darkblue[1]

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 3769–3779

MetaMixSpeech: Meta Task Augmentation for Low-Resource Speech Recognition

Yaqi Chen[1], Hao Zhang[1], Wenlin Zhang[1,2], Xukui Yang[1,2], Dan Qu[1,2,*], Yinghao Zhai[1],

[1]Information Engineering University, ZhengZhou, China
[2]Laboratory for Advanced Computing and Intelligence Engineering, Wuxi, China

Abstract

Meta-learning has proven to be a powerful paradigm for effectively improving the performance of low-resource speech recognition by learning generalizable knowledge across multiple tasks. However, multilingual meta learning also faces challenges such as task overfitting and learner overfitting, thereby reducing its ability to generalize to new tasks. To address these issues, we augment the meta-training task with "more data" during both training and evaluation phases. Concretely, we propose an interpolation-based task augmentation method called MetaMixSpeech, which includes both support augmentation and query augmentation. MetaMixSpeech enhances task diversity by linearly combining perturbed features from the support and query sets and performing the same linear interpolation on their corresponding losses. Experimental results on the FLEURS and Common Voice datasets demonstrate that MetaMixSpeech achieves a 6.35 % improvement in Word Error Rate (WER) compared to meta-learning approaches, effectively mitigating the overfitting problem and showcasing superior generalization across diverse datasets and language families.

1 Introduction

Automatic Speech Recognition (ASR) has revolutionized various aspects of people's lives, delivering remarkable success in several widely spoken languages (Radford et al., 2023; Zhang et al., 2023). However, there are more than 7,000 languages in the world, and it is estimated that 94% of them are spoken by fewer than 1 million people (Lewis, 2009). These languages are categorized as low-resource languages due to limited labeled data availability. Compared with commonly spoken languages, low-resource languages lack the necessary transcribed speech data, pronunciation dictionaries, and language scripts, making it challenging to build a usable speech recognition system. Despite these challenges, it is imperative to research efforts on low-resource languages and preserve their cultural heritage.

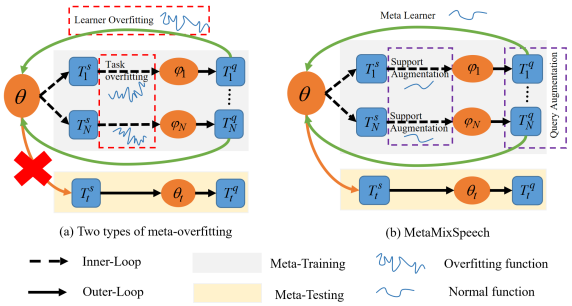


Figure 1: (a) task overfitting in the task learner φ while acquiring task-specific knowledge, and learner overfitting in the meta-learner θ while learning cross-task knowledge. This leads to poor generalization from meta-training to meta-testing. (b) Our proposed MetaMixSpeech employs support augmentation and query augmentation to resist overfitting.

A lot of progress has been made in low-resource speech recognition, which includes efforts like data augmentation (Park et al., 2019), transfer learning (Hu et al., 2019) and multilingual learning (Hou et al., 2020). Recently, a novel paradigm known as meta-learning has been investigated for its potential in enhancing low-resource speech recognition (Hsu et al.; Wang et al., 2023; Singh et al., 2022; Hou et al., 2022). Among these, gradient-based meta-learning algorithms (Finn et al., 2017) have gained widespread adoption in the speech domain due to their flexibility and generalization capabilities.

Gradient-based meta-learning algorithms train models at two levels. In the inner loop, the model undergoes a gradient descent on a small amount of training data (the support set), enabling the task learner to adapt to the task. In the outer loop, the adapted model is evaluated on the query set of the task, and the evaluation loss is optimized to im-

*Corresponding author

prove the model's generalization capability across the meta-training tasks. The key to meta-learning lies in optimizing the model initialization to maximize its generalizability, measured by the adapted model's performance on the query set.

However, when data is limited, meta learning face the risks of memory overfitting (Yin et al., 2020) and learner overfitting (Rajendran et al., 2020), which have been analyzed in computer vision applications. Unlike computer vision, multilingual meta-learning does not encounter the issue of memory overfitting due to several reasons. First, the labels of sequence generation tasks are not fixed, ensuring tasks' exclusivity. Second, the number of languages per episode is fixed, preventing different categories' learning within an episode.

However, we analyze that multilingual meta-learning still faces two types of overfitting risks, as shown in Figure.1: (1) Task-overfitting, where repeated learning of the same languages in each episode causes the task learner ϕ to overfit the support set and fails to generalize to the query set. (2) Learner-overfitting, where distributional differences between the source and target languages lead the model θ to overfit the meta-training tasks and subsequently fail to generalize to new tasks. Both types of overfitting significantly impair the generalization from meta-training to meta-testing.

