Low-Rank and Sparse Model Merging for Multi-Lingual Speech Recognition and Translation

Anonymous ACL submission

Abstract

Language diversity presents a significant challenge in speech-to-text (S2T) tasks, such as automatic speech recognition and translation. Traditional multi-lingual multi-task training approaches aim to address this by jointly optimising multiple speech recognition and translation tasks across various languages. While models like Whisper, built on these strategies, demonstrate strong performance, they still face issues of high computational cost, language interference, suboptimal training configurations, and limited extensibility. To overcome these challenges, we introduce LoRS-Merging (low-rank and sparse model merging), a novel technique designed to efficiently integrate models trained on different languages or tasks while preserving performance and reducing computational overhead. LoRS-Merging combines low-rank and sparse pruning to retain essential structures while eliminating redundant parameters, mitigating language interference, and enhancing extensibility. Experimental results across 10 languages demonstrate that LoRS-Merging significantly outperforms multi-lingual multi-task training, sequential training, and other merging methods, achieving over 20% improvement in normalised performance. Our findings suggest that model merging, particularly LoRS-Merging, is a scalable and effective complement to traditional multi-lingual training strategies for S2T applications[1].

1 Introduction

Language diversity poses a significant challenge in speech-to-text (S2T) tasks, such as automatic speech recognition (ASR) (Prabhavalkar et al., 2023) and speech translation (ST) (Xu et al., 2023). With over 7,000 languages spoken worldwide, developing robust S2T systems that generalise across varied linguistic structures remains a fundamental research goal (Liu and Niehues, 2024; Cheng et al., 2023; Sun et al., 2023; Saif et al., 2024; Wang et al., 2021; Le et al., 2021). The transition from pipeline systems to end-to-end (E2E) models (Chan et al., 2016; Gulati et al., 2020; Barrault et al., 2023) has marked a paradigm shift in S2T tasks, enabling direct speech-to-text mapping across multiple languages within a unified framework. A prominent example is Whisper (Radford et al., 2023), an advanced multi-lingual speech model trained on a large-scale, diverse dataset covering multiple languages and tasks. Despite these advances, existing multi-lingual models still encounter significant challenges in scalability, efficiency, and performance trade-offs.

To address these challenges, multi-lingual training strategies (Saif et al., 2024; Xiao et al., 2021; Bai et al., 2018) have been adopted, aiming to enhance model generalisation across languages. These approaches typically rely on joint optimisation of diverse S2T tasks across multiple languages, leveraging shared representations to improve performance. Nevertheless, multi-lingual training is subject to inherent limitations, including substantial training costs, complex model configurations, and limited access to training data across multiple languages and tasks. Moreover, when handling new languages, the training methods typically require training from scratch, leading to poor extensibility.

To mitigate these issues, this paper proposes to use model merging (Ilharco et al., 2023; Yang et al., 2024a; Khan et al., 2024) to integrate models trained on different languages or tasks while maintaining performance and reducing computational overhead. Model merging merges the parameters of multiple separate models with different capabilities to build a universal model. With its high flexibility, model merging enables the seamless incorporation of new languages or tasks without the need for retraining the entire model. Additionally, since model merging allows models for different languages or tasks to be trained independently, it

[1]The detailed data and code will be released at [URL]

can effectively alleviate negative transfer issues (Wang et al., 2019; Zhang et al., 2023b; Wang et al., 2020b) commonly observed in multi-lingual training. This training independence also enables optimal training configurations for each language or task to improve performance, instead of the unified settings required in multi-lingual training.

Moreover, we propose **Lo**w-**R**ank and **S**parse model **Merging** (LoRS-Merging), which uses a low-rank component to capture the compact structure and a sparse component to capture the scattered details in the weights. LoRS-Merging retains effective parts of structure and details while reducing redundant parts to reduce language interference. Specifically, coarse-grained singular value pruning is used to retain the low-rank structure, while fine-grained magnitude pruning is used to remove redundant details. The main contribution of this paper can be summarised as follows.

- To the best of our knowledge, this work is the first to explore model merging for speech-to-text models. Specifically, we treat speech tasks (recognition and translation) and different languages as two separate merging levels and explore different hierarchies for model merging.
- We propose LoRS-Merging, which exploits the low-rank structure and sparsity of model weights to minimise model redundancy and language conflicts as well as providing an efficient way to incorporate new knowledge from a task- or language-specialised model.
- Experiments are performed across 10 languages, where LoRS-Merging significantly outperforms multi-lingual multi-task training, sequential training, and other merging methods, achieving over 20% improvement in normalised performance.

2 Related Work

2.1 Multi-Lingual ASR and ST

Multi-lingual speech models inherently face a trade-off between knowledge sharing and negative interference. Early studies adopted hand-picked sub-network sharing strategies, such as language-specific decoders (Dong et al., 2015), attention heads (Zhu et al., 2020), and layer norm/linear transformation (Zhang et al., 2020). Recent research has shifted toward approaches such as mixture-of-experts (Kwon and Chung, 2023; Wang et al., 2023), adapters (Le et al., 2021; Kannan et al., 2019), and pruning (Lu et al., 2022; Lai et al., 2021). To enhance multi-lingual representation learning, language tokens (Johnson et al., 2017), embeddings (Di Gangi et al., 2019) or output factorisations (Zhang et al., 2023a) are introduced to encode language identity, helping the model distinguish between languages.

The more effective approach is to adopt multi-lingual training strategies, such as multi-objective optimisation (Saif et al., 2024; Zhang et al., 2022), adversarial learning (Xiao et al., 2021), meta learning (Hsu et al., 2020), and reinforcement learning (Bai et al., 2018). Moreover, large-scale pretraining by leveraging massive amounts of multi-lingual and multi-task data enables models to learn robust and transferable representations across languages, e.g. Whisper (Radford et al., 2023), SeamlessM4T (Barrault et al., 2023), and AudioPaLM (Rubenstein et al., 2023). LoRS-Merging, as an efficient post-training method proposed in this paper, further advances multi-lingual ASR and ST based on pretrained speech models.

