Received August 9, 2020, accepted September 2, 2020, date of publication September 14, 2020,
date of current version September 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3023783

End-to-End Amdo-Tibetan Speech Recognition Based on Knowledge Transfer

XIAOJUN ZHU[1,2,3] **AND HEMING HUANG**[1,2]

[1]School of Computer Science and Technology, Qinghai Normal University, Xining 810008, China
[2]MOE Key Laboratory of Tibetan Information Processing, Xining 810008, China
[3]School of Electronic and Information Engineering, Lanzhou City University, Lanzhou 730070, China

Corresponding author: Heming Huang (huanghm@qhnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62066039 and Grant 61662062, and in part by the Natural Science Foundation of Qinghai Province under Grant 2016-ZJ-904.

ABSTRACT The end-to-end speech recognition technology solves the problem that each component is independent and models cannot be jointly optimized in the traditional speech recognition model. It incorporates such components as the acoustic model, language model, and decoding unit of the hybrid model into a single neural network, that can avoid the inherent defects of multiple modules and greatly reduces the complexity of the speech recognition model. In this research, an Amdo-Tibetan speech recognition system is constructed based on Listen, Attend and Spell (LAS) model by the end-to-end speech recognition technology. It can realize the direct conversion from Amdo-Tibetan speech sequence to the corresponding character sequence and greatly reduces the difficulty of building the Amdo-Tibetan speech recognition model. To further improve the performance of the proposed system, the following improvements have been made: firstly, the Multi-Head Attention mechanism is introduced to improve the alignment accuracy between state vectors of decoder and encoder; secondly, the label smoothing technique is adopted to solve the problem of over-fitting; thirdly, an N-gram language model is combined with the LAS model to increase the accuracy of speech recognition and the maximum mutual information (MMI) criterion is employed for discriminative training; and finally, transfer learning is utilized to overcome the problem of insufficient training data. Experimental results show that the proposed model can significantly enhance the performance of Amdo-Tibetan speech recognition.

INDEX TERMS End-to-end, LAS model, transfer learning, low resource language speech recognition, Amdo-Tibetan.

I. INTRODUCTION

The era of artificial neural networks research has ushered since American neurophysiologist Warren McCulloch and mathematician Walter Pitts presented the concept of artificial neural network and its mathematical model in their joint work in 1943. After decades of development, it has been successfully applied to pattern recognition, automatic control, signal processing, artificial intelligence, and other research fields [1]–[4].

With the wide application of artificial neural networks in Automatic Speech Recognition (ASR), Deep Neural Networks-Hidden Markov Model (DNN-HMM) has become a typical model in ASR [5], [6]. It significantly improves the performance of the ASR system, but its training process relies heavily on the initial frame level label obtained by the traditional Gaussian Mixture Model-Hidden Markov Model (GMM-HMM). This leads to additional training steps [7]. Furthermore, DNN-HMM requires other resources such as pronunciation dictionary and speech context dependent tree during the decoding process. It is difficult to obtain these resources [8], [9], especially for low-resource languages. Therefore, it is not easy, for Tibetan, to construct a speech recognition system based on DNN-HMM [10].

The emergence of end-to-end technology simplifies the construction process of speech recognition system and reduces the complexity of speech recognition model by incorporating such components as acoustic model, language model, and decoding unit of the hybrid model into a single neural network [11]. In recent years, great progresses have been achieved in end-to-end speech recognition. In 2015, Baidu developed an end-to-end speech recognition system

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

named Deep-Speech2 which can recognize Chinese and English speech simultaneously. In 2017, Google developed an end-to-end architecture based on the LAS model with a 5.6% word error rate (WER). In 2018, Facebook developed an end-to-end speech recognition system based on Convolutional Neural Networks (CNN). Its WER has reduced to 5% and the speed of training has also been improved effectively [12]–[14].

The successful application of end-to-end technology in rich-resource language speech recognition has led to its application in Tibetan speech recognition. Wang *et al.* [15] implement an end-to-end Tibetan speech recognition system based on Connectionist Temporal Classification (CTC). They combine existing linguistic knowledge with the end-to-end acoustic model, and it greatly improves the discrimination ability and robustness of acoustic models. Huang *et al.* [16] apply the Recurrent Neural Network (RNN) and CTC in the Tibetan acoustic modeling, and they obtain higher training speed and decoding efficiency in maintaining the same recognition performance. Zhao *et al.* [17] establish a Tibetan multi-task recognition framework based on WaveNet-CTC. It identifies Tibetan speech recognition, speaker recognition, and dialect recognition simultaneously in an end-to-end network and achieves better performance than the task-specific model.

Tibetan speech recognition research is different from that of Chinese or English. Both Chinese and English have their own standard pronunciation, i.e. Mandarin and Received Standard English respectively. However, for Tibetan, there is no standard pronunciation. It has three major dialects, namely Ü-Tsang, Kham, and Amdo. The three dialects are unified in their writing, but their pronunciation are differ greatly [18]–[20]. Each dialect has roughly the same number of speakers and the area of geographical distribution is roughly equal. Therefore, Tibetan speech recognition must be specific to one of the three dialects. However, most of the academic reports mentioned above mainly focusing on the speech recognition of the Ü-Tsang dialect and few reports focus on the other two dialects. More seriously, the other two dialects have fewer research achievements in phonetics and there is no professional speech sample database available.

This article takes the Tibetan Amdo dialect, abbreviated as Amdo-Tibetan hereafter, as the research object and mainly focuses on the Sequence to Sequence (Seq2Seq) model based on the attention mechanism. After analyzing the pronunciation characteristics and determining the modeling unit of the Amdo-Tibetan, an efficient end-to-end speech recognition system is proposed based on the Listen, Attend and Spell (LAS) model. It can directly convert from a speech sequence to the corresponding character sequence, and its training process is much more efficient than the traditional model. In this study, various methods are employed to optimize the performance of the LAS model. For example, the Multi-Head Attention mechanism is introduced to improve the alignment accuracy between state vectors of decoder and encoder, the label smoothing and discriminative training technique is adopted to optimize the training process of the model, an N-gram language model is combined with the LAS model to increase the accuracy of speech recognition, and transfer learning is utilized to overcome the problem of insufficient training data. The proposed model is verified on the Amdo-Tibetan corpus recorded by our laboratory. Experimental results show that the end-to-end model proposed in this work can significantly improve the performance of Amdo-Tibetan speech recognition.

