*Computers, Materials & Continua*
DOI: 10.32604/cmc.2022.027092
***Article***

Tech Science Press

# Cross-Language Transfer Learning-based Lhasa-Tibetan Speech Recognition

**Zhijie Wang[1], Yue Zhao[1,*], Licheng Wu[1], Xiaojun Bi[1], Zhuoma Dawa[2] and Qiang Ji[3]**

[1]School of Information Engineering, Minzu University of China, Beijing, 100081, China
[2]School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China, Beijing, 100081, China
[3]Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA
*Corresponding Author: Yue Zhao. Email: zhaoyueso@muc.edu.cn
Received: 10 January 2022; Accepted: 30 March 2022

**Abstract:** As one of Chinese minority languages, Tibetan speech recognition technology was not researched upon as extensively as Chinese and English were until recently. This, along with the relatively small Tibetan corpus, has resulted in an unsatisfying performance of Tibetan speech recognition based on an end-to-end model. This paper aims to achieve an accurate Tibetan speech recognition using a small amount of Tibetan training data. We demonstrate effective methods of Tibetan end-to-end speech recognition via cross-language transfer learning from three aspects: modeling unit selection, transfer learning method, and source language selection. Experimental results show that the Chinese-Tibetan multi-language learning method using multi-language character set as the modeling unit yields the best performance on Tibetan Character Error Rate (CER) at 27.3%, which is reduced by 26.1% compared to the language-specific model. And our method also achieves the 2.2% higher accuracy using less amount of data compared with the method using Tibetan multi-dialect transfer learning under the same model structure and data set.

**Keywords:** Cross-language transfer learning; low-resource language; modeling unit; Tibetan speech recognition

## 1 Introduction

With the training data of fewer than 30 h, the speech recognition system usually is called a low-resource system. Because it lacks sufficient training data of target languages, the low-resource system has a poor performance on recognition accuracy [1]. In recent years, with the development of deep learning technology, speech recognition technology has become increasingly mature and widely used in major languages such as Chinese and English. In contrast, Tibetan speech recognition technology has received some limitations, such as lack of speech data and linguistic resources. In view of the above problems, this paper uses the transfer learning method to add Chinese and English data in training data which can make up for the lack of Tibetan data and also uses end-to-end model, because it avoids the need for linguistic resources like dictionaries and phonetic knowledge.

The latest research results show that transfer learning has received extensive attention from researchers and has been used in many fields with satisfactory results. For instance, Zia et al. [2] use transfer learning to classify citrus plant diseases. Faisal et al. [3] and Reda et al. [4] studied Corona Virus Disease 2019 (Covid-19) diagnosis technology based on transfer learning. Fu et al. [5] completed the detection of malware based on the Long-Short Term Memory (LSTM) model through the transfer learning method. Xu et al. [6] use this method to recognize weeds. Besides, transfer learning has increasingly been used in speech recognition fields to improve the performance of acoustic models for low-resource language. Yang [1] transferred the weights of the first five hidden layers of Deep Neural Networks-Hidden Markov Model (DNN-HMM) trained with Chinese speech data for low-resource Uyghur speech recognition, which reduced the Word Error Rate (WER) to 18.75%. Li [7] used the Deep Feedforward Sequential Memory Networks-Connectionist Temporal Classification (DFSMN-CTC) acoustic model pre-trained with 10,000 h of Chinese data and fine-tuned with Uyghur speech data, which decreased the WER of Uyghur speech recognition to 5.87%.

