ATTENTION-BASED WAVENET-CTC FOR TIBETAN MULTI-DIALECT MULTITASK SPEECH RECOGNITION

Jianjian Yue
School of Information and Engineering
Minzu University of China
Beijing, China
yuejian99@163.com

Yue Zhao (correspongding author)
School of Information and Engineering
Minzu University of China
Beijing, China
zhaoyueso@muc.edu.cn

Xiaona Xu
School of Information and Engineering
Minzu University of China
Beijing, China
xu_xiaona@163.com

Licheng Wu
School of Information and Engineering
Minzu University of China
Beijing, China
wulicheng@mail.tsinghua.edu.cn

Xiali Li
School of Information and Engineering
Minzu University of China
Beijing, China
xiaer_li@163.com

Bo Liu
School of Information and Engineering
Minzu University of China
Beijing, China
liuboer@163.com

Qiang Ji
Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
Troy NY 12180-3590, USA
qji@rpi.edu

***Abstract*—To date, Tibetan has very limited resources for conventional automatic speech recognition. It lacks of enough data, sub-word units, lexicons, and word inventories for some dialects. In this paper, we present an end-to-end model, attention-based WaveNet-CTC, to perform simultaneous Tibetan speech content recognition and dialect identification. This model avoids processing the pronunciation dictionary and word segmentation for new dialects while allowing for training of multi-dialect speech content recognition and dialect identification in a single model. We introduced the attention mechanism into WaveNet to capture the context information and relations of discontinuous speech frames among different dialects and tasks. And the dialect information is used in the output for multitask learning. The experimental results show that attention-based WaveNet-CTC has better performance compared with WaveNet-CTC and task-specific models for the multi-dialect multitask speech recognition.**

Keywords—multi-dialect speech recognition, multitask learning, WaveNet-CTC model, attention mechanism, Tibetan language

I. INTRODUCTION

Tibetan is one of the most widely used minority languages in China. It is also used in parts of India, Bhutan, and Nepal. During the long-term development of the Tibetan language, different dialects have been formed. The language is divided into three major dialects in China: Ü-Tsang, Kham, and Amdo. Tibetan dialects are pronounced very differently in different regions, but the written characters are unified across regions. Since the Lhasa of Ü-Tsang dialect is standard Tibetan speech, there is more research on its linguistics, speech recognition, and corpus establishment than on those of other dialects [1–8].

Recently, the field of speech recognition has been in the midst of a paradigm shift: end-to-end neural networks are challenging the dominance of hidden Markov models as a core technology [9-13]. End-to-end automatic speech recognition has more advantages than conventional DNN/HMM systems, especially for low-resource languages, because it avoids the need for linguistic resources like dictionaries and phonetic knowledge [14]. Hence, it is of great significance to make the process faster and easier using an end-to-end model. The work in [14] adopted the listen, attend, and spell (LAS) model for 7 English dialects, and it showed good performance. Similar work in [15] with multitask end-to-end learning for 9 Indian languages obtained the largest improvement by conditioning the encoder on the speech language identity. These studies showed that the attention-based encoder-decoder of end-to-end model can contribute to handling the variations between different dialects by learning and optimizing a single neural network.

However, Connectionist Temporal Classification (CTC) for end-to-end has its advantage of training simplicity, and is one of the most popular methods used in speech recognition [16,17]. The work in [18] directly incorporated attention modeling within CTC framework to address high word error rates (WER) for a character-based end-to-end. But for Tibetan multi-dialect speech recognition, Tibetan character is a two-dimensional planar character, which is written in Tibetan letters from left to right, but there is a vertical superposition in syllables, so a word-based CTC is more suitable for modelling. In our work, we try to introduce attention mechanism in encoder for CTC-based end-to-end model [19,20]. The attention is used in encoder to capture the context information and relations of discontinuous speech frames among different dialects for distinguishing dialect content and identity.

WaveNet is a deep generative model with very large receptive fields, it can model the long-term dependency on speech data [21]. It has been efficiently applied for speech generation and text-to-speech. A generative model can capture the underlying data distribution as well as the mechanisms to generate data; we believe that such abilities are crucial for shared representation across speech data from different dialects. Further, WaveNet can give the prediction distribution for speech data conditioned on all previous input, so dialect information can be used as additional label outputs during training to improve recognition performance in different dialects. So in this paper, we used WaveNet as encoder for CTC-based end-to-end model and incorporated attention

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

modeling within WaveNet for improving Tibetan multi-dialect multitask speech recognition.

