From Child Adaptation to Adult Retention: Merging Specialized Arabic–English ASR Models Across Architectures

Anonymous ACL submission

Abstract

Automatic speech recognition (ASR) systems exhibit persistent performance disparities across age groups and speaker nativity, with children's speech remaining a systematically underrepresented and challenging domain, even in high-resource languages. Existing adaptation strategies predominantly rely on fine-tuning, which often induces catastrophic forgetting and degrades performance on adult speech; these limitations are further amplified in bilingual children's ASR, where robust cross-language generalization is required. In this work, we explore weight-space model merging as a principled framework for age-robust and language-inclusive speech modeling. Starting from a shared multilingual base model, we fine-tune complementary child-adapted checkpoints and merge them using balanced weighting to preserve adult representations while incorporating age- and language-specific adaptations. Across all benchmarks, model merging consistently improves recognition accuracy for children while retaining or in some cases improving—adult performance, outperforming fine-tuning and joint training baselines. These results demonstrate that model merging provides a scalable and data-efficient alternative to fine-tuning for inclusive ASR across age groups, speaker nativity (L1 and L2 Arabic speakers), and languages (Arabic and English).

1 Introduction

Automatic Speech Recognition (ASR) systems have achieved strong performance on adult speech across multiple languages, largely due to the availability of large-scale training data and robust model architectures. In contrast, children's speech still remains challenging. Compared to adult speech, children's speech exhibits distinct acoustic and linguistic characteristics, including higher pitch, shorter vocal tracts, increased pronunciation variability, and frequent disfluencies, all of which contribute to degraded recognition performance (Wu et al., 2019; Bhardwaj et al., 2022). These challenges are further amplified by the limited availability of annotated children's speech data.

To overcome, the widely used approach involves fine-tuning large, pre-trained adult models on available children's data. While this often yields substantial gains on child-specific benchmarks like MyST (Pradhan et al., 2024; Linguistic Data Consortium, 2021) and CMU Kids (Jain et al., 2023), it introduces a critical representational trade-off. Standard fine-tuning often results in *catastrophic forgetting*, where the model's adaptation to child-specific acoustic shifts causes a drift in the underlying representations that support adult speech recognition. This limitation is particularly severe in bilingual or multilingual settings, where the model must simultaneously encode age-related acoustic variance and distinct language-specific structures (e.g., Arabic and English).

Previous efforts to mitigate this forgetting have largely focused on continual learning frameworks, utilizing regularization techniques like Elastic Weight Consolidation (EWC) to preserve knowledge from the original domain (Sadhu and Hermansky, 2020; Ahadzi et al., 2025). While effective, these methods require modifying the training objective and maintaining complex protocols during the adaptation phase. In contrast, a parallel line of research in vision and NLP offers a more flexible paradigm: *weight-space model merging*. Techniques such as model soups and task arithmetic demonstrate that averaging or interpolating the weights of independently fine-tuned models can improve generalization and combine task-specific behaviors without additional training or inference costs (Wortsman et al., 2022; Ilharco et al., 2023).

In this study, we shifts the focus from simple adaptation to a more fundamental question: "*how can child-specific acoustic shifts be integrated into a model without much compromising the in-*

tegrity of its previously learned adult representations?" Hence, rather than just finetuning with child-specific ASR data, we propose to decouple adaptation from retention using weight-space model merging.

Starting from a bi/multilingual base ASR, we derive multiple child-adapted checkpoints through targeted fine-tuning to capture complementary inductive biases including language-invariant adult phonetics, language-specific child acoustic patterns, and cross-lingual age-related characteristics. We then merge these checkpoints directly in weight space using various techniques, Linear Interpolation (LERP) , and TIES-merging (Yadav et al., 2023). This process produces a unified model that incorporates the nuances of children's speech while preserving high-fidelity adult performance across both Arabic and English.

We demonstrate the broad utility and architecture-agnostic nature of this approach by experimenting with three distinct architectural frameworks: (i) Multilingual Encoder-Decoder using Whisper (Radford et al., 2023); (ii) Bilingual Encoder-Decoder for Arabic-English ASR; and (iii) Bilingual Encoder-CTC for Arabic-English ASR; as the foundation model. Our evaluation is uniquely comprehensive, spanning both native and non-native speech across two languages. We utilize the established English MyST corpus (Pradhan et al., 2024) and introduce a new dataset of Arabic native (**AraKidsL1**) and non-native (**AraKidsL2**) children's speech. Crucially, we measure knowledge retention by benchmarking against original adult domains, including MGB-2 (Arabic) and LibriSpeech (English). By evaluating across these diverse architectures and datasets, we show that weight-space merging can acts as an implicit regularizer, aligning age-dependent acoustic patterns while safeguarding the linguistic structures required for robust, inclusive ASR.