Improving meta-generalization is challenging. In the field of image processing, methods such as meta-regularizers (Lee et al., 2020) and constrained search spaces (Yin et al., 2020), as well as active data augmentation (Rajendran et al., 2020; Ni et al., 2020), have been employed to improve meta-generalization. Compared to regularizers that restrict the flexibility of the inner loop, task augmentation methods (Liu et al., 2020) are more flexible and effective. For instance, Yao et al. (Yao et al., 2020) proposed a task augmentation method that improves reliance on the support set by mixing it with the query set, thereby mitigating memory overfitting. However, this approach is not suitable for multilingual meta-learning for not suffering from the problem of memory overfitting.

To address the risks of task overfitting and learner overfitting in multilingual meta-learning, we propose a flexible and effective task augmentation method called MetaMixSpeech. As shown in Figure.1, this method includes support augmentation and query augmentation, which conduct feature space manifold expansion by linearly combining perturbed features or hidden representations from the support and query sets and performing the same linear interpolation on their corresponding losses. It can effectively mitigate the two types of meta-overfitting and enhance the generalization capability of meta-learning methods.

The main contributions of this paper are:

- We analyze two types of meta-overfitting in multilingual meta learning, and propose MetaMixSpeech for task augmentation, which can improve meta-generalization. To the best of our knowledge, this is the first time a task augmentation method has been proposed for multilingual meta-learning.

- Extensive experiments on FLEURS and Common Voice illustrated that MetaMixSpeech exhibits a 6.35 % improvement in WER compared to meta-learning approaches, showing strong generalization across diverse datasets and language families.

2 Methods

2.1 Model Structure

Recently, self-supervised learning (SSL) (Baevski et al., 2020; Chen et al., 2022) has achieved significant advancements, enabling and bootstrapping ASR applications in low-resource languages. Due to the substantial storage and training costs associated with fine-tuning, adapters (Thomas et al., 2022) were proposed as an alternative approach using a lightweight neural network integrated at each layer of the pre-trained model to adapt to the low-resource downstream target language. Furthermore, Otake et al. (Otake et al., 2022) recently developed a new adapter structure to make full use of the feature representation from low to high levels in self-supervised models, which achieves superior performance in ASR. Our model architecture and adapter structure follow (Otake et al., 2022).

The adapter structure consists of two parts: Layer adapters (L-adapters) and Encoder adapters (E-adapters). E-adapters are embedded in each encoder layer, and L-adapters directly connect each encoder layer to the top layer. Each E-adapter consists of a two-layer fully connected (FC) layer with layer normalization (LN) and a skip connection, and each L-adapter consists of a fully connected (FC) layer followed by non-linear activation (Act) and layer normalization (LN). As shown in Figure.2, the modules in red are learnable and the

modules in gray are frozen. Additionally, LN and the head are also learnable.

Suppose that the output of the n-th encoder layer is $\mathbf{h_n}$, the L-adapter $f_l(\cdot)$ is applied to it to obtain adapted representations as $\mathbf{a}_n = f_l(\mathbf{h}_n)(n = 1, 2, \ldots L)$. The final model representation is a weighted sum of the adapted representations $\mathbf{h}^* = \sum_{n=1}^{L} w_n \mathbf{a}_n$, where w_n are learnable weights.

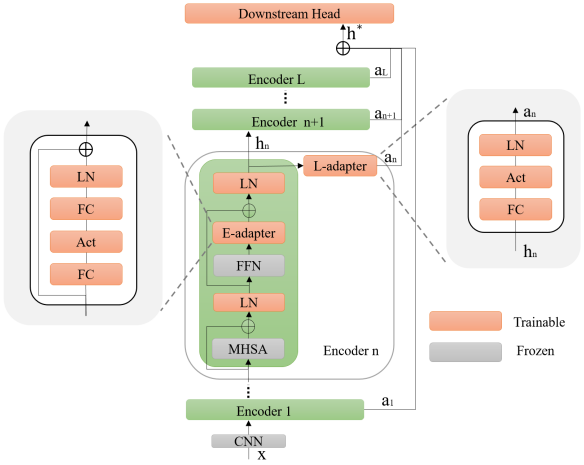


Figure 2: Architecture of the adapter module and its integration with the SSL model. We add the E-adapter module in each Encoder layer, and the L-adapter module to connect the output of each Encoder layer with the downstream head. The E-adapter consists of a bottleneck that contains few parameters relative to the attention and feedforward layers in the original model, and also contains a skip connection. The L-adapter comprises a non-linear layer to adapt downstream tasks. During adapter tuning, the red modules are trained on the downstream data, and the gray modules are frozen.

2.2 Meta Learning

Meta learning has proven to be a powerful paradigm for transferring the knowledge from source languages to facilitate the learning of low-resource target languages.

Consider the meta-training dataset as a set of N languages $D = D_i(i = 1, ...N)$, where each language D_i consists of speech-text pairs. Unlike traditional machine learning, meta-learning uses tasks as its training samples and acquires generic meta-knowledge over numerous training episodes. In each episode, N tasks are sampled from N languages to form a batch. For the i-th language, a task T_i is sampled from D_i, and is divided into two subsets: the support set T_i^s for fast adaptation and the query set T_i^q for evaluation. During pre-training and fine-tuning, the parameters of the self-supervised model θ_W are kept frozen. The meta-learning algorithm trains the adapter module to obtain a good initialization θ_M for quick adaptation to low-resource target languages.

