Interspeech 2022
18-22 September 2022, Incheon, Korea

Finer-grained Modeling units-based Meta-Learning for Low-resource Tibetan Speech Recognition

Siqing Qin[1], *Longbiao Wang*[1*], *Sheng Li*[2], *Yuqin Lin*[1], *and Jianwu Dang*[1,3]

[1] Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] National Institute of Information and Communications Technology, Kyoto, Japan
[3] Japan Advanced Institute of Science and Technology, Ishikawa, Japan

longbiao_wang@tju.edu.cn

Abstract

Tibetan is a typical under-resourced language due to its relatively smaller population. Although a character-based end-to-end (E2E) automatic speech recognition (ASR) model with transfer learning and multilingual training strategies has mitigated the problem of low resources, it often meets overfitting problem. Recently meta-learning performs great in solving overfitting problem. However, the widely-used coarse-grained modeling units are not significantly correlated to their pronunciation, which limits the performance improvement of the low-resource ASR system. Furthermore, meta-learning consists of a meta-training period and fast self-adaption on the target language, and the past meta-training period is lack target language-specific information. Therefore, we propose a novel E2E low-resource Lhasa dialect ASR model based on the finer-grained modeling units and transfer learning with reference to the properties of Chinese Pinyin. Chinese Pinyin and Tibetan decomposed radicals are more related to pronunciation than characters are, which can compensate for more acoustic information in low-resource situations. Furthermore, Tibetan modeling units are utilized in both meta-training and fast self-adaption processes to offer language-specific information to solve the low-resource problem. Experiments show that our proposed method achieves a 54.9% relative character error reduction rate than the baseline system.

Index Terms: Automatic speech recognition, modeling units, end-to-end model, Tibetan, low-resource

1. Introduction

Tibetan speech recognition technique has attracted increasing attention since it has not been developed well due to the low resource problem. The Lhasa dialect, as the central Tibetan dialect, has many Tibetan manuscripts with a long history. Therefore, applying natural language processing and speech recognition techniques to the Lhasa dialect would be meaningful. In the past decade, the end-to-end (E2E) ASR system emerged [1, 2]. This framework can directly recognize speech features as text without a lexicon. Different kinds of E2E models have been proposed, e.g., connectionist temporal classification (CTC) [3], attention-based encoder-decoder E2E models [4], and joint CTC-Attention [5]. Recently, the E2E transformer model [6] was proposed to address neural machine translation and was applied to ASR tasks [7], which achieved promising performance as expected. However, the E2E model is data-hungry, so the ASR performance is not good for low-resource languages.

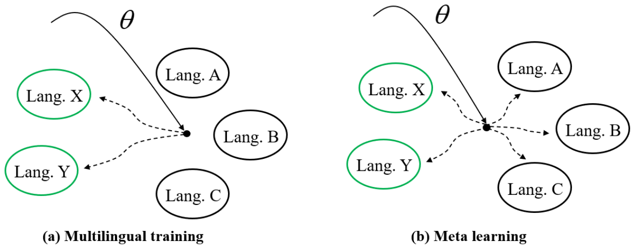


Figure 1: *The comparison between multilingual pretraining and meta-learning pretraning [12]. θ denotes the system parameters.*

The transfer learning-based multilingual training method proposed in the low-resource machine translation field has been used to improve the low-resource ASR performance by initializing with high-resource languages [8, 9, 10]. Recently, the meta-learning [11] has significantly advanced the low-resource problem by fast self-adapting. Meta-learning is proved that it is more beneficial for the ASR model than the transfer learning-based pretraining method [12] as shown in Figure 1. It is also proved that meta-learning can solve overfitting problem [12]. In this paper, we regard meta-learning as two separated processes, namely meta-training on source languages and fast self-adaption on the target language. However, the meta-training period is always lacking target language-specific knowledge, which is negative to improve the low-resource situation. Furthermore, the widely used coarse-grained modeling units always have a large quantity, which leads to the out-of-vocabulary (OOV) problem, especially under low-resource circumstances.

