Computers, Materials & Continua
DOI: 10.32604/cmc.2022.027591
Article

Tech Science Press

Research on Tibetan Speech Recognition Based on the Am-do Dialect

Kuntharrgyal Khysru[1,*], Jianguo Wei[1,2] and Jianwu Dang[3]

[1]Key Laboratory of Artificial Intelligence Application Technology State Ethnic Affairs Commission,
Qinghai Minzu University, Xining, 810007, China
[2]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin, 300072, China
[3]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
*Corresponding Author: Kuntharrgyal Khysru. Email: gtj186@tju.edu.cn

Received: 21 January 2022; Accepted: 30 March 2022

Abstract: In China, Tibetan is usually divided into three major dialects: the Am-do, Khams and Lhasa dialects. The Am-do dialect evolved from ancient Tibetan and is a local variant of modern Tibetan. Although this dialect has its own specific historical and social conditions and development, there have been different degrees of communication with other ethnic groups, but all the abovementioned dialects developed from the same language: Tibetan. This paper uses the particularity of Tibetan suffixes in pronunciation and proposes a lexicon for the Am-do language, which optimizes the problems existing in previous research. Audio data of the Am-do dialect are expanded by data augmentation technology combining noise and reverberation, and the morphological characteristics and characteristics of the Tibetan language are further considered. According to the particularity of Tibetan grammar, grammatical features are used to optimize grammatical relationships and are combined with a language model, and the Am-do dialect is scored and rescored. Experimental results show that compared with the baseline, our proposed new lexicon and data augmentation technology yields a relative increase of approximately 3% in character error rates (CERs) and a relative increase of 3%–19% in the recognition rate of acoustic models and language models.

Keywords: Am-do dialect; acoustic model; language model; rescoring

1 Introduction

The language spoken by Tibetans belongs to the Sino-Tibetan language family, the Tibetan branch of the Burmese-Tibetan language family [1]. Tibetan speakers are distributed in Tibet, Qinghai, Sichuan, Gansu, Yunnan and other regions of China. Linguists generally believe that the Tibetan language can be divided into three dialects: the U-Tsang, Am-do and Kham dialects [2].

In 2018, Xiaohui Huang, Jing Li and others implemented an end-to-end Tibetan speech acoustic model using recurrent neural networks and connection timing classification algorithms [3]. Yan et al. of Northwest University for Nationalities used time-delay neural networks and long short-term memory networks to establish a Tibetan acoustic model [4,5]. In 2019, Song Wang of Northwest University for Nationalities used the method of combining long short-term memory networks and connection timing classification to carry out end-to-end acoustic modeling and speech recognition of the Lhasa dialect within the U-Tsang dialect [6]. Yue Zhao and others at the Central University for Nationalities used a framework consisting of end-to-end speech recognition and the WaveNet-connectionist temporal classification\1 (CTC) method to train a multitask system that can complete dialect recognition, speech recognition, and speaker recognition tasks simultaneously [7–9].

In 2016, Jian Li of Tianjin University and others used a four-tone system to establish the Lhasa dialect phoneme set based on the U-Tsang dialect tonal pronunciation characteristics; they combined it with a deep neural network to propose a tonal information-based phoneme set—the Lhasa dialect acoustic modeling method in the U-Tsang dialect [10,11]. For the Tibetan U-Tsang dialect, Lixin Pan of Tianjin University and others proposed adopting the "Tibetan component" as the modeling unit and using the end-to-end transformer model based on a self-attention mechanism and combined it with migration based on the pronunciation and character characteristics of the Tibetan language. Learning and other technologies were used to train an end-to-end Lhasa dialect speech recognition model [12,13]. In 2020, Jianjian Le of the Minzu University of China studied three types of Tibetan dialects simultaneously, and based on the TensorFlow deep learning framework, used the attention mechanism-based WaveNet-CTC method to build a Tibetan multidialect multitask recognition system [14]. In the same year, Jingwen Sun used a hybrid CTC/attention-based modeling method for the Tibetan Am-do dialect to establish a Tibetan speech recognition model [15].