Present gradient-based meta-learning techniques can be described as bilevel optimization problems, utilizing an episodic training paradigm to train the entire model. The two-level meta-learning framework (Finn et al., 2017) can be characterized as follows:

$$\min_{\theta_M} \sum_{i=1}^{N} \mathcal{L}^{meta}(\theta_W, \omega^{*(i)}(\theta_M); T_i^q), \quad (1)$$

$$s.t. \;\; \omega^{*(i)}(\theta_M) = \arg\min \mathcal{L}(\theta_W, \theta_M; T_i^s). \quad (2)$$

Here, $\mathcal{L}^{meta}$ and $\mathcal{L}$ represent the meta loss (in the outer loop) and the task loss (in the inner loop), respectively. In particular, the inner loop Eq.(2) is designed to learn a language-specific task learner $\omega^{*(i)}(\theta_M)$ for each task using the support set T_i^s, while the outer loop Eq.(3) learns meta-knowledge from these task learners with the query set T_i^q.

Due to the challenge of computing the second-order derivatives and storing the Hessian matrix in Eq.(1), we employ first-order model-agnostic meta-learning (FOMAML) (Finn et al., 2017), which omits the calculation of the second-order gradient by approximation. In this condition, we can formulate the objective as follows:

$$\min_{\theta_M'} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}^{meta}(\theta_W, \theta_M'; T_i^q), \quad (3)$$

$$s.t. \;\; \theta_M' = \theta_M - \alpha \nabla_{\theta_M} \mathcal{L}(\theta_M; T_i^s). \quad (4)$$

2.3 MetaMixSpeech

As mentioned earlier, we can observe that meta learning may exhibit learner overfitting on meta-training tasks by Eq.(1) during the outer loop. Moreover, it may cause task overfitting over the specific task by Eq.(2) during the inner loop. Both types of meta-overfitting significantly impact the generalization from meta-training to meta-testing.

To mitigate these overfitting issues, we introduce a novel meta task augmentation strategy named MetaMixSpeech. This approach generates additional data by mixing the perturbed support sets and query sets. In speech recognition, direct mixing is not feasible due to the variable lengths of labels. Hence, we utilize the MixSpeech (Meng et al., 2021) method for mixing.

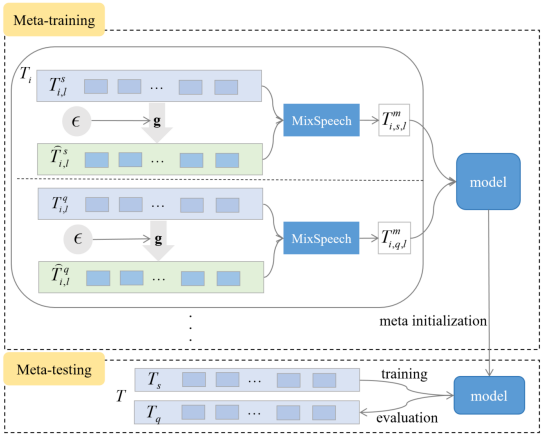


Figure 3: Framework of MetaMixSpeech based on bilevel optimization meta learning algorithm. In the meta-training stage, MetaMixSpeech selects the l-th feature respresentation $X^s_{i,l}$ and $X^q_{i,l}$ for augmentation. Then we denote the augmented task as $T^s_{i,l} = (X^s_{i,l}, Y^s_i)$ and $T^q_{i,l} = (X^q_{i,l}, Y^q_i)$. First, it applies the mapping function g with the random factor ϵ to these tasks, resulting in perturbed sets $\widetilde{T^i_{s,l}}$ and $\widetilde{T^i_{q,l}}$, respectively. Then we generate virtual examples using MixSpeech, yielding a new augmented support set $T^m_{i,s,l}$ and query set $T^m_{i,q,l}$ for training the meta initialization. In the meta-testing stage, we finetune the meta-initialized model using the training data T_s of the new language and evaluate the updated model by the test data T_q.

Assume that the speech recognition model $f(\cdot)$ is parameterized by θ and contains L layers in total. The hidden layer representation of a sample at the l-th layer is denoted as $f_{\theta^l}(X)(0 \le l \le L-1)$, where $f_{\theta^0}(X) = X$.

