Hindawi
Complexity
Volume 2020, Article ID 8894566, 10 pages
https://doi.org/10.1155/2020/8894566

Research Article

Multitask Learning with Local Attention for Tibetan Speech Recognition

Hui Wang, Fei Gao, Yue Zhao, Li Yang, Jianjian Yue, and Huilin Ma

School of Information Engineering, Minzu University of China, Beijing 100081, China

Correspondence should be addressed to Fei Gao; 18301390@muc.edu.cn

Received 22 September 2020; Revised 12 November 2020; Accepted 26 November 2020; Published 18 December 2020

Academic Editor: Ning Cai

In this paper, we propose to incorporate the local attention in WaveNet-CTC to improve the performance of Tibetan speech recognition in multitask learning. With an increase in task number, such as simultaneous Tibetan speech content recognition, dialect identification, and speaker recognition, the accuracy rate of a single WaveNet-CTC decreases on speech recognition. Inspired by the attention mechanism, we introduce the local attention to automatically tune the weights of feature frames in a window and pay different attention on context information for multitask learning. The experimental results show that our method improves the accuracies of speech recognition for all Tibetan dialects in three-task learning, compared with the baseline model. Furthermore, our method significantly improves the accuracy for low-resource dialect by 5.11% against the specific-dialect model.

1. Introduction

Multitask learning has been applied successfully for speech recognition to improve the generalization performance of the model on the original task by sharing the information between related tasks [1–9]. Chen and Mak [6] 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. Krishna et al. [7] proposed a hierarchical multitask model, and the performance differences between high-resource language and low-resource language were compared. Li et al. [8] and Toshniwal et al. [9] introduced additional information of language ID to improve the performance of end-to-end multidialect speech recognition systems.

Tibetan is one of minority languages in China. It has three major dialects in China, i.e., Ü-Tsang, Kham, and Amdo. There are also several local subdialects in each dialect. Tibetan dialects pronounce very differently, but the written characters are unified across dialects. In our previous work [10], Tibetan multidialect multitask speech recognition was conducted based on the WaveNet-CTC, which performed simultaneous Tibetan multidialect speech content recognition, dialect identification, and speaker recognition in a single model. WaveNet is a deep generative model with very large receptive fields, and it can model the long-term dependency of speech data. It is very effective to learn the shared representation from speech data of different tasks. Thus, WaveNet-CTC was trained on three Tibetan dialect data sets and learned the shared representations and model parameters for speech recognition, speaker identification, and dialect recognition. Since the Lhasa of Ü-Tsang dialect is a standard Tibetan speech, there are more corpora available for training than Changdu-Kham and Amdo pastoral dialect. Although two-task WaveNet-CTC improved the performance on speech recognition for Lhasa of Ü-Tsang dialect and Changdu-Kham dialect, the three-task model did not improve performance for all dialects. With an increase in task number, the speech recognition performance degraded.

To obtain a better performance, attention mechanism is introduced into WaveNet-CTC for multitask learning in this paper. Attention mechanism can learn to set larger weight to more relevant frames at each time step. Considering the computation complexity, we conduct a local attention using a sliding window on the whole of speech feature frames to create the weighted context vectors for different recognition tasks. Moreover, we explore to place a local attention at the different positions within WaveNet, i.e., in the input layer and high layer, respectively.

The contribution of this work is three-fold. For one, we propose the WaveNet-CTC with local attention to perform multitask learning for Tibetan speech recognition, which can automatically capture the context information among different tasks. This model improves the performance of the Tibetan multidialect speech recognition task. Moreover, we compared the performance of local attention inserted at different positions in the multitask model. The attention component embedded in the high layer of WaveNet obtains better performance than the one in the input layer of WaveNet for speech recognition. Finally, we conduct a sliding window on the speech frames for efficiently computing the local attention.

The rest of this paper is organized as follows: Section 2 introduces the related work. Section 3 presents our method and gives the description of the baseline model, local attention mechanism, and the WaveNet-CTC with local attention. In Section 4, the Tibetan multidialect data set and experiments are explained in detail. Section 5 describes our conclusions.

2. Related Work

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. Das et al. [11] directly incorporated attention modelling within the CTC framework to address high word error rates (WERs) for a character-based end-to-end model. But, in Tibetan speech recognition scenarios, the Tibetan character is a two-dimensional planar character, which is written in Tibetan letters from left to right, besides there is a vertical superposition in syllables, so a word-based CTC is more suitable for the end-to-end model. In our work, we try to introduce attention mechanism in WaveNet as an encoder for the CTC-based end-to-end model. The attention is used in WaveNet to capture the context information among different tasks for distinguishing dialect content, dialect identity, and speakers.

In multitask settings, there are some recent works focusing on incorporating attention mechanism in multitask training. Zhang et al. [12] proposed an attention mechanism for the hybrid acoustic modelling framework based on LSTM, which weighted different speech frames in the input layer and automatically tuned its attention to the spliced context input. The experimental results showed that attention mechanism improved the ability to model speech. Liu et al. [13] incorporated the attention mechanism in multitask learning for computer vision tasks, in which the multitask attention network consisted of a shared network and task-specific soft-attention modules to learn the task-specific features from the global pool, whilst simultaneously allowing for features to be shared across different tasks. Zhang et al. [14] proposed an attention layer on the top of the layers for each task in the end-to-end multitask framework to relieve the overfitting problem in speech emotion recognition. Different from the works of Liu et al. and Zhang et al. [13, 14], which distributed many attention modules in the network, our method merely uses one sliding attention window in the multitask network and has its advantage of training simplicity.

3. Methods

3.1. Baseline Model. We take the Tibetan multitask learning model in our previous work [10] as the baseline model as shown in Figure 1, which was initially proposed for Chinese and Korean speech recognition from the work of Xu [15] and Kim and Park [16]. The work [10] integrates WaveNet [17] with CTC loss [18] to realize Tibetan multidialect end-to-end speech recognition.