2.2 Model Merging

Model merging (Yang et al., 2024a; Khan et al., 2024) is an efficient post-training technique that integrates knowledge from models trained on different domains. One stream of research focuses on the loss landscape geometry (Khan et al., 2024) and studies the linear mode connectivity (LMC) (Frankle et al., 2020; Draxler et al., 2018) property that demonstrates the existence of a linearly connected path between local minima within the same loss basin. Many studies (Nagarajan and Kolter, 2019; Izmailov et al., 2018; Frankle et al., 2020) indicate that if two neural networks share part of their optimisation trajectory, such as different finetuned models from the same pretrained model, they typically satisfy LMC, allowing interpolation without sacrificing accuracy and forming the basis of our model merging method. For local minima in different loss basins, inspired by the permutation invariance (Entezari et al., 2021) of neural networks, neuron alignment techniques (Ainsworth et al., 2023; Singh and Jaggi, 2020; Tatro et al., 2020) can be used to place them into the same basin, thereby reducing merging loss.

Another stream considers the model spaces, including activation spaces and weight spaces. Research on activation spaces seeks to align the output representations or loss of the merged model with those of each single model as closely as possible (Yang et al., 2024b; Wei et al., 2025; Xiong et al., 2024). Studies based on weight spaces aim to localise effective parameters or remove redundant parameters to resolve task interference. TALL-masks (Wang et al., 2024) and Localise-and-Stitch (He et al., 2024) optimise binary masks to localise sparse and effective task-specific parameters. TIES-Merging (Yadav et al., 2024) and DARE (Yu et al., 2024) perform magnitude or random pruning on each single model to reduce redundancy at the detailed parameter level. TSV-M (Gargiulo et al., 2024), on the other hand, adopts singular value pruning to reduce redundancy at the structural level. In contrast, LoRS-Merging explores weight space merging by considering not only the detailed parameter redundancy as well as maintaining the effective structure of the weight space.

3 Methodology

3.1 Preliminaries

3.1.1 Task Arithmetic

Among diverse model merging methods, Task Arithmetic (TA) (Ilharco et al., 2023) has become a fundamental technique in this field due to its simplicity and effectiveness. TA introduces the concept of "task vector", defined as the delta parameter derived by subtracting pretrained weights from finetuned weights. By performing simple arithmetic operations on task vectors, TA enables task learning, forgetting, and analogising.

Assume that $\theta = \{W_l\}_{l=1}^{L}$ represents the parameters of the model, where W_l is the weight of l-th layer, and L is the total number of layers. Given a pretrained model θ_0 and a model θ_i finetuned on task t_i, the task vector is computed as $\tau_i = \theta_i - \theta_0$. Multiple task vectors can be summed to form a multi-task model, expressed as $\theta_{\text{merged}} = \theta_0 + \lambda \sum_{i=1}^{n} \tau_i$, where λ is a scaling coefficient for the task vectors.

3.1.2 Pruning

Given that neural networks are typically over-parameterised and exhibit high redundancy, a considerable number of neurons or connections can be pruned without affecting accuracy (LeCun et al., 1989). In model merging, pruning methods can reduce redundant parameters to mitigate task interference, thereby improving the merging performance.

Magnitude Pruning (MP) is an unstructured pruning method that prunes connections based on the magnitude of parameters as a measure of importance. Specifically, MP prunes the parameters according to a specific ratio p, as follows.

$$M_{ij} = \begin{cases} 1 & \text{if } |w_{ij}| \in \text{top } p\% \\ 0 & \text{o.w.} \end{cases} \tag{1}$$

$$W_{\text{pruned}} = M \odot W \tag{2}$$

where $W, M \in \mathbb{R}^{d \times k}$, and $\odot$ denotes the element-wise multiplication. However, MP only focuses on the redundancy at the parameter level, overlooking the crucial structural information, which may lead to the disruption of the weight structure.

Singular Value Pruning (SVP) is a structured pruning method that removes smaller singular values and their corresponding singular vectors. In particular, SVP retains only the top r singular values while discarding the others.

$$W = U \Sigma V^T \tag{3}$$

$$W_{\text{pruned}} = U_r \Sigma_r V_r^T \tag{4}$$

where $U \in \mathbb{R}^{d \times d}$ and $V \in \mathbb{R}^{k \times k}$ are the left and right singular vector matrices of W, and U_r, V_r denote their first r columns. Although SVP preserves a compact weight structure, its coarse pruning granularity makes it challenging to reduce redundancy at a fine-grained parameter level.

3.2 Model Merging for Speech Models

The model merging process for speech model on S2T tasks with LoRS-Merging as an example is shown in Fig. 1, which comprises four steps. In step 1, a suitable pretrained speech model is selected. In step 2, for each target language and target task combination, e.g. Catalan ASR, the pretrained model is finetuned with the task-language-specific data and the delta weight is obtained. In step 3, weight pruning is applied to remove redundant and conflicting delta parameters. In step 4, task arithmetic is applied to combine pruned delta weights into each single merged matrix and hence obtain the merged model.

Model merging allows new language or task knowledge to be integrated into the model in a flexible post-training manner. Compared to multilingual or multi-task training methods, model merging is a simpler and more efficient approach, enabling the seamless incorporation of new languages or tasks without the need for retraining. Additionally, due to its training independence, it mitigates language conflicts and provides optimal training configurations for each language or task to improve performance. Compared to sequential training,

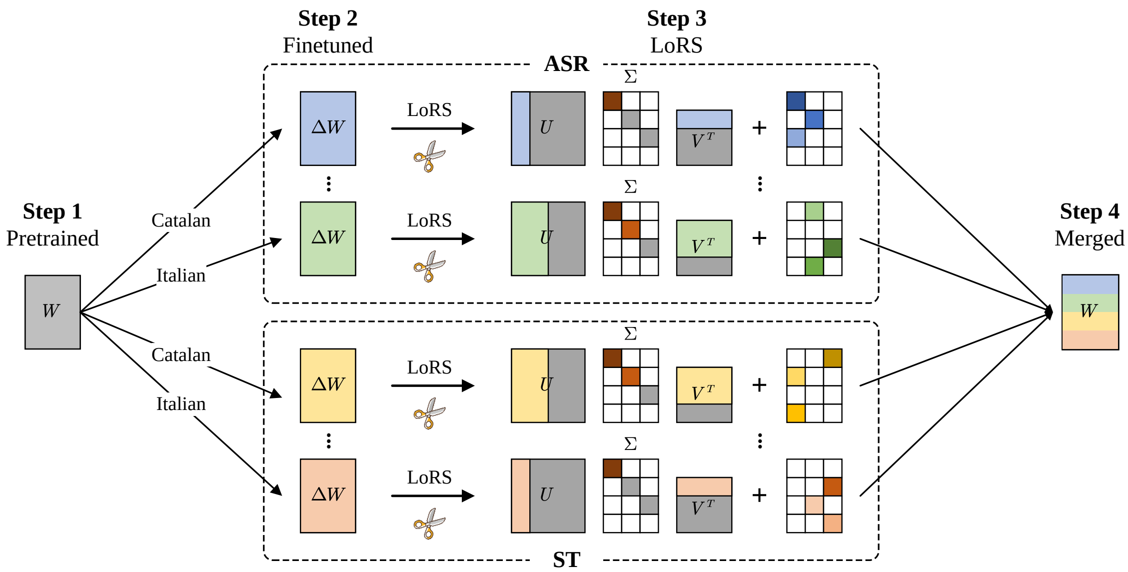

Figure 1: Model merging process with the proposed LoRS-Merging for speech models on multi-lingual ASR and ST tasks. In step 1, a suitable pretrained speech model is selected. In step 2, the pretrained model is finetuned with the task-language-specific data. In step 3, apply LoRS to the delta parameters to reduce model redundancy. In step 4, merge the delta parameters to get a multi-lingual and multi-task merged model.