Our main contribution includes the following:

1. We presented a post-hoc, training-free framework to balance "adaptation-retention" trade-off in children's ASR using weight-space model merging for knowledge integration.

2. We validate the effectiveness of such framework across three distinct ASR architectures – Multilingual Encoder-Decoder (Whisper), Bilingual Encoder-Decoder, and Bilingual Encoder-CTC, showcasing the architecture-agnostic robustness.

3. We introduce and release **AraKids** – a new dataset comprising both native and non-native Arabic children's speech. This addresses a significant gap in the community and serves as a challenging benchmark for cross-nativity ASR.

4. We evaluate and compare multiple existing merging strategies for their ability to align age-related acoustic shifts in a bilingual (Arabic-English) context.

2 Related Work

2.1 Children's ASR

Children's ASR remains difficult because child speech differs from adult speech acoustically and linguistically (e.g., higher pitch, shorter vocal tracts, greater pronunciation variability, and more disfluencies). A widely used approach is to adapt adult-trained acoustic or end-to-end ASR models with limited child data, and prior work shows that fine-tuning large pretrained/foundation models can yield substantial gains, especially with multilingual pretraining (Wu et al., 2019; Bhardwaj et al., 2022). Recent studies further demonstrate strong improvements when adapting Whisper-style or self-supervised models on common benchmarks such as MyST and CMU Kids (Jain et al., 2023; Linguistic Data Consortium, 2021). However, many adaptation studies primarily report target-domain gains and provide limited analysis of whether adult-domain performance is retained after fine-tuning.

2.2 Forgetting, Continual Learning, and Training-Free Alternatives

Sequential domain adaptation can trigger *catastrophic forgetting*, motivating continual/lifelong learning methods for ASR via regularization, replay, or constrained optimization (Sadhu and Hermansky, 2020; Houston and Kirchhoff, 2020). For children's ASR, online adaptation studies show that regularization-based approaches (e.g., EWC, Synaptic Intelligence) can better preserve earlier capabilities than naive fine-tuning (Ahadzi et al., 2025), but they add training-time complexity. In parallel, weight-space *model merging* offers a training-free route to combine fine-tuned behaviors: model soups and task-arithmetic-style operations can improve robustness when checkpoints share an initialization and remain compatible (Wortsman et al., 2022; Ilharco et al., 2023; Ortiz-Jimenez et al., 2023; Sung et al., 2023). Although less explored in speech than in vision/NLP, recent ASR

results indicate that merging independently fine-tuned checkpoints can improve robustness without extra inference cost (Ducorroy and Riad, 2025), making it a practical option when multiple adaptation runs are already available.

2.3 Positioning and Contributions

We position our work as a "training-free" paradigm adaption that shifts the burden of knowledge retention from the training phase to a post-hoc weight-space operation. While standard fine-tuning achieves high accuracy on children's speech at the expense of adult performance, and regularization methods like EWC mitigate this by introducing training-time complexity, we show model merging method achieves a similar balance with negligible computational overhead. Furthermore, we expand the scope of current research by investigating merging across distinct architectures; specifically Encoder-Decoder and CTC-based ASR model and applying these techniques to a bilingual Arabic-English context. By introducing a new Arabic native and non-native children's corpus, we are also able to evaluate the limits of cross-lingual and cross-age generalization in a way that remains largely unexplored in existing ASR literature.

3 Methodology

We aim to develop an inclusive ASR model that excels in the highly variable domain of Arabic and English children's speech while maintaining the comparable performance benchmarks established for adult speakers. We define this objective as an optimization problem in the shared latent parameter space such that the single model θ can accurately transcribe speech across distinct dimensions of variability: *(i)* **Speaker age** (child *vs.* adult); *(ii)* **Nativity** (native *vs.* non-native); and *(iii)* **Bilingualism** (Arabic *vs* English).

Base and Naive Fine-tuned Models: We define the base model (θ_{base}) as the original pre-trained ASR model, which is specialized in adult speech. Following, we designed monolingial and bilingual child-speech specialized models, θ_{child}, by fine-tuning the base with children training data $\ddot{D}^{L}_{child}$, where $L \in \{\text{Ar, En, Ar+En}\}$.

$$\theta_{\text{base}} \xrightarrow{\text{FT}^{L}} \theta^{L}_{\text{child}}$$

The shift in knowledge during this adaptation can also be encoded by the task vector, $\tau^{L} = \theta^{L}_{child} - \theta_{base}$, which isolates the parameter-level updates necessary for child-specific acoustics.