WaveNet contains the stacks of dilated causal convolutional layers as shown in Figure 2. In the baseline model, the WaveNet network consists of 15 layers, which are grouped into 3 dilated residual blocks of 5 layers. In every stack, the dilation rate increases by a factor of 2 in every layer. The filter length of causal dilated convolutions is 2. According to equations (1) and (2), the respective field of WaveNet is 46:

$$\text{Receptive\_field}_{\text{block}} = \sum_{i=1}^{n} \left(\text{Filter}_{\text{length}} - 1\right) \times \text{Dilation}_{\text{rate}_i} + 1. \tag{1}$$

$$\text{Receptive field}_{\text{stacks}} = S \times \text{Receptive field}_{\text{block}} - S + 1. \tag{2}$$

In equations (1) and (2), S refers to the number of stacks, $\text{Receptive\_field}_{\text{block}}$ refers to the receptive field of a stack of dilated CNN, $\text{Receptive field}_{\text{stacks}}$ refers to the receptive field of some stacks of dilated CNN, and $\text{Dilation}_{\text{rate}_i}$ refers to the dilation rate of the i-th layer in a block.

WaveNet also uses residual and parameterized skip connections [19] to speed up convergence and enable training of much deeper models. More details about WaveNet can be found in [17].

Connectionist temporal classification (CTC) is an algorithm that trains a deep neural network [20] for the end-to-end learning task. It can make the sequence label predictions at any point in the input sequence [18]. In the baseline model, since the Tibetan character is a two-dimensional planar character as shown in Figure 3, the CTC modeling unit for Tibetan speech recognition is Tibetan single syllable, otherwise a Tibetan letter sequence from left to right is unreadable.

3.2. Local Attention Mechanism. Since the effect of each speech feature frame is different for the target label output at current time, considering the computational complexity, we introduce the local attention [21] into WaveNet to create a weighted context vector for each time i. The local attention places a sliding window with the length $2n$ centered around the current speech feature frame on the input layer and before the softmax layer in WaveNet, respectively, and repeatedly produces a context vector C_i for the current input (or hidden) feature frame $x(h)_i$. The formula for C_i is shown

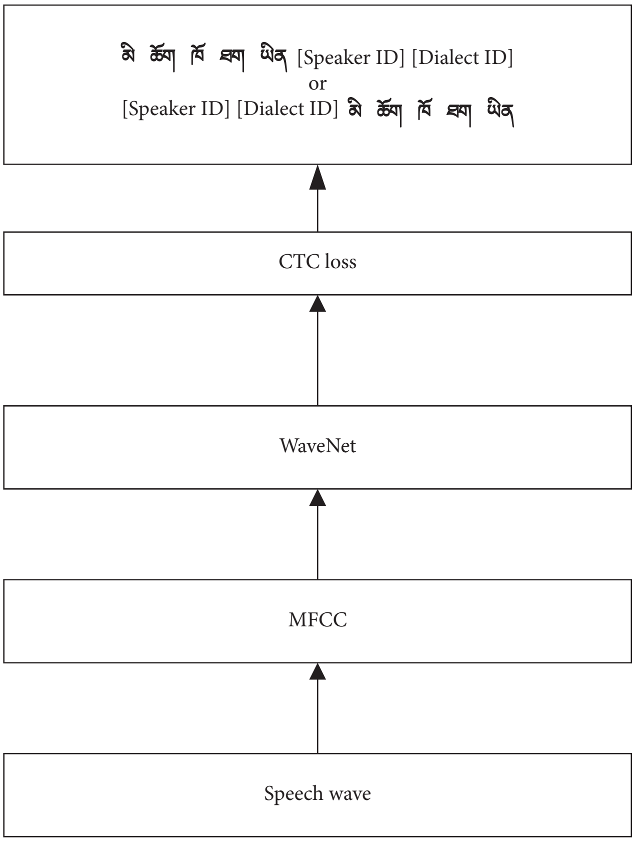
མི ཚོགས ཁོ ཐག ཡིན [Speaker ID] [Dialect ID]
or
[Speaker ID] [Dialect ID] མི ཚོགས ཁོ ཐག ཡིན
CTC loss
WaveNet
MFCC
Speech wave

Figure 1: The baseline model.

in equation (3), and the schematic diagram is shown in Figure 4:

$$C_i = \sum_{j=i-n, j\neq i}^{i+n} \alpha_{i,j} \cdot x(h)_j, \qquad (3)$$

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 follows:

$$\alpha_{i,j} = \frac{\exp\left(\text{Score}\left(x(h)_i, x(h)_j\right)\right)}{\sum_j \exp\left(\text{Score}\left(x(h)_i, x(h)_j\right)\right)}. \qquad (4)$$

It captures the correlation of speech frame pair $(x(h)_i, x(h)_j, j \neq i)$. The attention operates on n frames before and after the current frame. Score (.) is an energy function, whose value is computed as equation (5) by the MLP which is jointly trained with all the other components in an end-to-end network. Those $x(h)_j, j \neq i$ that get larger scores would have more weights in context vector C_i.

$$\text{Score}\left(x_i, x_j\right) = v_a^T \tanh\left(W_a\left[x(h)_i; x(h)_j\right]\right). \qquad (5)$$

Finally, $x(h)_i$ is concatenated with C_i as the extended feature frame and fed into the next layer of WaveNet as shown in Figures 5 and 6. The attention module is inserted in the input layer in Figure 5 referred as Attention-WaveNet-CTC. The attention module is embedded before the softmax layer in Figure 6 referred as WaveNet-Attention-CTC.