II. Related work

Deep learning has had a huge impact on speech recognition since Hinton et al. first introduced deep neural networks (DNNs) into this field. Researchers have applied various neural network models, such as recurrent neural networks (RNNs), long short-term memory (LSTM), and CNNs, to the field to greatly improve system performance [9,10,12]. However, these models divide the whole speech recognition system into separate components, such as acoustic models and language models; each component is trained separately, and then they are combined to perform recognition tasks. There are some obvious shortcomings to this approach: The creation of language resources is very time-consuming in the pipeline of a speech recognition system, especially for low-resource languages; expert knowledge is required to construct a pronunciation dictionary, which brings implicit difficulty for building a speech recognition system; there is no unified optimization function to optimize the components as a whole.

Based on this, the researchers have proposed direct speech-to-label mapping, namely, an end-to-end model [19, 22-25]. There are two major types of end-to-end architectures for ASR: The first one is based on connectionist temporal classification (CTC) [16], which aims to map the unsegmented data to the modeling unit sequence, sum all the probabilities of the legal sequences, and take the one of largest probability as the output sequence [17,19]. The other architecture is an attention-based encoder–decoder model. Compared with the standard encoder–decoder model, the attention mechanism can assign different weights to each part of input X and extract more related and important information, leading to more accurate predictions without bringing calculation and storage overload [26-28]. In the work [29], attention was directly incorporated in CTC network for small output unit, such as character-based CTC speech recognition. In this paper, we will try to use attention mechanism as an independent component to training a word-based CTC end-to-end model for multi-dialect multitask speech recognition.

In Tibetan speech recognition, researchers have paid more attention to the Lhasa of Ü-Tsang dialect. The recent work in [5] applied 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-Hidden Markov Models. The work in [30] 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 introduced 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 improved the training and decoding efficiency of the model. The application of an attention in CTC-based end-to-end model or hybrid CTC/attention mechanism for Tibetan speech recognition has not been seen yet. As for the speech recognition for other Tibetan dialects, a few related studies only focused on endpoint detection, speech feature extraction, and isolated word recognition [6–8,31] due to the lack of language resources for Kham and Amdo dialects. With regard to the topic of Tibetan dialect speech recognition or dialect identification, to our knowledge, there is almost no relevant research.

Inspired by the idea of the multitask processing mechanism of the brain, many researchers have conducted work on the application of a multitask framework to speech recognition. The work in [32] used the multitask 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 phoneme model of each language. The work in [33] integrated speaker recognition and speech recognition into a multitask learning framework using a recursive structure, which used the output of one task as additional information for another task to supervise individual task learning. In [34], the researchers conducted joint learning of accent recognizer and multi-dialect acoustic models to improve the performance of acoustic models using the same features. All these works demonstrate the effectiveness of multi-task mechanism.

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

III. Data And Methods

A. Data

Our experimental data is from an open and free Tibetan multi-lingual speech data set TIBMD@MUC, which can be downloaded from [35].

The text corpus consists of two parts. One is 1396 spoken language sentences selected from the book "Tibetan Spoken Language" [36] 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 Table Ⅰ, which is about 10 hours data. The training data consists of 4.40 hours Lhasa-Ü-Tsang, 1.21 hours Changdu-Kham, and 3.28 hours Amdo pastoral dialect, and their corresponding texts contain 1205 syllables. We collect 0.49 hours Lhasa-Ü-Tsang, 0.26 hours Changdu-Kham, and 0.37 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.

B. Method

1) WaveNet-CTC

The WaveNet model is composed of stacked dilated causal convolutional layers. The network models the joint probability of a waveform $\mathrm{x} = \{x_1, \cdots, x_T\}$ as a product of conditional probabilities, as shown in Eq. (1).

$$p(x) = \prod_{t=1}^{T} p(x_t | x_1, \cdots, x_{t-1}) \quad (1)$$

It is more efficient than standard causal convolution layers for increasing the receptive field, since the filter is applied over an area larger than its length by skipping input values with a certain step.

