


METHODOLOGY

Open Access



Exploring task-diverse meta-learning on Tibetan multi-dialect speech recognition

Yigang Liu^{1,2}, Yue Zhao^{1,2*} , Xiaona Xu^{1,2}, Liang Xu^{1,2}, Xubei Zhang³ and Qiang Ji⁴

Abstract

The disparities in phonetics and corpuses across the three major dialects of Tibetan exacerbate the difficulty of a single task model for one dialect to accommodate other different dialects. To address this issue, this paper proposes task-diverse meta-learning. Our model can acquire more comprehensive and robust features, facilitating its adaptation to the variations among different dialects. This study uses Tibetan dialect ID recognition and Tibetan speaker recognition as the source tasks for meta-learning, which aims to augment the ability of the model to discriminate variations and differences among different dialects. Consequently, the model's performance in Tibetan multi-dialect speech recognition tasks is enhanced. The experimental results show that task-diverse meta-learning leads to improved performance in Tibetan multi-dialect speech recognition. This demonstrates the effectiveness and applicability of task-diverse meta-learning, thereby contributing to the advancement of speech recognition techniques in multi-dialect environments.

Keywords Tibetan language, Multi-dialect speech recognition, Task-diverse meta-learning

1 Introduction

During the long process of historical development, Tibetan language has gradually evolved into three major dialects: the Ü-Tsang dialect, the Amdo dialect, and the Kham dialect. Among these, the Ü-Tsang dialect holds the status of the standard Tibetan, thereby garnering more attention and research in the domain of speech recognition [1–4]. Conversely, the research and exploration of the other two dialects are relatively limited. Presently, the open speech data for the three Tibetan dialects amounts to approximately 80–100 h, whereas major languages like Mandarin Chinese and English boast vast

corpus surpassing 10,000 h. As a result, the resources for Tibetan speech remain woefully inadequate. Because notable discrepancies in research progress have occurred across the three primary dialects, how to effectively perform speech recognition of Tibetan multi-dialect speech becomes a challenging problem.

In the case of exceedingly scarce speech data, the efficacy of pre-training and fine-tuning techniques exhibits a modest impact. Furthermore, as opposed to the Ü-Tsang dialect, the standard corpus of the Amdo and Kham dialects is inadequate. Consequently, in the domain of speech recognition of the Amdo and Kham dialects, fine-tuning the model with a very small amount of target data does not lead to more desirable results. So, it is necessary to explore new method aimed at enhancing the performance of recognition models. Meta-learning is usually used as a potential solution to the question of “How to make models more efficient for new data?”. Meta-learning allows models to fit new data rapidly or with a few steps, which is compatible with the challenges encountered in the domain of low-resource language speech recognition [5–7].

*Correspondence:

Yue Zhao

zhaoyueso@muc.edu.cn

¹ Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China, Beijing 100081, China

² School of Information Engineering, Minzu University of China, Beijing 100081, China

³ Linguistics & Computer Science, Boston University, Boston 02215, USA

⁴ Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA

Finn et al. [8] proposed a meta-learning algorithm known as model-agnostic meta-learning (MAML), which can apply diverse learning scenarios, including classification, regression, and reinforcement learning. The distinctive feature of MAML lies in its model-agnostic nature. It can apply to any gradient descent-trained model. The experiment results show that MAML has remarkable capabilities, particularly excelling in the domain of small-sample classification, where it has the highest performance. Furthermore, this algorithm has also achieved great outcomes in the domain of small-sample regression tasks, further proving its effectiveness. Gu et al. [9] used meta-learning on the neural machine translation for the low-resource language. In their work, they used a comprehensive framework that encompassed eighteen different languages as source tasks. To overcome the inherent challenges posed by input-output disparities across different languages, they used a universal vocabulary representation. Moreover, they selected five different languages from the source task as target tasks to evaluate the effectiveness of their meta-learning approach. The experimental results showed the superiority of the meta-learning method over transfer learning. The findings highlighted the remarkable performance gains achieved through meta-learning, thereby emphasizing its potential as a potent technique for enhancing low-resource language translation tasks. Huang et al. [10] used meta-learning as a training algorithm in the domain of multi-speaker speech synthesis. This approach allows the model to be quickly adapted to speech synthesis, even when confronted with limited data for each speaker. Through a series of experiments, the results showed the capabilities of meta-learning in synthesizing highly authentic speech from a relatively smaller number of samples. Furthermore, the speed of adaptability was fast, facilitating rapid adjustments to individual speaker characteristics. These findings underscore the efficacy of meta-learning as a powerful method in multi-speaker speech synthesis. When the current better speech emotion recognition model is fed with multi-language data, the model's recognition performance is poor [11]. Training corpus is a big problem when dealing with low-resource languages. The authors solved this problem by introducing meta-learning. The experimental results show that the model with the addition of meta-learning performs better, and meta-learning is more advantageous when dealing with multi-language problems.