4. Experiments

4.1. Data. Our experimental data are from an open and free Tibetan multidialect speech data set TIBMD@MUC [10], in which the text corpus consists of two parts: one is 1396 spoken language sentences selected from the book "Tibetan Spoken Language" [22] written by La Bazelen and the other part contains 8,000 sentences from online news, electronic novels, and poetry of Tibetan on internet. All text corpora in TIBMD@MUC include a total of 3497 Tibetan syllables.

There are 40 recorders who are 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, 16 bit quantization accuracy, and wav format.

Our experimental data for multitask speech recognition are shown in Table 1, which consists of 4.4 hours Lhasa-Ü-Tsang, 1.90 hours Changdu-Kham, and 3.28 hours Amdo pastoral dialect, and their corresponding texts contain 1205 syllables for training. We collect 0.49 hours Lhasa-Ü-Tsang, 0.19 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 128 ms window with 96 ms overlaps.

The experiments are divided into two parts: two-task experiments and three-task experiments. Three dialect-specific models and a multi-dialect model without attention are trained on WaveNet-CTC.

In WaveNet, 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.

4.2. Two-task Experiment. For two-task joint recognition, the performances of the dialect ID or speaker ID at the beginning and at the end of output sequence were evaluated, respectively. We set $n = 5$ frames before and after the current frame to calculate the attention coefficients for attention-based WaveNet-CTC, which are referred to as Attention (5)-WaveNet-CTC and WaveNet-Attention (5)-CTC, respectively, for the two architectures in Figures 5 and 6. Compared with the calculation of the attention coefficient of all frames, the calculation speed of local attention has been improved quickly, which is convenient for the training of models.

The speech recognition result is summarized in Table 2. The best model is the proposed WaveNet-Attention-CTC with the attention embedded before the softmax layer in WaveNet and dialect ID at the beginning of label sequence. It outperforms the dialect-specific model by 7.39% and 2.4%, respectively, for Lhasa-Ü-Tsang and Changdu-Kham and gets the SER close to the dialect-specific model for Amdo Pastoral, which has the highest ARSER (average relative syllable error rate) for three dialects. The model of dialectID-speech (D-S) in the framework of WaveNet-Attention-CTC is effective to improve multilinguistic speech content recognition. Speech content recognition is more sensitive to the

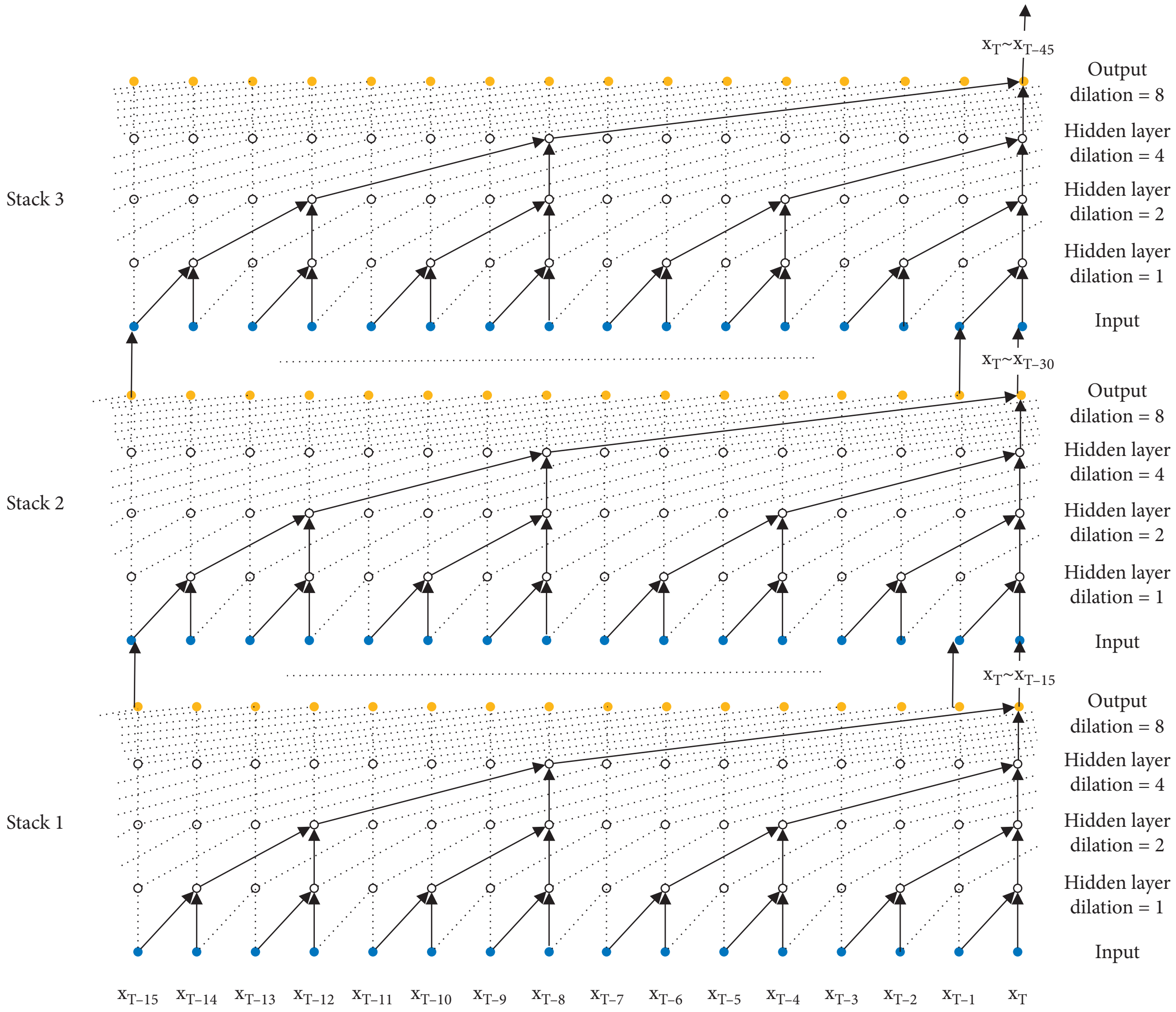
$x_T \sim x_{T-45}$
Output dilation = 8
Hidden layer dilation = 4
Hidden layer dilation = 2
Hidden layer dilation = 1
Input
Stack 3
$x_T \sim x_{T-30}$
Output dilation = 8
Hidden layer dilation = 4
Hidden layer dilation = 2
Hidden layer dilation = 1
Input
Stack 2
$x_T \sim x_{T-15}$
Output dilation = 8
Hidden layer dilation = 4
Hidden layer dilation = 2
Hidden layer dilation = 1
Input
Stack 1
x_{T-15} x_{T-14} x_{T-13} x_{T-12} x_{T-11} x_{T-10} x_{T-9} x_{T-8} x_{T-7} x_{T-6} x_{T-5} x_{T-4} x_{T-3} x_{T-2} x_{T-1} x_T