Catastrophic Forgetting: We define the catastrophic forgetting gap, $\Delta_{forgot}(\theta_{*}, \theta_{base})$, as the model performance degradation on the original source domain (Adult ASR) when evaluating a model adapted to the target domain (child ASR).

Model Merging Objective: The primary objective is to derive a generalized (inclusive) θ_m model, that achieve performance parity with specialized child model ($\theta_m \approx \theta^{L}_{child}$) while minimizing $\Delta_{forgot}(\theta_m, \theta_{base})$. We achieve this by combining the specialized models using a functional merging operator, $(Merge(.))$.

$$\theta_m = \theta_{base} + Merge(\theta_1, \cdots, \theta_N)$$

where N represents the number of specialized models (e.g., Arabic-child, English-child, Bilingual-child) being integrated into the base model.

Linear Interpolation (LERP): Linear Interpolation (LERP) operator, Merge_L, allows a fine-grained balance between the base model's stability and the specialists' expertise. Unlike uniform merging, LERP introduces a global scaling factor, λ, to control the overall intensity of the adaptation and individual mixing weights (α_i) to prioritize specific tasks.

$$\text{Merge}_{\text{L}}(\theta_1, \ldots, \theta_N) = \lambda \sum_{i=1}^{N} \alpha_i \, \theta_i.$$

where $\alpha \in [0, 1]$ and $\sum \alpha = 1$; for $\lambda \in [0, 1]$ and $\lambda \to 0$ values makes the θ_M bias to the θ_{base}, decreasing Δ_{forgot}, whereas, $\lambda \to 1$ maximizes the child-speech adaptation.

TIES-Merging: To reduce the interference of different task vectors,τ^{i} to move same weight in conflicting direction, we opt for TIES, which include three stages – Trim, Elect Sign and Disjoint Merge. Trim stage first retain the top $k\%$ weight values with hightest magnitude, setting rest to zero to reduce the parameter noise. The Elect sign stage calculate a consensus sign vector γ, based on the direction that possesses the highest cumulative magnitude across all N task vectors. Following, only the selected parameters across task vectors that aligns with γ are averaged, filtering out conflicting updates.

4 AraKids for L1 and L2 Speakers

Overview: We present **AraKids**, an Arabic children's speech corpus designed to benchmark ASR robustness across *nativity* (L1 vs. L2), *regional grouping*, and *age*. AraKids contains 64,705 utterances from 181 speakers (109.18 hours). It comprises two subsets: **AraKidsL1**, native Arabic children grouped by Arabic macro-regions (Egyptian, Gulf, Levantine, North African), and **AraKidsL2**, non-native Arabic children grouped by speakers' broad origin regions (Africa, Asia, Americas, Europe) (Table 2). We plan to release AraKids publicly in future work. To the best of our knowledge, AraKids is among the first large-scale Arabic children's speech dataset introduced specifically to support ASR research with explicit L1/L2 coverage and broad regional diversity, and we plan to release AraKids publicly in future work.

Prompts and text sources: Prompts are read and curated to be age-appropriate. For **AraKidsL1**, sentences are selected from school-grade materials to match speakers' expected literacy level. For **AraKidsL2**, Arabic teachers adapt and calibrate prompts to span multiple proficiency levels while controlling lexical and syntactic complexity. All prompts are fully diacritized.

Speaker ages: AraKids train covers ages 7–18, split into 7–12 (71 speakers, 38.80 hours) and 13–18 (110 speakers, 70.38 hours). For a focused evaluation on younger children, the test set is restricted to ages 7–12.

Splits and leakage control: We enforce **speaker-disjoint** train/test partitioning and **text-disjointness** at the prompt level (no overlap in recorded prompt sentences between train and test), reducing memorization effects and measuring generalization to unseen speakers and unseen prompt texts. In addition, we report prompt inventory statistics: across **train** (L1+L2), AraKids contains 8 unique `ScriptIDs` (book source) and 3,194 unique `UtteranceIDs` (sentence), while **test** (L1+L2) contains 5 unique `ScriptIDs` and 380 unique `UtteranceIDs`.

Benchmark test set: We construct a region-balanced test set of 2,000 utterances by sampling 250 utterances per region across the 8 regions (4 AraKidsL1 macro-regions + 4 AraKidsL2 origin regions). This yields 1,000 utterances for AraKidsL1 (26 speakers, 94 minutes) and 1,000 utterances for AraKidsL2 (40 speakers, 2.43 hours). Overall, the test set totals 3.99 hours, with average utterance durations of 5.64s (L1) and 8.74s (L2) (Table 1).