For the i-th task T_i, it has a support set $T^s_i = (X^s_i, Y^s_i) = \{(x^s_{i,k}, y^s_{i,k})\}^{N_s}_{k=1}$ and a query set $T^q_i = (X^q_i, Y^q_i) = \{(x^q_{i,k}, y^q_{i,k})\}^{N_q}_{k=1}$. Here, N_s and N_q denote the number of samples in the support set and the query set, respectively. First, we use an independent random factor ϵ and a mapping function $g : \epsilon, X, Y \to \widetilde{X}, \widetilde{Y}$, which are combined to create a perturbed set of data: $\widetilde{X}, \widetilde{Y} = g(\epsilon, X, Y)$. We assume that all $(x, y) \in (X, Y)$ are also in $(\widetilde{X}, \widetilde{Y})$. An example of this could be perturbing the order of x and y without changing their values. Therefore, we can obtain a perturbed support set and query set $\widetilde{T^s_i} = (\widetilde{X^s_i}, \widetilde{Y^s_i}) = \{(\widetilde{x^s_{i,k}}, \widetilde{y^s_{i,k}})\}^{N_s}_{k=1}$ and $\widetilde{T^q_i} = (\widetilde{X^q_i}, \widetilde{Y^q_i}) = \{(\widetilde{x^q_{i,k}}, \widetilde{y^q_{i,k}})\}^{N_q}_{k=1}$.

Considering that task overfitting and learner overfitting are caused by the limited data available for each task, the implementation of MetaMixSpeech contains two parts: support augmentation and query augmentation, as illustrated in Figure. 3.

To address the problem of task overfitting, we employ support augmentation to increase the training samples for the specific task. Given the feature of the l-th layer for X^s_i and its perturbed version $\widetilde{X^s_i}$, we mix the feature representation and the corresponding loss to generate intermediate samples. This can be denoted as:

$$\begin{aligned} X^m_{i,s,l} &= \boldsymbol{\lambda} f_{\theta^l}(X^s_i) + (\mathbf{I} - \boldsymbol{\lambda}) f_{\theta^l}(\widetilde{X^s_i}), \\ \mathcal{L}^m_{i,s} &= \boldsymbol{\lambda} \mathcal{L}(R^l_{i,s}, Y^s_i) + (\mathbf{I} - \boldsymbol{\lambda}) \mathcal{L}(R^l_{i,s}, \widetilde{Y^s_i}). \end{aligned} \tag{5}$$

where $\boldsymbol{\lambda} = diag(\{\lambda_j\}^{N_s}_{j=1})$, and each coefficient $\lambda_j \sim Beta(\alpha, \beta)$. $X^m_{i,s,l}$ represents the mixed features of the l-th layer for the support set of the i-th task. These mixed features continue to pass through the remaining $L - l$ layers of the model to obtain the output $R^l_{i,s} = f_{\theta^{L-l}}(\mathrm{X}^m_{i,s,l})$. The final loss function $\mathcal{L}^m_{i,s}$ is a linear combination of the losses between the model outputs and the corresponding labels.

To address the problem of learner overfitting, we propose query augmentation to increase the samples of meta-evaluation, which helps increase the generalization of meta-learner. Similarly, when we select the l-th layer for augmentation, given the query set of the i-th task X^q_i and its perturbed version $\widetilde{X^q_i}$, we conduct MixSpeech to generate intermediate samples. This can be denoted as:

$$\begin{aligned} X^m_{i,q,l} &= \boldsymbol{\lambda} f_{\theta^l}(X^q_i) + (\mathbf{I} - \boldsymbol{\lambda}) f_{\theta^l}(\widetilde{X^q_i}), \\ \mathcal{L}^m_{i,q} &= \boldsymbol{\lambda} \mathcal{L}(R^l_{i,q}, Y^q_i) + (\mathbf{I} - \boldsymbol{\lambda}) \mathcal{L}(R^l_{i,q}, \widetilde{Y^q_i}). \end{aligned} \tag{6}$$

where $X^m_{i,q,l}$ represents the mixed features of the l-th layer for the query set of the i-th task, and $R^l_{i,q} = f_{\theta^{L-l}}(\mathrm{X}^m_{i,q,l})$.

Support augmentation can enhance performance within similar language families by strengthening intra-task learning, while query augmentation can improve generalization across different language families by providing broader meta-evaluation. Consequently, given the frozen parameters θ_W and trainable parameters θ_M, the objective function of MetaMixSpeech is formulated as:

$$\begin{aligned} &\min_{\theta_M} \sum_{i=1}^{N} \mathbb{E}_{\lambda \sim \mathrm{Beta}} \mathbb{E}_{l \sim L} \left[\mathcal{L}^m_{i,q}(\theta_W, \omega^{*(i)}(\theta_M); T^m_{i,q,l}) \right], \\ &\text{s.t.} \quad \omega^{*(i)}(\theta_M) = \arg\min_{\theta_M} \mathbb{E}_{l \sim L} \left[\mathcal{L}^m_{i,s}(\theta_W, \theta_M; T^m_{i,s,l}) \right]. \end{aligned} \tag{7}$$

In fact, MetaMixSpeech is a model-agnostic algorithm that can be applied to any model or meta-learning framework for meta-task augmentation.