The aforementioned studies collectively highlight the feasibility and efficacy of adopting meta-learning methods as a means to tackle challenges associated with limited data. However, a noteworthy observation across these studies is that the source tasks and target tasks are similar, thereby falling to fully capitalize on the

inherent advantages of meta-learning. In this paper, we aim to expand the task types during source task selection. In addition to incorporating existing high-resource language speech recognition tasks, we introduce novel dimensions in the selection process, including Tibetan dialect ID recognition and Tibetan speaker recognition. It takes full advantage of meta-learning's ability to learn generalized knowledge from multiple tasks.

This paper's contributions are twofold. First, we integrate dialect identification (ID) recognition and speaker recognition into the meta-training framework of meta-learning. Previous meta-learning research predominantly focused on aligning meta-training tasks with the target task, overlooking the holistic properties inherent to meta-learning. This expansion of meta-training encompasses additional dimensions crucial for comprehensive meta-learning, thereby enriching the meta-learning process and advancing its efficacy. The second contribution lies in using multi-task pre-training as an initialization step for meta-learning. This method facilitates a more favorable starting point for the meta-learning process, enhancing its efficiency and effectiveness. By leveraging the insights gained from multi-task pre-training, the meta-learning framework is equipped with a robust foundation, thereby optimizing its performance and potential for subsequent task adaptation.

The structure of this paper is as follows: Sect. 2 introduces related work, including the multi-task technique and the current research status of meta-learning for Tibetan speech recognition. Section 3 introduces the meta-learning method and our multi-task meta-learning. Section 4 describes the experimental data, setting, and results. Section 5 offers a discussion and summary of this work.

2 Related work

The development of Tibetan dialect speech recognition has been hindered by a lack of linguistic knowledge and corpus data. Presently, due to widely use of the Ü-Tsang dialect, more linguistic and phonological research has been carried out and more research results have been achieved. In contrast, the Amdo and Kham dialects suffer from a considerable scarcity of linguistic knowledge and corpus data, which severely diminishes the feasibility of constructing speech recognition systems for these dialects. However, multi-task and multi-language learning are common methods for enhancing the performance of multi-dialect language recognition, particularly for low-resource languages.

Kannan et al. [12] addressed the problem of cross-language training data imbalance by constructing an end-to-end multi-language speech recognition system. The experimental results showed that the end-to-end

multi-language model had a lower error rate than the mono-language model. For the Japanese multi-dialect, Imaizumi et al. [13] used multi-task learning for dialect ID recognition and dialect recognition based on an end-to-end model. The experimental results showed that the recognition results of the multi-task learning model were better than those of the traditional single-task model. In the work [14], Siddharth Dalmia et al. showed that multi-task learning can improve the recognition performance of low-resource languages. Zhao et al. [15] used the multi-task learning method with shared hard-parameter to build a multi-task WaveNet-CTC model. The experimental results show that the multi-task model has high recognition accuracy for three Tibetan dialects, surpassing the single-dialect baseline model. Dan et al. [16] used a soft-parameter shared multi-task learning method to build a multi-task Transformer model. The experimental results show that the recognition accuracy of this model is higher than that of a single-dialect model and a hard-parameter shared multi-dialect model. However, multi-task learning may lead to interference among tasks. The weights of each task need be carefully designed. This in turn will increase the difficulty of training the model. But meta-learning is unbiased towards all tasks and does not weigh tasks. It just improves the learning ability of the model through different tasks.

The core idea of meta-learning is to make machine learning models better adapt to new tasks or new domains with a few labeled samples by learning how to learn [17–19]. By pre-training the model's initial parameters on different tasks, these initial parameters can learn some internal features that are more suitable for the delivery. When the model encounters a new task, the loss function of the new task will be more sensitive to the initial parameters, so that even a small amount of gradient update can make the model quickly converge, to achieve fast learning. In 2020, Hsu et al. [20] proposed to use meta-learning for low-resource automatic speech recognition. Their experiment uses six high-resource languages as source languages and four low-resource languages as target languages. The experimental results show that meta-learning performs better than pre-training on all target languages with different combinations of source languages. At present, there are relatively little researches on Tibetan speech recognition based on meta-learning. Dhanva Eledath et al. [21] used model-agnostic meta-learning to train the Conformer model, leading to enhanced speech recognition accuracy across various low-resource English accents. Qin et al. [22] used meta-learning based on fine-grained modeling units for speech recognition of Lhasa-Tibetan. The experimental results show that meta-learning is better than multi-language pre-training in the low-resource language Tibetan. So

meta-learning has a unique advantage for low-resource languages.