Figure 2: 3 stacks of 5 dilated causal convolutional layers with filter length 2.

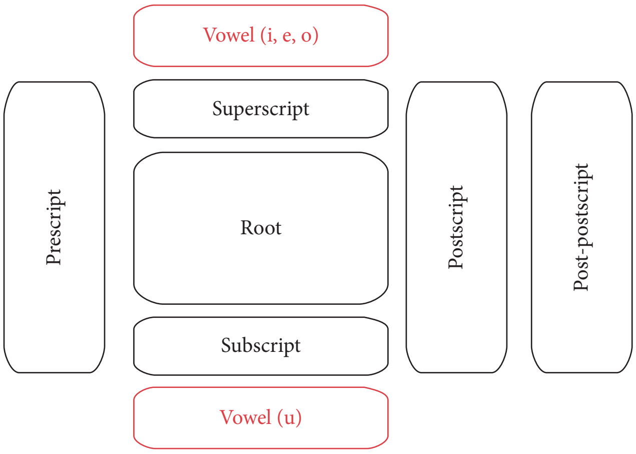

Figure 3: The structure of a Tibetan syllable.

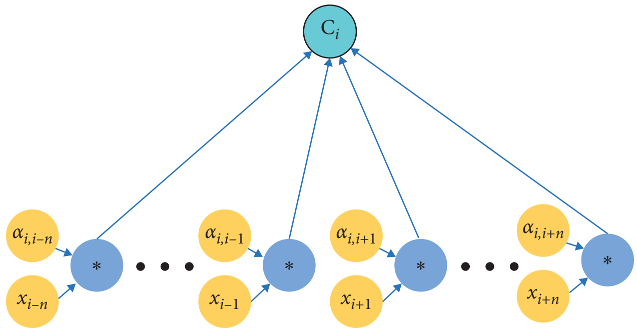

FIGURE 4: Local attention.

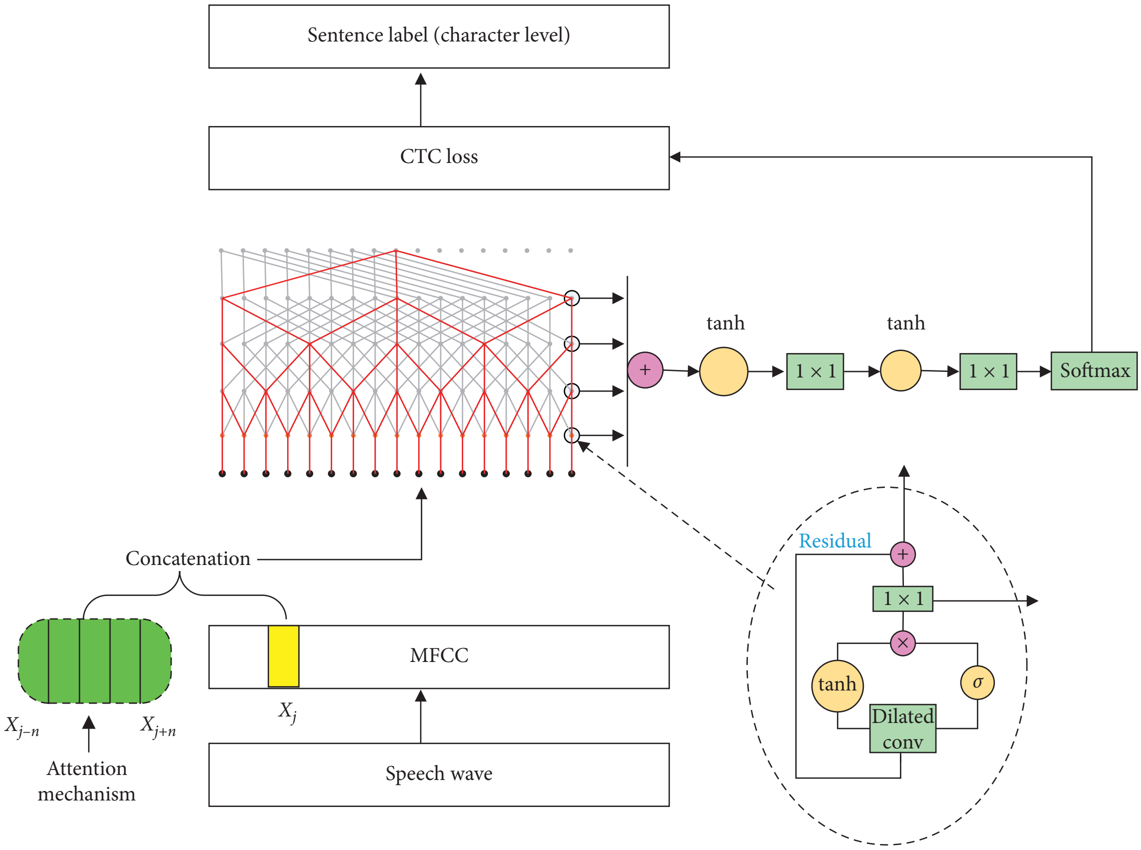

FIGURE 5: The architecture of attention-WaveNet-CTC.

recognition of dialect ID than speaker ID. The recognition of dialect ID helps to identify the speech content. However, the attention inserted before the input layer in WaveNet resulted in the worst recognition, which shows that raw speech feature cannot provide much information to distinguish the multitask.