In our previous work [13], a Tibetan radical-based ASR system was proposed to solve the out-of-vocabulary (OOV) problem. Most E2E ASR systems are based on coarse-grained modeling units, such as Chinese or Tibetan characters, which are always biased towards offering linguistic information, but not significantly correlated to their pronunciation. Since Tibetan and Chinese belong to the same language family, Tibetan radical structure is more similar to Chinese syllable structure in pronunciation aspect. Inspired by this knowledge, it is possible to utilize the correlation of Chinese syllables and Tibetan radicals to solve the low-resource problem of Tibetan, where the Chinese syllable is represented by Chinese Pinyin. Therefore, in this study, we solve the low-resource problem in two aspects for Lhasa dialect ASR model: (1) Utilizing the relation of Chinese Pinyin and Tibetan radical to make model learn more knowledge across similar languages and modeling units to

[*] corresponding author.

10.21437/Interspeech.2022-10015

tackle the low-resource problem. (2) We regard meta-learning as two phases, and the first meta-training is meant to get better parameters for fast self-adapting to the target task. So the performance of meta-training has a large impact on the performance of the final ASR model. We propose an initialization strategy for the first meta-training phase of meta-learning to replace random initialization to make model be trained well for the next stage. On this basis, the multi-level modeling units (character-level and radical or syllable-level) and meta-learning are combined to offer Tibetan language-specific information in meta-training process.

2. Related Work

Meta-learning method has achieved good results in K-shot learning [14, 15, 16], machine translation [17], dialogue generation [18], speaker adaption [19] and other fields. Model-agnostic meta-Learning (MAML) is a universal and model-independent meta-learning method [20]. It solved the problem of how to handle different model architectures and different problem settings. The key idea of MAML is to obtain the initial parameters through pre-training so that the model parameters have the best performance on the new task after passing through one or more gradient update steps. Recently, MAML has make a great progress in ASR [12]. Use f to denote model, which can map observation x to output y. During meta-learning, the model is trained to be able to adapt to a large or infinite number of tasks that are under distribution $p(\mathcal{T})$.

Formally, a model is represented by a parametrized function f_θ with parameters θ. When adapting to a new task $\mathcal{T}_i$, the model's parameters are updated to θ_i'. θ_i' is computed using gradient descent updates on task $\mathcal{T}_i$:

$$\theta_i' = \theta - \alpha \bigtriangledown_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta) \tag{1}$$

The step size α can be fixed as a hyperparameter. More concretely, the meta-objective is as follows:

$$\min_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'}) = \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha \bigtriangledown_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)}) \tag{2}$$

The meta-optimization across tasks is performed via stochastic gradient descent (SGD), such that the model parameters θ are updated as follows:

$$\theta \leftarrow \theta - \beta \bigtriangledown_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'}) \tag{3}$$

3. Proposed method

This paper treats bilingual finer-grained modeling units as source languages to improve the target Tibetan ASR system. And a transfer learning-based initialization strategy is proposed. On this basis, the multilevel Tibetan modeling units are combined with meta-learning. In this section, proposed methods will be introduced in detail.

3.1. Finer-grained Modeling Units for Tibetan Speech Recognition

A Tibetan character can be further segmented into a sequence of subcharacter tokens (radicals) [13]. The vertically stacking components in a character are separated and treated as individual units and the boundary mark between two consecutive characters as shown in Figure 2. This subcharacter unit set then consists of 56 Tibetan components and a boundary marker. In [13], the performance of Tibetan radical-based ASR systems is far better than Tibetan character-based systems, which proves that using radicals as modeling units can solve the OOV problem to some extent. In this paper, the radical-based Lhasa dialect E2E ASR system is one of our baselines.

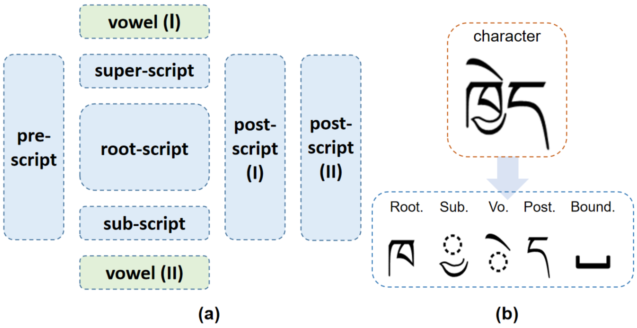
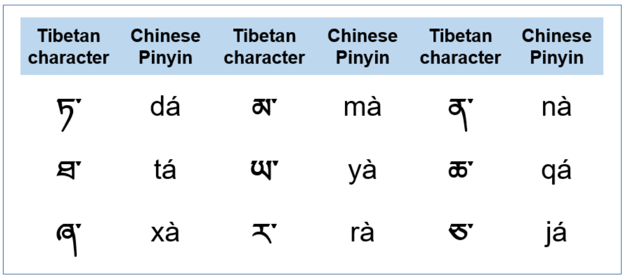