TABLE Ⅰ. TIBETAN MULTI-DIALECT EXPERIMENTAL DATA STATISTICS

Dialect	Training data (hours)	Training utterances (#)	Test data (hours)	Test utterances (#)
Lhasa-Ü-Tsang	4.40	6678	0.49	742
Changdu-Kham	1.21	1153	0.26	284
Amdo pastoral dialect	3.28	4649	0.37	516
Total	8.89	12480	1.12	1542

Stacking a few blocks of dilated causal convolutional layers creates a very large receptive field 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 a receptive field of size of 2^{12}.

WaveNet uses the same gated activation unit as the gated PixelCNN [37]. Its activation function is shown in Eq. 2.

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

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

WaveNet uses residual and parameterized skip connections to speed up convergence and enable training of much deeper models. More details on WaveNet can be found in [21].

We adopt the architecture of Speech-to-Text-WaveNet [38] for Tibetan multi-dialect 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 accelerate the process of mapping speech to a text sequence.

2) Attention mechanism

To automatically catch the difference of context information among different dialect speech, attention layer is proposed to create context vector of speech frame as a part of input for WaveNet-CTC. This layer produces an "attention range" when it produces output, which means that the next output should focus on which parts of the input sequence, and then produce the next output according to the region of interest, so to repeat. The schematic diagram of the attention layer is as Fig. 1.

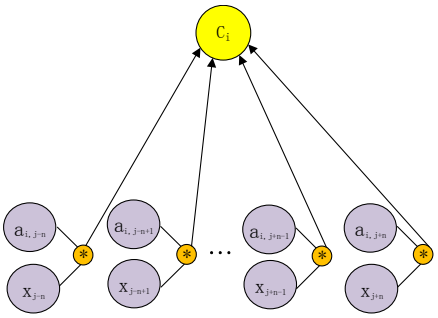

Fig. 1. Attention layer.

After introducing the attention mechanism, the formula for calculating the context vector C_i of the speech frame X_i is as equation (3).

$$C_i = \sum_{j \neq i} \alpha_{i,j} \cdot X_j \quad (3)$$

where $\alpha_{i,j}$ is the attention weight, subject to $\alpha \geq 0$ and $\sum_j \alpha_{i,j} = 1$ through softmax normalization. The $\alpha_{i,j}$ calculation method is as equation (4).

$$\alpha_{i,j} = \frac{\exp\left(Score(X_i, X_j)\right)}{\sum_j \exp\left(Score(X_i, X_j)\right)} \quad (4)$$

It represents the correlation of speech frame pair $(X_i, X_j, j \neq i)$. Considering the computation complex, the attention operates on a sliding window of frames before and after the current frame. $\mathrm{Score}(\cdot)$ is an energy function, which value is computed as equation (5) by the MLP which is jointly trained with all the other component in end-to-end network. Those $X_j, j \neq i$ that have large score have more weights in context vector C_i.

$$\mathrm{Score}(X_i, X_j) = v_a^T \tanh\left(W_a[X_i \oplus X_j]\right) \quad (5)$$

Finally, X_i is concatenated with C_i as the extended speech frame vector X_i'. X_i' is then fed into the WaveNet-CTC as shown as Fig. 2.

We also explore another position of the attention mechanism embedded, which is before the last layer of WaveNet. The architecture is shown in Fig. 3. The corresponding formula for calculating the context vector C_i of the semantic encoding C_i is as equation (6):

$$C_i = \sum_{j \neq i} \alpha_{i,j} \cdot h_j \quad (6)$$

where h_j is the feature vector of the WaveNet output. The $\alpha_{i,j}$ calculation method is as equation (7).

$$\alpha_{i,j} = \frac{\exp\left(Score(h_i, h_j)\right)}{\sum_j \exp\left(Score(h_i, h_j)\right)} \quad (7)$$

$\alpha_{i,j}$ represents the correlation of feature vector pair $(h_i, h_j, j \neq i)$ in WaveNet.