Algorithm 1 Meta-training Process of MetaMixSpeech

Require: Self-supervised model f_θ, Beta distribution parameters α ,β; Task distribution $p(T)$
Initialize adapter parameters θ_M
while not converge **do**
Sample a batch of tasks $\{T\}_{i=1}^{n} \sim p(T)$
for all T_i **do**
Sample support set T_i^s and query set T_i^q from T_i
Sample $\lambda_j \sim Beta(\alpha, \beta)$ and mixed layer l
Generate the augmented support set using Eq.(5)
Compute the adapted parameters $\omega^{(i)}(\theta_M)$ using Eq.(4)
Compute the mixed query loss using Eq.(6)
end for
Construct the meta loss $\mathcal{L}^{meta}$ using Eq.(7)
Update the adapter parameters θ_M using Eq.(3)
end while

3 Experiment Settings

3.1 Datasets

We used data from Mozilla's Common Voice Corpus (Ardila et al., 2020) and the FLEURS dataset (Conneau et al., 2022). From the Common Voice Corpus, we selected eight languages as source tasks: Italian, German, Swedish, English, French, Dutch, Russian, and Portuguese, each with about 5 hours of training data.

To evaluate generalization of different datasets, we selected eight target languages from the Common Voice Corpus: Breton (Br), Irish (Ga-IE), Romanian (Ro), Odia (Or), Sorbian (Hsb), Arabic(Ar), Ukrainian (Uk) and Czech (Cs). And ten target languages from the FLEURS dataset: Croatian (Hr), Maltese (Mt), Vietnamese (Vi), Finnish (Fi), Urdu (Ur), Nepali (Ne), Malayalam (Ml), Malay (Ms), Galician (Gl) and Kazakh (Ka). Furthermore, to assess generalization across various language families, we chose distinct language families for fine-tuning: Western Europe, Central Asia, South Asia and South-East Asia. We adhered to the validation and test splits as defined in the official dataset documentation. Detailed statistics of the datasets are presented in Table 1.

Table 1: Statistics of languages (h) in the Common Voice and FLEURS datasets.

Common Voice	Br	Ga-IE	Ro	Or
	2.84	2.10	3.04	0.45
	Hsb	Ar	Uk	Cs
	1.48	7.87	17.35	20.66
FLEURS	Hr	Mt	Vi	Fi
	11.00	9.87	9.01	8.81
	Ur	Ne	Ml	Ms
	9.39	6.30	9.63	9.55
	Gl	Ka		
	6.66	5.06		

3.2 Implementation Details

We utilized the WavLM Base model (Chen et al., 2022) as the self-supervised model. Specifically, we employed the wavlm-base-plus[1] version, which has 94.70M parameters. The adapter configuration was consistent with that described in (Otake et al., 2022). We trained the model for 120 epochs with a batch size of 64. During the adaptation process, we fine-tuned the adapter for 100 epochs with a batch size of 8 for FLEURS datasets and a batch size of 32 for Common Voice datasets. We set an early stop strategy for three times during training. We used the Adam optimizer for both the inner loop and outer loop, with a learning rate of 1e-3. The word error rate (WER) served as our evaluation metric. The hyperparameters α and β for the Beta distribution were set to 0.5. And the best performance was achieved when l was set to 0. The proportion of one batch of data to train using MetaMixSpeech is denoted as τ, which was set to 15% as default.

For each target language, we considered the following baseline approaches: (i) FT-Full: Optimizing all model parameters except for the feature extractor; (ii) Vanilla-ASR: Training the adapters with randomly initialized parameters; (iii) Multi-ASR: Pretraining the adapters by multilingual learn-

[1]https://huggingface.co/microsoft/wavlm-base-plus

Table 2: Word error rate (WER 100%) on ten target languages of the FLEURS datasets using various adaptation methods. Here, the terms "MixS" and "MixQ" refer to only support augmentation or query augmentation were applied.

Methods	Hr	Mt	Gl	Fi	Ml	Ne	Ur	Ms	Vi	Ka	Avg.
FT-Full	**72.81**	100.00	100.00	100.00	100.00	**75.23**	100.00	77.04	100.00	96.59	92.16
Vanilla-ASR	83.36	89.28	98.36	79.22	79.66	76.22	100.00	77.33	61.81	98.31	84.36
Multi-ASR	79.90	80.77	75.16	78.46	83.88	82.87	74.31	77.75	69.10	96.34	79.85
Meta-ASR	98.23	**74.32**	66.99	71.72	80.48	77.50	77.05	76.91	67.79	83.50	77.45
MixS	77.49	78.46	76.68	80.69	80.38	78.72	75.12	81.37	60.55	75.25	76.47
MixQ	76.60	74.88	67.53	67.76	**75.85**	79.14	76.33	78.56	65.92	71.63	73.42
MetaMixSpeech	79.39	78.09	**64.83**	**67.48**	76.59	**77.09**	**72.96**	**67.42**	**58.35**	**68.83**	**71.10**

ing using eight source languages; (iv) Meta-ASR: Pretraining the adapters by meta learning using eight source languages.

