Pseudo2Real: Task Arithmetic for Pseudo-Label Correction in Automatic Speech Recognition

Yi-Cheng Lin* **Yu-Hsuan Li Liang*** **Hsuan Su** **Tzu-Quan Lin**
Shang-Tse Chen **Yun-Nung Chen** **Hung-yi Lee**†
National Taiwan University, Taipei, Taiwan
†NTU Artificial Intelligence Center of Research Excellence, Taipei, Taiwan
{f12942075, r14922013, hungyilee}@ntu.edu.tw

Abstract

Robust ASR under domain shift is crucial because real-world systems encounter unseen accents and domains with limited labeled data. Although pseudo-labeling offers a practical workaround, it often introduces systematic, accent-specific errors that filtering fails to fix. We ask: How can we correct these recurring biases without target ground truth? We propose a simple parameter-space correction: in a source domain containing both real and pseudo-labeled data, two ASR models are fine-tuned from the same initialization, one on ground-truth labels and the other on pseudo-labels, and their weight difference forms a correction vector that captures pseudo-label biases. When applied to a pseudo-labeled target model, this vector enhances recognition, achieving up to a 35% relative Word Error Rate (WER) reduction on AFRISPEECH-200 across ten African accents with the Whisper TINY model.

1 Introduction

ASR technologies are increasingly deployed across various domains, including smart assistants, medical transcription, and emerging low-resource accents or languages (Zhang et al., 2023). However, when models encounter speech from new domains, labeled training data is often scarce or completely unavailable (Mai and Carson-Berndsen, 2022; Damianos et al., 2025). Collecting high-quality transcriptions in these settings is costly, time-consuming, and sometimes infeasible due to privacy or legal constraints (Bäckström, 2025; Shoemate et al., 2022).

A common approach is to generate pseudo-labels using existing ASR models (Likhomanenko et al., 2023; Moritz et al., 2021). However, pseudo-labels inherit the teacher model's systematic biases, such as under-recognizing rare words, accent-driven substitutions, or domain-specific mis-segmentations (Higuchi et al., 2022). When used for adaptation, these errors can accumulate and degrade real-world performance (Prakash et al., 2025), revealing the need for methods that automatically identify and correct structured pseudo-label errors without relying on target-domain ground truth.

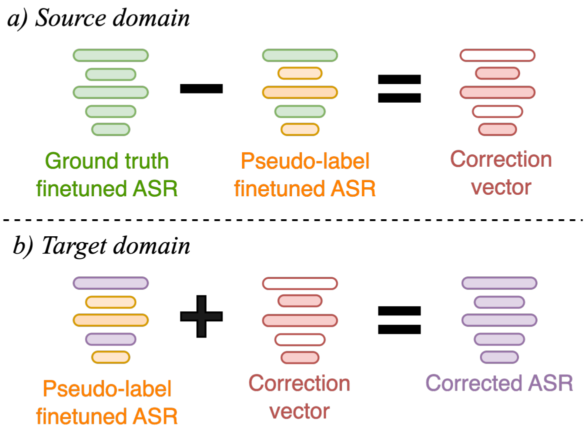

Figure 1: **Overview of Pseudo2Real.** a) In the source domain, two ASR models are fine-tuned from the same pretrained initialization: one using ground-truth transcripts and one using pseudo-labels. Their parameter difference defines a correction vector that captures systematic pseudo-labeling biases. b) In a new target domain, this correction vector is added to a pseudo-label fine-tuned model to produce a corrected ASR that better aligns with real-label performance. Color semantics: green = source-domain (ground-truth) knowledge, orange = pseudo-label noise, and purple = target-domain knowledge.

In this work, we ask: **How can we mitigate systematic error patterns in ASR pseudo-labeling when no ground truth annotations are available in the target domain?** Prior efforts address this challenge indirectly. Teacher-student self-training improves with confidence filtering and agreement checks (Flynn and Ragni, 2024; Kim and Lee, 2025), yet these strategies suppress noise without correcting the structured biases from the teacher. Iterative schedules such as Noisy Student (Chen

*These authors contributed equally.