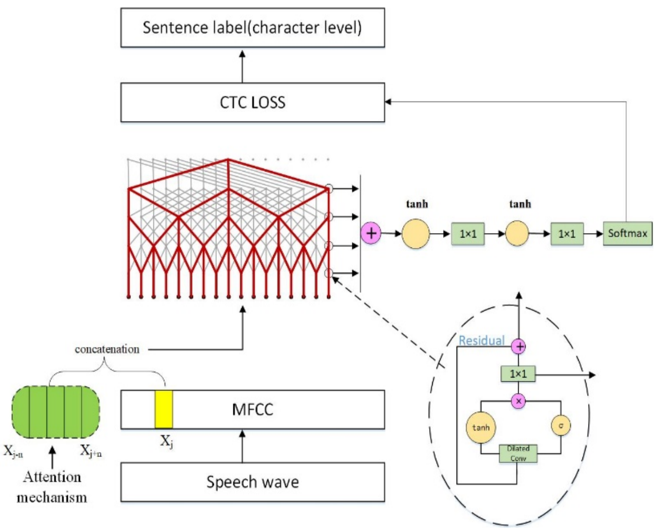

Fig. 2. The architecture of Attention-WaveNet-CTC.

3) Multitask learning

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 as shown as Fig. 4 [31]. 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 modeling unit of the end-to-end model because the output is not a recognized Tibetan character sequence. So, a syllable of Tibetan characters is naturally selected as the CTC modeling unit.

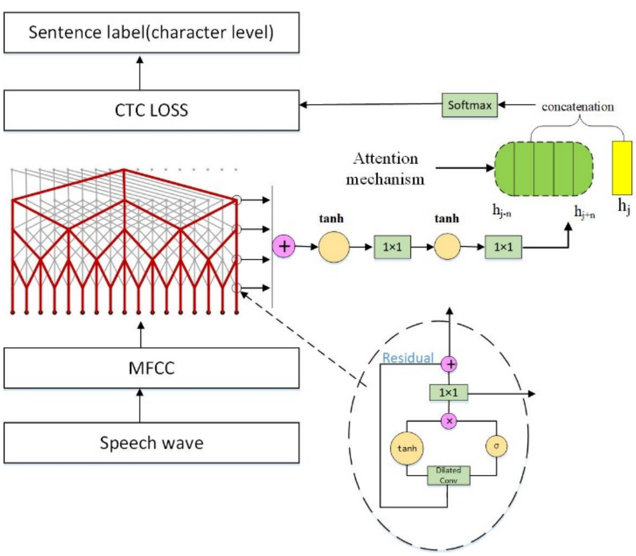

Fig. 3. The architecture of WaveNet-Attention-CTC.

A simple approach to train the end-to-end multitask speech recognition model is to directly expand the Tibetan character sequence with dialect symbols as output targets. We explored the impact of dialect ID on recognition performance at different locations. We evaluated two ways to add the dialect information into the label sequence. One was to add the symbol to the beginning of the target label sequence, like "ID ཐུགས་རྗེ་ཆེ" The other was to add the symbol at the end of the label sequence, like "ཐུགས་རྗེ་ཆེ ID". This kind of end-to-end model based on a whole WaveNet-CTC for Tibetan multitask recognition is shown as Fig. 6.

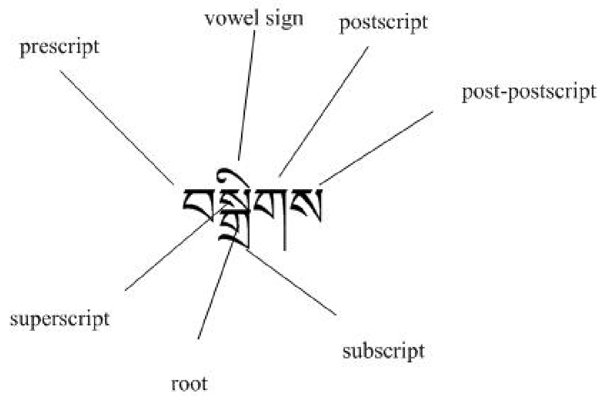

Fig. 4. The structure of a Tibetan syllable.

ང་ལ་སྒོར་བརྒྱད་ཡོད།

Fig. 5. A Tibetan sentence (meaning "I have eight bucks").