4 Experiment Results

4.1 Main Result

Results on the FLEURS dataset. Table 2 shows the performance of various methods on ten languages using a 5%-shot subset of the FLEURS datasets. First, using FT-Full failed to achieve convergence under this extremely low-resource scenario, with the WER for most languages remaining around 100%. We infer that the model may overfit under this low-resource setting. In contrast, employing a randomly initialized adapter (Vanilla-ASR) allowed the model to converge for all languages except Urdu (Ur), demonstrating the adapter's effectiveness. Second, it is evident that Multi-ASR outperforms Vanilla-ASR in most languages due to the benefits of multilingual pretraining. However, Meta-ASR surpasses Multi-ASR in most languages, showcasing superior performance in fast learning for low-resource settings.

Additionally, we investigated the effects of applying only support augmentation (MixS) or query augmentation (MixQ). Experimental results demonstrate that both augmentation strategies are effective. Notably, query augmentation has a greater impact compared to support augmentation, suggesting that learner overfitting is a more critical issue than task overfitting in this context. This is likely because the languages come from various language families, and the primary role of the meta-learner is to enhance the model's generalization across multiple languages. Furthermore, our proposed MetaMixSpeech outperforms other methods in most languages, achieving an additional 6.35% improvement in WER over Meta-ASR. This underscores its superior generalizability in adapting to low-resource languages. These findings also indicate that when query augmentation is applied, the benefits of support augmentation become evident, highlighting the crucial role of the meta-learner in influencing meta-generalization.

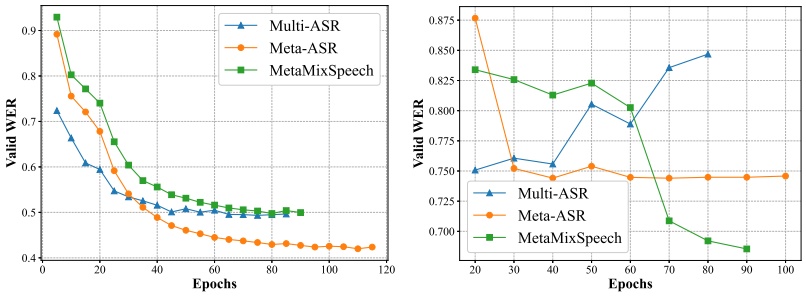
Results on the Common Voice dataset. In addition, we evaluated the performance of these methods on eight target languages in the Common Voice dataset and compared our results with those from previous works (Hou et al., 2022, 2021). As shown in Table 3, our baseline surpasses prior approaches, which can be attributed to the employment of the pre-trained self-supervised model. Moreover, pre-trained methods such as Multi-ASR and Meta-ASR consistently outperform Vanilla-ASR overall. And Meta-ASR shows superior performance compared to Multi-ASR, which is consistent with the observations made on the FLEURS dataset.

However, we find that for some languages (e.g., Ga-IE, Cs), the performance of Multi-ASR and Meta-ASR is inferior to that of Vanilla-ASR. This suggests an overfitting issue with the pre-training methods, which affects their generalization to target tasks. Additionally, the overfitting in Meta-ASR is less pronounced compared to Multi-ASR, as observed in languages like Hsb and Uk. However, MetaMixSpeech effectively mitigates the overfitting issue and achieves the best performance across most languages, demonstrating its robustness and superior capability in adapting to low-resource languages.

Extensibility to Other Different Language Families. Multilingual pre-training is highly dependent on the similarity between the pre-training languages and the target languages. The languages used in our pre-training mainly belong to the West-European language family. To analyze the gener-

Table 3: Word error rate (WER 100 %) on eight target languages of the Common Voice dataset using various adaptation methods.

Methods	Br	Ga-IE	Ro	Or	Hsb	Ar	Cs	Uk	Avg.
MAML-ASR (Hou et al., 2021)	80.70	68.10	66.00	64.80	75.60	/	/	/	/
Reptile-ASR (Hou et al., 2021)	79.90	67.00	64.30	64.10	75.70	/	/	/	/
Meta-Adapter (Hou et al., 2022)	58.49	/	44.59	/	/	46.82	37.13	49.36	/
SimAdapter+ (Hou et al., 2022)	59.14	/	47.29	/	/	46.39	34.72	47.41	/
Vanilla-ASR	52.24	**32.32**	41.54	62.76	67.28	42.90	21.42	24.21	43.08
Multi-ASR	49.42	37.49	37.90	47.07	70.15	36.60	23.73	25.56	40.99
Meta-ASR	46.72	33.05	36.45	45.62	64.20	37.92	22.82	23.99	38.84
MetaMixSpeech	**45.52**	34.05	**36.35**	**41.22**	**55.17**	**33.60**	**20.65**	**22.22**	**36.09**



(a) Source Languages (b) Target Languages

Figure 4: Average WER on source and target languages under different pretraining epochs using various adaptation methods.

alization of our proposed methods across different language families, we observe the languages in Table 2: Malayalam (Ml), Nepali (Ne), Urdu (Ur), Malay (Ms), Vietnamese (Vi), and Kazakh (Ka), which come from other language families: South Asia, Central-Asia, and Southeast Asia. Experiment results show that MetaMixSpeech effectively mitigates meta-overfitting problems and shows superior generalization and extensibility across different language families, especially in Malay (Ms), Vietnamese (Vi), and Kazakh (Ka).