For dialect ID recognition, in Table 3, we can see that the model with attention mechanism added before the softmax layer performs better than which is added in the input layer, and the dialect ID at the beginning is better than that at the end. From Table 2 and Table 3, it can be seen that the dialect ID recognition influences the speech content recognition.

We also test the speaker ID recognition accuracy for the two-task models. Results are listed in Table 4. It is worth noting that the Attention-WaveNet-CTC model performs poorly on both tasks of the speaker and speech content recognition. Especially in the speaker identification task, the recognition rate of the speakerID-speech model in all three dialects is very poor. Among the Attention-WaveNet-CTC models, it can be seen that the modelling ability of two models of the dialectID-speech and speakerID-speech model shows big gap, which means the Attention-WaveNet-CTC architecture cannot learn effectively the correlation among multiple frames of acoustic feature for multiple classification tasks. In contrast, the WaveNet-Attention-CTC model has a

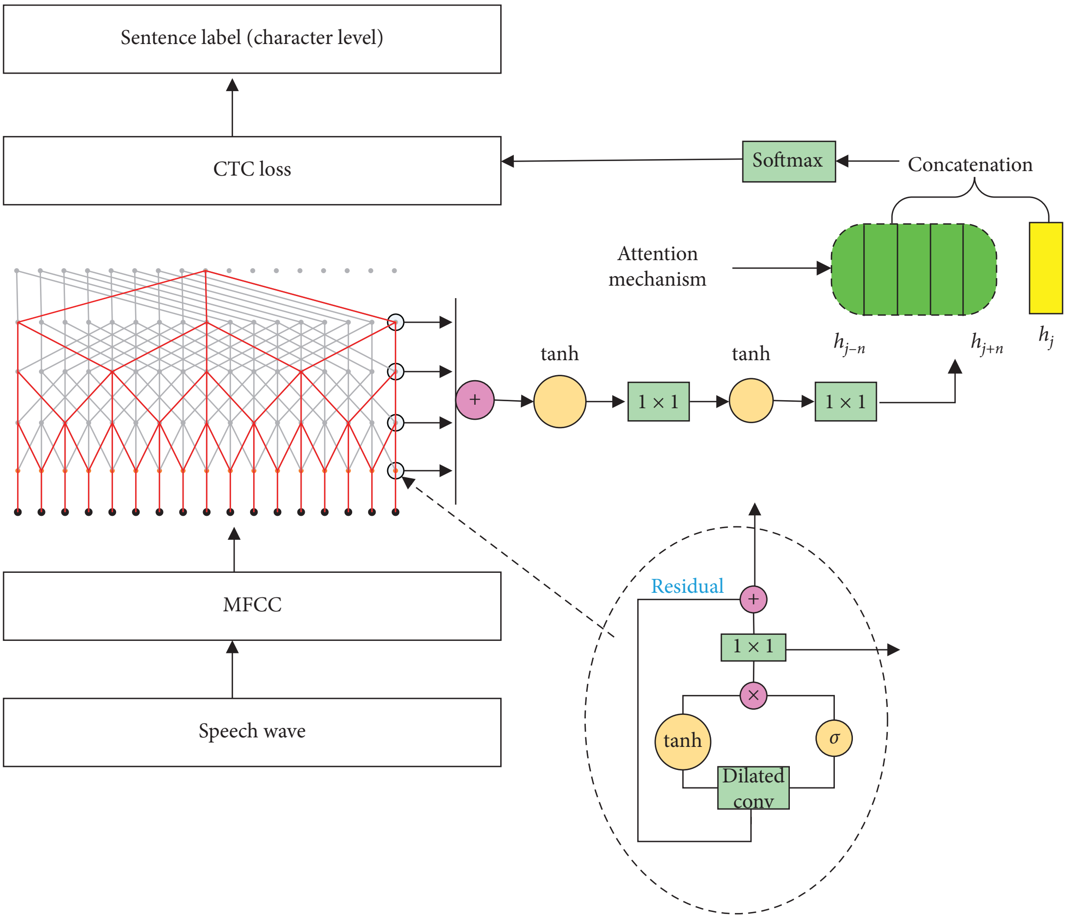

Figure 6: The architecture of WaveNet-attention-CTC.

Table 1: The experimental data statistics.

Dialect	Training data (hours)	Training utterances	Test data (hours)	Test utterances	Speaker
Lhasa-Ü-Tsang	4.40	6678	0.49	742	20
Changdu-Kham	1.90	3004	0.19	336	6
Amdo pastoral	3.28	4649	0.37	516	14
Total	9.58	14331	1.05	2110	40

much better performance on the two tasks. The attention embedded before the softmax layer can find the related and important frames to lead to high recognition accuracy.

4.3. Three-task Experiment. We compared the performances of two architectures, namely, Attention-WaveNet-CTC and WaveNet-Attention-CTC on three-task learning with the dialect-specific model and WaveNet-CTC, where we evaluated $n = 5$, $n = 7$, and $n = 10$, respectively, for the attention mechanism. The results are shown in Table 5.