In Tibetan speech recognition, transfer learning has also been used to improve the accuracy. Kang [8] pre-trained Visual Geometry Group Net (VGGNet) on two Chinese corpora THCHS-30 and ST-CMDS, to get better initialization weights, and then he used the Pre-training + Bi-LSTM + CTC model as the Amdo-Tibetan speech recognition model. The experimental results show that the CER is reduced to 26.6% by using the transfer learning method and decoding with CTC. Yan et al. [9] transferred the hidden layers of the Chinese speech recognition model to the Lhasa-Tibetan model. Based on the semi-orthogonal factorization of the acoustic model of Time-Deep Neural Networks-Hidden Markov Model (TDNN-HMM), as a result, the Lhasa-Tibetan CER is reduced by 14.74%. The existing works not only use high-resource languages as source languages, such as Chinese and English, but also use a low-resource language similar to the target language. Yue [10] applied the Lhasa-Tibetan dialect model as the initial model and fine-tuned the model with Amdo-Tibetan dialect data. The experimental results showed that the CER was reduced by 4% compared with the dialect-specific model using only Amdo-Tibetan dialect data. The study [11] based on Wavenet-CTC end-to-end model studied the multi-dialect and multi-task transfer learning and used Amdo-Tibetan dialect as training data to improve the accuracy of Lhasa-Tibetan speech recognition. As a result, the CER was reduced by 2.7%.

Modeling units are essential for transfer learning. Wang et al. [12] used phonemes as modeling units and evaluated the different phoneme sets. Their work found that the phoneme set with consonant suffix and long vowel as the modeling unit is better than other phoneme sets. For Tibetan-Mandarin bilingual speech recognition, Wang et al. [13] used characters instead of phonemes as the modeling unit for the end-to-end speech recognition model. This method provides a direction for the low-resource languages which lack pronunciation dictionaries to build a speech recognition model.

Inspired by the above works, this paper explores the Lhasa-Tibetan end-to-end speech recognition model based on transfer learning in three folds: transfer learning method, modeling units, and source language selection. First, multi-language learning is compared with pre-training technology to build the transfer model. Second, in terms of the modeling units, four kinds of modeling units are evaluated, i.e., the Latin letter set, the multi-language character set, the Chinese syllable and the Tibetan letter set, and the Latin and Tibetan letter set. Finally, for source data, Chinese, English, and Chinese-English mixed data are used as the source languages for transfer learning. By studying different kinds of combinations of the above three, we optimize the Lhasa-Tibetan speech recognition model using a small amount of Tibetan training data.

The rest of this paper is organized as follows: Section 2 introduces the data sets used in this paper and audio data processing process. We detail the technical principles used in this paper and the processing of the text data used in the experiments in Section 3. Experiments are explained in detail and their results are discussed in Section 4. Finally, we describe our conclusions in Section 5.

## 2 Data

### *2.1 Data Sets*

Lhasa-Tibetan dialect data comes from an open Tibetan multi-dialect speech data set, TIBMD@MUC [14], which is used in the work of [10,11]. The text data of this data set consists of two parts: one part is 1369 sentences of Tibetan spoken language selected from the book of "Spoken Tibetan Language", and the other part is 8000 sentences of news, electronic novels, and Tibetan poetry collected on the Internet. The recorders are some college students. The Lhasa dialect data is divided into 2.7 h of training data and 0.58 h of test set.

The English data comes from LibriSpeech Automatic Speech Recognition (ASR) corpus. This data set is a large corpus containing about 1000 h of English speech. The data comes from audio books from the LibriVox project. In this paper we use 34.5 h of data as training data. The Chinese data comes from the open-source THCHS-30, and the training data duration is 31.5 h. The text of THCHS-30 is selected from large-capacity news, and most of the people involved in the recording are college students who can speak fluent Mandarin. Data statistics are shown in Tab. 1.

**Table 1:** Statistics on three languages

| | Training data (h) | Training utterances | Testing data (h) | Testing utterances |
|---|---|---|---|---|
| English | 34.5 | 10000 | 0 | 0 |
| Chinese | 31.8 | 12495 | 0 | 0 |
| Lhasa-Tibetan | 2.7 | 3157 | 0.58 | 732 |

### *2.2 Data Processing*

All audio files in three languages are converted into Windows Audio Volume (WAV) format with 16KHz sampling rate and 16-bit quantization accuracy. In addition, 39 Mel Frequency Cepstrum Coefficient (MFCC) features of each observation frame are extracted from speech data using a 25 ms window with 10 ms overlaps.