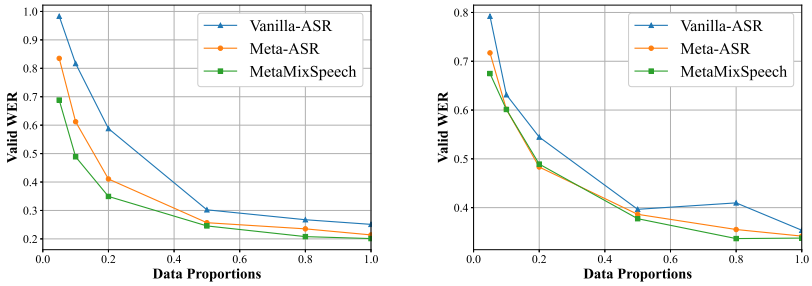
In summary, MetaMixSpeech demonstrates robust performance across different datasets and different language families, exhibiting excellent generalization.

4.2 Discussion

4.2.1 Different Pretraining Epochs

Our analysis revealed that Meta-ASR may have limitations concerning both learner overfitting and task overfitting. To investigate this issue, we analyzed the average WER of source languages at different pre-training epochs using various pre-training methods, as shown in Figure 4a. Experimental results show that Meta-ASR exhibited a lower WER in fitting the pre-training languages compared to other methods, indicating an over-learning tendency. In contrast, Multi-ASR converges more quickly but its convergence performance is limited. Moreover, through task augmentation, MetaMixSpeech effectively reduced Meta-ASR's over-learning on source languages, leading to earlier convergence.

However, the performance on source languages alone may not fully reflect the overfitting problem. Therefore, we further evaluated the average performance of different methods on three target languages (Finnish, Galician, and Kazakh) using 5% of the data at different pre-training epochs, as shown in Figure.4b. We observed that although Multi-ASR did not exhibit over-learning on source languages, it experienced overfitting on the target languages after the 20th epoch. This suggests that the Multi-ASR paradigm is not as robust as Meta-ASR, as it overfits quickly and more severely. In contrast, Meta-ASR's performance remained relatively stable with increasing epochs, avoiding a high WER. However, MetaMixSpeech demonstrated even better performance than Meta-ASR, indicating its effectiveness in adapting target languages and mitigating meta-overfitting. This highlights MetaMixSpeech's superior capability in en-



(a) Kazakh (b) Finnish

Figure 5: Valid WER curves under different proportions of data using different adaptation methods.

hancing model's generalization and robustness.

4.2.2 Different Proportions of Data

To explore the relationship between the methods' performance and the amount of training data, we sampled different proportions of data from Kazakh and Finnish for adaptation. The results are shown in Figure 5. MetaMixSpeech consistently achieves the best performance across different data proportions, demonstrating its effectiveness. Notably, MetaMixSpeech using only 10% of the data achieves comparable performance to Vanilla-ASR using 30% of the data in Kazakh, highlighting its effectiveness in low-resource scenarios.

For Kazakh, MetaMixSpeech consistently maintains the best performance, indicating that the overfitting issue always exists regardless of the amount of data. This is likely because Kazakh does not belong to the same language family as the source languages (West European), leading to serious learner overfitting that cannot be diminished even as the data increases. For Finnish, the performance difference between MetaMixSpeech and other methods decreases when the data proportions increases to 10%. This reduction in the performance gap can be attributed to Finnish belonging to the Western European language family. As it is more similar to the source languages, the overfitting problem manifests as task overfitting, which becomes less pronounced with the availability of more data.

4.2.3 Learning Curves of Fine-tuning Target Languages

To explore the rapid adaptation process of our method, we fine-tuned the Finnish language using only 5% of the data. As is shown in Figure.6, experimental results indicate that Vanilla-ASR converge to suboptimal performance eventually. Multi-ASR and Meta-ASR do not converge in the first 20 epochs, then rapidly decline, and finally converge to a relatively good performance. We attribute this behavior to warmup mechanisms, which require a larger learning rate to break free from local optima. Overall, Meta-ASR achieves faster and better performance than Multi-ASR, showing the fast learning ability of meta-learning. However, it is evident that MetaMixSpeech breaks free from local optima more quickly and achieves superior performance compared to Meta-ASR within the first 20 epochs. This demonstrates its fast learning ability and superiority in low-resource scenarios.