Figure 2: *The decomposition strategy of Tibetan character.*



Tibetan character	Chinese Pinyin	Tibetan character	Chinese Pinyin	Tibetan character	Chinese Pinyin
ད་	dá	མ་	mà	ན་	nà
ཐ་	tá	ཡ་	yà	ཆ་	qá
ཤ་	xà	ར་	rà	ཇ་	já

Figure 3: *Some examples of Tibetan letters with corresponding Chinese Pinyin.*

Finer-grained bilingual modeling units: Pinyin is an official Chinese character phonetic Latinization plan [21]. The Tibetan alphabet has 30 basic letters, and every letter has corresponding Chinese Pinyin [22]. This paper regarded one letter in different positions (up or down) as various radicals, so the number of radicals is expanded to 56. In Figure 3, nine examples of Tibetan letters are shown. These two finer-grained units have high correlation and are more correlated with pronunciation, which is better to offer more acoustic knowledge to solve the low-resource problem. In addition, Chinese is in the same language family as Tibetan, so the ASR model learns the information across two languages. In this paper, a novel bilingual ASR system with the finer-grained and bilingual modeling units is proposed, and multilevel modeling units are investigated to improve the performance of the low-resource Lhasa dialect E2E ASR system. The multilevel modeling units mean the trainable units contain two different granularities, like character and Pinyin in Chinese or character or character and radical in Tibetan.

Combination of multilevel modeling units with meta-learning: The meta-learning consists of two steps, namely, meta-training and fast self-adaption. On the basis of solving the overfitting problem, the meta-learning-based ASR model can learn target language-specific information in meta-training process to alleviate the low-resource problem by combining it with multilevel modeling units as shown in Figure 4. In this paper, the Chinese character, Pinyin, and Tibetan character are used as source tasks, while Tibetan radical is used as target task. The

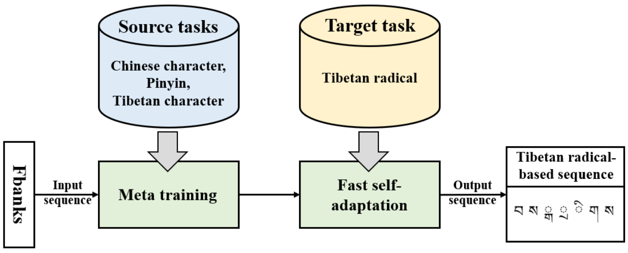


Figure 4: *The combination of meta-learning and multilevel modeling units.*

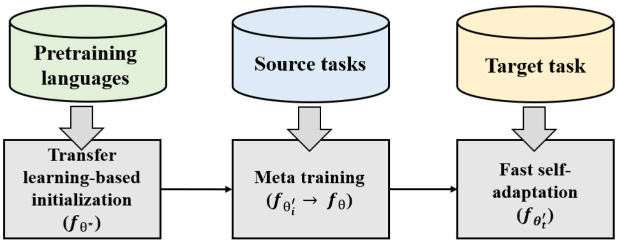

Pretraining languages
Source tasks
Target task
Transfer learning-based initialization (f_{θ^*})
Meta training ($f_{\theta_i'} \to f_\theta$)
Fast self-adaptation ($f_{\theta_t'}$)

Figure 5: *Transfer learning-based meta learning method.*

proposed method can offer Tibetan language-specific information in both meta-training and fast self-adaption processes.

3.2. Transfer learning-based Meta Training for Tibetan Speech Recognition

In ASR field, the different languages are treated as different tasks. Given the source tasks $\mathcal{T} = \{T_1, T_2, ..., T_K\}$, MAML can learn initial parameters θ from $\mathcal{T}$ to fast adapt to the target task T_t with parameters θ_t' [12]:

$$\theta_t' = Learn(T_t, \theta) = Learn(T_t, MetaLearn(\mathcal{T})) \quad (4)$$

$Learn$ represents the learning process using cross-entropy loss. $MateLearn$ represents the process of meta-learning.