We can see that the three-task models have worse performance compared with the two-task model, and WaveNet-Attention-CTC has lower SERs for Lhasa-Ü-Tsang and Amdo Pastoral against the dialect-specific model, but for Changdu-Kham, a relative low-resource Tibetan dialect, the model of dialectID-speech-speakerID (D-S-S2) based on the framework of WaveNet-Attention (10)-CTC achieved the highest recognition rate in all models, which outperforms the dialect-specific model by 5.11%. We analyzed the reason that maybe is the reduction of generalization error of the multitask model with the number of learning tasks increasing. It improves the recognition rate for small-data dialect, however not for big-data dialects. Since ASER reflects the generalization error of the model, D-S-S2 of WaveNet-Attention (10)-CTC has highest ASER in all models, which shows it has better generalization capacity. Meanwhile, WaveNet-Attention (10)-CTC achieved the better performance than WaveNet-Attention (5)-CTC and WaveNet-Attention (7)-CTC for speech content recognition as shown in Figure 7, where the syllable error rates declined with the number of n increasing for three dialects, and Changdu-Kham's SER has a quickest descent. We can conclude that attention mechanism needs a longer range to distinguish more tasks, and it pays more attention on the

Table 2: Syllable error rate (%) of two-task models on speech content recognition.

Architecture	Model	Lhasa-Ü-Tsang		Changdu-Kham		Amdo Pastoral		
		SER[1]	RSER[2]	SER	RSER	SER	RSER	ASER[3]
Dialect-specific model		28.83		62.56		**17.6**		
WaveNet-CTC		29.55	−0.72	62.83	−0.27	33.52	−15.92	−5.63
WaveNet-CTC with dialect ID or speaker ID (baseline model)	D-S[4]	32.84	−4.01	68.58	−6.02	33.00	−15.40	−8.48
	S-D[5]	26.80	2.03	64.03	−1.47	30.79	−13.09	−4.21
	S-S1[6]	27.21	1.62	64.17	−1.61	29.68	−12.08	−4.02
	S-S2[7]	28.13	0.7	62.43	0.13	28.04	-10.44	−3.20
Attention (5)-WaveNet-CTC	D-S	52.19	−23.36	65.24	−2.68	50.22	-32.62	−19.55
	S-D	55.16	−26.33	67.78	−5.22	55.23	-37.63	−23.06
	S-S1	77.42	−48.59	85.44	−22.88	82.08	-64.48	−45.32
	S-S2	83.32	−54.49	89.15	−26.94	81.47	-63.87	−48.43
WaveNet-Attention (5)-CTC	D-S	**21.44**	**7.39**	**60.16**	**2.40**	20.46	**−2.86**	**2.31**
	S-D	23.79	5.04	62.96	−0.4	24.15	−6.55	−0.64
	S-S1	34.86	−6.03	63.36	−0.8	40.10	−22.50	−9.78
	S-S2	34.83	−6.00	62.70	−0.14	37.63	−20.03	−8.72

[1]SER: syllable error rate, [2]RSER: relative syllable error rate, [3]ARSER: average relative syllable error rate, [4]D-S: the model trained using the transcription with dialect ID at the beginning of target label sequence, like “A བ྄གས ར྄ ` ཆ ` ,” [5]S-D: the model trained using the transcription with dialect ID at the end of target label sequence, [6]S-S1: the model trained using the transcription with speaker ID at the beginning of target label sequence, and [7]S-S2: the model trained using the transcription with speaker ID at the end of target label sequence.

Table 3: Dialect ID recognition accuracy (%) of two-task models.

Architecture	Model	Lhasa-Ü-Tsang	Changdu-Kham	Amdo Pastoral
DialectID model		97.88	92.24	97.9
WaveNet-CTC with dialect ID	D-S	98.57	95.23	**99.6**
	S-D	99.01	97.61	99.41
Attention (5)-WaveNet-CTC	D-S	**100**	89.28	94.52
	S-D	0	0	0
WaveNet-Attention (5)-CTC	D-S	**100**	**98.8**	99.41
	S-D	**100**	94.04	98.06

Table 4: Speaker ID recognition accuracy (%) of two-task models.

Architecture	Model	Lhasa-Ü-Tsang	Changdu-Kham	Amdo Pastoral
SpeakerID model		67.75	93.13	95.31
WaveNet-CTC with speaker ID	S-S1	68.32	92.85	**97.48**
	S-S2	**71.15**	95.23	96.12
Attention (5)-WaveNet-CTC	S-S1	0	0	0
	S-S2	60.64	77.38	85.85
WaveNet-Attention (5)-CTC	S-S1	70.35	92.85	**97.48**
	S-S2	69.40	**100**	96.70

low-resource task. It is also observed that WaveNet-Attention (5)-CTC has better performance than Attention (5)-WaveNet-CTC, which demonstrates again that the attention mechanism placed in the high layer can find the related and important information which leads to more accurate speech recognition than when it is put in the input layer.

From Tables 6 and 7, we can observe that models with attention have worse performance than the ones without attention for dialect ID recognition and speaker ID recognition, and longer attention achieved the worse recognition for the language with large data. It also shows that in the case of more tasks, the attention mechanism tends towards the low-resource task, such as speech content recognition.

In summary, combining the results of the above experiments, whether two task or three task, the multitask model can make a significant improvement on the performance of the low-resource task by incorporating the attention mechanism, especially when the attention is applied to the high-level abstract features. The attention-based multitask model can achieve the improvements on speech recognition for all dialects compared with the baseline model. With an increase in the task number, the multitask model needs to increase the range for attention to distinguish multiple dialects.

Table 5: Syllable error rate (%) of three-task models on speech content recognition.