## 3 Method

### *3.1 Source Language Selection*

There are many language families in the world, such as Sino-Tibetan, Indo-European, Semitic, etc. Among them, Indo-European and Sino-Tibetan are utilized most extensively [15]. English belongs to the former, while Chinese and Tibetan constitutes part of the latter.

For a language, pronunciation, grammar, and vocabulary are three elements of speech, and they are indispensable and interdependent [16]. In terms of phonetics, each syllable in the Sino-Tibetan language family has a fixed tone, distinguishing the meaning of vocabulary and grammar. However, most languages of the Indo-European language family have no tones, and only a few languages

have simply tones. About grammar, the Sino-Tibetan language family is an analytic language, which expresses grammatical relationships with word order and function words. Usually, the word order is relatively fixed. However, the Indo-European language family is an inflectional language, and there are many variations of verbs, nouns, and adjectives. These tortuous variations represent different meanings of sentences [17]. For vocabulary, most vocabularies have precise definitions in Sino-Tibetan. However, in Indo-European languages, the same word may mean verbs, nouns, and adjectives.

In summary, compared with English, Chinese has more similarities with Tibetan. Therefore, Chinese should be used as the source language of transfer learning for the Tibetan speech recognition model. In experiences, we evaluate the performance of Chinese, English, and Chinese-English mixed as the source language.

### *3.2 Modeling Units*

The end-to-end model integrates the traditional acoustic and language models into a single model with no lexicon. In our work, we adopted four kinds of modeling units for cross-language transfer learning end-to-end model. They are the Latin letter set, the multi-language character set including Tibetan characters, Chinese characters, and English words, the Chinese Pinyin and Tibetan letter set, and the Latin and Tibetan letter set.

For the Latin letter set, Tibetan characters are converted to Latin letter sequences by Wylie transliteration. Chinese characters are transcribed by Pinyin with no tone. After this, the texts of sentences are uniformly expressed in Latin letters for Chinese, English, and Lhasa-Tibetan. Meanwhile, in English text, we also convert the uppercase letters to their corresponding lowercase letters. Some punctuation marks in the sentence are deleted. Besides, there are some other processing, like that the text "I'm" is changed to "I am", and "H•M" is changed to "h m". As for the Lhasa-Tibetan text, the mark " ' " is replaced with "f", "+" with "q", and "." with "v". To be more precise, these substitutes of Latin letters have never appeared in the Wylie transliteration text for Lhasa-Tibetan.

Multi-language character set consists of Chinese characters, English words and Tibetan characters from all experimental texts. For Chinese Pinyin and Tibetan letter set, Chinese text is transcribed by Pinyin with no tone, and Tibetan characters are rewritten horizontally in Tibetan letters from left to right. This method is from the work of [13].

For Latin and Tibetan letter set, Chinese characters and English words are expressed by Latin letters, and Tibetan characters are transcribed in Tibetan letters. The Tibetan text example is shown in Fig. 1.

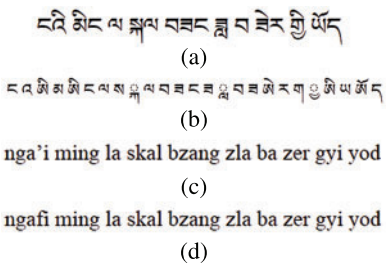


**Figure 1:** (a) Tibetan original text. (b) Tibetan letter text. (c) Tibetan Wylie transliteration text. (d) Tibetan Wylie transliteration text after processing

### *3.3 Transfer Learning Methods*

Transfer learning is based on the common sense between the source domain and target domain. Therefore, it can assist the learning of the target domain and avoid time consumption or unideal performance caused by learning from scratch [18]. According to the existence of labels in the source and target data, transfer learning is categorized as follows: fine-tuning, multi-task learning, domain-adversarial training, and zero-shot learning.