Findings of the Association for Computational Linguistics: ACL 2026, pages 1160–1175
July 2-7, 2026 ©2026 Association for Computational Linguistics

et al., 2023) and moving-average teacher updates (Zhang et al., 2024) improve pseudo-labels but require multiple passes and careful tuning, and they still propagate the teacher's recurring mistakes.

We propose Pseudo2Real, a parameter-space correction that operates without target labels, as depicted in Figure 1. In source domains that provide both real transcripts and pseudo-labels, we fine-tune two models from the same backbone and form a correction vector as their difference. This vector captures systematic discrepancies introduced by pseudo-labeling. When adapting to a new target domain, the model fine-tuned on pseudo-labeled audio is adjusted by adding a scaled version of this vector, yielding a corrected ASR system that better aligns with real-label performance. Our extensive experiments on the AfriSpeech-200 dataset demonstrate that Pseudo2Real achieves consistent gains across ten African accents and multiple Whisper model sizes, including up to 35% relative WER reduction on Whisper TINY. Our main contributions include:

1. We introduce **Pseudo2Real**, an effective parameter-space correction that mitigates systematic pseudo-label errors.
2. We extend it to **Pseudo2Real-SC**, which leverages speaker clustering to compute subgroup-specific correction vectors, thereby further enhancing robustness.
3. We demonstrate substantial performance improvements across accents and model scales, analyze the effect of scaling factors and the number of clusters, and provide insights into how structured pseudo-label biases can be corrected directly in parameter space.

2 Related work

2.1 Pseudo-labeling for ASR unsupervised domain adaptation

Work in ASR widely adopts teacher–student self-training to exploit unlabeled target audio (Flynn and Ragni, 2024). A first line of research uses a strong teacher to generate pseudo-labels, then trains a student on these labels. The quality of pseudo-labels can be improved through iterative self-training approaches such as Noisy Student (Park et al., 2020; Singh et al., 2023; Ahmad et al., 2024), where the teacher is repeatedly updated on its own pseudo-labels with noise injection and augmentation, yielding large relative WER gains in ASR adaptation. A second line focuses on label quality control. Confidence filtering and agreement checks between models are used to downweight or discard unreliable segments before student training, which prevents error amplification in the target domain (Kim and Lee, 2025; Zhu et al., 2023; Likhomanenko et al., 2023). A third line refines the teacher itself during adaptation (Rao et al., 2023). For example, KAIZEN updates the teacher as an exponential moving average of the student, yielding stronger pseudo-labels and improved unsupervised adaptation (Manohar et al., 2021).

Our work differs in purpose and mechanism. Instead of only filtering or iterating on noisy pseudo-labels, we learn a reusable correction in parameter space. We construct a vector that captures the discrepancy between models trained on synthetic and real speech in auxiliary domains, then add this vector to a pseudo-label adapted model in a new domain to mitigate systematic error patterns without any target labels. This complements prior pseudo-label pipelines and can be combined with confidence filtering or iterative self-training.

2.2 Task arithmetic in speech

Recent works have explored task vectors (task arithmetic) (Ilharco et al., 2023) as a means to transfer capabilities between models (Li et al., 2025; Lin et al., 2025a; Huang et al., 2024; Ritter-Gutierrez et al., 2025a,b). In ASR, Task Vector Algebra shows that difference vectors between models trained on related settings can enable zero-shot domain adaptation and task analogy for low-resource scenarios (Ramesh et al., 2024). Extending this idea, Kang et al. (2024) demonstrates that multilingual ASR can be controlled or composed across languages via simple vector addition or negation, while Nagasawa et al. (2025) shows that combining task vectors from related high-resource languages improves low-resource ASR through cross-lingual transfer. LoRS-Merging (Zhao et al., 2025) merges language- or task-specific deltas using low-rank and sparse decomposition to enhance multilingual ASR without retraining. Building on this paradigm, SYN2REAL (Su et al., 2024) defines a vector between ASR models fine-tuned on authentic versus synthetic speech and applies it to bridge the gap in acoustic signal distributions.