Multi-task learning achieves performance gains for each of these individual tasks. However, the source and target tasks in the above meta-learning are speech recognition tasks. This way does not fully utilize the characteristic of meta-learning to acquire general knowledge from multiple tasks and the advantages of multi-task learning. Therefore, in this paper, the source tasks include not only speech recognition tasks but also two different types of tasks: Tibetan speaker recognition and Tibetan dialect ID recognition.

3 Method

3.1 Meta-learning

Unlike multi-task learning, meta-learning places a strong emphasis on a model's learning capacity rather than its performance on each specific task. Meanwhile, by enabling the model to swiftly adapt to new tasks, meta-learning has a great advantage over pre-training, particularly in facing low-resource multi-dialect. The adaptability of the model is greatly enhanced, enabling it to respond more rapidly and effectively to new problems. The learning process maximizes the sensitivity of the loss function to the model's parameters for each new task. As the sensitivity increases, even minute adjustments to the parameters trigger substantial feedback in the loss function, thereby facilitating the model's rapid and accurate parameter adjustments in response to these heightened feedback signals. This mechanism enables the model to fine-tune itself efficiently, ultimately leading to superior performance on various tasks, especially in low-resource and diverse linguistic contexts.

Meta-learning can be divided into two main processes: the meta-training process and the adaptive process. The meta-training process is the most essential feature that distinguishes meta-learning from traditional machine learning. The training unit of traditional machine learning is a single data, while the training unit of meta-training is a single task. Each task has a support set and a query set. The meta-training process is shown in Fig. 1.

$T = \{T_1, T_2, \dots, T_n\}$ represents the set of training tasks. The task T_k is selected from the set, for which the support set is D_k^{tr} and the query set is D_k^{te} . The learning rate is η' , then the parameter θ'_k in the task T_k is updated in the way shown in Eq. (1):

$$\theta'_k = \theta'_{k-1} - \eta' \nabla_{\theta} L(\theta'_{k-1}, D_k^{tr}) \quad (1)$$

At this point, the overall parameters of the meta-learning are updated with the D_k^{te} of the task T_k . The meta-learning rate is η . The parameters of the meta-learning are updated in the way shown in Eq. (2):

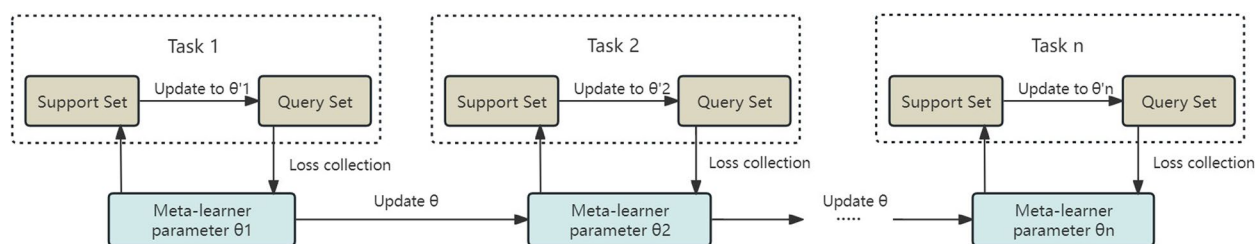


Fig. 1 The meta-training process

$$\theta_k = \theta_{k-1} - \eta \nabla_{\theta} L(\theta_{k-1}, D_k^{te}) \quad (2)$$

The update of meta-learning consists of an outer loop and an inner loop. The inner loop is the parameter update within each task. The outer loop is the parameter update performed between source tasks, i.e., the parameter update of meta-learning. Once the source tasks are all trained, the model is fine-tuned with the support set of the target task and the model is then tested on the query set.

3.2 Task-diverse meta-learning

Meta-learning can help the model make better use of the small amount of data, allowing the model to quickly adapt to changes in different dialects and speakers. Through meta-learning, the model can adapt and adjust the parameters of the model faster when it receives a new dialect, thus improving the overall recognition performance. The key idea of meta-learning is to enable the model to learn quickly on a new task through patterns and knowledge learned from different tasks. One of its strengths is very suitable for low-resource speech recognition. However, existing source task choices of meta-learning for Tibetan speech recognition are all speech recognition tasks, which do not fully utilize the inherent advantages of meta-learning.