When the amount of target data is small and the amount of source data is large, fine-tuning method usually is used for training model. It pre-trains a model using source data and then puts the target data into the model to fine-tune the parameters.

For the fine-tuning method, it only pays attention to the recognition effect of the target language but does not attend to the performance on the source data. Multi-task learning is different from fine-tuning method. It completes multiple tasks in a model simultaneously and aims to improve the accuracy of each task. Multi-task learning is often applied for multi-language speech recognition. Dong et al. [19] found that the hidden feature layer of the multilingual speech recognition model contains some common features of human language. Huang et al. [20] used European languages to improve Chinese speech recognition. Through multi-task learning, the recognition performance of 50 h of Chinese data is the same as that of 100 h alone.

Domain-adversarial training is mainly used for the case that the data of target tasks is similar to the source tasks, such as speech recognition tasks with target data containing the noise and source data with noiseless data. Zero-shot learning is mainly used for two domains with quite different data. For example, the source picture has a cat and dog, and the target picture is a monkey.

In speech recognition, transfer learning is generally used for learning acoustic model because all languages have similarities in acoustic features. In this paper, we adopt an end-to-end model for transfer learning speech recognition. End-to-end model integrates the traditional acoustic model and language model into a single model, which does not need to create a pronunciation dictionary and a separate language model. Therefore, it is especially suitable for languages with a lack of language knowledge and data. We compare the fine-tuning method with the multi-language learning method for the end-to-end model to evaluate which way is better to transfer both acoustic and linguistic knowledge.

The fine-tuning method first uses a large amount of source language data for the pre-training end-to-end model and then retrains the pre-trained model with Tibetan data, as shown in Fig. 2. The multi-language learning method trains an end-to-end model using the joint data of Chinese, English, and Tibetan data, as shown in Fig. 3. We adopt the WaveNet-CTC as the end-to-end model in two transfer learning methods, as shown in Fig. 4.