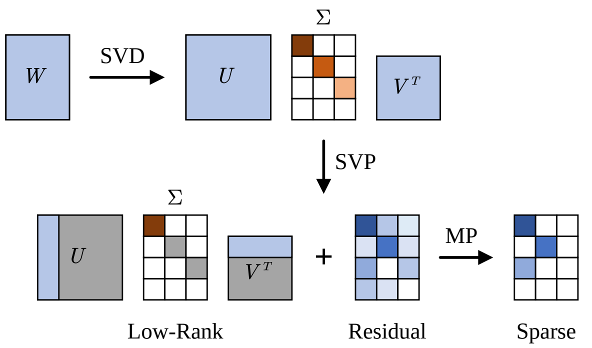

Figure 2: Illustration of LoRS-Merging method in detail. SVD stands for singular value decomposition and SVP for singular value pruning. MP is magnitude pruning operating on residual of the original weight matrix and the low-rank matrix.

model merging eliminates the need for additional training data to avoid catastrophic forgetting. Our experiments thoroughly demonstrate these benefits.

3.3 Low-Rank and Sparse Model Merging

The weights of neural networks contain information on both structure and details. Structural information is coherent, compact, and coarse-grained, whereas detail information is incoherent, scattered, and fine-grained. Both structural and detail information include effective and redundant parts. To reduce redundant parts in both the structure and detail aspects of the weights while retaining effective parts, the LoRS-Merging method is introduced as shown in detail in Fig. 2, which exploits the combination of low-rank structure by SVP and sparsity by MP. SVP performs coarse-grained pruning at the structure level, while MP enables fine-grained pruning at the detail level.

In the implementation, we approximate the original weights as the sum of a low-rank component and a sparse component, where the low-rank component captures the compact structure, and the sparse component captures the scattered details, as shown in Eqn. (5).

$$W \approx L + S \tag{5}$$

where L represents the low-rank component, and S represents the sparse component. Specifically, L is the low-rank matrix obtained by retaining the top r singular values and their corresponding singular vectors from W:

$$L = U_r \Sigma_r V_r^T \tag{6}$$

and S is the sparse matrix obtained by performing MP on the residual of W and L:

$$S = M \odot (W - L) \tag{7}$$

To simplify the description, we refer to this entire process as LoRS$(\cdot)$. In this manner, SVP decouples the structure and details of the weight, preserving a compact structure while allowing fine-grained MP to remove redundant parts in the details.

For each model finetuned on single specific language or task data, we apply LoRS$(\cdot)$ to its task vector as a preprocessing step to reduce language or task interference in model merging. A multi-lingual or multi-task model can be achieved through simple merging, expressed as:

$$\theta_{\text{merged}} = \theta_0 + \lambda \sum_{i=1}^{n} \text{LoRS}(\tau_i) \tag{8}$$

4 Experimental Setup

4.1 Data

CoVoST-2 (Wang et al., 2020a) is a large-scale multi-lingual ST corpus based on Common Voice. It covers translations from English into 15 languages and from 21 languages into English, with a total of 2,880 hours of speech from 78k speakers. We selected 5 high-resource languages and 5 low-resource languages as two language sets to investigate their ASR tasks and the from X to English ST tasks. The high-resource language set includes Catalan (ca), German (de), Spanish (es), French (fr), and Italian (it), while the low-resource language set includes Indonesian (id), Dutch (nl), Portuguese (pt), Russian (ru), and Swedish (sv). Due to the more abundant data in the high-resource language set, our main experimental results are obtained on the high-resource language set, while the low-resource language set serves as an auxiliary evaluation set. To balance the amount of data across different languages, we fixed the duration of traning data for each language, with 5 hours for the high-resource language set and 1 hour for the low-resource language set. The dev and test sets of both language sets are 1 hour in duration.

4.2 Model and Training Specifications

Whisper (Radford et al., 2023) is a general-purpose multi-lingual ASR and ST model, a Transformer-based model trained on 680k hours of diverse audio. We chose the small version as the foundation model for the experiments because it achieves a good balance between performance and cost. It has 244 million parameters, with the encoder and decoder each consisting of 12 Transformer blocks. The weight matrices of the attention layers are all 768 by 768, and the MLP layers are 768 by 3072.

For each language-specific or task-specific finetuned model, we use a different, optimal learning rate for each during training, and these models are subsequently used for model merging. Finetuning involves updating all parameters. We choose Adam as the optimiser, set the batch size to 8, the accumulation iterations to 4, and train for 10 epochs. The beam size for decoding is set to 20 across all languages and tasks. We use Sclite and SacreBLEU tools to score the ASR and ST results, respectively. In addition, we perform statistical significance testing using a paired bootstrap test with 1,000 resampling iterations, each sampling 300 examples with replacement, and report the results in the caption of each table. See Appendix A for more details on the experimental setup. Our experiments are performed on a single RTX 4090 GPU where training on one language and one task with 5 hours of speech data requires 1 hour.

4.3 Baseline and Merging Methods

We use the **pretrained model** as the baseline and **multi-lingual multi-task training** as the stronger baseline, which is trained on data mixed from both multi-lingual and multi-task sets. Note that finetuned models are typically available, so model merging requires no finetuning and only adjusts merging coefficients on a small development set. Even if finetuning is needed, the complex training configurations of multi-lingual and multi-task training require more hyperparameter tuning steps. Overall, model merging consumes significantly fewer computational resources than multi-lingual and multi-task training.

In addition to LoRS-Merging, we investigate the following model merging methods:

Weight Averaging (WA) (Wortsman et al., 2022) merges multiple single models by a uniform averaging of their weights.