In recent years, deep learning has been applied to study Tibetan language models. Tongtong Shen et al. proposed a Tibetan variable unit research method based on a recurrent neural network (RNN) [16] that relieves the data in a certain sense. The sparseness problem and this method have achieved better results than traditional methods [17]. Using Tibetan morphological features and grammatical relationships, Khysru et al. proposed a language model based on grammatical and morphological features, which not only alleviated the data sparseness problem but also solved the grammatical relationship existing in sentences [18,19].

A detailed analysis of the Tibetan Am-do lexicon tonal characteristics is an important Tibetan automatic speech recognition (ASR) application task. However, it is difficult to use tonal information because the Am-do dialect has multiple tonal modes, and there are controversies in research. Few studies have focused on modeling the Am-do dialect tonal information for speech recognition purposes. Therefore, we studied the effect of tonal information on Am-do Tibetan speech recognition performance. Since there is no definite tonal pattern in Am-do, in this study, we used a four-tone pattern and modeled it based on a comparison of tonal information.

The contributions of our work are summarized as follows:

An Am-do dialect lexicon is proposed that combines the Am-do dialect pronunciation characteristics and uses the special features of the Tibetan suffix in pronunciation.

Second, the Am-do dialect audio data are expanded by using data augmentation technology, which adds noise and reverberation, and by using the augmented data to train the network parameters to improve acoustic model performance.

Third, considering the morphological characteristics and grammatical relationship of the Tibetan language in the model, the Am-do dialect is scored and rescored.

2 Phone Set and Corpus

A Tibetan character is composed of several different "components", which are somewhat similar to the combination of radicals in Chinese characters, but a Tibetan character can be composed of up to 7 such components. Some documents refer to such parts as Tibetan letters. In a Tibetan character, according to the position occupied, the alphabet can be divided into the root, superscript, subscript, prefix, suffix, and farther suffix. In addition, there can be a vowel symbol. Among them, the base character is the core, and other Tibetan letters are overwritten and appended based on the base character to form a complete Tibetan character, as shown in Fig. 1.

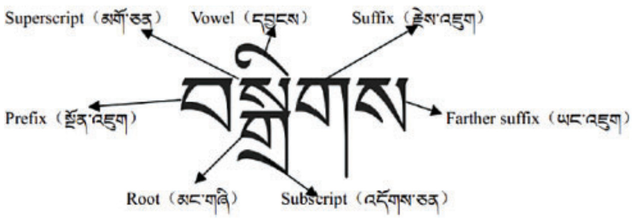

Figure 1: Most important components of a character in tibetan: prefix, root (vowel, superscript and subscript) and suffix (farther suffix)

There may be several methods for choosing acoustic representations for the Am-do dialect, such as syllable-based phone sets or initial/final-based phone sets. In our experiment, we chose the initial/final phone set and expanded the finals into tone-specified finals. The phone set is built by referring to previous phonological studies of the Am-do spoken language [20]. There are 48 initial consonants and 57 final units without considering the tones. All the initials and finals are listed in Tab. 1 below.

Table 1: Am-do dialect phone set

Am-do Tibetan initials							
p	c	ts	tɕ	ɳ	s	h	
p^h	c^h	ts^h	tɕ^h	ŋ	ɬ	ɹ	
t	k	tʂ	m	l	ʂ	w	
t^h	k^h	tʂ^h	n	g	ɕ	j	
b	hts	dz	dʐ	z	ʑ	d	
ɟ	ndʑ	x^h	hʋ	dʑ	dʒ	ht	
nd	hw	ʁ	mb	hɳ	f		
Am-do Tibetan finals							
i	y	o:	io	an	oŋ	op	u?
e	as	il	im	on	uŋ	up	ir
a	ɛ	ul	em	yn	ip	i?	er
ɛ:	i:	ø :	am	iŋ	ep	us	ar
o	e:	ak	om	eŋ	ap	a?	or
u	a:	iu	um	aŋ	en	o?	ur
ad	id	ud	ed	od	ik	e?	ø?
er							