Transfer learning aims to capitalize on previously acquired knowledge and experiences to enhance performance in novel tasks or domains. However, the efficacy of transfer learning hinges on the degree of similarity between the source and target tasks. When faced with substantial disparities between tasks, the applicability of learned knowledge to the new task diminishes, thereby posing challenges to effective knowledge transfer. So, meta-learning and transfer learning diverge in selecting source data. Transfer learning focuses on utilizing source data that closely mirrors the characteristics of the target data, fostering a more seamless transition during the training process. This similarity between the source and target datasets enhances the effectiveness of the training, as the model can leverage existing

knowledge to learn new tasks more efficiently. In contrast, meta-learning operates without stringent requirements regarding source and target data similarity. Instead, it embraces a diverse array of source data, aiming to identify the most suitable parameters for rapid adaptation to varying tasks and contexts. This flexibility allows meta-learning algorithms to explore a wide range of data types and structures, enabling the model to generalize better across different scenarios. While transfer learning relies on source data similarity for effective training, meta-learning thrives on diversity, leveraging various data sources and skillfully applied parameters to optimize adaptation and mitigate overfitting.

3.3 Pre-training initialization for task-diverse meta-learning

The initial parameters of a deep learning model have an important impact on the performance and convergence speed of the model. Inappropriate initial parameters may lead to problems such as the model not converging properly, falling into local optimal solutions, vanishing or exploding gradients, etc. It affects the training effect and generalization ability of the model. So, the initial parameters of the model have a vital influence on the final effect of the model.

In previous research related to meta-learning, the initial parameters of the model before meta-training are generally obtained by random initialization. The disadvantage of random initialization is that the parameters may be inappropriate. Random initialization does not ensure that each neuron receives the appropriate initial weights, which may result in consistently smaller or larger outputs for some neurons, making the model's performance ability insufficient. Generally, pre-training involves the initial training of a model on a large-scale dataset, followed by fine-tuning on the target task. This method circumvents the need to train the model from scratch and significantly reduces the demand for labeled data.

4 Experiment

4.1 Experimental data

The experimental datasets in this paper are obtained from the Open Speech and Language Resources (OpenSLR), as shown in Table 1. Since Tibetan is a low-resource language, the existing corpus is limited. However, there are a large scale of Chinese and English corpus, so we use more Chinese and English data than Tibetan data for meta-learning.

The first dataset comes from the Aishell-1 Chinese dataset [23]. It is a Chinese speech corpus covering 11 fields such as smart homes, autonomous driving, and industrial production. We select 80,604 sentences as the training set with a total duration of 100.42 h and 4551 sentences as the test set with a total duration of 5.67 h. The audio files are converted into Windows Audio Volume (WAV) format with 16 KHz sampling rate and 16-bit quantization accuracy.

The second dataset is the LibriSpeech English dataset [24]. We select 28,539 sentences as the training set with a total duration of 100.59 h and 1530 sentences as the test set with a total duration of 5.39 h. The audio files are converted into Windows Audio Volume (WAV) format with 16 KHz sampling rate and 16-bit quantization accuracy.

The third dataset is the TIBMD@MUC Tibetan dataset [25]. It is a multi-dialect Tibetan dataset. To compare the results with the works of [15, 16], the same data were used for the experiments. So, we select 20594 Ü-Tsang sentences as the training set with a total duration of 23.74 h and 2575 Ü-Tsang sentences as the test set with a total duration of 2.98 h. There are 33 Ü-Tsang dialect native speakers including 13 males and 20 females. We select 17286 Amdo sentences as the training set with a total duration of 18.83 h and 2161 Amdo sentences as the test set with a total duration of 2.36 h. There are 19 Amdo dialect native speakers including 10 males and 9 females. We select 2376 Kham sentences as the training set with a total duration of 2.61 h and 269 Kham sentences as the test set with a total duration of 0.33 h. There are 10 Kham dialect native speakers including 7 males and 3 females.

The audio files are converted into Windows Audio Volume (WAV) format with 16KHz sampling rate and 16-bit quantization accuracy.

4.2 Experimental settings

The recognition model is based on the Conformer model [26], an architecture that has achieved remarkable results in the field of speech recognition in recent years. It combines the self-attention mechanism and the convolutional neural network. It has strong capabilities of modeling and context understanding. For low-resource languages, the introduction of the Conformer model can help the model to better capture the features and structures in the speech signal, thus achieving more accurate recognition.