Task Arithmetic (TA) (Ilharco et al., 2023) uses a scaling factor, estimated on a small development set, to weight multiple task vectors.

TIES-Merging (Yadav et al., 2024) merges single models via Trim, Elect, and Disjoint Merge steps to reduce parameter interference.

DARE (Yu et al., 2024) drops and rescales delta parameters to mitigate parameter interference.

TSV-M (Gargiulo et al., 2024) proposes Task Singular Vectors and reduces structural redundancy to reduce task interference.

In addition, we report the normalised performance difference defined in Eqn. (9).

$$\Delta_{\text{norm}} = \frac{|M - M_{\text{pretrained}}|}{|M_{\text{finetuned}} - M_{\text{pretrained}}|} \times 100\% \quad (9)$$

where M is the performance metric of the target system, $M_{\text{pretrained}}$ and $M_{\text{finetuned}}$ are for the pretrained (baseline performance) and finetuned (topline performance) systems respectively. *Note that Δ_{norm} better reflects the performance gains for model merging since the pretrained system already achieves competitive performance.*

Table 1: Multi-lingual ASR model merging. Avg. denotes average WER. $*$ LoRS-Merging outperforms all others in Δ_{norm} by >20% ($p < 0.05$).

System	WER↓						
	ca	de	es	fr	it	Avg.	Δ_{norm}
Pretrained	20.6	19.6	14.7	24.5	19.4	19.88	-
Finetuned	19.5	19.7	14.4	22.1	19.2	19.05	100.0%
Multi-lingual training	17.1	21.8	15.1	22.6	21.9	19.69	22.9%
Sequential training	20.6	19.6	14.6	24.4	19.4	19.84	4.8%
Weight Averaging	19.1	19.1	14.2	24.5	20.3	19.55	39.8%
Task Arithmetic	19.1	18.8	13.9	24.0	19.8	19.23	78.3%
TIES-Merging	19.3	19.3	13.9	23.8	18.1	18.99	107.2%
DARE	18.9	18.9	13.9	23.8	19.8	19.16	86.7%
TSV-M	19.5	19.5	14.1	23.5	18.4	19.10	94.0%
LoRS-Merging	18.9	18.8	13.9	23.6	18.1	**18.77**	**133.7%**

Table 2: Multi-lingual ST model merging. Avg. denotes average BLEU. $*$ LoRS-Merging outperforms all others in Δ_{norm} by >20% ($p < 0.05$).

System	BLEU↑						
	ca	de	es	fr	it	Avg.	Δ_{norm}
Pretrained	21.1	24.1	28.6	26.8	26.8	25.48	-
Finetuned	22.6	24.6	29.2	27.2	27.3	26.18	100.0%
Multi-lingual training	21.4	24.4	28.8	26.8	27.2	25.72	34.3%
Sequential training	21.5	24.3	28.9	26.9	27.3	25.78	42.9%
Weight Averaging	22.3	24.1	28.6	27.2	26.9	25.82	48.6%
Task Arithmetic	22.1	24.3	28.9	27.3	26.8	25.88	57.1%
TIES-Merging	22.1	24.7	29.0	27.1	26.9	25.96	68.6%
DARE	22.1	24.5	28.9	27.2	26.8	25.90	60.0%
TSV-M	22.0	24.6	29.0	27.3	26.8	25.94	65.7%
LoRS-Merging	22.4	24.8	28.9	27.6	27.0	**26.14**	**94.3%**

Table 3: Multi-task model merging performed on each language independently. Avg. denotes average WER/BLEU. $*$ LoRS-Merging outperforms all others in Δ_{norm} by >20% ($p < 0.05$). Per-language results are shown in Appendix C.

System	Avg. WER↓	Δ_{norm}	Avg. BLEU↑	Δ_{norm}
Pretrained	19.88	-	25.48	-
Finetuned	19.05	100.0%	26.18	100.0%
Multi-task training	19.00	106.0%	25.90	60.0%
Sequential training	18.95	112.0%	26.12	91.4%
Weight Averaging	18.84	125.3%	26.18	100.0%
Task Arithmetic	18.76	134.9%	26.30	117.1%
TIES-Merging	18.60	154.2%	26.38	128.6%
DARE	18.71	141.0%	26.28	114.3%
TSV-M	18.70	142.2%	26.40	131.4%
LoRS-Merging	**18.39**	**179.5%**	**26.56**	**154.3%**

5 Evaluation Results and Analysis

5.1 Multi-Lingual Model Merging

First, we investigate the merging of finetuned models for different languages on the same task, which corresponds to *multi-lingual single-task* learning.

Language knowledge interference yields imbalanced improvements: Table 1 shows the multi-lingual results of the ASR task with the high-resource language set. On average, multi-lingual training slightly improves the pretrained model but significantly underperforms the finetuned models and merging methods. This may be due to negative interference between the knowledge of different languages, leading to gradient conflicts during training (Wang et al., 2020b). From a per-language perspective, it is observed that ca and fr achieve the largest improvements during finetuning while still showing significant improvements in multi-lingual training, whereas languages with smaller improvements during finetuning exhibit a substantial performance drop in multi-lingual training, even worse than the pretrained model. This indicates a strong language conflict in multi-lingual training, with ca and fr dominating. Additionally, we observe that the optimal learning rates for finetuned models vary significantly across languages (see Appendix A), while the unified learning rate configuration required by multi-lingual training prevents each language from reaching its optimal performance. Moreover, the substantially inferior performance of sequential training indicates the presence of catastrophic forgetting.

Model merging mitigates language conflicts: In contrast, model merging methods show significant improvements across almost all languages, demonstrating reduced conflict and better stability. Among model merging methods, TA outperforms WA due to its flexible scaling factor. TIES-Merging and DARE reduce redundancy at the detail level, while TSV-M addresses redundancy at the structural level, thereby achieving obvious improvements over TA. Furthermore, LoRS-Merging reduces redundancy at the detail level while preserving critical structures, resulting in the best performance.

Table 2 provides the multi-lingual results on ST task with the high-resource language set. The main conclusion is consistent with the ASR task: model merging methods still significantly outperform multi-lingual training and sequential training, with LoRS-Merging achieving the best performance, demonstrating superior multi-lingual and continual learning capabilities.

5.2 Multi-Task Model Merging

Next, we merge finetuned models for different tasks (ASR and ST) with the same language which corresponds to *multi-task single-language* learning.

ASR and ST tasks for the same language can mutually benefit from each other: Table 3 presents the multi-task results with the high-

Table 4: Multi-lingual multi-task model merging. Avg. denotes average WER/BLEU.