In this study, the gaussian mixture model (GMM) acoustic model was used to verify the lexicon validity. Based on the lexicon of the Am-do dialect, the Kaldi speech recognition toolbox first extracted the mel frequency cepstral coefficient (MFCC) + pitch acoustic features of the training set and the test set. Here, we used 13-dimensional MFCC features plus 3-dimensional pitch features for a total of 16 dimensions and then randomly selected 20,000 sentences from the training set as a subset to train a monophone acoustic model. Then, the trained monophone acoustic model was used to align the speech data of the entire training set. After that, the triphone acoustic model was trained based on the aligned data. This research trained 5 triphone models. The first-order difference and second-order difference information of the acoustic features were added and recorded as the tri1 model; linear discriminant analysis (LDA) and maximum likelihood linear regression (MLLR) were used as two linear transformations for the features to create the tri2 model; speaker adaptation was added to the tri3 model (using LDA, MLLT + SAT); the tri4 model built a larger SAT model by adjusting parameters; and finally, the "quick fast training" script in Kaldi was used to train a larger scale GMM model, denoted the tri5p model [21–23].

The new lexicon is based on the Am-do dialect characteristics, and the old lexicon is the one proposed as the forget gate. The experimental results (in Tab. 2) show that the recognition effect of the GMM acoustic model built by the Am-do dialect lexicon used in this paper is slightly better than that of the model built by the previously proposed lexicon. The analysis from this result shows that the new Am-do dialect lexicon established in this study has no performance degradation compared with the previously proposed lexicon, and it shows that the lexicon used in this article was modified on some of the Tibetan text entries and characters. The expansion of text pronunciation items is more scientific, and it has certain reliability in constructing the Am-do dialect speech pronunciation recognition system.

The results show that the GMM acoustic model based on the two lexicons shows little performance difference in decoding on the test set. The Am-do dialect lexicon used in this article is based on the original and further expands the pronunciation annotations of some Tibetan characters, so it can also play a better role in covering Tibetan characters in daily Tibetan language. Therefore, this research determined that this Am-do dialect lexicon is available for acoustic modeling in the Am-do dialect. In the following experiments, this article builds an acoustic model of the Am-do dialect and the entire speech recognition system based on the Am-do dialect lexicon.

Table 2: Lexicon experiment comparison

Lexicon	CER%	SER%
New lexicon	27.10	80.36
Old lexicon	27.70	82.33

The audio database used in this article is a comprehensive Tibetan corpus. Each speaker had approximately 1,000 pieces of voice data. The Tibetan (Am-do dialect) speakers involved in the recording were 114 university students, including 58 males and 56 females ranging in age from 18 to 23. The data were recorded in a relatively quiet environment. The sentences in the corpus were all commonly used sentences in daily life; thus, the entire sample was biased toward daily spoken language. Each sound bite was relatively short, generally within ten seconds. Audio data collection used a mono 16 kHz sampling rate and 16-bit quantization, and data were saved in the WAV file format (in Tab. 3).

Table 3: Basic information on the tibetan audio data

Number of Speakers		Speech signal		Number of text	
Male	Female	Sampling Rate	Quantification precision	Train set	Test set
56	54	16KHZ	16-bit	110,000	3,996

The database used to build the model of the Am-do dialect speech recognition system was divided into a training set and a testing set. The testing set selected 4 speakers, two males and two females, and 3,996 sentences for approximately 5.13 h. The Am-do dialect sentences of the remaining 110 speakers were used as the training set, and the total time was approximately 142.77 h. We used the text corpus [18–20].

3 Data Enhancement and Models

In this research, based on the HMM-DNN model architecture, the two neural network models, time-delay neural network (TDNN) and factorized TDNN (TDNN-F), combined with the hidden Markov model were used to model the phoneme units of the Am-do dialect. The experiment explored the performances of the TDNN and TDNN-F models in Am-do dialect speech recognition and studied the influence of network structure changes on the final recognition effect by adjusting the parameters. Finally, this study chose the best-performing model used on the test set as the acoustic model of the Am-do dialect speech recognition system built in this article.

