cross-entropy is used to represent the loss value. We uses a multi-objective learning framework (MOL), which combines cross-entropy of both CTC and attention to improve the robustness[16], as follows:

$$L_{MOL} = \alpha L^{ctc} + (1-\alpha) L^{att} \tag{14}$$

An adjustment parameter α linearly interpolate both objective functions in (14) and satisfies $0 \le \alpha \le 1$. The smaller the cross-entropy of MOL, the closer the predicted value is to the real value. When α equals to 0, the experiment is based on the CTC architecture only. Conversely, when α equals to 1, the experiment is based on the attention architecture only.

It is worth mentioning that we have adopted the form of joint decoding by CTC and Attention-based methods in a algorithm named one-pass beam search. The purpose of using this algorithm is to eliminate misalignment and improve accuracy.

III. EXPERIMENTS

A. Ando Dialect Corpus Construction

Tibetan Lhasa dialect and Ando dialect have different pronunciation, but their characters are the same. Tibetan belongs to alphabetic writing and phonemes are the basic unit. Tibetan characters are composed of one or more phonemes according to certain rules. Therefore, we select 8000 common Tibetan sentences for text corpus construction from Tibetan web pages, books, daily Tibetan expressions and so on. These texts are averagely divided into eight groups, each independently recorded by an Ando speaker. We require the speakers to speak clearly and pronounce fluently, whose age is between 18 and 30.

We numbered each sentence regularly, such as “1-fash-b”. Among them, “1” refers to the text number, “fash” stands for the name of the speaker, and “b” represents Ando dialect. What is more, the texts have been processed. For example, the syllable-dividing symbols “·” in each Tibetan sentence are replaced by spaces and single pendulum characters “།” are deleted.

In particular, two text files need to be prepared, one is the path of all corpus, the other is the text content corresponding to all speech files.

B. Ando Dialect Recognition

Totaling 8000 Tibetan Ando dialect corpus were used in the experiment. Our Tibetan Ando dialect corpus was recorded using a phone device in office environments, based on a collection of monologue speech data. The length of corpus is 12.1 hours, of which 9 hours are used for training set and the rest for testing set. The sampling rate is 16 kHz. 80 mel-scale filterbank coefficients with pitch features to obtain a total of 83 feature values per frame as an input feature vector[17]. The total of training epoch is set to 17.

We tested six groups of experiments with different α : 0.0, 0.2, 0.4, 0.6, 0.8 and 1.0 for training and decoding, and a four-layer BLSTM is used for the training networks. The experimental results are as follows:

Table I End-to-end Tibetan Ando dialect speech recognition results in hybrid CTC\attention architecture.

α	Corr/%	Sub/%	Del/%	Ins/%	Len/h
0.0	63.8	28.6	7.8	13.0	12.1
0.2	64.5	27.1	8.6	11.1	12.1
0.4	63.9	28.5	7.9	11.1	12.1
0.6	63.3	29.0	8.0	11.7	12.1
0.8	62.1	29.3	7.9	13.2	12.1
1.0	61.6	30.0	8.8	12.9	12.1

As summarized in Table I, the character speech recognition task is evaluated in terms of character recognition rate for the proposed hybrid CTC/attention end-to-end ASR for Tibetan Ando dialect speech recognition task. The best recognition rate achieved 64.5% while α is equal to 0.2 without using any linguistic resources include pronunciation lexicon and language model. At this time, substitution error (Sub) is 27.1%, deletion error (Del) is 8.6%, Insert error (Ins) is 11.1% and corpus length (Len) is 12.1 h.

The recognition rate of the experiment based on the CTC architecture only is 63.8%. The recognition rate of the experiment based on the attention architecture only is 61.6%. The experimental results verify the superiority of the hybrid model.