System	WER↓							BLEU↑						
	ca	de	es	fr	it	Avg.	Δ_{norm}	ca	de	es	fr	it	Avg.	Δ_{norm}
Pretrained	20.6	19.6	14.7	24.5	19.4	19.88	-	21.1	24.1	28.6	26.8	26.8	25.48	-
Finetuned	19.5	19.7	14.4	22.1	19.2	19.05	100.0%	22.6	24.6	29.2	27.2	27.3	26.18	100.0%
ML and MT training	20.5	19.7	14.6	24.5	19.4	19.86	2.4%	21.3	24.3	28.3	27.1	26.9	25.58	14.3%
ML and MT Task Arithmetic	18.9	19.2	14.1	23.7	18.4	18.96	110.8%	22.2	24.4	29.0	27.3	26.9	25.96	68.6%
ML and MT LoRS-Merging	18.7	19.1	14.0	23.8	18.0	18.82	127.7%	22.2	24.8	29.0	27.5	27.0	26.10	88.6%
MT training	17.0	19.7	14.4	24.2	19.4	19.00	-	22.3	24.6	28.7	27.0	26.9	25.90	-
↪ + ML Task Arithmetic	18.1	19.0	14.2	24.5	20.6	19.37	61.4%	22.7	24.7	28.6	27.3	26.5	25.96	68.6%
↪ + ML LoRS-Merging	18.1	19.0	14.1	24.2	20.3	19.23	78.3%	22.4	24.5	29.1	27.6	26.7	26.06	82.9%
ML training	17.1	21.8	15.1	22.6	21.9	19.69	-	21.4	24.4	28.8	26.8	27.2	25.72	-
↪ + MT Task Arithmetic	17.1	18.5	13.3	22.7	18.0	18.00	226.5%	22.6	25.0	29.2	27.5	26.9	26.24	108.6%
↪ + MT LoRS-Merging	16.9	18.3	13.3	22.4	17.8	**17.82**	**248.2%**	22.8	25.2	29.3	27.6	27.0	**26.38**	**128.6%**

resource language set. In general, multi-task training performs similarly to finetuned models on ASR but is a lot worse on ST. This is likely due to the substantial differences in optimal hyper-parameter configurations between the two tasks. Sequential training performs similarly to finetuned models overall, as it also benefits from training independence. Model merging methods clearly outperform finetuned models, which not only demonstrates their effectiveness but also shows the mutual benefits between ASR and ST. In terms of performance gains, the improvement in ASR is greater than in ST. We attribute this to the fact that ASR is inherently simpler than ST and can be viewed as a step in the ST task. Furthermore, as before, model merging methods combined with pruning further improve performance, and the proposed LoRS-Merging achieves the best performance across the table.

5.3 Multi-Lingual Multi-Task Model Merging

Then, we investigate the merging of finetuned models for both different languages and tasks, which correspond to *multi-lingual (ML) and multi-task (MT)* learning. Specifically, we explore 4 different training and merging settings:

ML and MT training: Finetuning on all languages and both tasks jointly.

ML and MT merging: Finetuning on each language for each task separately and merging all.

MT training and ML merging: Finetuning both tasks jointly for each language, and merging models from different languages.

ML training and MT merging: Finetuning on all languages jointly for each task, and merging models from different tasks.

Table 4 displays the multi-lingual and multi-task results with the high-resource language set. Multi-lingual and multi-task training shows little improvement over the pretrained model, due to language interference during training and the use of a unified training configuration for all languages and tasks. Nevertheless, the performance of multi-lingual and multi-task merging is on par with that of finetuned models, further underscoring the superiority of model merging. ML training followed by MT merging achieves the best performance, even significantly outperforming finetuned models. Although we did not observe the same phenomenon on the low-resource language set, this suggests the potential of using a combination of training and merging to achieve better performance. We provide additional experiments on the low-resource language set in Appendix B to demonstrate the robustness and generalisability of model merging and LoRS-Merging.

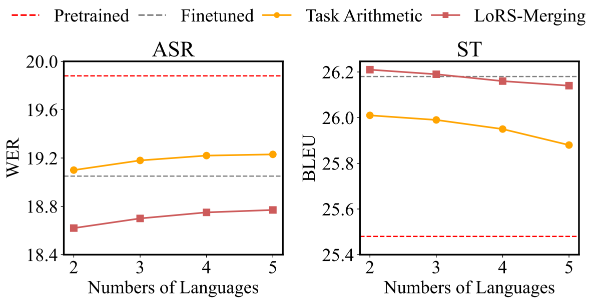

Figure 3: WER and BLEU against the number of languages. Performance is averaged across all languages and all training runs of language combinations.

5.4 Effect of Numbers of Languages

To further demonstrate the robustness of LoRS-Merging to language selection, experiments are performed using different numbers of languages. Figure 3 shows the average performance across all languages and all training runs with possible combinations of 2, 3, 4 or 5 languages.

LoRS-Merging improvements are consistent across different numbers of languages: As the number of languages increases, the performance of both TA and LoRS-Merging degrades due to negative interference between languages. LoRS-

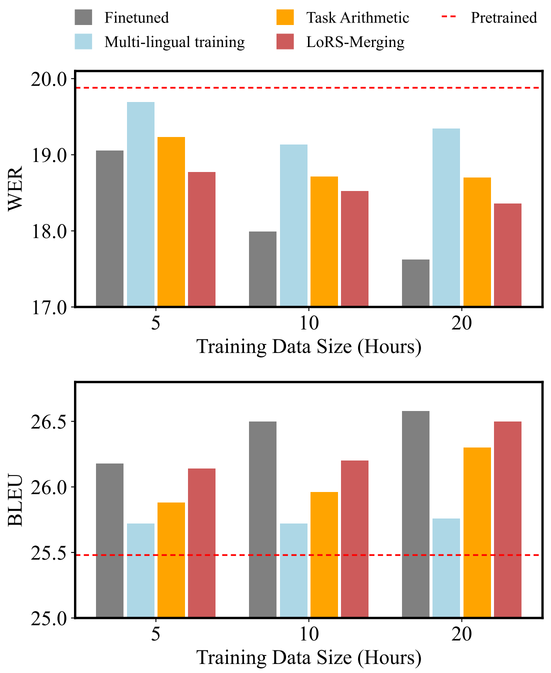

Figure 4: Performance variation against different training data sizes (number of hours for each language) on ASR (top) and ST (bottom) tasks.