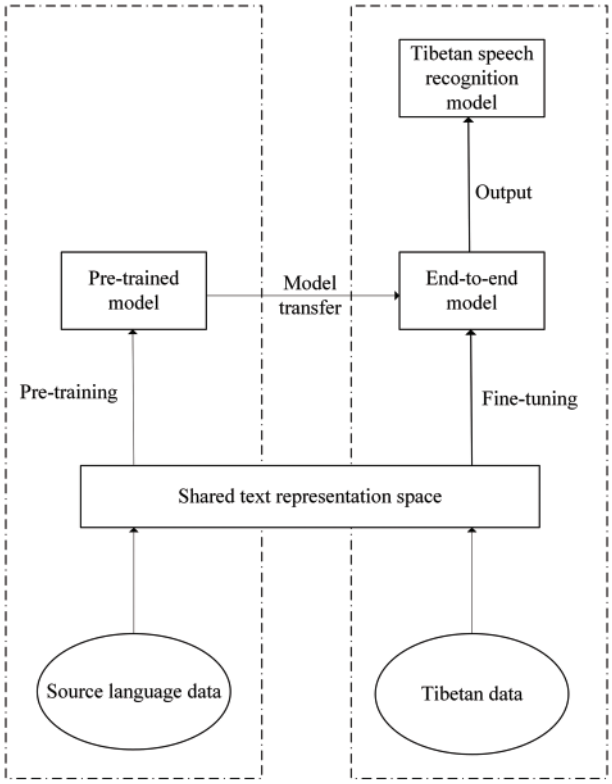


**Figure 2:** Tibetan speech recognition model based on fine-tuning transfer learning

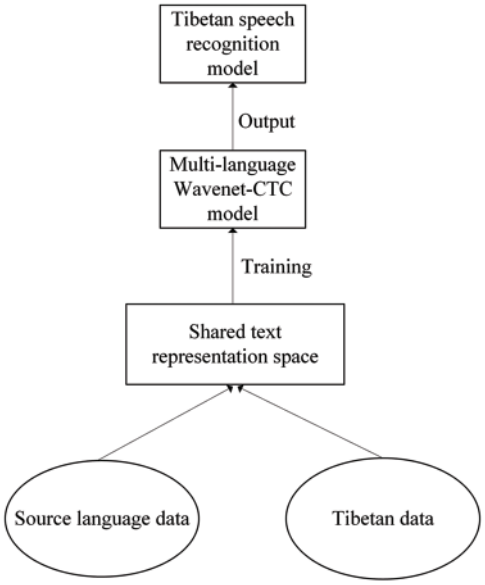


**Figure 3:** Tibetan speech recognition model based on multi-language learning

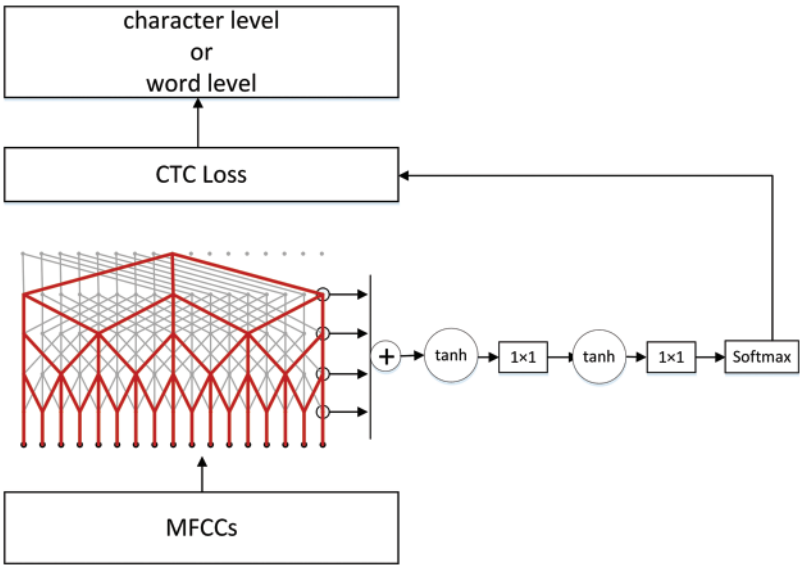


**Figure 4:** WaveNet end-to-end speech recognition model

## 4 Experiments

### *4.1 Experimental Setting*

In this paper, the WaveNet network has 15 layers, consisting of 3 dilated residual blocks stacks with 5 layers in each block. The dilation rates from the first layer to the fifth layer are 1,2,4,8,16. The filter size of causal dilated convolutions is 7. The number of hidden units in the gating layers is 128, the number of hidden units in the residual connection is 128, and the learning rate is $2 \times 10^{-4}$. The model has about 44 million parameters. Due to different modeling units, the number of model parameters is not exactly the same in each experiment. All models are trained for 50 epochs with the adaptive moment estimation (Adam) optimizer with a batch size of 10 on a Linux system loaded with two Nvidia RTX 2070 Super GPUs. All experiments are carried out under the above experimental configuration to ensure fairness when comparing experimental results.

In the experimental evaluation, the fine-tuning model uses Chinese, English, and Chinese-English mixed data as pre-training data, and Tibetan data is as fine-tuning data. The multi-language learning model uses Chinese-Tibetan mixed data, English-Tibetan mixed data, and Chinese-English-Tibetan mixed data as training data. Besides, there are two baseline models for comparison. One is the model which only uses 2.7 h Lhasa-Tibetan speech data as training data in our experiments. The other is the one from work [11] which used the Tibetan multi-dialect data including 4.4 h Lhasa-Tibetan for training. It is the multi-dialect learning model based on the same WaveNet-CTC structure, hyerparameters and Tibetan data set with our method.