3.1 Data Augmentation Technology

In the acoustic model training process in this study, two data augmentation techniques were used for acoustic data: the speech rate perturbation of the speech data and the processing method for adding noise and reverberation to the original speech data to generate noise and voice data of reverberation. Since velocity disturbance is a more commonly used data augmentation technique in speech recognition systems, this study did not set up a comparative experiment for velocity disturbance. The following is only an experimental exploration of the data augmentation technology of adding noise and reverberation.

This research used the method proposed in [24] to add noise and reverberation to the training set of the Am-do dialect speech dataset. The noise audio and reverberation audio dataset used in this paper is called the RIRs NOISE dataset. The noisy part is the point source noise dataset. There were a total of 843 noisy sentences, and each sentence was used as foreground noise or background noise of the speech signal. The reverberation dataset consisted of three parts: the RWCP auditory scene dataset, the 2014 REVERB challenge dataset, and the Aachen impulse response dataset. A total of 325 sentences had real reverberation data. In addition, there was a simulated reverberation audio dataset, SimulatedRIRs, which used a 16-kHz audio sampling rate to simulate impulse responses in rooms of different sizes. The rooms were divided into small, medium, and junior middle room types. Small rooms had areas from 1 to 10 square meters, medium rooms had areas from 10 to 30 square meters, and large rooms had areas from 30 to 50 square meters. This part of the dataset was set to be randomly sampled using these room sizes.

In this study, two TDNN models were trained using the Am-do dialect speech data training set: one used only the original clean data at 3 times the speed of perturbation technology for data augmentation to train the TDNN network, which was established in the previous section of this article. The other model combined the original clean audio data with noise and reverberation at a ratio of 1:1 and used the combined dataset to train the TDNN acoustic model. The TDNN structure used in this experiment was the structure introduced in the previous section: a total of 16 hidden layers, each with 625 neurons, and subsampling technology. To train the two models as consistently as possible on factors other than the data type, the learning rates and minibatch size settings in this study were the same, and the epoch of the TDNN experimental training that only used clean data was set to use clean data twice as much as the model trained with the noise reverberation data to ensure that both models were trained to the greatest extent.

3.2 Morphological Language Model

In this paper, based on the relationship between the morphological structure and the grammatical relationship in the Tibetan language and according to Tibetan morphological verb characteristics, we proposed adjusting the discriminant weight online. To further understand the morphological verbs during the test, we adjusted the discriminant weight online, namely, $\mathrm{P}(\mathrm{c}_{\mathrm{i}}|\mathrm{v}_{\mathrm{i}-1})$, to improve prediction accuracy. Specifically, the RNNLM output was assigned a threshold ε. Then, we generated a binary output

$$\overline{P_{RNN}}\left(\mathrm{w}_i|v_{i-1}\right) = P_{RNN}\left(\mathrm{w}_i|v_{i-1}\right) > \varepsilon. \tag{1}$$

$$\tilde{P}\left(c_i|v_{i-1}\right) = \frac{\tilde{P}_{RNN}\left(w_i|v_{i-1}\right)}{P\left(w_i|c_{i,}v_{i-1}\right)}. \tag{2}$$

We combined it with $\tilde{P}\left(c_i|v_{i-1}\right)$ to generate a new$P\left(c_i|v_{i-1}\right)$

$$\overline{P}\left(c_i|v_{i-1}\right) = P\left(c_i|v_{i-1}\right) + \alpha\tilde{P}\left(c_i|v_{i-1}\right). \tag{3}$$

where α represents the combined weight of $\tilde{P}\left(c_i|v_{i-1}\right)$. We denote this method RNNLM_tuning discriminative weights (_TDW).