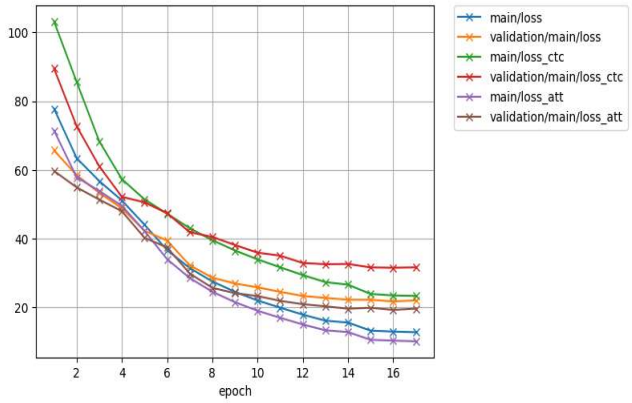

Fig. 2 Convergence curve of hybrid CTC/Attention architecture training and validation data.

Fig. 2 shows the learning convergence curves for the training sets (main/loss) and the validation sets (validation/main/loss) of CTC, Attention and hybrid CTC/Attention mode over the training epochs while α is equal to 0.2 separately. The horizontal axis is the number of training epoch (Epoch) and the vertical axis is the loss value (Loss). As seen from the Fig. 2, when the number of training epoch reaches 15, the loss value of train sets is about 10% and the loss value of validation sets is about 21%, and almost does not decrease with the increase of epoch. The value of loss is calculated by (14).

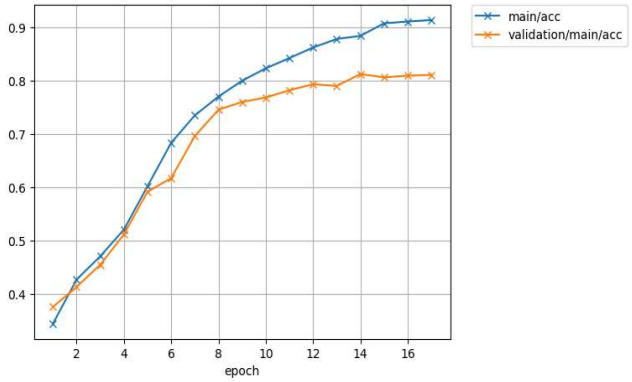

Fig.3 Accuracy of training and verification data in hybrid CTC/attention architecture.

Fig. 3 shows the learning curves of the accuracy for the training sets (main/acc) and validation sets (validation/main/acc) of Tibetan Ando dialect speech corpus over the training epochs. The horizontal axis is the number of training epoch (Epoch) and the vertical axis is the accuracy value (Acc). As seen from the Fig. 3, when the number of training epoch reaches 15, the accuracy of train sets is about 91% and the accuracy of validation sets is about 81%, and almost does not increase with the increase of epoch.

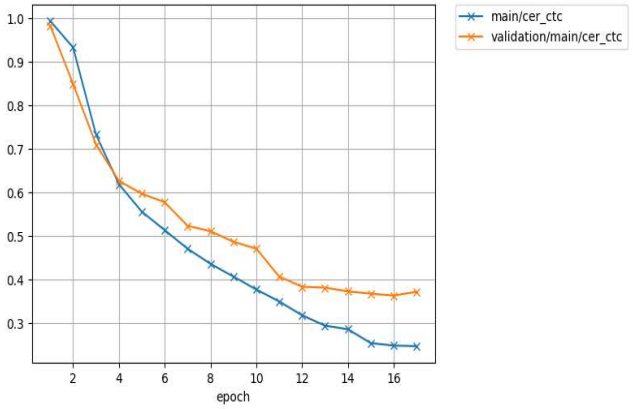

Fig. 4 Character error rate of training and verification data in hybrid CTC/attention architecture

Fig. 4 shows the character error rate for the training sets (main/loss) and the validation sets (validation/main/loss) over the training epochs while α is equal to 0.2. The horizontal axis is the number of training epoch (Epoch) and the vertical axis is the character error rate (CER). As seen from the Fig. 4, when the number of training epoch reaches 15, character error rate of train sets is about 24% and character error rate of validation sets is about 36%, and almost no longer changes with the increase of epoch.