### *4.2 Experimental Evaluation Criteria*

In this paper, we use edit distance to calculate the recognition error rate. The error rate(ER) is expressed as Eq. (1).

$$\mathrm{ER} = \frac{\mathrm{dic}\ (\mathrm{L}_1, \mathrm{L}_2)}{\mathrm{len}\ (\mathrm{L}_2)} \times 100\% \tag{1}$$

The $L_1$ represents the predicted text, $L_2$ represents the original text, dic(.) is the function to calculate the edit distance between texts, and len (.) is the function to calculate the sequence length. The Letter Error Rate (LER)is the letter-level ER of text transcribed by Latin letters, the Syllable Error Rate (SER) is single syllable ER of the text transcribed by Latin letters, the Tibetan CER is the character-level ER of text transcribed by Tibetan characters, and the Tibetan letter error rate is the letter-level ER of text transcribed by Tibetan letters.

### *4.3 Experimental Results and Analysis*

The experimental results of baseline models, the multi-language learning method, and the fine-tuning method are shown in Tabs. 2–4, respectively.

**Table 2:** The error rates of baseline models for Lhasa-Tibetan speech recognition

| Modeling unit | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Latin letter set | | Multi-language character set | Chinese Pinyin and Tibetan letter set | | Latin and Tibetan letter set | |
| Model | LER (%) | SER (%) | Tibetan CER (%) | Tibetan LER (%) | Tibetan CER (%) | Tibetan LER (%) | Tibetan CER (%) |
| Lhasa-Tibetan specific model | 28.7 | 55.8 | 53.4 | 46.7 | 71.2 | 46.7 | 71.2 |
| Multi-dialect model in work [11] | - | - | 29.5 | - | - | - | - |

**Table 3:** The error rates of multi-language learning model for Lhasa-Tibetan speech recognition

| Modeling unit | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Latin letter set | | Multi-language character set | Chinese Pinyin and Tibetan letter set | | Latin and Tibetan letter set | |
| Model | LER (%) | SER (%) | Tibetan CER (%) | Tibetan LER (%) | Tibetan CER (%) | Tibetan LER (%) | Tibetan CER (%) |
| Chinese-Tibetan | 28.9 | 58.4 | **27.3** | **30.2** | **54.5** | **29.3** | **53.2** |
| English-Tibetan | 30.2 | 58.3 | 57.6 | - | - | 33.1 | 59.3 |

(Continued)

**Table 3:** Continued

| Modeling unit | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Latin letter set | | Multi-language character set | Chinese Pinyin and Tibetan letter set | | Latin and Tibetan letter set | |
| Model | LER (%) | SER (%) | Tibetan CER (%) | Tibetan LER (%) | Tibetan CER (%) | Tibetan LER (%) | Tibetan CER (%) |
| Chinese-English-Tibetan | 28.6 | 56.9 | 58.9 | - | - | 30.8 | 55.6 |

**Table 4:** The error rates of fine-tuning model for Lhasa-Tibetan speech recognition

| Modeling unit | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Latin letter set | | Multi-language character set | Chinese Pinyin and Tibetan letter set | | Latin and Tibetan letter set | |
| Model | LER (%) | SER (%) | Tibetan CER (%) | Tibetan LER (%) | Tibetan CER (%) | Tibetan LER (%) | Tibetan CER (%) |
| Chinese data pre-training | 26.5 | 52.1 | 28.0 | 32.8 | 58.1 | 30.7 | 55.7 |
| English data pre-training | 27.4 | 54.2 | 36.4 | - | - | 30.6 | 54.7 |
| Chinese-English mixed data pre-training | **25.9** | **51.8** | 29.8 | - | - | 40.0 | 65.3 |