The rest of the paper is organized as follows. Section II introduces the pronunciation characteristics and modeling units of Amdo-Tibetan. Section III introduces the LAS model and its optimization. Section IV describes the transfer learning in the Amdo-Tibetan speech recognition system. The experimental setups and results are described in section V. Finally, conclusions are made in section VI.

II. PRONUNCIATION CHARACTERISTICS AND MODELING UNIT OF TIBETAN

A. CHARACTERISTICS OF TIBETAN PRONUNCIATION

Language and phonetics are closely related. Therefore, it is necessary to introduce briefly Tibetan grammar such as the rules for the formation of words and the structural relations between words and sentences. These are the cornerstone and prerequisite for the study of Tibetan phonetics.

In Tibetan alphabet, there are 30 consonants and 4 vowels [21]. The 30 consonants are ཀ་ཁ་ག་ང་ཅ་ཆ་ཇ་ཉ་ཏ་ཐ་ད་ན། པ་ཕ་བ་མ། ཙ་ཚ་ཛ་ཝ། ཞ་ཟ་འ་ཡ། ར་ལ་ཤ་ས། ཧ་ཨ།, and the 4 vowels are ཨི་ཨུ་ཨེ་ཨོ།. Consonants, or be joined together with other consonants and/or vowels, usually form a word either horizontally or vertically. A Tibetan word is also called a syllable. Tibetan grammar prescribes a set of strict rules for letter stitching, so as to form Tibetan syllable of different lengths and forms. For example, a syllable may consist of a single consonant, a consonant and a vowel, or several consonants and a vowel. In each syllable, there is a basic consonant, called root consonant; other consonant would be prefix, superscript, subscript, suffix, or farther suffix, according to their relative position to the root consonant. There is a separator "་" between any two syllables. A vertical terminator "།" or "༎" indicates the end of a sentence [22], [23]. The relationship between letters, syllables and sentences in Tibetan is shown in Fig. 1.

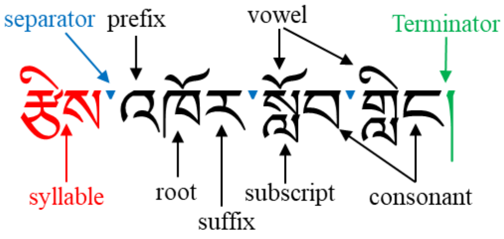

FIGURE 1. Examples of Tibetan letters, syllables, and sentence.

The pronunciation unit of Tibetan is a syllable, and the pronunciation of each syllable is determined by the phonetic alphabet of the syllable [24].

B. TIBETAN MODELING UNIT

For the Tibetan speech recognition, selection of the modeling unit is one of the most important problems, and different modeling units have different requirements for the size of training data. In previous studies of Chinese and English speech recognition, researchers have considered modeling units of different granularity such as word, syllables, Initial/Final, phoneme, etc. The greater the granularity, the better the accuracy will be. Correspondingly, the training process will be more difficult, and more training corpus will be needed.

The modeling units selected for each language are different according to their pronunciation characteristics. For example, Chinese is a monosyllabic language and syllables are usually selected as the modeling unit; English is a polysyllabic language and phonemes are often used as modeling units [25], [26]. For Tibetan speech recognition, there is no authoritative conclusion on the granularity of the modeling unit. Monosyllable is not a good choice if there is no adequate training corpus, especially for low-resource language such as Tibetan. Therefore, this article selects phoneme as the modeling units of Amdo-Tibetan speech recognition system.

III. MODEL AND OPTIMIZATION

A. LISTEN, ATTEND AND SPELL MODEL

The LAS model was firstly proposed by William Chan *et al.* in 2016 [27]. It is a neural networks speech recognizer that transcribes the sequence of speech features directly into the character sequence without acoustic model such as Hidden Markov Models (HMMs) or other components of traditional speech recognizers. The LAS model is a Seq2Seq model based on the attention mechanism. Its goal is to maximize the conditional probability of the output character sequence under the given conditions for speech inputs [28], [29]. The model is trained directly with the input speech feature sequence and its corresponding character sequence. Label alignment information is not required during training. In the inference stage, the probability of the output sequence calculated by the chain rule of probability is defined as

$$P(y|x) = \prod_i P(y_i|x, y_{<i}), . \tag{1}$$

where $x = (x_1, x_2, \cdots, x_m)$ represents the filter bank spectra features of input sequence while $y = (y_1, y_2, \cdots, y_n)$ represents the output sequence of characters.

The LAS model consists of Listener and Speller. The Listener is a typical encoder structure, whose key operation is 'Listen' and its function is to convert the original signal x into a high level representation $h = (h_1, h_2, \cdots, h_u)$. The operation is defined as

$$h = Listen\left(x, \theta_{Lis}\right), \tag{2}$$

where θ_{Lis} denotes the parameters of the Listener.