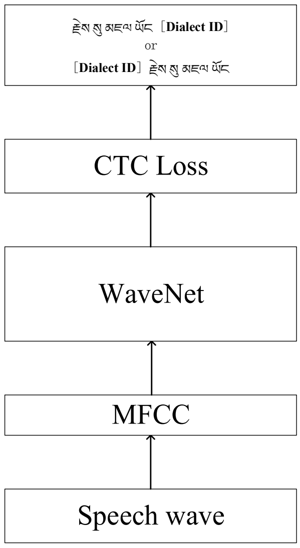

Fig. 6. End-to-end model based on a whole WaveNet-CTC for Tibetan multitasks recognition.

IV. Experiments

The experiments are divided into three parts: single-task experiments, multi-dialect speech content recognition experiments and multiple task experiments. Single-task experiments include speech content recognition and dialect ID recognition. Multi-dialect speech content recognition experiments evaluate the syllable error rate (SER) of WaveNet-CTC model and attention-based WaveNet-CTC model. Multitask experiments evaluate SER and recognition accuracy of dialect ID for the attention-based WaveNet-CTC and WaveNet-CTC.

The WaveNet network consists of 15 layers, which are grouped into 3 dilated residual block stacks of 5 layers, and in each layer, original input was added with the output of a residual block and taken as new input into the next residual block to enhance the data abstraction of different depth levels of the network. 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 learning rate is 2×10^{-4}. The number of hidden units in the residual connection is 128.

We evaluated n=5 frames before and after the current frame to calculate the attention coefficient. Compared with the calculation of the attention coefficient of all frames, the calculation speed has been improved quickly, which is convenient for the training of models.

A. Single-task experiment

We used WaveNet-CTC to train three dialect-specific models for a single task of speech content recognition. The single-task model for dialect identification was trained on LSTM and softmax. The results of the single task are listed in Table Ⅱ and Table Ⅲ.

TABLE Ⅱ. SER (%) OF SINGLE-TASK MODELS BASED ON WAVENET-CTC FOR SPEECH CONTENT RECOGNITION

	Lhasa-Ü-Tsang	**Changdu-Kham**	**Amdo Pastoral**
dialect-specific	28.83	62.56	17.60

TABLE Ⅲ. RECOGNITION ACCURACY (%) OF SINGLE TASK FOR DIALECT ID

	Lhasa-Ü-Tsang	**Changdu-Kham**	**Amdo Pastoral**
dialect ID	97.88	92.24	97.09

B. Multi-dialect speech content recognition experiments

We train three models on joint speech data of three dialects for multi-dialect speech content recognition, which are WaveNet-CTC, Attention-WaveNet-CTC and WaveNet-Attention-CTC. The results are shown in Table Ⅳ. First, it can be seen in Table Ⅳ that the performance of both WaveNet-CTC model and the WaveNet-CTC with the attention are worse than that of the specific dialect model for speech content recognition.

TABLE Ⅳ. SER (%) OF MULTI-DIALECT MODELS FOR SPEECH CONTENT RECOGNITION

	Lhasa-Ü-Tsang	**Changdu-Kham**	**Amdo Pastoral**
dialect-specific model	**28.83**	**62.56**	**17.60**
WaveNet-CTC	29.55	62.83	33.52
Attention-WaveNet-CTC	55.22	63.50	39.52
WaveNet-Attention-CTC	38.80	64.97	36.24

C. Multitask experiments

For multitask recognition, we conducted the experiments based on different architectures: WaveNet-CTC, Attention-WaveNet-CTC, WaveNet-Attention-CTC. And the performances of the dialect ID at the beginning and the end of output sequence were evaluated respectively. The results are tabulated in Table Ⅴ and Table Ⅵ. From these two tables, we can see that the dialectID-speech model in WaveNet-Attention-CTC framework performed best, which reduced the SER by 7.39% and 2.4% over the dialect-specific model for Lhasa-Ü-Tsang and Changdu-Kham, and gets the close SER with dialect-specific model for Amdo Pastoral, and got the highest accuracy of dialect ID recognition for Lhasa-Ü-Tsang and Changdu-Kham. It also shows that speech content recognition is sensitive to the recognition of dialect ID in attention-based WaveNet-CTC model.