Training and testing on a server equipped with the NVIDIA Quadro RTX 4000 were done. The Conformer model is built on the Pytorch framework, and it is trained with 80 epochs. The batch size is 10. For optimization, the Adam algorithm with gradient clipping is used. The learning rate is $\eta' = 0.005$; the meta-learning rate is $\eta = 0.01$. The label-smoothing technology is used during training, and the smoothing parameter $\alpha = 0.1$.

Our model is not limited to speech recognition in the selection of source tasks for meta-learning and makes full use of the mutual benefit relationship between multiple tasks by introducing Tibetan speaker identification and Tibetan dialect ID recognition. The structure of the task-diverse meta-learning method is shown in Fig. 2. This comprehensive method enhances the system's ability to understand and model diverse speech data. Through dialect ID recognition, the model can distinguish the differences among different dialects and perform more targeted training and optimization for a specific dialect. Through speaker recognition, the model can learn the characteristics of different speakers. The model can be better trained to accommodate differences between dialects and speakers. These additional source tasks help provide more speech features and contextual information to further improve the accuracy and robustness of the recognition.

We choose to use the pre-trained model as the initialization for meta-learning, which provides a set of high-quality initial parameters for the model. It greatly reduces the randomness of the parameters and provides a clear direction for subsequent model training. The structure of the model is shown in Fig. 3. Our model realizes the meta-learning Tibetan multi-dialect speech recognition with pre-training initialization.

In our work, we choose English word recognition, Chinese character recognition, and Chinese pinyin recognition as the pre-training tasks. Here, we do not need to fine-tune the pre-trained model on the target language; just fine-tune the meta model after the meta-training is over. This pre-training method can provide a clear direction for the

Table 1 The statistics of dataset

Dataset	Speech utterances		Duration (hours)		
	Training data	Test data	Training data	Test data	
Chinese	80604	4551	100.42	5.67	
English	28539	1530	100.59	5.39	
Tibetan	Ü-Tsang	20594	2575	23.74	2.98
	Amdo	17286	2161	18.83	2.36
	Kham	2373	269	2.61	0.33

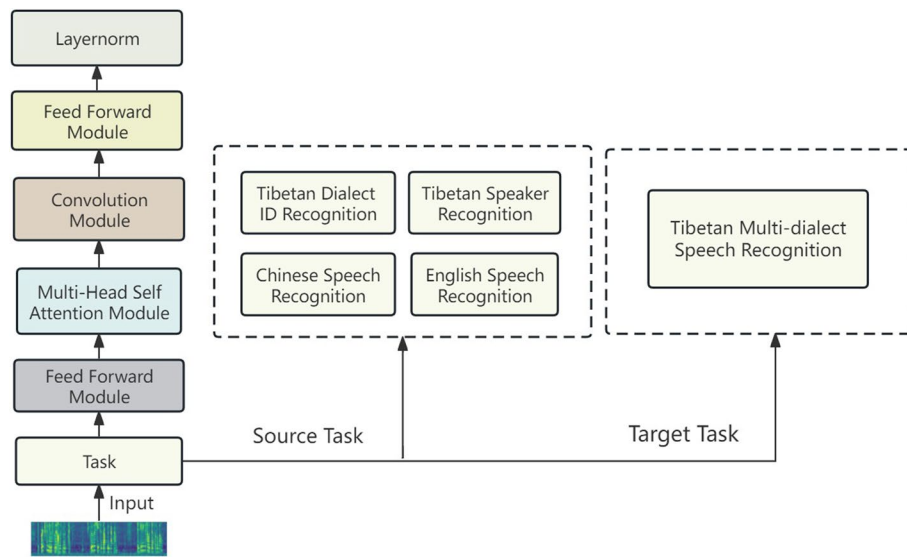
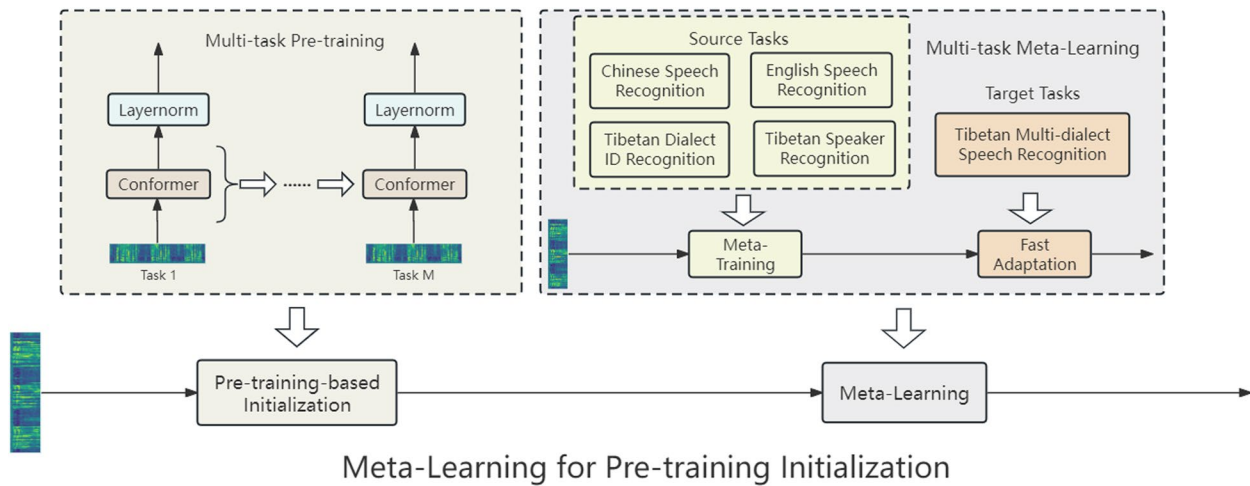