Merging consistently outperforms TA in both ASR and ST tasks. Notably, in the ASR task, it even surpasses the finetuned models. This is primarily because finetuned models contain substantial redundancy (see Fig. 5), whereas LoRS-Merging reduces redundancy through pruning, leading to significant performance improvements. Additionally, we observe that the optimal learning rate for the finetuned ASR model is significantly larger compared to the ST task. This may lead to overfitting in ASR. In contrast, LoRS-Merging improves generalisation through model merging, thus outperforming the finetuned models for the ASR task.

5.5 Effect of Language Data Scale

We then demonstrate the robustness of merging methods to different training data sizes for both tasks. Fig. 4 shows the WER (top) and BLEU (bottom) scores for ASR and ST at different data scales, respectively. As the data scale increases, the performance of multi-lingual training does not always improve. This is because the multi-lingual capabilities of pretrained models are already near convergence, and only meticulous training can further improve performance. Increased training data amplifies both language interference and the negative effects of uniform training configurations, thereby offsetting the gains from increased data. Furthermore, the performance loss of model merging increases with data scale, compared to finetuned models. It can be explained by the fact that larger training data tends to increase the divergence in the optimisation trajectories of different finetuned models, resulting in the breakdown of linear mode connectivity, which leads to a greater performance loss. Moreover, LoRS-Merging still achieves obvious and stable improvement compared to TA.

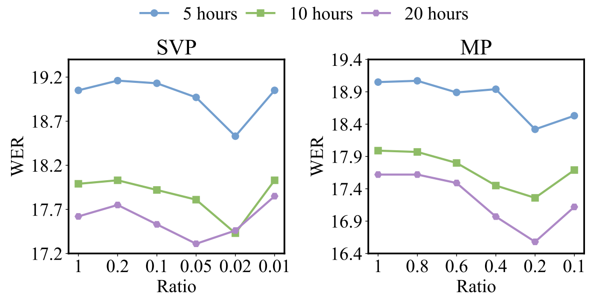

Figure 5: Model performance against the retain ratio in SVP (left) and MP (right) for ASR finetuned models. Three different training data sizes are used.

5.6 Analysis of Model Redundancy

Furthermore, we justify the necessity of SVP and MP to remove model redundancy by showing the model performance against the pruning ratio of finetuned models for ASR as shown in Fig. 5. As shown, both SVP and MP significantly improve the performance of finetuned models, indicating the presence of substantial redundancy in the structure and details of the finetuned models, respectively. The model performance reaches the best at a high pruning level, indicating that the redundancy is particularly large for ASR. We observed a much smaller redundancy in ST, which also explains the observation that LoRS-Merging achieves more salient improvement on ASR than ST. Moreover, redundancy increases with training data, possibly due to the accumulation of gradient noise during training. MP achieves greater performance gains than SVP, indicating more redundancy at the detail level, which is better addressed by fine-grained MP.

6 Conclusion

This paper explores model merging for multi-lingual ASR and ST on pretrained speech models and proposes the LoRS-Merging approach. LoRS-Merging combines low-rank and sparse pruning to retain essential structures and reduce redundant parameters. Experiments across 10 languages show that LoRS-Merging effectively alleviates language interference and significantly outperforms multi-lingual multi-task training, sequential training, and other merging methods.

7 Limitations

There are three main limitations of this work. First, as a common limitation of all model merging methods, the same model structure is required across all tasks and languages. This is less of a concern under the current trend of using the same Transformer structure, but methods need to be developed in the future to accommodate subtle structural differences. Second, reasonably-sized training sets are required for each language, and low-resource languages may suffer from reduced improvements. Third, this work mainly explores the two most popular S2T tasks. Other possible tasks can be explored in future work, including spoken language understanding and speaker adaptation.

References

Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. 2023. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*.

He Bai, Yu Zhou, Jiajun Zhang, Liang Zhao, Mei-Yuh Hwang, and Chengqing Zong. 2018. Source-critical reinforcement learning for transferring spoken language understanding to a new language. In *International Conference on Computational Linguistics*, pages 3597–3607.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4960–4964.

Yong Cheng, Yu Zhang, Melvin Johnson, Wolfgang Macherey, and Ankur Bapna. 2023. Mu2slam: multitask, multilingual speech and language models. In *International Conference on Machine Learning*, pages 5504–5520.

Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2019. One-to-many multilingual end-to-end speech translation. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 585–592.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 1723–1732.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. 2018. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning*, pages 1308–1317.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. 2021. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. 2020. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pages 3259–3269.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. 2024. Task singular vectors: Reducing task interference in model merging. *arXiv preprint arXiv:2412.00081*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech*, pages 6–10.

Yifei He, Yuzheng Hu, Yong Lin, Tong Zhang, and Han Zhao. 2024. Localize-and-stitch: Efficient model merging via sparse task arithmetic. *Transactions on Machine Learning Research*.

Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee. 2020. Meta learning for end-to-end low-resource speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7844–7848.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *International Conference on Learning Representations*.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. In *International Conference on Uncertainty in Artificial Intelligence*, pages 876–885.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Anjuli Kannan, Arindrima Datta, Tara N Sainath, Eugene Weinstein, Bhuvana Ramabhadran, Yonghui Wu, Ankur Bapna, Zhifeng Chen, and Seungji Lee. 2019. Large-scale multilingual speech recognition with a streaming end-to-end model. In *Interspeech*, pages 2130–2134.

Arham Khan, Todd Nief, Nathaniel Hudson, Mansi Sakarvadia, Daniel Grzenda, Aswathy Ajith, Jordan Pettyjohn, Kyle Chard, and Ian Foster. 2024. Sok: On finding common ground in loss landscapes using deep model merging techniques. *arXiv preprint arXiv:2410.12927*.

Yoohwan Kwon and Soo-Whan Chung. 2023. Mole: Mixture of language experts for multi-lingual automatic speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Cheng-I Jeff Lai, Yang Zhang, Alexander H Liu, Shiyu Chang, Yi-Lun Liao, Yung-Sung Chuang, Kaizhi Qian, Sameer Khurana, David Cox, and Jim Glass. 2021. Parp: Prune, adjust and re-prune for self-supervised speech recognition. In *Advances in Neural Information Processing Systems*, pages 21256–21272.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 817–824.

Yann LeCun, John Denker, and Sara Solla. 1989. Optimal brain damage. In *Advances in Neural Information Processing Systems*, pages 598–605.

Danni Liu and Jan Niehues. 2024. Recent highlights in multilingual and multimodal speech translation. In *International Conference on Spoken Language Translation*, pages 235–253.