In the meta-training process, samples from a source task i from $\mathcal{T}$ is divided into training set T_i^{tr} and test set T_i^{te}. First, T_i^{tr} is used to simulate the language-specific training process to obtain θ_i'. This process is optimized by the gradient descent method, and the loss function is the cross-entropy loss. Then, θ_i' is evaluated on the test set T_i^{te} by the meta-objective shown in Equation (2) and updated with meta-gradient shown in Equation (3). This process is carried out for each source task to obtain a great initialization parameter θ. Finally, θ is fine-tuned by samples from target task T_t to obtain θ_t'. The test set of the target task is used for the final performance evaluation.

As mentioned above, the meta training process will obtain initialization parameters θ, which is crucial for fast adaptation and system performance. So it can be regarded as a key part of the entire model training. In the previous meta-learning, the model always has a random initialization to obtain θ^* before meta training. After meta training procedure, θ^* will be updated to meta-initialization parameters θ to be fine-tuned with target task T_t and obtain θ_t'. On the contrary, It is no doubt that this random initialization process will greatly affect meta-trained model parameters and then affect the performance of the target task. In this paper, the transfer learning-based high-resource language initialization strategy instead of random initialization was utilized to obtain θ^* as shown in Figure 5, which makes a significant improvement. The character-level and radical-level modeling units were regarded as one of the source tasks and the target task, respectively.

Table 1: *ASR performance (CER%) with different settings of modeling units based on joint-training on Lhasa dialect (**LA**) and Mandarin (**CH**) selected from the AISHELL-1 corpus ("**c**" indicates a Tibetan character and Chinese character-based model, "**r**" indicates a Tibetan radical-based model and "**p**" indicates a Chinese Pinyin-based model).*

	Source Languages	Target Languages	LA-TST (CER%)
Bilingual Training	\	LA (**c**) (baseline 1)	36.02
	\	LA (**r**) (baseline 2)	34.96
	CH (**c**)	LA (**c**)	32.94
	CH (**p**)	LA (**c**)	32.91
	CH (**p**)	LA (**r**)	**30.44**
	\	LA (**c+r**) (baseline 3)	29.61
	CH (**c**)	LA (**c+r**)	28.84
	CH (**p**)	LA (**c+r**)	**28.59**
	CH (**c+p**)	LA (**c+r**)	28.81
Meta Learning	Source Task	Target Task	LA-TST (CER%)
	CH (**c**)	LA (**c**)	28.65
	CH (**p**)	LA (**c**)	27.74
	CH (**p**)	LA (**r**)	17.07
	CH (**c**)+LA (**c**)	LA (**r**)	16.63
	CH (**p**)+LA(**c**)	LA (**r**)	16.59
	CH (**c**)+CH (**p**)+LA (**c**)	LA (**r**)	**16.24**

4. Experiments

In this section, we evaluate a set of models built using our proposed method.

4.1. Dataset

The Lhasa dialect speech corpus contains 33.2 hours of speech data corresponding to more than 38,700 sentences collected from 14 male and 11 female Lhasa Tibetan native speakers. The recording script is mainly composed of declarative sentences covering a wide range of topics. The speech signal is sampled at 16 kHz with 16-bit quantization. The training set contains 30.4-hour speech data. The development set and testing set contain 1.1-hour and 1.7-hour speech data, respectively. To train a balanced bilingual ASR system, we select partial speech data from the AISHELL-1 corpus (CH) [23]. The training set contains 27 hours of speech data, and the test set includes 4.1 hours of speech data, with three female and three male speakers.

4.2. The E2E Baseline ASR Systems for Lhasa Dialect

In this study, ASR tasks are based on the attention-based transformer model (ASR-Transformer) [6]. In this paper, the 6×6 Encoder-Decoder with 8-head Attention structure is used, and the hidden units are 512. First, we built a monolingual E2E transformer speech recognition system with only the Lhasa dialect based on the character-level and radical-level modeling units, respectively. In this paper, the actual acoustic feature dimension is 480 with the stitching frame method.