Split	#Spk	#Utt	Hours
Train (AraKidsL1)	85	29,222	42.27
Train (AraKidsL2)	96	35,483	66.91
Train (Total)	181	64,705	109.18
Test (AraKidsL1)	26	1,000	1.57
Test (AraKidsL2)	40	1,000	2.43
Test (Total)	66	2,000	3.99

Table 1: AraKids split overview. The test set is region-balanced by sampling 250 utterances per region across the 8 regions and is restricted to the 7–12 age group.

Region	#Spk	#Utt	Hours
AraKidsL1			
Egyptian	29	9,478	14.49
Gulf	18	6,796	10.11
Levantine	22	6,873	8.75
North African	16	6,075	8.92
AraKidsL2			
Africa	14	5,339	11.03
Americas	16	6,000	12.44
Asia	32	11,796	22.01
Europe	34	12,348	21.43
Total	**181**	**64,705**	**109.18**

Table 2: AraKids train distribution by region. AraKidsL1 regions are Arabic macro-regions; AraKidsL2 regions are speakers' origin regions.

5 Experimental Setup

5.1 ASR Model Architectures

We evaluate weight-space merging across three representative ASR architectures to verify that our approach is not tied to a specific modeling or decoding paradigm. Throughout the paper, we denote the encoder–decoder family as **ED**, the encoder–CTC family as **ECTC**, and Whisper Large v3 as **WL**.

5.2 ASR Model Architectures

Bilingual Encoder-CTC (E-CTC): Our encoder–CTC backbone is an encoder-only bilingual (Arabic–English) ASR model adapted from the OWSM-CTC architecture. It consists of (i) a speech encoder built on top of a frozen Arabic-centric self-supervised front-end whose embeddings are processed by a stack of L transformer encoder layers, and (ii) a transformer text-history encoder whose representations are injected into selected speech-encoder layers via cross-attention to condition recognition on previous context.

Training uses self-conditioned CTC by applying auxiliary CTC losses at intermediate layers in addition to the final-layer CTC loss, averaged to improve convergence. For long-form audio, we employ parallel chunk-wise greedy decoding with overlapping 30-second windows, followed by overlap alignment and confidence-based stitching to produce the final transcript. We tokenize transcripts using a 20,002 BPE vocabulary, and apply SpecAugment during training.

Bilingual Encoder–Decoder (ED): Our bilingual encoder–decoder (**ED**) baseline is an ESPnet Conformer–Transformer model trained on Arabic–English speech only. It uses an XLSR-53 self-supervised front-end (via S3PRL) with multi-layer feature extraction, followed by a 12-block Conformer encoder and a 6-block Transformer decoder, jointly optimized with a hybrid CTC/attention objective ($\lambda_{\text{CTC}} = 0.3$). We tokenize transcripts using a 10k BPE vocabulary, and apply SpecAugment during training.

Multilingual Encoder-Decoder (WL): For the multilingual encoder–decoder setting, we use **Whisper Large v3 (WL)** (Radford et al., 2023), a large-scale ASR model pretrained on diverse multilingual and multitask speech data. WL serves as a strong architecture-independent baseline for our merging experiments due to its robust multilingual representations and widely adopted encoder–decoder formulation.

5.3 Datasets and Benchmarks

We evaluate adaptation to children's speech and retention on adult speech using a consistent benchmark suite spanning both domains.

Child Specialized Model Training Data: We build child-specialized models by fine-tuning on AraKids (Arabic children; Table 1), and MyST (Pradhan et al., 2024) with 171 hours of training audio. For all child-training datasets and all model architectures, we apply SpecAugment (Park et al., 2019) and speed perturbation (Ko et al., 2015) to improve robustness.

Children's Benchmarks: We measure in-domain performance on two children-focused test sets: AraKids Test (Table 1) and MyST clean test set (Pradhan et al., 2024).

Adult Retention Benchmarks: To quantify retention of adult ASR capability, we evaluate on LibriSpeech test-clean (Panayotov et al., 2015) and the MGB-2 Arabic multi-dialect broadcast benchmark (Ali et al., 2016). These adult benchmarks serve as out-of-domain anchors, enabling us to measure how well merging preserves general-purpose performance while improving children's ASR.