From Tabs. 2 and 4, we can find that the fine-tuning transfer learning model performs better than the Lhasa-Tibetan specific model. The model with Chinese data as pre-training data achieves the best recognition accuracy in Tab. 4, which is 1.5% higher than multi-dialect model. It also can be seen in Tabs. 2 and 3 that the performances of Chinese-Tibetan multi-language learning models are also better than the Lhasa-Tibetan specific model, except for the Latin letter set as the modeling unit. Using Chinese-English mixed data, the fine-tuning model achieves the best performance on the Latin letter set as a modeling unit. Compared with the Lhasa-Tibetan specific model, the LER and SER of the fine-tuning model are reduced by 2.8% and 4%, respectively. However, the multi-language learning model has better performance than the fine-tuning model on other modeling units.

The Chinese-Tibetan multi-language learning model in Tab. 3 achieves the best Tibetan CER among all models. Based on the multi-language character set, the Tibetan CER dropped by 26.1%

and 2.2%, respectively, compared with the Lhasa-Tibetan specific model and the multi-dialect model. Furthermore, on the modeling unit of the Chinese Pinyin and Tibetan letter set, the one of Latin and Tibetan letter set, the Tibetan CER of the Chinese-Tibetan multi-language model has decreased by 16.7% and 18%, respectively, compared with the Lhasa-Tibetan specific model. The end-to-end multi-language learning model using Chinese-Tibetan multi-language character set as the modeling unit is better than other methods. It achieves the lowest CER at 27.3%. Under the same model and experimental environment, it is 2.2% accuracy higher than the model using the multi-dialect learning method in work [11]. Therefore, compared with the work [11], the method in this paper achieved the higher accuracy using less amount of data.

The above experimental results show that Chinese as a source language is more suitable than English as a transfer language for Tibetan. This is because there are more similarities in pronunciation between Chinese and Tibetan languages. English belongs to the Indo-European language family, but Chinese and Tibetan both belong to the Sino-Tibetan language family, the same language family can share more knowledge in transfer learning. Meanwhile, the experimental results also show that the multi-language learning model, using Chinese-Tibetan joint training data and multi-language character sets as the modeling unit, has the lowest Tibetan character error rate. These show that the multi-language learning model shares both the acoustic features of speech and language knowledge of the grammar, vocabulary, et al. between languages. Although Lhasa-Tibetan speech recognition based on multi-dialect transfer learning method in the work of [11] has a good performance, other Tibetan dialects, such as Amdo-Tibetan, also lack of speech corpus, which are much smaller than Chinese speech corpus, so they cannot contribute more to improve Lhasa-Tibetan speech recognition using transfer learning.

## 5 Conclusion

Under limited target data, Lhasa-Tibetan end-to-end speech recognition methods based on cross-language transfer learning are explored in this paper. Three aspects of modeling unit selection, transfer learning method, and source language selection are discussed. From the analysis of the experimental results, the end-to-end multi-language learning model using Chinese-Tibetan bilingual character set as the modeling unit is better than other methods. It achieves the lowest CER at 27.3%, and also has the largest CER reduction compared to the Lhasa-Tibetan specific model with a 26.1% reduction. Compared with the work [11], it achieves 2.2% higher accuracy using less amount of data. The experiments show that our method can learn both shared acoustic features and shared language knowledge between Chinese and Lhasa-Tibetan. Therefore, Chinese language is more suitable than English to establish a Tibetan speech recognition model as a cross-language transfer learning. Future work will explore the accuracy of Tibetan speech recognition using phoneme sets and even more finely divided articulatory feature sets as modeling units.

**Acknowledgement:** Thanks for the help of Professor Zhao Yue and all students in the research group.

**Funding Statement:** This work was supported by three projects. Zhao Y received the Grant with Nos. 61976236 and 2020MDJC06. Bi X J received the Grant with No. 20&ZD279.

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

## References


[1] X. B. Yang, "Low-Resource Uyghur Speech Recognition System Design and Implementation," M.S. dissertation, Northwest University for Nationalities, China, 2021.