In addition, the effect of suffixes on function words, namely, the Tibetan radical suffix unit (TRSU), which affects sentence semantics, was considered. We enhanced the suffix weight based on the RNNLM to more accurately connect the suffix to the function word.

$$c_t = f\left(W_{IC}x_t + W_{SC}s_t\right) \tag{4}$$

$$h_t = f\left(W_{CH}c_t + W_{HH}h_{t-1}\right) \tag{5}$$

$$o_t = g\left(W_{HO}h_t\right) \tag{6}$$

In Eq. (4), W_{IC} is the character weight, and $\mathbf{W}_{\mathrm{SC}}$ is the suffix feature weight. The activation function inputs the character probabilities and merges the suffix feature weights. W_{SC} is used to process suffix information in Tibetan grammar. The suffix information is connected with the function word so that the sentence can accurately express the meaning. Eq. (5) calculates the hidden layer h_t. From the start time to time c_t, all advance information is saved. The hidden layer adds weight W_{HH} at time h_{t-1}. The activation function retains contextual information, and the network outputs the result.

4 Experiment

As a minority language in China, the Tibetan language circulates only in Tibetan areas, where the official language is Chinese. Therefore, the corpus available on the internet is limited, and research personnel are rare. Our text corpus cited the literature [18–20].

4.1 Acoustic Model

After training the two TDNN acoustic models, the Am-do dialect lexicon used in this article and the same language model, the two speech recognition systems were decoded and tested on the test set. The results are shown in Tab. 4.

Table 4: Experimental comparison of the validation set and test set with noise added to the data

Data	Model	CER(%)	SER (%)
validation set	TDNN(clean)	15.27	62.69
	TDNN(clean_rvb)	13.62	60.04
Test set	TDNN(clean)	52.17	90.64
	TDNN(clean_rvb)	23.20	73.90

In Tab. 4, “clean” denotes that only the original clean data were used for training, and “clean + rvb” denotes that the original data plus the dataset with noise and reverberation were used for training. The results show that the model recognition effect of the TDNN acoustic model trained with noise and reverberation data on the test set was better than that of the model trained with only the original data. This shows that data augmentation methods such as adding noise and reverberation have a certain improvement effect on the HMM-TDNN Tibetan acoustic model built in this article.

Tab. 4 shows that the model trained with only clean speech data and the model trained with clean data plus noise and reverberation demonstrated small differences in performance on the original test set. However, the test set with noise and reverberation showed great performance differences. The performance of the HMM-TDNN acoustic model trained using noise and reverberation data augmentation technology was significantly better than that of the acoustic model trained using the original data. The character error rate (CER) results obtained on the test set reached an absolute improvement of 28.97%.

The above comparative experiments show that the data augmentation technology of “velocity disturbance + noise and reverberation” used to construct the Am-do dialect speech recognition system in this paper optimizes the acoustic model performance and is also oriented toward application scenarios. Considering the Tibetan speech recognition system is even more necessary. In real speech recognition system application scenarios, it is often difficult to achieve a sound environment, such as a recording studio. Human life scenes are generally full of various noises, and reverberation occurs in rooms, cars and cockpits. Therefore, the value of such a method is not only the improvement of the experimental results and the model for scientific research but also optimization for the application of the speech recognition system in real life. Thus, this research is a great help in building a Tibetan Am-do dialect speech recognition system that can be used in real applications in the future.

In this paper, the L2 regularization method was used in the TDNN-F network [25], where the coefficients were set to 0.01 for each layer of the network and 0.002 for the output layer. After the training was completed, the same language model used previously was combined to perform composition decoding, and then tests were performed on the test sets with and without noise reverberation. The experimental results are shown in Tab. 5. This experiment also added a set of experiments for a TDNN acoustic model with 1,536 hidden neurons with the same parameter settings as those of the previous TDNN network.