The Listener consists of a pyramid bidirectional long short-term memory (pBLSTM) networks, as shown in Fig. 2. In each successive stacked pBLSTM layer, the time resolution is reduced by a factor of 2 [30]. This structure is very effective in speech recognition. Speech is a continuous signal, and the speech signals that input to the model may be hundreds or thousands of frames at a time. Meanwhile, the difference between the adjacent frames is not very obvious because of the overlap. Therefore, this structure will not affect the accuracy of speech signal encoding, and subsequent attention models may also extract relevant information from a smaller number of times steps. In addition, it reduces the calculation complexity and accelerates the convergence of the model. The output of pBLSTM at the *i-th* step, from the *j-th* layer, is computed as

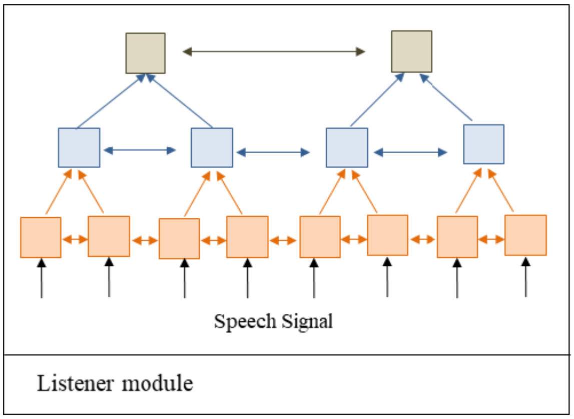

FIGURE 2. Schematic diagram of listener module.

$$h_i^j = pBLSTM\left(h_{i-1}^j, \left[h_{2i}^{j-1}, h_{2i+1}^{j-1}\right]\right). \tag{3}$$

The Speller is a decoder network based on attention mechanism. During the decoding process, the Speller predicts the subsequent graphemes based on the probability of all graphemes obtained previously, i.e. $P(y_i|x, y_{i-1}, \cdots, y_1)$. The Key operation of Speller is to decode the output of the Listener into a sequence of characters. The operation is defined as

$$P(y|x) = Speller(h; \theta_{Spl}), \tag{4}$$

where θ_{Spl} denotes the parameters of the Speller, which consists of the parameters of attention mechanism and decoder. Specifically, at each step, the attention mechanism generates attention vector according to the matching degree between the state vector of decoder and encoder output at first; then, based on the state vector of decoder and attention vector, the Speller predicts the next character.

The Listener and the Speller can be trained jointly and its objective function is defined as

$$\max_\theta \sum_i \log P\left(y_i|x, y^*_{<i}; \theta\right), \tag{5}$$

where θ denotes the parameters of the LAS model, $y^*_{<i}$ is the ground truth of the previous characters. It should be noted that the objective function is the maximum logarithmic likelihood estimation.

B. DECODING STRATEGY

The decoding process of the LAS model aims at finding the most possible character sequence for a given acoustic input; the operation is defined as

$$\hat{y} = \arg\max_{y} \log P(y|x). \tag{6}$$

The decoding process of the LAS model, which is based on output, is different from that of weighted finite-state transducer (WFST), which is based on the frames [31]–[33]. When the Speller starts decoding, the first output token is <sos>, which means the beginning of the sentence. When the output encounters the token <eos>, it indicates the end of decoding. The decoding process is performed with the beam search strategy [34]. In other words, following each output step, the Speller only keeps the top N paths with the highest probability among all decoding paths. It is noted that N is an adjustable parameter, and the N paths with the highest scores are called the N-best list.

In the process of decoding, the beam search strategy controls the convergence speed of the model well. Compared with the greedy search, it increases the diversity of the generated sequence to a certain degree and prevents the error in one step of decoding process from continuing in the subsequent steps. However, the similarity between the sequences generated by the beam search strategy is very high. Moreover, with the increase of beam size, the memory occupancy rate will increase, and the sequence generation speed will slow down. Therefore, the determination of beam size is the key issue for the beam search strategy.

C. MULTI-HEAD ATTENTION MECHANISM

The Multi-Head Attention (MHA), first proposed by Google in 2017, achieves ideal effect in machine translation [35]. The essence of the conventional attention mechanism is a mapping function of query and key-value pairs. The output of function is the weighted sum of the value vectors, where the weight is calculated by the similarity between the query and the corresponding key vectors, which is expressed as

$$Attention(Q, K, V) = F(Q, \mathrm{K})V, \tag{7}$$

where Q denotes a query vector, K and V denotes a set of key-value pairs.

In the MHA mechanism, the query vector and a set of key-value pairs are linearly mapped multiple times respectively [36], [37]. Each mapping can generate a different attention distribution, and its results are calculated by the Scaled Dot-Product Attention mechanism [35]. All output vectors of the Scaled Dot-Product Attention mechanism are spliced, and the results are linearly mapped again. The result of the mapping is the output vector of the Multi-Head Attention mechanism [35]. The Schematic diagram of the computational process is shown in Fig. 3.

The MHA extends the traditional attention mechanism to have multiple heads, so that information extraction can be carried out in different subspaces of hidden representation. Considering the corresponding coefficients of encoding results and decoding states from a multi-dimensional perspective, the results are more accurate and representative. Therefore, the Multi-Head Attention mechanism can improve the recognition performance of the model.

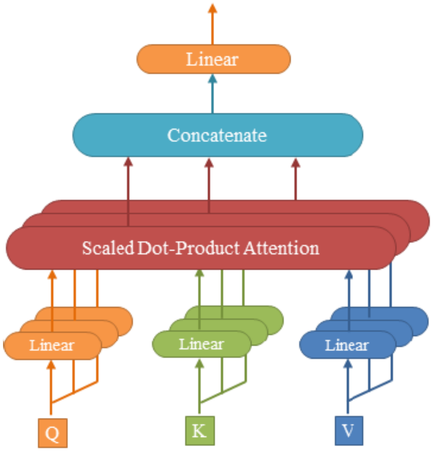

FIGURE 3. Mechanism of Multi-Head Attention.