5.4 Naive Fine-tuning Parameters

As a reference adaptation strategy, we perform *naive fine-tuning* of each baseline model on the children training data using standard supervised ASR objectives. To limit overfitting and to keep adaptation budgets comparable across architectures, we fine-tune the ESPnet models (**ED** and **ECTC**) for **5 epochs** each, while **WL** is fine-tuned for **1 epoch**.

For **ECTC**, we optimize the E-CTC model with self-conditioned intermediate CTC supervision (intermediate layers at $\{6, 9, 11\}$) and a text-history prompt encoder. We use AdamW with a peak learning rate of 2.5×10^{-4} and a piecewise linear warmup schedule, batch size 32 with gradient accumulation of 2, mixed-precision training (AMP), and SpecAugment; the self-supervised HArnESS front-end is kept frozen.

For **ED**, we fine-tune the ED under a hybrid CTC/attention objective ($\lambda_{\text{CTC}} = 0.3$). Optimization uses Adam with learning rate 2×10^{-4} and WarmupLR (15k warmup steps), SpecAugment, and gradient accumulation of 16 with numel-based batching. The XLSR-53 front-end is frozen during training.

For **WL**, we fine-tune the WL using the HuggingFace `Seq2SeqTrainer` setup with learning rate 10^{-5}, warmup of 500 steps, per-device batch size 32 with gradient accumulation of 2, FP16 training, and checkpoint selection based on validation WER.

5.5 Model Merging Parameters

We instantiate the merging operator $Merge(\cdot)$ using two weight-space methods: LERP and TIES (Yadav et al., 2023). In all cases, we merge *task-specialized* checkpoints into a single inclusive model while controlling the retention–adaptation trade-off through mixing weights (LERP) or sparse, sign-consensus task-vector aggregation (TIES). Unless stated otherwise, merges follow the checkpoint precision of each model family (ESPnet in `float32`, Whisper in `float16`). For each merging technique, we manually swept the main hyperparameters (e.g., interpolation weights, density, and update weights) and selected the reported configuration based on the best average WER computed

across *all* evaluation test sets.

LERP. For LERP, we use convex weights $\{\alpha_i\}$ over the participating checkpoints (and set the global scaling $\lambda = 1$ as per our implementation), yielding a weighted average in parameter space. For **ED**, we merge the encoder and decoder *separately* using the same weights: $\alpha_{\text{base}} = 0.6$ and $\alpha_{\text{Ar}} = \alpha_{\text{Ar+En}} = 0.2$. For **WL**, we merge the encoder with $\alpha_{\text{base}} = 0.6$ and $\alpha_{\text{Ar+En}} = 0.4$. For **ECTC**, we merge the encoder with $\alpha_{\text{base}} = 0.6$, $\alpha_{\text{Ar}} = 0.2$, and $\alpha_{\text{Ar+En}} = 0.2$.

TIES. For TIES, we use the baseline checkpoint as θ_{base} and merge one or more specialists by applying (i) trimming via a density parameter (retaining the top-magnitude fraction of each task update), (ii) electing a consensus sign, and (iii) disjointly averaging only sign-aligned updates (Yadav et al., 2023). We enable `normalize=true` and `int8_mask=true` in all TIES runs. For **ED**, we apply TIES on the encoder (baseline as `base_model`) with per-expert settings `density`$\in \{0.8, 0.6\}$ and `weight`$\in \{0.6, 0.4\}$. For **WL**, we apply TIES between the baseline and the 1-epoch fine-tuned checkpoint with `density`=0.7 and `weight`=0.7 (baseline as `base_model`). For **ECTC**, we apply TIES using the baseline encoder as `base_model` and two specialists (θ^{Ar+En}_{child} and Aθ^{Ar}_{child}), with `density`$\in \{0.8, 0.6\}$ and `weight`$\in \{0.6, 0.4\}$.

ED merging scope (what is merged vs. kept from baseline). For the **ED** architecture, merging is applied *only* to the main encoder and decoder tensors. We explicitly do *not* merge auxiliary components and statistics: the frozen self-supervised front-end is kept from the baseline, the **CTC head** is inherited from the baseline (`ctc.safetensors`), and **batch-normalization statistics** are also kept from the baseline (`bn_stats.safetensors`). In addition, we retain a baseline `rest.safetensors` block (i.e., the remaining non-encoder/decoder parameters, including feature-extraction/extractor-side and other auxiliary modules) and exclude small tensors (`small_tensors.safetensors`) from merging.