Fig. 2 The structure of task-diverse meta-learning. The inputs to the model are Fbank features. The initial parameters are randomly initialized. The four source tasks are Tibetan Dialect ID Recognition, Tibetan Speaker Recognition, Chinese Speech Recognition, and English Speech Recognition, respectively



Meta-Learning for Pre-training Initialization

Fig. 3 The task-diverse meta-learning with pre-training initialization. The inputs to the model are Fbank features. The initial parameters are provided by the pre-trained model. The four source tasks are Tibetan Dialect ID Recognition, Tibetan Speaker Recognition, Chinese Speech Recognition, and English Speech Recognition, respectively

following meta-learning and improve the training efficiency of the model.

The metric for experimental evaluation in this paper is Tibetan character error rate (CER). The method of calculating CER is shown in Eq. (3).

The meta-learning tasks include English word recognition, Chinese character recognition, Tibetan speaker recognition, and Tibetan dialect ID recognition. This can enrich the variety of source tasks. The meta model can get

$$\text{Character Error Rate} = \frac{\text{Insertions} + \text{Substitutions} + \text{Deletions}}{\text{Total Words}} \times 100\% \tag{3}$$

more knowledge from different tasks, which strengthens the model's learning and adaptive capabilities.

4.3 Experimental results and analysis

In Table 2, the experimental results are shown for the comparison of several methods on the Tibetan multi-dialect dataset. We compare our Task-Diverse Meta-Learning Conformer without pre-training with the Hard-MTL Transformer [15], the Single-Dialect Transformer, Multi-Dialect Transformer, Soft-MTL Transformer, and Adaptive Soft-MTL Transformer [16]. Meanwhile, we also compare the pre-training and fine-tuning Conformer, which is first pre-trained in Chinese and English languages, followed by fine-tuning on the Tibetan multi-dialect dataset. The Task-Diverse Meta-Learning without pre-training is our model.

From Table 2, we can see that the Single-Dialect Transformer has the highest error rates among all three dialects. The effect of multi-task learning is obviously better than single-task learning, and the mutually beneficial relationships between multiple tasks benefits each task. The results of our proposed method are relatively good, but not the best, so we need to continue to explore it with pre-training process.

Previous meta-learning was initialized by random initialization. To avoid the drawbacks associated with random initialization, we initialize the model by using pre-training. The model is firstly trained on the Chinese and English, and then the trained model parameters are

Table 2 The error rate of Tibetan multi-dialect speech recognition

Model	CER(%)		
	Ü-Tsang	Amdo	Kham
Single-Dialect Transformer [16]	14.96	8.90	107.73
Multi-Dialect Transformer [16]	8.96	4.26	24.33
Pre-Training Fine-Tuning Conformer	9.74	4.81	41.25
Hard-MTL Transformer [15]	9.53	4.74	40.64
Soft-MTL Transformer [16]	7.86	3.07	29.40
Adaptive Soft-MTL Transformer [16]	5.67	2.70	26.48
Task-Diverse Meta-Learning Conformer without Pre-training	6.57	3.11	26.88

The bold values represent the lowest error rate

used as meta-learning initial parameters instead of the previous random initialization. The error rates of the two models on the Tibetan multiple dialects are shown in Table 3.