D. LABEL SMOOTHING REGULARIZATION

In the classification problem, the label of a sample is usually encoded with one-hot format. More specially, N categories are encoded using n-bit state registers. Each category has its own register bit and only one bit is valid at each time. The labels with one-hot format give the probability distribution of the real samples, and it is very convenient for us to calculate the cross-entropy. However, some problems will be introduced when the model prediction probability is used to fit the real probability of one-hot format. For example, it leads models to over-believing prediction results and reduces the generalization ability. The label smoothing technique is adopted to solve such problems. Specifically, a fixed probability distribution, independent of the input sequence, is introduced to smooth the real probability. The operation is expressed as

$$Q'(y|x) = (1-e) \cdot Q(y|x) + e \cdot U(k), \tag{8}$$

where $Q(y|x)$ denotes the actual probability distribution of the sample labels, e is a smoothing factor, and $U(k)$ denotes a fixed and known probability distribution. A new distribution $Q'(y|x)$ is formed by mixing the real distribution $Q(y|x)$ and the fixed distribution $U(k)$ according to the weight of $1-e$ and e. This operation is equivalent to adding noise to the actual distribution of sample labels. In other words, when the label is 1, we no longer treat 1 as a training objective but replace it with a number closer to 1. The same is true when the label is 0. Using the label smoothing technology to train the model can alleviate the problem of over-fitting caused by one-hot encoding and improve the generalization ability of the model [38].

In the experimental process of this article, the LAS model adopts the cross-entropy as the objective function. Its optimization goal is to make the predicted probability distribution approaching the real probability distribution as much as possible. In this case, the cross-entropy optimization makes the model over-focused on the category with larger probability, the small probability samples are gradually ignored by the model. Therefore, a certain degree of deviation is generated and the generalization ability of the model is reduced. The Amdo-Tibetan is a low-resource language, and the training data is inadequate. This is more likely lead to over-fitting. Therefore, the label smoothing technique is introduced in the training process of the LAS model.

E. EXTERNAL LANGUAGE MODEL

The Speller can learn the language information from training data; however, the amount of training data is insufficient for the training of the language model. Fortunately, combining the LAS model with an external language model could increase the accuracy of speech recognition [39], [40]. In this article, an N-gram language model is chosen as an external language model. Specifically, the decoding path is decided by the probability score of the LAS model and that of N-gram language model. Namely

$$Score = S_{LAS} + \alpha \cdot S_{LM}, \tag{9}$$

where α is an adjustable parameter, and it represents the proportion of the score of the N-gram language model in all scores.

F. DISCRIMINATIVE TRAINING

The maximum likelihood function is usually used as the objective function in the LAS model. However, in the actual optimization process, such an optimal objective may not achieve the expected results. Specifically, in the training process, when the value of the loss function is smaller, the WER is not necessarily lower. Discrimination training can alleviate this problem by using the solution of the traditional speech recognition system for reference [41]–[43].

In this article, the Maximum Mutual Information (MMI) criterion is used for discriminative training. The optimization objective of the MMI criterion is a fractional value in which the numerator is the probability of the correct prediction while the denominator is the sum of the probabilities of all the wrong predictions. In the optimization process, the numerator is maximized and the denominator is minimized. In this way, minimizing the value of the objective function can lead to minimizing the WER. During the actual optimization process, it is infeasible to calculate for all the error probabilities in the denominator, therefore, the N-best list is used to estimate all error probabilities. The loss function of MMI criterion is calculated as

$$L_{MMI}^{N-best}(x, y) = \frac{P(y|x)}{\sum\limits_{y_i \in N-best \& y_i \neq y} P(y_i|x)}. \tag{10}$$

To get better training effect, MMI criterion and cross-entropy are interpolated to the loss function. The loss function after interpolation is

$$L_{MMI} = \sum_{(x,y)} (1-\lambda) L_{MMI}^{N-best}(x, y) + \lambda L_{CE}. \tag{11}$$

G. SUMMARY OF MODEL COMPLEXITY

The complexity of artificial neural network includes space complexity and time complexity. The former determines the training/prediction time of the neural network. The latter determines the number of parameters in the neural network.

In this study, an Amdo-Tibetan speech recognition system is established based on the LAS model. Its encoder module is constituted of the pBLSTM networks, as described in section III. This structure greatly reduces the number of parameters and, thereby, reduces the spatial complexity of the model. In the meantime, the beam search strategy is adopted in the decoding process of the model, and the time complexity of the model is reduced by controlling the value of beam size. In the two ways, the complexity of the LAS model is reduced and the convergence speed of the model is well controlled.

IV. TRANSFER LEARNING

Ideally, the recognition ability of an ASR system can be achieved that of a human being. However, the construction of ASR system with good performance needs a large amount of training data. For the low-resource language, speech recognition faces the challenge of insufficient resources. Based on the similarity between data and tasks, transfer learning can apply models learned in the old domain to the learning process of the new problem. Therefore, transfer learning has been widely used to improve the performance of speech recognition for data deficient languages [44].

In 2014, Yosinsk of Cornell University carried out a study on the portability of deep neural networks based on ImageNet data sets [45], [46]. The results show that: (1) with the help of transfer learning, it is better to use an existing network than a neural network whose weights are randomly initialized and trained with a small amount of data; (2) fine-tuning in neural network parameters can achieve better results of transfer learning. Based on the above conclusions, transfer learning is employed to improve the performance of Amdo-Tibetan speech recognition system in this study. Specifically, the pre-trained model with abundant Chinese and English resources is applied to the Amdo-Tibetan speech recognition through transfer learning. Fig. 4 illustrates the flow chart of the transfer learning in this study.

Transfer learning brings several advantages to the proposed model: firstly, the model is pre-trained by using the source language with rich corpus such as Chinese and English, which ensures the recognition effect of the model; secondly, the corpus of the target language: Amdo-Tibetan in this study is used to fine-tune the trained model. It is equivalent to expanding the Amdo-Tibetan corpus and contributes to the robustness

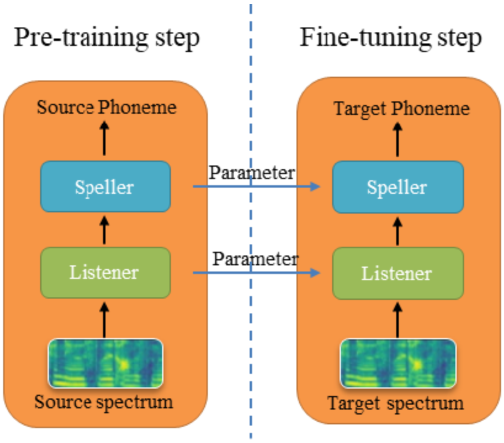

FIGURE 4. Schematic diagram of transfer learning.

of the model; finally, the training model is no longer trained from the beginning and thus reduces the training time.