TABLE Ⅴ. SER (%) OF TWO-TASK MODELS FOR SPEECH CONTENT RECOGNITION

Architecture	**Model**	**Lhasa-Ü-Tsang**	**Changdu-Kham**	**Amdo Pastoral**
dialect-specific model		28.83	62.56	**17.6**
WaveNet-CTC without dialect ID		29.55	62.83	33.52
WaveNet-CTC with dialect ID	dialectID-speech	32.84	68.58	33
	speech-dialectID	26.8	64.03	30.79
Attention-WaveNet-CTC	dialectID-speech	52.19	65.24	50.22
	speech-dialectID	55.16	67.78	55.23
WaveNet-Attention-CTC	dialectID-speech	**21.44**	**60.16**	20.46
	speech-dialectID	23.79	62.96	24.15

TABLE Ⅵ. DIALECT ID RECOGNITION ACCURACY (%) OF TWO-TASK MODELS

Architecture	**Model**	**Lhasa-Ü-Tsang**	**Changdu-Kham**	**Amdo Pastoral**
DialectID model		97.88	92.24	97.09
WaveNet-CTC with dialect ID	dialectID-speech	98.57	95.23	**99.6**
	speech-dialectID	99.01	97.61	99.41
Attention-WaveNet-CTC	dialectID-speech	**100**	89.28	84.52
	speech-dialectID	**100**	79.76	90.31
WaveNet-Attention-CTC	dialectID-speech	**100**	**98.8**	99.41
	speech-dialectID	**100**	94.04	98.06

V. CONCLUSIONS

This paper proposes an attention-based multitask recognition mechanism, integrating speech recognition and dialect identification into a unified neural network for Tibetan multi-dialect multitask scenarios. The experimental results show that the attention mechanism embedded into WaveNet can capture the context information and relations of discontinuous speech frames for multitask learning.

ACKNOWLEDGMENT

This work is supported by national natural science foundation of china (61976236), and MUC 111 project.

REFERENCES

[1] Y.Zhang, "Research on Tibetan Lhasa dialect speech recognition based on deep learning," Master, Northwest Normal University, 2016.

[2] S. Yuan, W. Guo, and L. Dai, "Speech Recognition Based on Deep Neural Networks on Tibetan Corpus," Pattern Recognition and Artificial Intelligence, vol. 28, no. 03, pp. 209–213, 2015.

[3] C. Pei, "Research on Tibetan Speech Recognition Technology Based on Standard Lhasa," Master, Tibet University, 2009.

[4] L. I. Guan-Yu and M. Meng, “Research on Acoustic Model of Large-vocabulary Continuous Speech Recognition for Lhasa Tibetan,” Computer Engineering, vol. 38, no. 5, pp. 189–191, Mar. 2012.

[5] Q. Wang, W. Guo, and C. Xie, “Towards End to End Speech Recognition System for Tibetan,” Pattern Recognition and Artificial Intelligence, vol. 30, no. 4, pp. 359–364, 2017.

[6] L. Cai and C. Zhao, “Method and Implementation of Endpoint Detection in Ando Tibetan Language,” Gansu Science and Technology, vol. 24, no. 05, pp. 46–47, 2008.

[7] L. Cai, “Study of Methods of Speech Features Extraction of Ando Tibetan,” Master, Qinghai Normal University, Xining, 2009.

[8] Q. Han and H. Yu, “Research on Speech Recognition for Ando Tibetan Based on HMM,” Software Guide, vol. 09, no. 7, pp. 173–175, 2010.

[9] A. Graves, A.-r. Mohamed, G. Hinton, “Speech Recognition with Deep Recurrent Neural Networks,” arXiv:1303.5778 [cs].

[10] J. T. Geiger, Z. Zhang, F. Weninger, B. Schuller, G. Rigoll, “Robust Speech Recognition using Long Short-Term Memory Recurrent Neural Networks for Hybrid Acoustic Modelling,” in INTERSPEECH 2014,15th Annual Conference of the International Speech Communication Association, Singapore, pp. 631–635,2014.

[11] X. Chen, X. Liu, Y. Wang, M. J. F. Gales, P. C. Woodland, “Efficient Training and Evaluation of Recurrent Neural Network Language Models for Automatic Speech Recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.24, no.11, pp. 2146–2157, 2016.