5.6 Evaluation Measures

We use Word Error Rate ($\mathcal{WER}(\theta, \mathcal{D}_*)$) as the primary metric to measure ASR performances. To ensure fair scoring under Arabic orthographic variability, we apply a lightweight normalization to both references and hypotheses before computing WER: we remove diacritics, normalize common letter variants (e.g., *alef* forms, *hamza* variants, *alef maksura* $\rightarrow$ *ya*, and *teh marbuta*), and lowercase Latin text. For **MyST**, we additionally exclude from WER computation segments whose reference is empty or contains only non-speech markers (e.g., `<silence>`, `<noise>`, `(silence)`), as well as unintelligible placeholders (e.g., `(())`); segments with empty hypotheses are also skipped. This prevents the metric from penalizing models for orthographic conventions or no-speech regions.

Following, to quantify the trade-offs between the inherent in model adaptation, we introduce two specific ratios: the *Retention Index* ($\mathcal{R}_{adult}$) and the Adaptation Recovery ($\mathcal{R}_{child}$).

Retention Index The Retention Index, $\mathcal{R}_{adult} = \frac{\mathcal{WER}(\theta_{base}, D^L_{adult})}{\mathcal{WER}(\theta_m, D^L_{adult}))}$ measures the extent to which the merged model θ_m preserves the original capabilities of the base model θ_{base} on adult benchmarks D^L_{adult} ($L \in Ar, En$). with $\mathcal{R}_{adult} < 1.0$ represent the forgetting cost occurred during the adaptation.

Adaptation Recovery The Adaptation Recovery, $\mathcal{R}_{child} = \frac{\mathcal{WER}(\theta_{ft}, D^L_{child})}{\mathcal{WER}(\theta_m, D^L_{child}))}$, measures the performance gap between a unified merged model θ_m and a child-specific ASR.

6 Results and Discussion

Child Fine-Tuning *vs.* Adult Pretraining Across all three architectures – Encoder-Decoder (ED), Encoder-CTC (ECTC), and WhisperLargev3 (WL) – fine-tuning the adult pretrained ASR model on child speech consistently improves performance on child-centric benchmarks (see Table 3). Relative to the adult-only baseline (θ_{base}), fine-tuning with *Arabic child data* yields clear WER reductions on AraKidsL1 and AraKidsL2, indicating effective adaptation to child acoustics and pronunciation patterns. Incorporating *bilingual Arabic+English child data* further strengthens this trend and leads to more balanced improvements across AraKids and MyST, particularly for encoder-decoder based architectures: ED and EL. However, these gains are accompanied by a generalization trade-off. While child-domain WER improves substantially, fine-tuned models may degrade on adult or out-of-domain benchmarks, especially on MGB2 and, to a lesser extent, Libri-C. This suggests that naive fine-tuning on niche data such as children's speech can partially overwrite representations learned from large-scale adult corpora.

Models	AraKidsL1	AraKidsL2	MyST	Libri-C	MGB2	Avg. WER
Encoder–Decoder						
θ_{base}	9.14	44.11	36.68	9.86	10.92	22.142
θ^{Ar}_{child}	6.41	26.76	44.23	12.21	10.7	20.062
θ^{En}_{child}	25.75	106.98	22.68	6.66	14.16	35.246
θ^{Ar+En}_{child}	6.37	22.71	22.52	6.66	12.97	14.312
Encoder–CTC						
θ_{base}	14.72	63.37	26.63	3.83	13.18	24.346
θ^{Ar}_{child}	13.11	16.32	77.32	18.64	23.63	29.804
θ^{En}_{child}	51.44	92.73	15.11	6.8	26.26	38.468
θ^{Ar+En}_{child}	12.45	24.24	15.9	6.66	21.2	16.09
Whisper Large v3						
θ_{base}	9.68	40.68	21.96	2.47	15.15	17.988
θ^{Ar+En}_{child}	8.7	11.84	14.47	3.11	23.52	12.328

Table 3: Reported WER for the base adult-specific θ_{base} model and specialized model with naive fine-tuning for 3 different ASR architecture.

Model Arch	MOperation	AraKidsL1	AraKidsL2	MyST	Libri-C	MGB2	Avg. WER
Encoder–Decoder (ED)							
ED	LERP	7.71	31.88	33.59	9.33	10.84	18.67
ED	LERP (Enc)	7.86	32.04	33.86	9.16	10.86	18.756
ED	TIES	7.14	26.76	28.67	9.23	11.26	16.612
ED	TIES (Enc)	7.35	27.01	28.25	8.16	10.86	16.326
Encoder–CTC (ECTC)							
ECTC	LERP	10.38	39.56	21.91	3.91	13.13	17.778
ECTC	TIES	11.07	18.94	18.68	5.66	18.71	14.612
Whisper-Large v3 (WL)							
WL	LERP	6.55	20.01	17.78	2.43	14.49	12.252
WL	LERP (Enc)	7.64	26.38	20.45	2.41	15.00	14.376
WL	TIES	8.48	11.37	14.36	2.91	22.68	11.96
WL	TIES (Enc)	7.16	19.83	17.55	2.61	17.01	12.832

Table 4: WER (%) across benchmarks for different merge operations. Within each architecture block (ED/ECTC/WL), the best (lowest) score in each column is highlighted in blue.