V. EXPERIMENT AND DISCUSSION

The experiments are mainly divided into two parts: the experiments related to the LAS model and those related to transfer learning. Concretely, the experiments related to the LAS model include performance testing for baseline model and effect testing for various optimization technologies. The experiments related to transfer learning include the tests to assess the influence of language similarity and different source language data volume on the performance of transfer learning.

A. DATABASE

The experiments are carried out on three corpora: English corpus, Chinese corpus, and Amdo-Tibetan corpus. In the experiments related to the LAS model, the Amdo-Tibetan corpus is used directly to train the model. In the experiments related to transfer learning, English and Chinese are source languages while Tibetan is the target language. During the training process, the corpus of the source language is used to pre-train the model while that of the target language is used to fine-tune the model.

The Amdo-Tibetan corpus used in this article is recorded by 12 speakers, and the sampling rate is 16 kHz with the mono channel. The corpus contains 8,400 pieces of speech with a total of 14 hours, and 7,500 sentences with a total of 13 hours are selected as the training set while the remaining 900 sentences with a total of 1 hour are used as the test set.

To decide which language is more suitable as the source language of transfer learning, a 150-hour Chinese corpus is selected from AISHELL and a 150-hour of English corpus is selected from LibriSpeech.

B. SETTING

Firstly, a baseline LAS model is implemented, on which the beam search strategy is tested to get the best decoding efficiency. Secondly, the model is optimized by several technologies such as the Multi-Head Attention mechanism, label smoothing, external language model, and discriminative training. Finally, the transfer learning experiment is carried out on the optimized model.

The Listener is structured with 512 nodes of 3 layers pBLSTM (i.e., 256 nodes per direction), which reduces the time resolution by 8 times. The Speller uses two layer one-way LSTM with 512 nodes each. The attention mechanism adopts feed-forward neural network. This model uses the cross-entropy function as loss function to update network parameters. In the training phase, the Adam algorithm is employed for optimization and the initial learning rate is set to 0.001.

The 40-dimensional log-Mel filter bank features are used as inputs feature. The frame length is 25ms and frame shift is 10ms. The adjacent ±2 frames are spliced for each frame. 200 dimensional concatenated features with a total of five frames are used as the input of the current frame. Phoneme error rate (PER) is regarded as the final evaluation criterion.

C. EXPERIMENT ON LAS MODEL

1) DECODING STRATEGY

The beam search strategy could speed up the convergence rate and improve the generalization ability of the LAS model. To this strategy, the determination of the beam size is a key issue. In this section, the beam search strategy is validated with the baseline LAS model. The influence of various beam sizes on decoding results is shown in Table 1 and Fig. 5.

TABLE 1. Results under different beam sizes.

Model	Baseline-LAS				
Strategy	Greedy Search	Beam Search			
Beam size	1	5	10	15	20
Test-PER	45.32	43.18	42.21	42.10	42.04

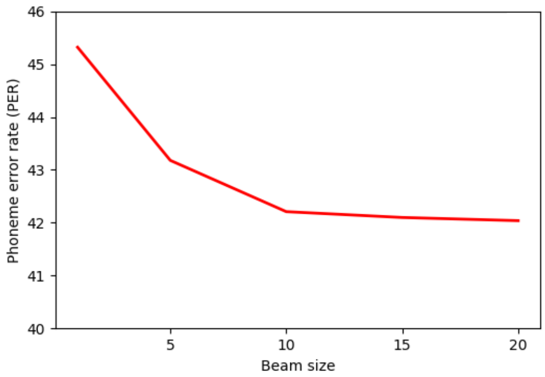

FIGURE 5. Phoneme error rates under different beam sizes.

The HMM framework or CTC framework for speech recognition often retain thousands of possible paths in the decoding process. It can be seen from Table 1 that, in the

LAS model, the beam size is much smaller than that in traditional speech recognition model and CTC model. The best recognition result can be obtained only by keeping the possible path of single digits. The sharp reduction of candidate paths simplifies the decoder framework and improves the decoding speed greatly. Meanwhile, it can be seen from Fig. 5 that the PER decreases sharply as the beam size is adjusted from 1 to 5 and 5 to 10; and it decreases gently as the beam size is adjusted from 10 to 20. Considering the balance between system performance and convergence time, it is appropriate to set the beam size to be 10 in subsequent experiments.

2) EFFECTIVENESS OF MULTI-HEAD ATTENTION

The effectiveness of the Multi-Head Attention mechanism, proposed in the section III of this article, is verified in this section.

Table 2 lists the experimental results on the baseline LAS model and LAS model with Multi-Head Attention. It can be seen that, the PER on the training set decreases gradually with the increase of the number of heads. The PER on the test set also decreases gradually at first, but it begins to increase when the number of headers reaches 8. The change curve of PRE is shown in Fig. 6. The basic reason is that expansion of parameters leads to over-fitting of the model.

TABLE 2. Results of Multi-Head Attention mechanism.