Table 5: Comparison of TDNN and TDNN-F acoustic model results

Model	CER(%)	SER(%)	CER-rvb(%)	SER-rvb(%)
TDNN-625	13.62	60.04	23.10	73.90
TDNN-1536	13.79	60.26	22.62	73.17
TDNN-F	13.68	59.78	21.90	71.72

The acoustic modeling performance of the TDNN-F model used in this research and the other two TDNN models on the test set were not very different. On the original test set without noise and reverberation, the test results of the TDNN-F acoustic model were between those of the other two models, and the performances of the three models were relatively close. On the test set with noise and reverberation, the performance of the TDNN-F acoustic model was slightly better than that of the other two models. Compared with the previously established TDNN acoustic model with a hidden-layer dimension of 625, the word error rate results were improved by approximately 5%.

These results show that the network structure of the TDNN-F model adds a low-dimensional bottleneck layer between the two hidden layers, thereby decomposing the weight matrix into two small-scale matrices and restricting one of the matrices to semiorthogonality. As a result, the modeling ability of the overall model did not significantly decrease.

However, the TDNN-F Am-do dialect acoustic model has fewer parameters than the TDNN model. Compared with the original two hidden layers of the TDNN with 625 neurons, the number of parameters was reduced by approximately 65.8%. Therefore, even if the number of hidden-layer neurons was expanded from 625 to 1,536 according to the TDNN-F model structure used in this article, the number of overall parameters was reduced by approximately 16.1%. That is, in the acoustic modeling task of the Am-do dialect, the TDNN-F acoustic model is more advantageous than the TDNN. While achieving comparable performance, the TDNN-F acoustic model has fewer parameters, which effectively reduces the space and time costs required to build the model. Furthermore, fewer model parameters can speed up the construction of decoded images and make it possible to embed the entire speech recognition system into the terminal device.

4.2 Language Model

In a speech recognition system, the quality of the language model is directly related to the performance of the entire recognition system. Even though the previous acoustic model is very accurate at the phoneme level, the speech recognition output must ultimately be implemented in the text sequence. To build an Am-do dialect speech recognition system with the best recognition effect possible, this research conducted a series of experimental studies on the Tibetan language model.

Tab. 6 shows the CER (%) and SER results of our latest data method. We used the radical-based uniform weight of Tibetan radicals (_TRU) method as the baseline [26–30]. The CER (%) of our method was approximately 8%–11.5% less than that of the n-gram method and 3.1%–4.4% less than that of the baseline_TRU method, and good SER results were achieved, which shows the effectiveness of our method.

Table 6: Latest language model method, %CER

Language model	CER(%)	SER(%)
N-gram(Mikolov et al.2012)	15.21	63.67
RNNLM (Mikolov et al.2012)	14.33	62.55
CharCNN (Kim et al., 2016)	14.25	62.32
_TRU(Shen et al.2017)	14.12	62.05
_TDW	**13.99**	**61.87**
TRSU	**13.45**	**61.23**

The RNNLMs and n-gram LMs have modeling characteristics of two essentially different LMs. RNNLMs typically use a fixed weight of linear interpolation in conjunction with the n-gram LM. Tab. 6 shows the result of our method and n-gram interpolation on the text dataset, referring to the value of λ in [31,32] ($\lambda = 0.5$).

We know that a lattice is a structure that is decoded once in the speech recognition process and contains many candidate results. Since the neural network uses historical information to predict the next word, reevaluating the grid decreases the search speed. Compared with the word structure of the case, N-best is more suitable for the model extension of long-distance information. This article uses N-best's intermediate results for rescoring, as shown in Tab. 7. We validated our model in the ASR experiment on the Am-do dialect audio dataset. The Tibetan CER (%) and SER (%) validation results show that our method has a good effect on weighting rare words in the TLM.

Table 7: Evaluation result of the CER (%) with N-best rescoring

Language model	CER(%)	SER(%)
N-gram(Mikolov et al.2012)	14.23	60.74
RNNLM (Mikolov et al.2012)	13.54	60.66
CharCNN (Kim et al., 2016)	13.26	58.73
_TRU(Shen et al.2017)	12.78	57.10
_TDW	**12.22**	**56.33**
TRSU	**11.97**	**56.03**

In summary, the weighting method based on morphological features and grammatical relationships affected the sentence semantics to a certain extent and achieved good results. Therefore, it is necessary to strengthen Tibetan morphological verbs and increase language knowledge to strengthen the semantic relationship of sentences.