[2] M. Zia, F. Ahmed, M. A. Khan, U. Tariq, S. S. Jamal *et al.,* "Classification of citrus plant diseases using deep transfer learning," *Computers, Materials & Continua*, vol. 70, no. 1, pp. 1401–1417, 2022.

[3] M. Faisal, F. Albogamy, H. ElGibreen, M. Algabri, S. Ahad *et al.,* "Covid-19 diagnosis using transfer-learning techniques," *Intelligent Automation & Soft Computing*, vol. 29, no. 3, pp. 649–667, 2021.

[4] A. Reda, S. Barakat and A. Rezk, "A transfer learning-enabled optimized extreme deep learning paradigm for diagnosis of covid-19," *Computers, Materials & Continua*, vol. 70, no. 1, pp. 1381–1399, 2022.

[5] Z. Fu, Y. Ding and M. Godfrey, "An LSTM-based malware detection using transfer learning," *Journal of Cyber Security*, vol. 3, no. 1, pp. 11–28, 2021.

[6] Y. Xu, Y. Zhai, B. Zhao, Y. Jiao, S. Kong *et al.,* "Weed recognition for depthwise separable network based on transfer learning," *Intelligent Automation & Soft Computing*, vol. 27, no. 3, pp. 669–682, 2021.

[7] Y. B. Li, "*Research on Uyghur Speech Recognition System Based on DFSMN-CTC Acoustic Model*," M.S. dissertation, Xinjiang University, China, 2020.

[8] J. Kang, "Design of End-to-End Amdo-Tibetan Speech Recognition System Based on Deep Learning," M.S. dissertation, Qinghai Normal University, China, 2021.

[9] J. H. Yan, "Research on Acoustic Model of Tibetan Lhasa Dialect Based on Lattice-Free MMI and Transfer Learning," M.S. dissertation, Northwest University for Nationalities, China, 2019.

[10] J. J. Yue, "Tibetan Multi-Task and Multi-Dialect Speech Recognition," M.S. dissertation, Minzu University of China, China, 2020.

[11] Y. Zhao, J. J. Yue, X. N. Xu, L. C. Wu and X. L. Li, "End-to-End-Based Tibetan Multitask Speech Recognition," *IEEE Access*, vol. 7, pp. 162519–162529, 2019.

[12] H. C. Wang, K. Khyuru, J. Li, G. Y. Li, J. W. 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 Processing Association Annual Summit and Conf. (APSIPA)*, Jeju, Korea, pp. 1–4, 2016.

[13] Q. N. Wang, W. Guo, P. X. Chen and Y. Song, "Tibetan-mandarin bilingual speech recognition based on end-to-end framework," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA ASC)*, Kuala Lumpur, Malaysia, pp. 1214–1217, 2017.

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

[15] Z. G. Li, "Language classification of nationalities in the world," *Journal of Guangxi University for Nationalities (Social Science Edition)*, no. 4, pp. 128–132, 1981.

[16] X. L. Ma, "Theories and methods of the study of Sino-Tibetan language family," *Minority Languages of China*, no. 4, pp. 5–9, 1996.

[17] H. Y. Yu, "A comparative analysis of Tibetan and Chinese English law structure–a simple example of the correspondence between parts of speech and sentence components in three languages," *Campus English*, no. 14, pp. 179–181, 2018.

[18] F. Z. Zhuang, P. Lu, Q. He and Z. Z. Shi, "Research progress of transfer learning," *Journal of Software*, vol. 26, no. 1, pp. 26–39, 2015.

[19] D. X. Dong, H. Wu, W. He, D. H. Yu and H. F. Wang, "Multi-task learning for multiple language translation," in *the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Processing*, Beijing, China, pp. 1723–1732, 2015.

[20] J. T. Huang, J. Y. Li, D. Yu, L. Deng and Y. F. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vancouver*, BC, Canada, pp. 7304–7308, 2013.