Model	Heads	Train-PER	Test-PER
Baseline-LAS	1	27.27	42.21
Multi-Head-LAS	2	26.81	41.16
Multi-Head-LAS	4	26.34	40.23
Multi-Head-LAS	8	25.96	40.88

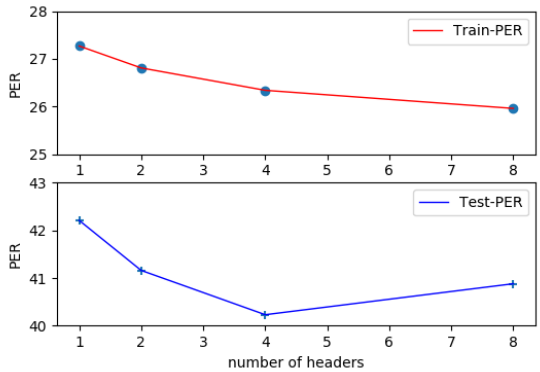

FIGURE 6. PER on training and test set under different number of head.

The advantage of the Multi-Head Attention mechanisms is that the correlation between the state vector of encoder and decoder is quantified from more dimensions. This makes the context vector generated by the attention module has more discriminant information. Nevertheless, it can make the model parameters too large, which makes the model difficult to converge or produce over-fitting. Therefore, increasing the dimension of representation subspace in a certain range can improve recognition performance. If the dimension is too high, it may lead to the risk of over-fitting.

In order to reduce the degree of over-fitting, the label smoothing regularization (LSR) is introduced in the experiments. Uniform distribution is adopted as the fixed probability distribution to realize the label smoothing strategy, and the smoothing ratio is set to be 0.1. The experimental results are shown in Table 3.

TABLE 3. Results of label smoothing regularization.

Model	Heads	Train-PER	Test-PER
Baseline-LAS+LSR	1	27.04	41.66
Multi-Head-LAS+LSR	2	26.48	40.75
Multi-Head-LAS+LSR	4	25.86	39.84
Multi-Head-LAS+LSR	8	25.47	40.42

Comparing the results of Table 2 and Table 3, it can be seen that the lower PER can be obtained by model with the label smoothing normalization under the same parameters. When the Multi-Head LAS model is set to 4 heads, the best PER is obtained on the test set. Therefore, the number of head should be set to 4 in subsequent experiments.

3) EXTERNAL LANGUAGE MODEL

In order to improve the recognition accuracy, an N-gram language model is combined with the LAS model. This section tests the effect of different weights of N-gram language model on the Multi-Head Attention LAS model with labels smoothing. The results are shown in Table 4.

TABLE 4. Result under different LM weights. MHLAS-LSR refers to Multi-Head Attention LAS model with labels smoothing.

Model	MHLAS-LSR	MHLAS-LSR +LM		
α	0	0.1	0.2	0.3
TEST-PER	39.84	39.17	39.09	39.26

In the HMM framework, the weight of the N-gram language model is generally between 10 and 20. In the CTC framework, the weight of N-gram language model is generally between 1.0 and 2.0. From Table 4, it can be seen that the weight of N-gram language model is in between 0.1 and 0.3. This indicates that, compared with traditional models, the LAS model is much less dependent on the language model. In addition, it can be seen that the recognition accuracy of the model is improved with the help of the external language model and the weight of the external language model affects the recognition accuracy. In the experiments, the best recognition result is obtained when the weight sets as 0.2.

4) DISCRIMINATIVE TRAINING

The loss function of the LAS model with discriminative training is calculated with Eq. (11), where λ and N-best list are the adjustable super parameters. Experiments in this section are designed to explore the influence of λ and the N-best list on discriminative training, as shown in Table 5 and Fig. 7.

TABLE 5. Result of discrimination training.

	λ	N-best	TEST-PER
MMI	0.1	**8**	**38.64**
		10	38.91
		12	39.03
	0.3	8	38.56
		10	**38.42**
		12	38.67
	0.5	8	39.04
		10	38.97
		12	**38.86**

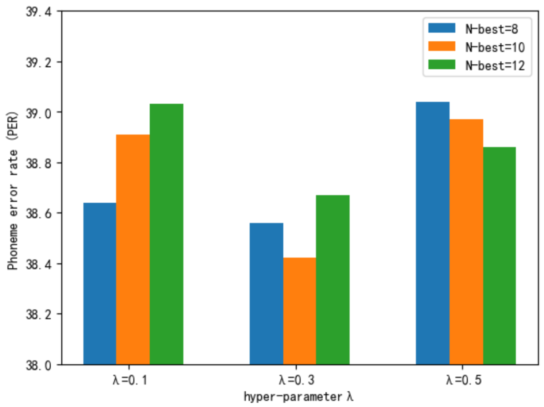

FIGURE 7. Phoneme error rate under different hyper-parameter during discriminative training.

Previous experiments show that good performance and decoding speed can be achieved when the beam size is set to 10. Therefore, for each value of λ, three beam sizes, namely 8, 10 and 12, are set respectively. From the results of Table 5 we can see that the recognition accuracy of the model has been improved by introducing discrimination training. Meanwhile, it can be seen from Fig. 7 that the accuracy of model varies with the proportion setting of the cross-entropy loss function, and the number of paths retained in the N-best list is also different with the value of λ. The best performance of discrimination training is obtained when the λ value is set to 0.3 and the N-best list is set to 10.

D. EXPERIMENT ON TRANSFER LEARNING

This section verifies the effectiveness of the transfer learning to Amdo-Tibetan speech recognition from two different perspectives, namely, language similarity and volume of the source language data. The best model, abbreviated as MHLAS-LLM, is selected to examine the effect of transfer learning, where MHLAS-LLM refers to the Multi-Head Attention LAS model with Labels Smoothing, external Language Model, and discriminative training.

1) THE INFLUENCE OF LANGUAGE SIMILARITY

In this section, three experiments are implemented to verify the influence of language similarity on transfer learning. The first experiment utilizes only the Amdo-Tibetan corpus to train the LAS model. In the second experiment, the LAS model is pre-trained with 150 hours of Chinese corpus, and then the model is fine-tuned with Amdo-Tibetan corpus. During the pre-training, syllables are selected as modeling units. In the third experiment, the LAS model is pre-trained with 150 hours of English corpus, and then the model is fine-tuned with Amdo-Tibetan corpus. Phonemes are selected as modeling units in the pre-training stage. The experimental results are shown in Table 6.