5 Conclusion

This article proposed a pronunciation dictionary for the Am-do dialect and verified the dictionary's validity. Through data enhancement and based on the HMM-DNN framework, three acoustic models, namely, the DNN, TDNN, and TDNN-F, were built. Based on the morphological characteristics and grammatical relationship of the Tibetan language, the TDW and TRSU methods were proposed and verified. The experimental results show that the character error rate (CER) of our pronunciation dictionary increased by 11%. This preliminary study shows that suffix information plays an important role in speech recognition in the Tibetan Am-do dialect. Second, the _TDW and TRSU methods achieved comparative effects on CER (%) and SER (%), verifying that the morphological features and grammatical relationships affected the semantics of Tibetan sentences. Therefore, we hope to consider linguistic knowledge in future research to better assist Tibetan speech recognition, speech synthesis and machine translation tasks.

Funding Statement: This work was supported by the Regional Innovation Cooperation Project of Sichuan Province (Grant No. 22QYCX0082).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Sun, *Research on Tibetan A-mdo dialect speech recognition based on deep learning*. Gansu, Northwest Normal University, 2020.

[2] R. Wei, “The phonetic research of Tibetan dialects in 70 years of New China,” *Tibet Science and Technology*, vol. 9, pp. 72–77, 2019.

[3] X. Huang and L. I. Jing, “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.

[4] J. Yan, H. Yu and G. Li, “Tibetan acoustic model research based on TDNN,” in *Asia-Pacific Signal and Information Proc. Association Annual Summit and Conf. (APSIPA ASC)*, USA, pp. 601–604, 2018.

[5] J. Yan, Z. Lv, S. Huang and H. Yu, “Low-resource Tibetan dialect acoustic modeling based on transfer learning,” in *Int. Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*, pp. 6–10, 2018.

[6] W. Song, *Tibetan Lhasa speech recognition system based on LSTM-CTC*. Gansu, Northwest University for Nationalities, 2019.

[7] 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.

[8] Y. Zhao, J. Yue, W. Song, X. Xu, X. Li *et al.,* “Tibetan multi-dialect speech and dialect identity recognition,” *Computers, Materials & Continua*, vol. 58, no. 2, pp. 1223–1235, 2019.

[9] M. H. Changrampadi, A. Shahina, M. B. Narayanan and A. N. Khan, “End-to-end speech recognition of tamil language,” *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 1309–1323, 2022.

[10] J. Li, H. Wang, L. Wang, J. Dang, K. Khuru *et al.,* “Exploring tonal information for Lhasa dialect acoustic modeling,” in *10th Int. Symp. on Chinese Spoken Language Proc. (ISCSLP)*, Tianjin, pp. 1–5, 2016.

[11] H. Wang, K. Khyuru, J. Li, G. Li, J. Dang *et al.,* “Investigation on acoustic modeling with different phoneme set for continuous Lhasa Tibetan recognition based on DNN method,” in *2016 Asia-Pacific Signal and Information Proc. Association Annual Summit and Conf. (APSIPA)*, Jeju, Island, pp. 1–4, 2016.

[12] L. Pan, S. Li, L. Wang and J. Dang, “Effective training end-to-end ASR systems for low-resource Lhasa dialect of Tibetan language,” in *2019 Asia-Pacific Signal and Information Proc. Association Annual Summit and Conf. (APSIPA ASC)*, Lanzhou, pp. 1152–1156, 2019.

[13] Y. Zhao, J. Yue, W. Song, X. Xu, L. Li *et al.,* “Tibetan multi-dialect speech recognition using latent regression bayesian network and end-to-end mode,” *Journal of Internet of Things*, vol. 1, no. 1, pp. 17–23, 2019.