Yizhou Lu, Mingkun Huang, Xinghua Qu, Pengfei Wei, and Zejun Ma. 2022. Language adaptive cross-lingual speech representation learning with sparse sharing sub-networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6882–6886.

Vaishnavh Nagarajan and J Zico Kolter. 2019. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 11611–11622.

Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:325–351.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

AFM Saif, Lisha Chen, Xiaodong Cui, Songtao Lu, Brian Kingsbury, and Tianyi Chen. 2024. M2asr: Multilingual multi-task automatic speech recognition via multi-objective optimization. In *Interspeech*, pages 1240–1244.

Sidak Pal Singh and Martin Jaggi. 2020. Model fusion via optimal transport. In *Advances in Neural Information Processing Systems*, pages 22045–22055.

Haoran Sun, Xiaohu Zhao, Yikun Lei, Deyi Xiong, et al. 2023. Towards a deep understanding of multilingual end-to-end speech translation. In *Findings of Conference on Empirical Methods in Natural Language Processing*, pages 14332–14348.

Norman Tatro, Pin-Yu Chen, Payel Das, Igor Melnyk, Prasanna Sattigeri, and Rongjie Lai. 2020. Optimizing mode connectivity via neuron alignment. In *Advances in Neural Information Processing Systems*, pages 15300–15311.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Annual Meeting of the Association for Computational Linguistics*, pages 993–1003.

Changhan Wang, Anne Wu, and Juan Pino. 2020a. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. 2024. Localizing task information for improved model merging and compression. In *International Conference on Machine Learning*, pages 50268–50287.

Wenxuan Wang, Guodong Ma, Yuke Li, and Binbin Du. 2023. Language-routing mixture of experts for multilingual and code-switching speech recognition. In *Interspeech*, pages 1389–1393.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. 2019. Characterizing and avoiding negative transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11293–11302.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020b. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Conference on Empirical Methods in Natural Language Processing*, pages 4438–4450.

Yongxian Wei, Anke Tang, Li Shen, Feng Xiong, Chun Yuan, and Xiaochun Cao. 2025. Modeling multi-task model merging as adaptive projective gradient descent. *arXiv preprint arXiv:2501.01230*.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998.

Yubei Xiao, Ke Gong, Pan Zhou, Guolin Zheng, Xiaodan Liang, and Liang Lin. 2021. Adversarial meta sampling for multilingual low-resource speech recognition. In *AAAI Conference on Artificial Intelligence*, pages 14112–14120.

Feng Xiong, Runxi Cheng, Wang Chen, Zhanqiu Zhang, Yiwen Guo, Chun Yuan, and Ruifeng Xu. 2024. Multi-task model merging via adaptive weight disentanglement. *arXiv preprint arXiv:2411.18729*.

Chen Xu, Rong Ye, Qianqian Dong, Chengqi Zhao, Tom Ko, Mingxuan Wang, Tong Xiao, and Jingbo Zhu. 2023. Recent advances in direct speech-to-text translation. In *International Joint Conference on Artificial Intelligence*, pages 6796–6804.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2024. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, pages 7093–7115.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024a. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*.

Enneng Yang, Li Shen, Zhenyi Wang, Guibing Guo, Xiaojun Chen, Xingwei Wang, and Dacheng Tao. 2024b. Representation surgery for multi-task model merging. In *International Conference on Machine Learning*, pages 56332–56356.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning*, pages 57755–57775.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639.

Chao Zhang, Bo Li, Tara Sainath, Trevor Strohman, Sepand Mavandadi, Shuo yiin Chang, and Parisa Haghani. 2022. Streaming end-to-end multilingual speech recognition with joint language identification. In *Interspeech*, pages 3223–3227.

Chao Zhang, Bo Li, Tara N. Sainath, Trevor Strohman, and Shuo yiin Chang. 2023a. Uml: A universal monolingual output layer for multilingual asr. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. 2023b. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10:305–329.

Yun Zhu, Parisa Haghani, Anshuman Tripathi, Bhuvana Ramabhadran, Brian Farris, Hainan Xu, Han Lu, Hasim Sak, Isabel Leal, Neeraj Gaur, et al. 2020. Multilingual speech recognition with self-attention structured parameterization. In *Interspeech*, pages 4741–4745.

A Details of the Experimental Setup

For multi-lingual and multi-task training, a uniform training configuration is used across all languages and tasks. For sequential training, considering that there are 5! = 120 possible sequences for 5 languages, and the optimal training configuration for the same language differs across sequences, the hyper-parameter search cost for sequential training is much higher than that for the model merging. To simplify the configuration, we select 5 sequences for the experiments, corresponding to all cyclic permutations of the language order ca-de-es-fr-it, and report the results from the sequence that yields the best average performance.

MP and SVP are applied to each linear layer. The detailed hyper-parameter settings for each language are shown in Table 5 for ASR and Table 6 for ST, respectively.

Table 5: ASR hyper-parameters for high-resource languages.

System	ASR				
	ca	de	es	fr	it
Finetuned					
learning rate	1×10^{-6}	5×10^{-8}	1×10^{-7}	1×10^{-6}	5×10^{-6}
Multi-lingual training					
learning rate			1×10^{-5}		
Task Arithmetic					
scaling factor λ			0.15		
LoRS-Merging					
scaling factor λ			0.15		
SVP ratio r	5%	3%	2%	1%	1%
MP ratio p	40%	60%	40%	10%	10%

Table 6: ST hyper-parameters for high-resource languages.

System	ST				
	ca	de	es	fr	it
Finetuned					
learning rate	1×10^{-6}	2×10^{-8}	2×10^{-8}	5×10^{-8}	5×10^{-8}
Multi-lingual training					
learning rate			5×10^{-9}		
Task Arithmetic					
scaling factor λ			0.15		
LoRS-Merging					
scaling factor λ			0.15		
SVP ratio r	5%	3%	5%	2%	1%
MP ratio p	60%	40%	20%	20%	20%

B Results of Low-Resource Language Set

The results of the low-resource language set are shown in this section. Specifically, Table 7 and 8 show the multi-lingual single-task training and merging for ASR and ST respectively.

Table 7: Multi-lingual ASR model merging with the low-resource language set. Avg. denotes average WER.