A well-trained Mandarin ASR model with a CER of 9% was used as an initializing model. This model was trained by 178 hours of speech data from the AISHELL-1 corpus [23]. Finally, our character-based baseline reduced the CER to 36.02%, and the radical-based baseline reduced the CER to 34.96%, as shown in Table 1. Furthermore, we built a self-fusion E2E ASR system with a CER of 29.61% in the same way as described in [13], shown in Table 1, as another baseline. It was jointly trained by Tibetan characters and Tibetan radicals for comparison with multilevel unit-based systems.

4.3. Bilingual Training-based Lasha Dialect E2E ASR Systems with Multi-level Unit

In this section, based on the multilingual transformer architecture, two different languages were jointly trained based on multi-level modeling units. All transcriptions have been marked with the corresponding language tags at the front. The Tibetan radicals and Chinese pinyin-based system (**LA(r)+CH(p)**) in Table 1 showed the best bilingual modeling performance in the single-level Tibetan modeling unit-based ASR systems, with a CER of 30.44%, which was relatively reduced by 15%. Therefore, by using the pronunciation similarity of these two units and their correlation with pronunciation, joint-training Tibetan radicals, and Chinese Pinyin can solve the low-resource problem to some extent.

In addition, the multi-level Tibetan modeling units are utilized in these experiments. In Table 1, the performance of multi-level Tibetan units-based ASR systems was shown to be better than others due to the sharing of information across different modeling units and the increase in available data. The character- and radical-based Lhasa dialect E2E ASR system jointly trained with Chinese Pinyin (**LA(c+r)+CH(p)**) performed best with CER of 28.59%, and relatively improved the performance by 3.4% and 20.6% compared with that of **LA(c+r)** and **LA(c)**, respectively. These results further proved that Pinyin could significantly improve the low-resource problem of the Lhasa dialect. However, the performance of **LA(c+r)+CH(c+p)** decreased little compared with **LA(c+r)+CH(p)**. We reason that the well-trained model may be more partial to the Chinese speech recognition task because of the inseparable correlation between Chinese characters and Chinese Pinyin. In general, the multi-level unit-based training can further solve the low-resource problem and improve the performance of Lhasa dialect ASR.

4.4. Meta Learning-based Lasha Dialect E2E ASR Systems with Multilevel Units

We used the 178h speech data of AISHELL-1 to pretrain the meta-training model. By using the initialization strategy, the CER was reduced to 17.52% as shown in Table 2 comparing with **LA(c+r) (baseline 3)**. So the rest experiments were all based on this method. Using Chinese characters as source task and Tibetan characters as target task (**Source: CH(c)+Target: LA(c)**) was the meta-learning baseline system with the CER of 28.65%. The same modeling unit settings were utilized to build ASR models for fairly comparing the meta-learning with joint training. Experiments in Table 1 showed the radical-based modeling unit as the target task (**Target: LA(r)**) significantly improved the ASR systems because the radical-based modeling units can solve the OOV problem [13].

With longitudinally comparing, the performances of **Source: CH(p)+Target: LA(c)** and **Source: CH(p)+Target: LA(r)** were better than these of **Source: CH(c)+Target: LA(c)** and **Source: CH(c)+Target: LA(p)**, respectively. There was a relative performance improvement by 40.4% compared with the meta-learning baseline. So the finer-grained modeling units were also beneficial for meta-learning. The ASR systems based on multi-level Tibetan modeling units (using **LA(c)** as one of the source tasks and **LA(r)** as target task) performed better than these based on single-level Tibetan modeling units (only using one kind of Tibetan modeling units as target task). It denoted that the language-specific information can improve the performance of low-resource ASR systems. The performance was improved perfectly by using Chinese characters, Chinese Pinyin, and Tibetan characters as source task (**CH(c)+CH(p)+LA(c)**) with CER of 16.24%, which improved the performance by 43.3% compared with the meta-learning baseline. So the meta-learning can learn more information across source tasks to fast self-adapt to the target task without the over-fitting problem.