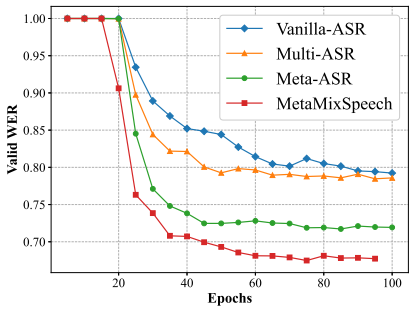


Figure 6: Valid WER of Finnish using 5% of data under different adaptation methods.

5 Conclusion

In this paper, we first analyze two overfitting issues in multilingual meta-learning: task overfitting and learner overfitting. Therefore, we introduce MetaMixSpeech, an innovative interpolation-based task augmentation method that enhances task diversity through support augmentation and query augmentation to overcome these problem. Experimental results on the FLEURS and Common Voice datasets demonstrate that MetaMixSpeech is highly effective, achieving superior performance and generalization across multiple language families.

Limitations

Although MetaMixSpeech performs superbly in a variety of languages, there still exist some limitations in our work: 1) We use eight languages with five hours each for meta-training. While in most settings, the data might be imbalanced for training, we have neglected the long-tail distribution effect of the data. 2) Our experiments are solely conducted on speech tasks and do not include various NLP tasks, even though they are all language-based.

To address these limitations, we plan to expand our future meta-training settings to account for long-tail distributions. Additionally, we intend to extend our method to other NLP tasks, like machine translation and language generation tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.62171470), Natural Science Foundation of Henan Province (No.232300421240, 252300420990).

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *LREC*, pages 4218–4222.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. FLEURS: few-shot learning evaluation of universal representations of speech. In *IEEE Spoken Language Technology Workshop, SLT*, pages 798–805. IEEE.

Chelsea Finn, P. Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*.

Wenxin Hou, Yue Dong, Bairong Zhuang, Longfei Yang, Jiatong Shi, and Takahiro Shinozaki. 2020. Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning. In *Interspeech*, pages 1037–1041.

Wenxin Hou, Yidong Wang, Shengzhou Gao, and Takahiro Shinozaki. 2021. Meta-adapter: Efficient cross-lingual adaptation with meta-learning. In *ICASSP*, pages 7028–7032. IEEE.

Wenxin Hou, Han Zhu, Yidong Wang, Jindong Wang, Tao Qin, Renjun Xu, and Takahiro Shinozaki. 2022. Exploiting adapters for cross-lingual low-resource speech recognition. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:317–329.

Jui-Yang Hsu, Yuan-Jui Chen, and Hung yi Lee. Meta learning for end-to-end low-resource speech recognition. *ICASSP 2020*, pages 7844–7848.

Ke Hu, Antoine Bruguier, Tara N. Sainath, Rohit Prabhavalkar, and Golan Pundak. 2019. Phoneme-based contextualization for cross-lingual speech recognition in end-to-end models. In *Interspeech 2019*, pages 2155–2159.

Haebeom Lee, Taewook Nam, Eunho Yang, and Sung Ju Hwang. 2020. Meta dropout: Learning to perturb latent features for generalization. In *ICLR*.

Paul M. A. Lewis. 2009. Ethnologue : languages of the world.

Jialin Liu, Fei Chao, and Chih-Min Lin. 2020. Task augmentation by rotating for meta-learning. *Arxiv*, abs/2003.00804.

Linghui Meng, Jin Xu, Xu Tan, Jindong Wang, Tao Qin, and Bo Xu. 2021. Mixspeech: Data augmentation for low-resource automatic speech recognition. In *ICASSP*, pages 7008–7012. IEEE.

Renkun Ni, Micah Goldblum, Amr Sharaf, Kezhi Kong, and Tom Goldstein. 2020. Data augmentation for meta-learning. *Arxiv*, abs/2010.07092.

Shinta Otake, Rei Kawakami, and Nakamasa Inoue. 2022. Parameter efficient transfer learning for various speech processing tasks. *ArXiv*, abs/2212.02780.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*, pages 2613–2617.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518.

Janarthanan Rajendran, Alexander Irpan, and Eric Jang. 2020. Meta-learning requires meta-augmentation. In *NeurIPS*.

Satwinder Singh, Ruili Wang, and Feng Hou. 2022. Improved meta learning for low resource speech recognition. In *ICASSP*, pages 4798–4802. IEEE.

Bethan Thomas, Samuel Kessler, and Salah Karout. 2022. Efficient adapter transfer of self-supervised speech models for automatic speech recognition. In *ICASSP*, pages 7102–7106. IEEE.

Qiuli Wang, Wen-Rui Hu, Lin Li, and Qingyang Hong. 2023. Meta learning with adaptive loss weight for low-resource speech recognition. *ICASSP*.

Huaxiu Yao, Long-Kai Huang, Linjun Zhang, Ying Wei, Li Tian, James Y. Zou, Junzhou Huang, and Zhenhui Jessie Li. 2020. Improving generalization in meta-learning via task augmentation. In *ICML*.

Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. 2020. Meta-learning without memorization. In *ICLR* ,.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, and Gary Wang. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *ArXiv*, abs/2303.01037.