TABLE 6. The influence of language similarity.

Model	MHLAS-LLM		
No.	1	2	3
Pre-training	Rand init.	Chinese	English
Fine-tuning	Amdo-Tibetan		
TEST-PER	38.42	35.78	36.64

Table 6 shows that the PER of the test set is the highest when the random initial network parameters are trained with Amdo-Tibetan data directly. For transfer learning, the PER decreases obviously no matter English or Chinese is used for pre-training. Compared with the direct training model of Amdo-Tibetan data, the pre-trained model using Chinese corpus has an absolute improvement of 2.64%. This indicates that transfer learning can effectively enhance the performance of the model. Moreover, the effect of pre-training using Chinese is better than that of English. To the target language i.e. Amdo-Tibetan, Chinese is more suitable for pre-training than English. The reason may be that Chinese and Tibetan have some similarities in syntactic structure, and English and Tibetan are quite different in grammar and pronunciation.

2) THE INFLUENCE OF DIFFERENT VOLUME

This section verifies the influence of the volume of source language data for Tibetan transfer learning in pre-training stage. Three experiments are organized and all of them use Chinese data to pre-train the LAS model. The model is pre-trained with 50 hours, 100 hours, and 150 hours of

Chinese corpus respectively at first, and then that is fine-tuned with Amdo-Tibetan corpus. The results are shown in Table 7.

TABLE 7. Quantitative influence of source language data.

Model	MHLAS-LLM		
Pre-training	Chinese		
Data size (hours)	50	100	150
Fine-tuned	Amdo-Tibetan		
TEST-PER	37.24	36.71	35.78

It can be inferred from Table 7 that, with the increase of Chinese pre-training data, there is a decline of PER of Tibetan speech recognition. The performance of the model with 150-hour corpus for pre-training is 1.46% higher than that of 50-hour corpus. At the same time, compared with the previous experiment, it comes to a conclusion that if there are higher similarities between the source language and the target language, the difficulties of transfer learning between the two languages can be reduced. It could achieve good results in transfer learning even utilizing less source language data.

VI. CONCLUSION

An end-to-end model was proposed for Amdo-Tibetan speech recognition as it can provide a simple and effective solution for the construction of Tibetan speech recognition system. And then, the performance of the system was optimized by several techniques such as Multi-Head Attention mechanism, label smoothing, external language model, and discriminative training. In the meantime, Amdo-Tibetan is a low-resource language; especially its linguistic resources and corpus are very limited. To solve this problem, transfer learning was introduced. By using Chinese and English corpus to pre-train the model, it reduced the impact on the model performance due to inadequate of Amdo-Tibetan training data. The experimental results showed that the proposed end-to-end model improved the recognition performance of Amdo-Tibetan significantly.

In the future, data enhancement methods will be explored to improve the performance of Amdo-Tibetan speech recognition. In addition, improving the robustness and generalization of the Amdo-Tibetan speech recognition system is also the future research direction.

ACKNOWLEDGMENT

The authors would like to thank Prof. Yasser Morgan from University of Regina and Prof. Henry Han from Fordham University for reviewing an early draft of this article. Thanks also to Huiyun Zhang and Jie Kang from Qinghai Normal University for their useful discussion.

REFERENCES

[1] Y. Wang, W. Zhou, J. Luo, H. Yan, H. Pu, and Y. Peng, "Reliable intelligent path following control for a robotic airship against sensor faults," *IEEE/ASME Trans. Mechatronics*, vol. 24, no. 6, pp. 2572–2582, Dec. 2019.

[2] M. Elforjani and S. Shanbr, "Prognosis of bearing acoustic emission signals using supervised machine learning," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5864–5871, Jul. 2018.

[3] V. Puri, S. Jha, R. Kumar, I. Priyadarshini, L. H. Son, M. Abdel-Basset, M. Elhoseny, and H. V. Long, "A hybrid artificial intelligence and Internet of Things model for generation of renewable resource of energy," *IEEE Access*, vol. 7, pp. 111181–111191, 2019.

[4] J. Maier, A. Naber, and M. Ortiz-Catalan, "Improved prosthetic control based on myoelectric pattern recognition via wavelet-based de-noising," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 506–514, Feb. 2018.

[5] R. Sahraeian and D. Van Compernolle, "Cross-entropy training of DNN ensemble acoustic models for low-resource ASR," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 1991–2001, Nov. 2018.

[6] J. Kang, W.-Q. Zhang, and J. Liu, "Gated convolutional networks based hybrid acoustic models for low resource speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 157–164.

[7] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "Flat-start single-stage discriminatively trained HMM-based models for ASR," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 1949–1961, Nov. 2018.

[8] J. Liu and W. Q. Zhang, "Research progress on key technologies of low resource speech recognition," *J. Data Acquisition Process.*, vol. 32, no. 2, pp. 205–220, 2017.

[9] J. Yi, J. Tao, Z. Wen, and Y. Bai, "Adversarial multilingual training for low-resource speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4899–4903.

[10] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1240–1253, Dec. 2017.

[11] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, "Sequence-based multi-lingual low resource speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4909–4913.

[12] G. Krishna, C. Tran, J. Yu, and A. H Tewfik, "Speech recognition with no speech or with noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1090–1094.

[13] L. Li, D. Wang, Y. Chen, Y. Shi, Z. Tang, and T. F. Zheng, "Deep factorization for speech signal," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5094–5098.

[14] F. Tao and C. Busso, "Gating neural network for large vocabulary audio-visual speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 7, pp. 1290–1302, Jul. 2018.

[15] Q. Wang, W. Guo, and C. Xie, "Towards end to end speech recognition system for tibetan," *Pattern Recognit. Artif. Intell.*, vol. 30, no. 4, pp. 359–364, Apr. 2017.