[14] L. Jianjian, *Tibetan multi-task and multi-dialect speech recognition*. Beijing, Central University for Nationalities, 2019.

[15] J. Sun, *Research on Tibetan A-mdo dialect speech recognition based on deep learning*. Northwest Normal University, 2020.

[16] T. Y.Emily, “Tibet and the problem of radical reductionism,” *Antipode*, vol. 41, no. 5, pp. 983–1010, 2009.

[17] T. Shen, L. Wang, X. Chen, K. Khysru and J. Dang, “Exploiting the tibetan radicals in recurrent neural network for low-resource language models,” in *Int. Conf. on Neural Information Proc.*, Guangzhou, Cham, pp. 266–275, 2017.

[18] K. Khysru, D. Jin, Y. Huang, H. Feng and J. Dang, “A Tibetan language model that considers the relationship between suffixes and functional words,” *IEEE Signal Processing Letters*, vol. 28, pp. 459–463, 2021.

[19] K. Kuntharrgyal, D. Jin and J. Dang, “Morphological verb-aware tibetan language model,” *IEEE Access*, vol. 7, pp. 72896–72904, 2019.

[20] Y. Daoqian, *Tibeto-chinese Lhasa vernacular dictionary (Tibetan)*. Beijing, The Ethnic Publishing House, 1983.

[21] G. Jyoshna, M. Zia and L. Koteswararao, “An efficient reference free adaptive learning process for speech enhancement applications,” *Computers, Materials & Continua*, vol. 70, no. 2, pp. 3067–3080, 2022.

[22] X. R. Zhang, W. F. Zhang, W. Sun, X. M. Sun and S. K. Jha, “A robust 3-D medical watermarking based on wavelet transform for data protection,” *Computer Systems Science & Engineering*, vol. 41, no. 3, pp. 1043–1056, 2022.

[23] X. R. Zhang, X. Sun, X. M. Sun, W. Sun and S. K. Jha, “Robust reversible audio watermarking scheme for telemedicine and privacy protection,” *Computers, Materials & Continua*, vol. 71, no. 2, pp. 3035–3050, 2022.

[24] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer and S. Khudanpu, “A study on data augmentation of reverberant speech for robust speech recognition,” in *2017 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, USA, pp. 5220–5224, 2017.

[25] W. Dong and F. Z. Thomas, “Transfer learning for speech and language processing,” in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA)*, Weedse, pp. 1225–1237, 2015.

[26] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký and S. Khudanpur, “Recurrent neural network based language model,” *Interspeech*, vol. 2, no. 3, pp. 1045–1048, 2010.

[27] T. Mikolov and Z. Geoffrey, “Context dependent recurrent neural network language model,” in *IEEE Spoken Language Technology Workshop (SLT)*, USA, pp. 234–239, 2012.

[28] S. Tongtong, W. Longbiao, C. Xie, K. Kuntharrgyal and D. Jianwu, “Exploiting the Tibetan radicals in recurrent neural network for low-resource language models,” in *Int. Conf. on Neural Information Processing*, Guangzhou, pp. 266–275, 2017.

[29] K. Yoon, Y. Jernite, D. Sontag and M. R.A., “Character-aware neural language models,” in *Thirtieth AAAI conf. on artificial intelligence*, USA, 2016.

[30] X. Chen, X. Liu, A. Ragni, Y. Wang and M. J. F. Gales, “Future word contexts in neural network language models,” in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Japan, pp. 97–103, 2017.

[31] X. Chen, X. Wang, X. Liu, M. J. F. Gales and P. C. Woodland, *Efficient GPU-based training of recurrent neural network language models using spliced sentence bunch*. Singapore, INTERSPEECH, pp. 641–645, 2014.

[32] X. Chen, X. Liu, Y. Qian, M. J. F. Gales and P. C. Woodland, “CUEDRNNLM—An open-source toolkit for efficient training and evaluation of recurrent neural network language models,” in *Acoust. Speech Signal Process (ICASSP)*, Shanghai, China, pp. 6000–6004, 2016.