In Table 3, compared with Pre-Training Fine-Tuning Conformer, the experimental results show that the contribution of meta-learning to Tibetan multi-dialect speech recognition is greater than that of pre-training. The task-diverse meta-learning with pre-training obviously has better recognition rates, with a 1.33% reduction in the error rate for the Ü-Tsang dialect, a 0.24% reduction in the error rate for the Amdo dialect, and a 3.37% reduction in the error rate for the Kham dialect. After pre-training, initialization can improve the training efficiency of the model instead of blindly random initialization, thus adapting to new tasks faster. In comparison with previous models, our proposed task-diverse meta-learning after pre-training achieves the best results on both Ü-Tsang and Kham dialects. On the Amdo dialect, our method is slightly worse than Adaptive Soft-MTL Transformer in CER.

Adaptive Soft-MTL Transformer is good at dealing with tasks with small differences, while meta-learning is more skilled at learning general knowledge from different tasks. Multi-task learning typically requires models to handle multiple tasks simultaneously, which can lead to interferences between tasks or require large amounts of data to balance the individual tasks. Meta-learning, on the other hand, focuses on training the model to learn how to learn compared to Adaptive Soft-MTL Transformer so that it can adapt to new tasks or new domains more quickly. So, our proposed model is generally better than Adaptive Soft-MTL Transformer.

5 Conclusion

This work aims to explore the application of task-diverse meta-learning methods to multi-dialect speech recognition in Tibetan. Traditional speech recognition models often face challenges of multi-dialect situations because there are significant differences between each dialect. These differences make it difficult for single-task models to adapt to the features of different dialects. To solve this problem, we introduce the idea of task-diverse meta-learning. Multi-task meta-learning can help the model

Table 3 The error rates of task-diverse meta-learning with and without pre-training

Model	CER(%)		
	Ü-Tsang	Amdo	Kham
Pre-Training Fine-Tuning Conformer	9.74	4.81	41.25
Task-Diverse Meta-Learning Conformer without Pre-training	6.57	3.11	26.88
Task-Diverse Meta-Learning Conformer with Pre-training	5.24	2.87	23.41

The bold values represent the lowest error rate

learn more general and robust representations, enabling the model to better adapt to the differences between dialects. By introducing Tibetan dialect ID recognition and Tibetan speaker recognition tasks as source tasks for meta-learning, we force the model to focus on the commonalities and differences between dialects, thus improving the model's performance on the Tibetan multi-dialect speech recognition task. Task-diverse meta-learning is better able to utilize the knowledge of the source task to guide the learning of the target task. This task-diverse meta-learning method provides the model with richer training signals and enhances the model's understanding of the variations and differences between dialects.

In addition, we also compared our approach with hard-parameter-sharing multi-task learning and soft-parameter-sharing multi-task learning models. The experimental results show that our method has significant improvement in the task of Tibetan multi-dialect speech recognition. Thus, our study provides an innovative method to solve the problem of multi-dialect speech recognition and lays the foundation for further improvement of speech recognition techniques in multi-dialect environments.

Authors' contributions

Conceptualization, Y.Z.; methodology, Y.Z.; software, Y.L.; validation, Y.Z., Y.L. and L.X.; formal analysis, Y.Z.; investigation, Y.Z. and Y.L.; resources, Y.Z.; data curation, Y.L., Y.Z. and L.X.; writing—original draft preparation, Y.L.; writing—review and editing, Y.Z., X.Z., and Q.J.; visualization, Y.L.; supervision, Y.Z.; project administration, Y.Z. and X.X.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the National Natural Science Foundation of China, grant number 61976236.

Availability of data and materials

Three datasets were used for this article. The first dataset comes from the Aishell-1 Chinese dataset. Its download link is as follows: <http://www.openslr.org/33/>; the second dataset is the LibriSpeech English dataset; its download link is as follows: <http://www.openslr.org/12/>; the third dataset is the TIBMD@MUC Tibetan dataset; its download link is as follows: <http://www.openslr.org/124/>.

Declarations

Competing interests

The authors declare no competing interests.