Table 2: *The comparison of random initialization and proposed initialization strategy in meta training process.*

Training set or Source task+Target task	initialization strategy	LA-TST (CER%)
LA (**c+r**)(baseline3)	yes	29.61
LA (**c**)+LA (**r**)	no	31.19
LA (**c**)+LA (**r**)	yes	**17.52**

With horizontally comparing with results of **Bilingual Training** and **Meta Learning** in Table 1, meta-learning-based ASR systems significantly outperformed the joint-training ones with the same modeling units settings. Overall, the meta-learning method improved the system performance by 54.9% compared with **LA(c) baseline 1**. In conclusion, meta-learning is a promising method in low-resource ASR tasks and better than the joint-training method. In addition, our proposed finer-grained bilingual modeling units-based method also obtained a satisfactory performance.

Therefore, the proposed method of combining meta-learning with multilevel modeling units proceeded with a great performance on low-resource Tibetan speech recognition. On the basis of meta-learning solving the overfitting problem, multilevel Tibetan modeling units let model learn sufficient Tibetan language-specific knowledge in meta-training period.

5. Conclusion and future work

In this paper, we focused on compensating for acoustic knowledge in the E2E ASR model and meta training ASR models based on transfer learning to improve the low-resource situation for the Lhasa dialect. To solve the low-resource data issue, we investigated multi-level modeling units-based joint training and meta training. Proposed finer-grained modeling units were useful for compensating for acoustic information. The similarity of Pinyin and Tibetan radical offered more trainable information across languages and modeling units. Furthermore, the Tibetan language-specific information in meta-training period improved the performance of low-resource ASR systems. Compared to the baseline, the performance of the best jointly learning-based E2E ASR system we proposed showed a 20.6% relative improvement, and that of the meta learning-based E2E ASR system showed a 54.9% relative improvement. Experiments show that our proposed methods can effectively model the low-resource Tibetan speech. This paper supports a new direction for low-resource language research. In future work, we will investigate the correlation between the source languages and target language to obtain promising performance.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62176182.

7. References

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[2] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[3] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.

[4] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *arXiv preprint arXiv:1506.07503*, 2015.

[5] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[7] Y. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao, J. Mahadeokar, H. Huang, A. Tjandra, X. Zhang, F. Zhang *et al.*, "Transformer-based acoustic modeling for hybrid speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6874–6878.

[8] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," *arXiv preprint arXiv:1604.02201*, 2016.

[9] V. M. Shetty, M. S. M. NJ, and S. Umesh, "Improving the performance of transformer based low resource speech recognition for indian languages," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8279–8283.

[10] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, "Sequence-based multi-lingual low resource speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4909–4913.

[11] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *arXiv preprint arXiv:2004.05439*, 2020.

[12] J.-Y. Hsu, Y.-J. Chen, and H.-y. Lee, "Meta learning for end-to-end low-resource speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7844–7848.

[13] L. Pan, S. Li, L. Wang, and J. Dang, "Effective training end-to-end asr systems for low-resource lhasa dialect of tibetan language," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1152–1156.

[14] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," *arXiv preprint arXiv:1807.05960*, 2018.

[15] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *arXiv preprint arXiv:1703.05175*, 2017.

[16] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, pp. 3630–3638, 2016.

[17] J. Gu, Y. Wang, Y. Chen, K. Cho, and V. O. Li, "Meta-learning for low-resource neural machine translation," *arXiv preprint arXiv:1808.08437*, 2018.

[18] F. Mi, M. Huang, J. Zhang, and B. Faltings, "Meta-learning for low-resource natural language generation in task-oriented dialogue systems," *arXiv preprint arXiv:1905.05644*, 2019.

[19] O. Klejch, J. Fainberg, and P. Bell, "Learning to adapt: a meta-learning approach for speaker adaptation," *arXiv preprint arXiv:1808.10239*, 2018.

[20] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[21] Z. Jinming, "The chinese pinyin system scheme: The cornerstone of teaching chinese as a second language," *Applied Linguistics*, vol. 4, 2009.

[22] A. Hu, M. Wang, and H. Yu, "Early processing research of tibetan two-syllable words' tone in lhasa," in *Journal of Physics: Conference Series*, vol. 1237, no. 3. IOP Publishing, 2019, p. 032002.

[23] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," 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)*. IEEE, 2017, pp. 1–5.