Arabic-only vs. bilingual child models and Model Merging Across merging setups, using the bilingual child model θ^{Ar+En}_{child} tends to outperform using Arabic-only child models θ^{Ar}_{child} (see Table 6). This pattern indicates that bilingual fine-tuning induces more robust intermediate representations that generalize better when merged with the adult anchor, reducing over-specialization to a single child-language distribution. At the same time, θ^{Ar}_{child} encodes Arabic-centric representations that remain useful to disentangle Ar-only vs. Ar+En directions in weight space. Based on the empirical comparison in Table 3, we therefore select θ^{Ar+En}_{child} and θ^{Ar}_{child} as primary merging candidates anchored with θ_{base}.

LERP vs. TIES. For our reported results, in Table 4, we observed LERP provides moderate improvements on children benchmark, but its recovery on adult benchmarks is remarkable. This is because the interpolation model gives much priority to the base (as anchor). In contrast, TIES consistently reduces average WER across architectures for children ASR benchmark while lacking in the retention capabilities, suggesting the operation is more effective for alignment child-adapted parameters, while retaining major portion of the results. The advantage of TIES is most pronounced for ED and WL, where it yields a stronger adaptation.

Architectural Considerations in Model Merging: For ED and WL architectures, we compare merging *encoder-only* parameters against the *full model* (encoder and decoder) in Table 4. For TIES, Encoder-only merging generally yields superior adult-domain retention – most notably on the MGB2 benchmark, though it may slightly reduce the competitive edge in child-domain specialization relative to full-model merging. This disparity in stability arises because decoders adapted to niche domains, such as children's speech, may collapse toward domain-specific token distributions, thereby reducing their capacity to retain the linguistic breadth acquired during large-scale adult pre-training. By preserving the adult decoder as a stable language model anchor, encoder-only merg-

ing avoids this failure mode while still benefiting from child-adapted acoustic representations in the encoder.

In contrast, for CTC-based architectures, merging the CTC linear projection layer alongside the encoder tends to reduce overall adaptation capability. Due to the limited capacity of the CTC head and its lack of an internal autoregressive language model, the weights are highly sensitive to distribution shifts. Consequently, merging these layers with the base model often dilutes the specialized child-speech distributions learned during fine-tuning, leading to a more significant performance degradation than that observed in ED decoders. This sensitivity is reflected in a simple CTC-head ablation: using the adult CTC head yields 14.61 WER, while swapping to child-specialized CTC heads results in 15.41 WER (CTC head of θ_{child}^{Ar+En}, +0.80) and 18.79 WER (CTC head of θ_{child}^{Ar}, +4.18).

Key Findings: From our exploratory experiments, we observed the following:

- **Adaptation-Generalization Trade-off:** While fine-tuning adult pretrained models consistently improves performance on child-centric benchmarks, it introduces a generalization trade-off where representations learned from large-scale adult corpora are partially overwritten by niche child-speech data.

- **Bilingual Robustness:** Bilingual child models (θ_{child}^{Ar+En}) produce more robust intermediate representations than monolingual experts. This reduces over-specialization and allows for better alignment when merged with adult anchors.

- **Operator Specialization:** The choice of merging operator determines the performance profile:
  - **LERP** excels at adult-domain retention by prioritizing the base anchor, making it ideal for stability.
  - **TIES** excels at child-domain adaptation, effectively aligning child-specific parameters to achieve lower WERs in complex architectures like ED and WL.

- **Decoder Stability vs. CTC Constraints:**
  - In **Encoder-Decoder** and **Whisper** architectures, *encoder-only merging* is the optimal strategy to prevent "decoder collapse," preserving the adult linguistic breadth while adapting to child acoustics.
  - In **CTC** architectures, the linear projection layer lacks the capacity to handle distribution merging, often leading to performance dilution when merged with the base model.

7 Conclusion

Weight-space model merging consistently improves children's ASR while retaining strong adult performance, offering a practical balance between adaptation and retention without additional training cost. Across architectures and benchmarks, merging reduces child-domain WER and mitigates the degradation typically observed after naive fine-tuning, supporting a single inclusive ASR model across age, nativity (L1/L2), and language (Arabic/English).