[16] X. Huang and J. Li, "The acoustic model for tibetan speech recognition based on recurrent neural network," *J. Chin. Inf. Process.*, vol. 32, no. 5, pp. 49–55, May 2018.

[17] W. Zhang, H. Yang, X. Bu, and L. Wang, "Deep learning for mandarin-tibetan cross-lingual speech synthesis," *IEEE Access*, vol. 7, pp. 167884–167894, 2019.

[18] 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.

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

[20] R. Li, X. Wang, S. H. Mallidi, S. Watanabe, T. Hori, and H. Hermansky, "Multi-stream end-to-end speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 646–655, 2020.

[21] C. Qin and L. Zhang, "Deep neural network based feature extraction for low-resource speech recognition," *Acta Automatica Sinica*, vol. 43, no. 7, pp. 1208–1219, 2017.

[22] Q. De, "A review of tibetan speech recognition," *J. Tibet Univ.*, vol. 25, pp. 192–195, May 2010.

[23] H. Huang, L. Ma, and W. Zhao, "Tibetan character set standard," in *Research on Handwritten Tibetan Character Recognition*, 1st ed. BeiJing, China: Science Press, 2016, pp. 4–9.

[24] Z. Cai and R. Cai, "Study on the distribution of tibetan font structure," *J. Chin. Inf. Process.*, vol. 30, no. 4, pp. 98–205, 2016.

[25] J. Yi, J. Tao, Z. Wen, and Y. Bai, "Language-adversarial transfer learning for low-resource speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 3, pp. 621–630, Mar. 2019.

[26] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 8614–8618.

[27] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4960–4964.

[28] T. Moriya, T. Tanaka, T. Shinozaki, S. Watanabe, and K. Duh, "Evolution-strategy-based automation of system development for high-performance speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 1, pp. 77–88, Jan. 2019.

[29] A. Tjandra, S. Sakti, and S. Nakamura, "End-to-end speech recognition sequence training with reinforcement learning," *IEEE Access*, vol. 7, pp. 79758–79769, 2019.

[30] Y. Takashima, R. Takashima, T. Takiguchi, and Y. Ariki, "Knowledge transferability between the speech data of persons with dysarthria speaking different languages for dysarthric speech recognition," *IEEE Access*, vol. 7, pp. 164320–164326, 2019.

[31] J. Kim, M. El-Khamy, and J. Lee, "Bridgenets: Student-teacher transfer learning based on recursive neural networks and its application to distant speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5719–5723.

[32] Y. Shi, J. Bai, P. Xue, and D. Shi, "Fusion feature extraction based on auditory and energy for noise-robust speech recognition," *IEEE Access*, vol. 7, pp. 81911–81922, 2019.

[33] S. Deena, M. Hasan, M. Doulaty, O. Saz, and T. Hain, "Recurrent neural network language model adaptation for multi-genre broadcast speech recognition and alignment," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 3, pp. 572–582, Mar. 2019.

[34] P. Agrawal and S. Ganapathy, "Modulation filter learning using deep variational networks for robust speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 244–253, May 2019.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, and Ł. Kaiser, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[36] B. Yusuf, B. Gundogdu, and M. Saraclar, "Low resource keyword search with synthesized crosslingual exemplars," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1126–1135, Jul. 2019.

[37] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 92–102, Apr. 2018.

[38] Y. Zheng, J. Tao, Z. Wen, and J. Yi, "Forward–backward decoding sequence for regularizing End-to-End TTS," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 12, pp. 2067–2079, Dec. 2019.

[39] Q. Zhang and J. H. L. Hansen, "Language/dialect recognition based on unsupervised deep learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 5, pp. 873–882, May 2018.

[40] M. Sundermeyer, H. Ney, and R. Schluter, "From feedforward to recurrent LSTM neural networks for language modeling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 3, pp. 517–529, Mar. 2015.

[41] Y.-H. Tu, J. Du, and C.-H. Lee, "Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 12, pp. 2080–2091, Dec. 2019.

[42] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, Mar. 2017.

[43] R. Hasan, H. Hussein, P. Lazaridis, S. Khwandah, M. Ritter, and M. Eibl, "Improvement of speech recognition results by a combination of systems," in *Proc. 23rd Int. Conf. Autom. Comput. (ICAC)*, Sep. 2017, pp. 1–4.

[44] G. Kim, H. Lee, B.-K. Kim, S.-H. Oh, and S.-Y. Lee, "Unpaired speech enhancement by acoustic and adversarial supervision for speech recognition," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 159–163, Jan. 2019.

[45] B. Myagmar, J. Li, and S. Kimura, "Cross-domain sentiment classification with bidirectional contextualized transformer language models," *IEEE Access*, vol. 7, pp. 163219–163230, 2019.

[46] D. Liu, J. Xu, P. Zhang, and Y. Yan, "Investigation of knowledge transfer approaches to improve the acoustic modeling of vietnamese ASR system," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 5, pp. 1187–1195, Sep. 2019.

XIAOJUN ZHU received the B.S. degree in automation from East China Jiaotong University, in 2000, and the M.S. degree in computer science and technology from Lanzhou Jiaotong University, in 2011. He is currently pursuing the Ph.D. degree in computer science and technology with the College of Computer Science and Technology, Qinghai Normal University, Xining, China.

His research interests include speech signal processing, deep learning, and Tibetan speech recognition.

HEMING HUANG received the B.S. degree in mathematics from Shaanxi Normal University, in 1992, the M.S. degree in computer science and technology from the College of Information Science and Technology, Lanzhou University, in 2004, and the Ph.D. degree in pattern recognition and intelligent system from the College of Automation, Southeast University, in 2014.

He is currently a Professor with the College of Computer Science and Technology, Qinghai Normal University. He edited a book titled *Research on Handwritten Tibetan Character Recognition* (2016). He has published more than 60 articles on major journals and conferences and holds two patents. His research interests include speech recognition, character recognition, and Tibetan information processing.