Received: 12 February 2024 Accepted: 25 June 2024

Published online: 17 July 2024

References

- N. Zhou, Research on Tibetan non-specific person continuous speech recognition based on deep learning. Master's thesis, Central University for Nationalities (2017)
- X. Huang, J. Li, Acoustic model for Tibetan speech recognition based on recurrent neural network. *J. Chin. Inf.* **32**(5), 189–191 (2018)
- Q. Wang, W. Guo, C. Xie, Tibetan speech recognition based on end-to-end technology. *Pattern Recognit. Artif. Intell.* **30**(4), 359–363 (2017)
- S. Yuan, W. Guo, L. Dai, Tibetan language recognition based on deep neural networks. *Pattern Recognit. Artif. Intell.* **28**(3), 209–213 (2015)
- S. Min, M. Lewis, L. Zettlemoyer et al., Metaicl: Learning to learn in context[J]. arXiv preprint arXiv:2110.15943 (2021)
- C. Finn, K. Xu, S. Levine, Probabilistic model-agnostic meta-learning[J]. *Adv. Neural Inf. Process. Syst.* **31**, (2018)
- L. Collins, A. Mokhtari, S. Shakkottai, Task-robust model-agnostic meta-learning[J]. *Adv. Neural Inf. Process. Syst.* **33**, 18860–18871 (2020)
- C. Finn, P. Abbeel, S. Levine, in *International conference on machine learning*. Model-agnostic meta-learning for fast adaptation of deep networks[C] (PMLR, 2017), pp. 1126–1135
- J. Gu, Y. Wang, Y. Chen et al., Meta-learning for low-resource neural machine translation[J]. arXiv preprint arXiv:1808.08437 (2018)
- S.F. Huang, C.J. Lin, D.R. Liu et al., Meta-TTS: Meta-learning for few-shot speaker adaptive text-to-speech[J]. *IEEE/ACM Trans. Audio Speech Lang. Process.* **30**, 1558–1571 (2022)
- A. Naman, C. Sinha, in *Machine Intelligence and Smart Systems: Proceedings of MISS 2021*. Fixed-MAML for few-shot classification in multilingual speech emotion recognition[M] (Springer Nature Singapore, Singapore, 2022), pp. 473–483
- A. Kannan, A. Datta, T.N. Sainath et al., Large-scale multilingual speech recognition with a streaming end-to-end model[J]. (2019). arXiv preprint arXiv:1909.05330
- R. Imaizumi, R. Masumura, S. Shiota et al., End-to-end Japanese multi-dialect speech recognition and dialect identification with multi-task learning[J]. *APSIPA Trans. Signal Inf. Process.* **11**(1), (2022)
- S. Dalmia, R. Sanabria, F. Metz et al., in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Sequence-based multilingual low resource speech recognition[C] (IEEE, 2018), pp. 4909–4913
- Y. Zhao, J. Yue, X. Xu et al., End-to-end-based Tibetan multitask speech recognition[J]. *IEEE Access* **7**, 162519–162529 (2019)
- Z. Dan, Y. Zhao, X. Bi et al. in *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24-25, 2022, Proceedings, Part I*. Multi-task learning with auxiliary cross-attention transformer for low-resource multi-dialect speech recognition[C] (Springer International Publishing, Cham, 2022), pp. 107–118
- D. Li, Y. Yang, Y.Z. Song et al., in *Proceedings of the AAAI conference on artificial intelligence*. Learning to generalize: Meta-learning for domain generalization[C], vol. 32(1) (2018)
- K. Yang, R. Wang, L. Wang, in *31st International Joint Conference on Artificial Intelligence (IJCAI-22)*. Metafinger: Fingerprinting the deep neural networks with meta-training[C] (2022)
- H. Yao, L.K. Huang, L. Zhang et al., in *International conference on machine learning*. Improving generalization in meta-learning via task augmentation[C] (PMLR, 2021), pp. 11887–11897
- J.Y. Hsu, Y.J. Chen, H. Lee, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Meta learning for end-to-end low-resource speech recognition[C] (IEEE, 2020), pp. 7844–7848
- D. Eledath, A. Baby, S. Singh, in *2024 National Conference on Communications (NCC)*. Robust speech recognition using meta-learning for low-resource accents[C] (IEEE, 2024), pp. 1–6
- S. Qin, L. Wang, S. Li et al., in *Proc. INTERSPEECH*. Finer-grained modeling units-based meta-learning for low-resource tibetan speech recognition[C] (2022)
- H. Bu, J. Du, X. Na et al., in *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline[C] (IEEE, 2017), pp. 1–5
- V. Panayotov, G. Chen, D. Povey et al., in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Librispeech: An ASR corpus based on public domain audio books[C] (IEEE, 2015), pp. 5206–5210
- Y. Zhao, X. Xu, J. Yue et al., An open speech resource for Tibetan multi-dialect and multitask recognition[J]. *Int. J. Comput. Sci. Eng.* **22**(2–3), 297–304 (2020)
- A. Gulati, J. Qin, C.C. Chiu et al. Conformer: Convolution-augmented Transformer for speech recognition[J]. arXiv preprint arXiv:2005.08100 (2020)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.