8 Limitations

While our results consistently show that weight-space model merging improves children's ASR without sacrificing adult retention, our hyperparameter exploration is limited. For each merging technique (e.g., LERP and TIES), we performed only a small manual sweep over the main parameters (interpolation weights, density, and update weights) and selected a single configuration based on the best average WER across all evaluation sets. We did not conduct an exhaustive or automated search (e.g., grid/random/Bayesian optimization), nor did we systematically study sensitivity to these parameters across architectures and datasets. As a result, the reported numbers may underestimate the best achievable performance, and a more comprehensive tuning study could further improve both adaptation and retention, or reveal more stable parameter regimes.

References

Edem Ahadzi, Vishwanath Pratap Singh, Tomi Kinnunen, and Ville Hautamäki. 2025. Continuous learning for children's asr: Overcoming catastrophic forgetting with elastic weight consolidation and synaptic intelligence. In *Proc. Interspeech*.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The MGB-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284. IEEE.

V. Bhardwaj and 1 others. 2022. Automatic speech recognition (asr) systems for children. *Applied Sciences*.

Alexandre Ducorroy and Rachid Riad. 2025. Robust fine-tuning of speech recognition models via model merging: application to disordered speech. In *Proc. Interspeech*.

Brady Houston and Katrin Kirchhoff. 2020. Continual learning for multi-dialect acoustic models. In *Proc. Interspeech*.

Gabriel Ilharco and 1 others. 2023. Editing models with task arithmetic. In *ICLR (OpenReview)*.

R. Jain, L. Barcovschi, and 1 others. 2023. Adaptation of whisper models to child speech recognition. In *arXiv preprint arXiv:2307.13008*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Interspeech 2015*, pages 3586–3589.

Linguistic Data Consortium. 2021. Myst (my science tutor) children's conversational speech. LDC Catalog.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. 2023. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *NeurIPS*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

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 2019*, pages 2613–2617.

Sameer Pradhan, Ronald A. Cole, and Wayne H. Ward. 2024. My science tutor (MyST)–a large corpus of children's conversational speech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12040–12045, Torino, Italia. ELRA and ICCL.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Samik Sadhu and Hynek Hermansky. 2020. Continual learning in automatic speech recognition. In *Proc. Interspeech*.

Yi-Lin Sung and 1 others. 2023. An empirical study of multimodal model merging. In *Findings of EMNLP*.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, and 1 others. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *arXiv preprint arXiv:2203.05482*.

F. Wu and 1 others. 2019. Advances in automatic speech recognition for child speech. In *Proc. Interspeech*.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*.

A Appendix

Arch.	Merge (*)	Adaptation Recovery (AR)				Retention Index (RI)		
		AraKidsL1	AraKidsL2	MyST	Avg.	Libri-C	MGB2	Avg.
ED	LERP	82,62%	71,24%	67,04%	73,63%	100%↑	**100%**↑	100%↑
	TIES	89,22%	84,87%	78,55%	**84,21%**	100%↑	96,98%	98,49%
ECTC	LERP	**61,37%**	57,41%	100%↑	72,92%	100%↑	**83,17%**	91,58%
	TIES	57,54%	100%↑	**100%**↑	**85,84%**	100%↑	58,36%	79,18%
WL	LERP	**97,25%**	100%↑	100%↑	**99,03%**	100%↑	**75,36%**	87,68%
	TIES	75,12%	100%↑	100%↑	**91,70%**	100%↑	48,15%	74,07%

Table 5: AR and RI summary across architectures and merge methods. Values $\geq$100% are capped at 100% and marked with ↑.

Technique	Setup	Children			Adult		Overall
		AraKidsL1	AraKidsL2	MyST	Libri-C	MGB2	Avg. WER
LERP	M0 + B	10,27	34,92	27,8	4,01	12,98	17,996
	M0 + C	11,65	46,72	19,18	3,97	13,47	18,998
	M0 + A + B	11,45	52,34	21,52	3,9	13,1	20,462
	M0 + B + C	10,38	39,56	21,91	3,91	13,13	**17,778**
	M0 + A + B + C	11,72	48,32	19,21	3,92	13,24	19,282

Table 6: Impact of Arabic-only vs. bilingual child specialists in merging. M0 denotes the adult baseline; A/B/C denote child-specialized checkpoints used as merging candidates (A: MyST-only, B: AraKids-only, C: AraKids+MyST bilingual). Average WER is computed over all reported test sets.