System	WER↓					
	id	nl	pt	ru	sv	Avg.
Pretrained	16.9	16.0	10.1	17.1	17.1	15.43
Finetuned	15.0	14.8	9.7	16.8	14.7	14.20
Multi-lingual training	16.7	15.5	10.0	17.0	16.6	15.14
Weight Averaging	15.7	15.2	10.1	17.1	15.8	14.77
Task Arithmetic	15.7	15.1	9.9	17.0	15.8	14.69
MP-Merging	15.7	15.1	10.0	16.7	15.7	14.63
SVP-Merging	15.7	15.1	9.9	16.9	15.7	14.65
LoRS-Merging	15.7	15.1	9.7	16.8	15.6	**14.57**

Table 8: Multi-lingual ST model merging with the low-resource language set. Avg. denotes average BLEU.

System	BLEU↑					
	id	nl	pt	ru	sv	Avg.
Pretrained	32.5	31.6	43.3	35.5	32.1	35.00
Finetuned	35.2	34.0	43.8	36.7	37.6	37.46
Multi-lingual training	32.3	33.2	43.5	35.4	34.3	35.74
Weight Averaging	33.6	32.2	43.2	35.3	34.2	35.70
Task Arithmetic	33.9	32.8	43.1	35.5	34.3	35.92
MP-Merging	33.8	32.8	43.5	35.8	34.0	35.98
SVP-Merging	33.6	32.6	43.4	35.6	34.3	35.90
LoRS-Merging	33.9	32.8	43.2	35.9	34.5	**36.06**

Then, Table 9 shows the multi-task single-language training and merging performance (c.f. compare to Table 3 for high-resource languages).

Last, Table 10 shows the results of multi-lingual and multi-task training and merging results for low-resource languages (compare to Table 4 for high-resource languages.). LoRS-Merging achieved the best performance across all merging and training methods in all tables.

C Detailed Results on Multi-Task Model Merging

Detailed per-language results of Table 3 are shown in Table 11.

Table 9: Multi-task model merging with the low-resource language set. WER/BLEU scores are averaged across languages.

System	WER↓						BLEU↑					
	id	nl	pt	ru	sv	Avg.	id	nl	pt	ru	sv	Avg.
Pretrained	16.9	16.0	10.1	17.1	17.1	15.43	32.5	31.6	43.3	35.5	32.1	35.00
Finetuned	15.0	14.8	9.7	16.8	14.7	14.20	35.2	34.0	43.8	36.7	37.6	37.46
Multi-task training	15.4	15.0	9.3	16.6	14.3	14.12	35.3	33.7	43.6	36.2	35.8	36.92
Weight Averaging	14.7	14.9	9.3	16.6	13.8	13.88	35.4	33.9	44.1	36.3	35.9	37.12
Task Arithmetic	14.6	14.9	9.3	16.5	14.0	13.88	35.3	33.8	44.3	36.1	36.4	37.18
MP-Merging	14.4	14.7	9.4	16.5	13.8	13.78	35.7	33.9	44.3	36.1	36.1	37.22
SVP-Merging	14.6	14.8	9.2	16.4	13.9	13.80	35.3	33.9	44.3	36.2	36.3	37.20
LoRS-Merging	14.4	14.7	9.2	16.4	13.8	**13.72**	35.6	33.9	44.3	36.3	36.4	**37.30**

Table 10: Multi-lingual multi-task model merging with the low-resource language set. WER/BLEU scores are averaged across languages.

System	WER↓						BLEU↑					
	id	nl	pt	ru	sv	Avg.	id	nl	pt	ru	sv	Avg.
Pretrained	16.9	16.0	10.1	17.1	17.1	15.43	32.5	31.6	43.3	35.5	32.1	35.00
Finetuned	15.0	14.8	9.7	16.8	14.7	14.20	35.2	34.0	43.8	36.7	37.6	37.46
ML and MT training	16.9	15.7	9.6	17.0	16.3	15.08	32.8	32.9	43.3	35.4	32.6	35.40
ML and MT Task Arithmetic	16.4	15.5	9.6	16.8	15.7	14.79	33.7	33.1	43.2	35.7	34.9	36.12
ML and MT LoRS-Merging	16.1	15.5	9.5	16.8	15.7	14.72	33.7	33.2	43.5	35.8	34.9	**36.22**
MT training	15.4	15.0	9.3	16.6	14.3	14.12	35.3	33.7	43.6	36.2	35.8	36.92
↪ + ML Task Arithmetic	16.0	15.5	9.5	16.9	15.4	14.66	34.1	32.8	43.7	35.6	33.3	35.90
↪ + ML LoRS-Merging	16.1	15.3	9.4	16.8	15.3	**14.57**	34.2	32.7	43.8	35.8	33.5	36.00
ML training	16.7	15.5	10.0	17.0	16.6	15.14	32.3	33.2	43.5	35.4	34.3	35.74
↪ + MT Task Arithmetic	17.1	15.5	9.5	17.0	15.5	14.89	32.1	33.1	43.6	35.7	33.6	35.62
↪ + MT LoRS-Merging	16.9	15.5	9.4	16.8	15.5	14.80	32.6	33.2	43.6	35.9	33.6	35.78

Table 11: Multi-task model merging with the high-resource language set. WER/BLEU scores are averaged across languages.

System	WER↓						BLEU↑					
	ca	de	es	fr	it	Avg.	ca	de	es	fr	it	Avg.
Pretrained	20.6	19.6	14.7	24.5	19.4	19.88	21.1	24.1	28.6	26.8	26.8	25.48
Finetuned	19.5	19.7	14.4	22.1	19.2	19.05	22.6	24.6	29.2	27.2	27.3	26.18
Multi-task training	17.0	19.7	14.4	24.2	19.4	19.00	22.3	24.6	28.7	27.0	26.9	25.90
Sequential training	16.7	19.4	14.3	24.6	19.4	18.95	22.9	24.8	28.7	27.3	26.9	26.12
Weight Averaging	17.1	19.6	13.9	23.7	19.6	18.84	22.9	24.4	29.0	27.7	26.9	26.18
Task Arithmetic	17.2	19.3	14.0	23.3	19.7	18.76	23.4	24.5	28.9	27.7	27.0	26.30
TIES-Merging	17.7	19.5	14.4	23.6	17.4	18.60	23.1	24.5	29.1	27.7	27.5	26.38
DARE	17.5	19.4	14.2	23.5	18.6	18.71	23.2	24.5	29.0	27.6	27.1	26.28
TSV-M	17.8	19.4	14.3	23.7	17.9	18.70	23.0	24.7	29.2	27.7	27.4	26.40
LoRS-Merging	17.3	19.4	14.1	23.1	17.7	**18.39**	23.3	24.6	29.3	28.0	27.6	**26.56**