Architecture	Model	Lhasa-Ü-Tsang		Changdu-Kham		Amdo Pastoral		
		SER	RSER	SER	RSER	SER	RSER	ASER
Dialect-specific model		**28.83**		62.56		**17.60**		
WaveNet-CTC with dialect ID and speaker ID (baseline model)	S-D-S	30.64	−1.81	64.17	−1.61	34.06	−16.46	−6.62
	D-S-S1	39.64	−10.81	65.10	−2.54	45.15	−27.55	−13.63
	D-S-S2	33.43	−4.60	64.83	−2.27	37.56	−19.96	−8.94
Attention (5)-WaveNet-CTC	S-D-S	48.69	−19.86	68.31	−5.75	63.22	−45.62	−23.74
	D-S-S1	52.57	−23.74	69.38	−6.82	71.42	−53.82	−28.13
	D-S-S2	49.10	−20.27	79.41	−16.85	61.09	−43.49	−26.87
WaveNet-Attention (5)-CTC	S-D-S	30.75	−1.92	69.51	−6.95	34.21	−16.61	−8.49
	D-S-S1	33.17	−4.34	69.51	-6.95	38.49	−20.89	−10.73
	D-S-S2	31.16	−2.33	69.25	−6.69	34.14	−16.54	−8.52
WaveNet-Attention (7)-CTC	S-D-S	30.39	−1.56	70.05	−7.49	32.7	−15.1	−8.05
	D-S-S1	35.28	−6.45	68.12	−5.56	38.03	−20.73	−10.81
	D-S-S2	32.58	−3.75	62.74	−0.18	37.16	−19.56	−7.83
WaveNet-Attention (10)-CTC	S-D-S	30.25	**−1.42**	69.25	−6.69	32.01	**−14.41**	−7.51
	D-S-S1	34.06	−5.23	70.05	−7.49	40.10	−22.50	−11.74
	D-S-S2	31.85	−3.02	**57.45**	**5.11**	33.65	−16.05	**−4.65**

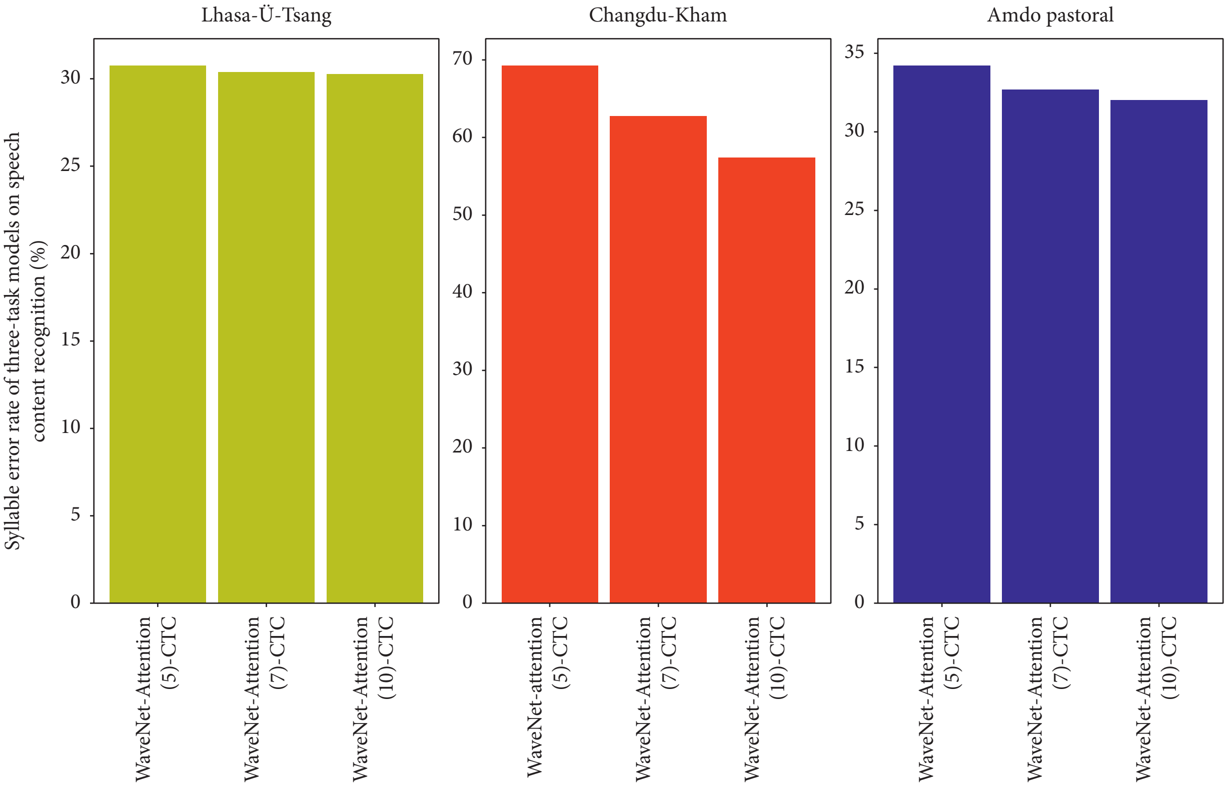

Figure 7: Syllable error rate of WaveNet-Attention-CTC for different lengths of the attention window.

Table 6: Dialect ID recognition accuracy (%) of three-task models.