[12] Y. Qian, M. Bi, T. Tan, K. Yu, “Very Deep Convolutional Neural Networks for Noise Robust Speech Recognition,” IEEE/ACM Trans. Audio, Speech & Language Processing, vol.24, no.12, pp. 2263–2276, 2016.

[13] A. Graves, N. Jaitly, A. Mohamed, “Hybrid speech recognition with Deep Bidirectional LSTM,” in 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 273–278, 2013.

[14] B. Li et al., “Multi-Dialect Speech Recognition with a Single Sequence-to-Sequence Model.,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, pp. 4749–4753, 2018.

[15] S. Toshniwal et al., “Multilingual Speech Recognition with a Single End-to-End Model,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, pp. 4904–4908, April 15-20, 2018.

[16] A. Graves, S. Fernández, F. J. Gomez, J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, pp.369–376, 2006.

[17] O. Siohan, “CTC Training of Multi-Phone Acoustic Models for Speech Recognition.,” in Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, pp.709–713,2017.

[18] S. Watanabe, T. Hori, and J. R. Hershey, “Language independent end-to-end architecture for joint language identification and speech recognition,” in 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 265–271, 2017.

[19] T. Hori, S. Watanabe, J. Hershey, “Joint CTC/attention decoding for end-to-end speech recognition,” presented at Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Vancouver, Canada, 2017.

[20] S. Kim, T. Hori, S. Watanabe, “Joint CTC-attention based end-to-end speech recognition using multi-task learning.,” presented at 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, 2017.

[21] A. van den Oord et al., “WaveNet: A Generative Model for Raw Audio,” CoRR, vol. abs/1609.03499, 2016.

[22] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case et al., “Deep Speech 2: End-to-End Speech Recognition in English and Mandarin,” in International Conference on Machine Learning, New York City, NY, USA, pp. 173–182, 2016.

[23] Y. Miao, M. Gowayyed, F. Metze, “EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding,” arXiv:1507.08240 [cs].

[24] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp. 4945–4949, 2016.

[25] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “Joint CTC-attention based end-to-end speech recognition using multi-task learning,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, pp. 4835–4839, 2017.

[26] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-Based Models for Speech Recognition. arXiv:1506.07503 [cs, stat].

[27] A. Zeyer, K. Irie, R. Schluter, and H. Ney, “Improved Training of End-to-end Attention Models for Speech Recognition,” in Interspeech 2018, Hyderabad, India, pp. 7–11, 2018.

[28] C. Weng, J. Cui, G. Wang, J. Wang, C. Yu, D. Su, and D. Yu, “Improving Attention Based Sequence-to-Sequence Models for End-to-End English Conversational Speech Recognition,” in Interspeech 2018, Hyderabad, India, pp. 761–765, 2018.

[29] X. Huang and J. Li, “The Acoustic Model for Tibetan Speech Recognition Based on Recurrent Neural Network,” Journal of Chinese Information Processing, vol.32, No.5, pp.49–55, 2018.

[30] J. Li, H. Wang, L. Bi, J. Dang, K. Khuru, and G. Lobsang, “Exploring tonal information for Lhasa dialect acoustic modeling,” in 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), TianJin, China, pp. 1–5, 2016.

[31] G. Li, H. Yu, T. F. Zheng, J. Yan, S. Xu, “Free linguistic and speech resources for Tibetan,” in 2017 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, Kuala Lumpur, Malaysia, pp.733–736, 2017.

[32] D. Chen, and B. Mak, “Multitask Learning of Deep Neural Networks for Low-resource Speech Recognition,” Trans. Audio, Speech and Lang. Proc., vol.23, No.7, pp.1172–1183, 2015.

[33] Z. Tang, L. Li, and D. Wang, “Multi-task Recurrent Model for Speech and Speaker Recognition,” arXiv:1603.09643 [cs, stat].

[34] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawajohnson, “Joint Modeling of Accents and Acoustics for Multi-Accent Speech Recognition,” ArXiv e-prints, 2018.

[35] https://pan.baidu.com/s/14CihgqjA4AFFH1QpSTjzZw.

[36] Bazelen, L., “Tibetan spoken language,” Minzu press, Beijing, 2005.

[37] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel Recurrent Neural Networks,” arXiv:1601.06759 [cs].

[38] http://github.com/CynthiaSuwi/Wavenet-demo.