IV. CONCLUSIONS

This paper applies an end-to-end speech recognition technology to the Tibetan Ando dialect speech recognition process. Compared with the conventional HMM/DNN-based speech recognition method in the previous research, end-to-end Tibetan Ando dialect speech recognition does not require complex data preparation and linguistic resources such as pronunciation dictionary, and language model. In the experimental, we chose BLSTM as the encoder network, and the training method adopt the multiobjective learning framework, and implement a end-to-end Tibetan Ando dialect speech recognition based on CTC/attention architecture. The optimal weight of CTC is found to be 0.2 by adjusting constantly. The recognition rate reached 64.5%. Future work will extract different speech features and Tibetan Ando dialect language model to the end-to-end Tibetan Ando dialect speech recognition in order to improve the recognition accuracy. Furthermore, adding contrast experiment based on DNN to perfect the experimental conclusion.

V. ACKNOWLEDGMENT

The research leading to these results was partly funded by the National Natural Science Foundation of China (Grant No.11664036), Natural Science Foundation of Gansu (Grant No. 1506RJYA126), High School Science and Technology Innovation Team Project of Gansu (2017C-03).

REFERENCES

[1] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling Using Deep Belief Networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 14-22, 2012.

[2] Y. Sheng-Long，G. Wu，and D. Li-Rong, "Speech Recognition Based on Deep Neural Networks on Tibetan Corpus," *J. Pattern Recognition & Artificial Intelligence*, vol. 28, pp. 209-213, 2015.

[3] A. Graves, S. Fernández, and F. Gomez, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," *International Conference on Machine Learning. ACM*, pp. 369-376, 2006.

[4] A Graves, and N Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," I*nternational Conference on Machine Learning*, pp.1764-1772, 2014.

[5] M. Yajie, G. Mohammad, and M. Florian, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," *Automatic Speech Recognition and Understanding*, pp. 167-174, 2016.

[6] M. Schuster, and K. K. Paliwal, "Bidirectional recurrent neural networks," *J. IEEE Transactions on Signal Processing*, vol. 45, 1997.

[7] G. S. Jumian, and G. A. Y. jing, "Tibetan Dialect Summary," *Hous e of M inority*, 2002.

[8] L. Yonghong, K. Jiangping, and Y. Hongzhi, "Rules for the auto-transformation of Tibetan text to IPA," *J. Journal of Tsinghua University*, 2008.

[9] D. Drolma, "Study on Tibetan Language Speech Recognition," *J. Journal of Tibet University*, 2010.

[10] H. Xiaohui, and L. Jing, "The Acoustic Model for Tibetan Speech Recognition Based on Recurrent Neural Network," *J. Journal of Chinese Information Processing*, 2018.

[11] G. Li, and M. Meng, "Research on Acoustic Model of Large-vocabulary Continuous Speech Recognition for Lhasa Tibetan," *J. Computer Engineering*, vol. 38, pp. 189-191, 2012.

[12] Q. Wang, W. Guo, and C. Xie, "Towards End to End Speech Recognition System for Tibetan," *J. Pattern Recognition and Artificial Intelligence*, 2015.

[13] W. Hui, Z. Yue, L. Xiao-feng, X. Xiao-na, Z. Nan, and X. Yan-min, "Deep Feature Learning for Tibetan Speech Recognition using Sparse Auto-encoder," *J. Journal of Northeast Normal University*, 2015.

[14] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4960-4964, 2016.

[15] J Chorowski, D Bahdanau, and D Serdyuk, et al, "Attention-Based Models for Speech Recognition," *J. Computer Science*, vol. 10, pp. 429-439, 2015.

[16] K. Suyoun, H. Takaaki, and W. Shinji, "Joint CTC-Attention based End-to-End Speech Recognition using Multi-task Learning," *J*. pp. 4835-4839, 2017.

[17] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," *J. IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 1240-1253, 2017.