Architecture	Model	Lhasa-Ü-Tsang	Changdu-Kham	Amdo Pastoral
DialectID model		97.88	92.24	97.9
WaveNet-CTC with dialect ID and speaker ID	D-S-S1	98.01	**98.8**	99.41
	D-S-S2	99.73	96.42	**99.61**
	S-D-S	99.25	95.23	99.03
Attention (5)-WaveNet-CTC	S-D-S	**100**	76.19	91.27
	D-S-S1	**100**	90.47	94.18
	D-S-S2	**100**	82.14	93.02
WaveNet-Attention (5)-CTC	S-D-S	**100**	89.28	93.79
	D-S-S1	**100**	85.71	93.79
	D-S-S2	**100**	95.23	94.18
WaveNet-Attention (7)-CTC	S-D-S	0	85.71	91.66
	D-S-S1	0	89.98	93.88
	D-S-S2	0	89.28	95.34
WaveNet-Attention (10)-CTC	S-D-S	0	85.71	95.54
	D-S-S1	0	94.04	93.99
	D-S-S2	0	0	0

Table 7: Speaker ID recognition accuracy (%) of three-task models.

Architecture	Model	Lhasa-Ü-Tsang	Changdu-Kham	Amdo pastoral
SpeakerID model		67.75	93.13	95.31
WaveNet-CTC with dialect ID and speaker ID	S-D-S	**72.91**	**98.8**	96.12
	D-S-S1	70.21	95.23	93.6
	D-S-S2	70.35	96.42	96.89
Attention (5)-WaveNet-CTC	S-D-S	61.08	83.33	89.53
	D-S-S1	62.12	83.33	87.01
	D-S-S2	61.99	84.52	90.11
WaveNet-Attention (5)-CTC	S-D-S	61.99	85.71	92.05
	D-S-S1	62.53	82.14	91.08
	D-S-S2	61.18	89.28	92.44
WaveNet-Attention (7)-CTC	S-D-S	60.91	85.71	91.66
	D-S-S1	62.04	84.31	92.01
	D-S-S2	58.49	86.90	90.69
WaveNet-Attention (10)-CTC	S-D-S	58.49	84.52	92.05
	D-S-S1	59.43	83.33	91.27
	D-S-S2	63.47	92.85	**97.86**

5. Conclusions

This paper proposes a multitask learning mechanism with local attention based on WaveNet to improve the performance for low-resource language. We integrate Tibetan multidialect speech recognition, speaker ID recognition, and dialect identification into a unified neural network and compare the attention effects on the different places in architectures. The experimental results show that our method is effective for Tibetan multitask processing scenarios. The WaveNet-CTC model with attention added into the high layer obtains the best performance for unbalance-resource multitask processing. In the future works, we will evaluate the proposed method on larger Tibetan data set or on different languages.

Data Availability

The data used to support the findings of this study are available from the corresponding author (1009540871@qq.com) upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Hui Wang and Yue Zhao contributed equally to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation under grant no. 61976236.

References

[1] Z. Tang, L. Li, and D. Wang, "Multi-task recurrent model for speech and speaker recognition," in *Proceedings of the 2016 Asia-Pacific signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–4, Jeju, South Korea, December 2016.

[2] O. Siohan and D. Rybach, "Multitask learning and system combination for automatic speech recognition," in *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 589–595, Scottsdale, AZ, USA, December 2015.

[3] Y. Qian, M. Yin, Y. You, and K. Yu, "Multi-task joint-learning of deep neural networks for robust speech recognition," in *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 310–316, Scottsdale, AZ, USA, December 2015.

[4] A. Thanda and S. M. Venkatesan, "Multi-task learning of deep neural networks for audio visual automatic speech recognition," 2020, http://arxiv.org/abs/1701.02477.

[5] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, April 2018.

[6] D. Chen and B. K.-W. Mak, "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, 2015.

[7] K. Krishna, S. Toshniwal, and K. Livescu, "Hierarchical multitask learning for ctc-based speech recognition," 2020, http://arxiv.org/abs/1807.06234.

[8] B. Li, T. N. Sainath, Z. Chen et al., "Multi-dialect speech recognition with a single sequence-to-sequence model," 2017, http://arxiv.org/abs/1712.01541.

[9] S. Toshniwal, T. N. Sainath, B. Li et al., "Multilingual speech recognition with a single end-to-end model," 2018, http://arxiv.org/abs/1711.01694.

[10] Y. Zhao, J. Yue, X. Xu, L. Wu, and X. Li, "End-to-end-based Tibetan multitask speech recognition," *IEEE Access*, vol. 7, pp. 162519–162529, 2019.

[11] A. Das, J. Li, R. Zhao, and Y. F. Gong, "Advancing connectionist temporal classification with attention," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, April 2018.

[12] Y. Zhang, P. Y. Zhang, and H. Y. Yan, "Long short-term memory with attention and multitask learning for distant speech recognition," *Journal of Tsinghua University (Science and Technology)*, vol. 58, no. 3, p. 249, 2018, in Chinese.

[13] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June2019.

[14] Z. Zhang, B. Wu, and B. Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019.

[15] S. Xu, "Speech-to-text-wavenet: end-to-end sentence level Chinese speech recognition using deepmind's wavenet," 2020, https://github.com/CynthiaSuwi/Wavenet-demo.

[16] Kim and Park, "Speech-to-text-WaveNet," 2016, https://github.com/buriburisuri/ GitHub repository.

[17] A. van den Oord, A. Graves, H. Zen et al., "WaveNet: a generative model for raw audio," 2016, http://arxiv.org/abs/1609.03499.

[18] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer, New York, NY, USA, 2012.

[19] M. Wei, "A novel face recognition in uncontrolled environment based on block 2D-CS-LBP features and deep residual network," *International Journal of Intelligent Computing and Cybernetics*, vol. 13, no. 2, pp. 207–221, 2020.

[20] A. S. Jadhav, P. B. Patil, and S. Biradar, "Computer-aided diabetic retinopathy diagnostic model using optimal thresholding merged with neural network," *International Journal of Intelligent Computing & Cybernetics*, vol. 13, no. 3, pp. 283–310, 2020.

[21] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2020, http://arxiv.org/abs/1508.04025.

[22] B. La, "Tibetan spoken language," in Chinese, 2005.