Our work adopts task arithmetic but targets a different problem: mitigating systematic pseudo-label errors in unsupervised ASR domain adaptation, rather than transferring general capabilities across languages, modalities, or text domains. The closest prior method is SYN2REAL (Su et al., 2024): it constructs a task vector between ASR models trained on real vs. synthetic audio to close the acoustic gap between TTS-generated and human speech, typically used for adapting to new text domains. In contrast, Pseudo2Real constructs a correction vector between models trained on real vs. pseudo-labels of the same real audio to mitigate systematic label-noise biases, and applies it to new acoustic domains such as accented speech. Furthermore, while SYN2REAL relies on *domain labels* to ensemble task vectors across multiple text domains, our Pseudo2Real-SC variant uses automatic speaker clustering to form multiple correction vectors, requiring no domain labels in the source data.

3 Methodology

3.1 Problem Formulation

We study acoustic domain adaptation for ASR, focusing on accent as the primary axis of domain variation. Let the source domain D_s consist of paired speech and text (S_s, T_s), and let the target domain D_t provide only unlabeled speech S_t. Ground-truth transcriptions T_t are unavailable due to annotation cost. A common strategy is to train a teacher ASR model on D_s, generate pseudo-labels $\hat{T}_t$ for S_t, and then train a student ASR model on $(S_t, \hat{T}_t)$.

This approach is effective but suffers from **systematic error propagation**: if the teacher consistently misrecognizes rare words or accent-specific patterns, these biases are inherited by the student. Confidence filtering or re-weighting pseudo-labels can reduce noise, but cannot correct the structured error patterns that arise from teacher model biases. Our goal is therefore to automatically detect and mitigate systematic pseudo-label errors without ground-truth annotations in D_t.

3.2 Pseudo2Real

We build on task arithmetic to design a parameter-space correction transferable across domains. A key enabling observation is *linear mode connectivity* (Frankle et al., 2020; Neyshabur et al., 2020): models fine-tuned from the same pre-trained initialization tend to remain within a shared low-loss region of parameter space, so that their weight differences can be interpreted as meaningful directions rather than arbitrary noise. This interpretation underlies the task-vector framework (Ilharco et al., 2023) and has been further supported by theoretical analyses showing that fine-tuning induces

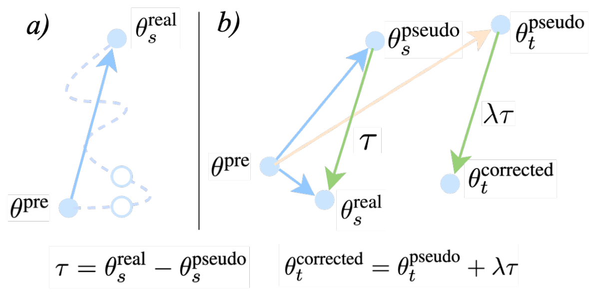

Figure 2: **Learning and applying correction vectors in parameter space.** *a)* A task vector is obtained by taking the difference between a pretrained model θ^{pre} and its fine-tuned version θ_s^{real} (or θ_s^{pseudo}). *b)* In the source domain, two models are fine-tuned from the same pretrained initialization θ^{pre}: one with real transcripts (θ_s^{real}) and one with pseudo-labels (θ_s^{pseudo}). Their difference defines the correction vector τ. In a new target domain, we first obtain θ_t^{pseudo} by fine-tuning on pseudo-labels, then apply the correction vector to yield the final model $\theta_t^{\text{corrected}}$.

weight-disentangled directions in parameter space, each governing a localized region of function space (Ortiz-Jimenez et al., 2023). Our key observation is that, in a source domain where both real labels and pseudo-labels are available, one can learn such a direction between models trained on the two label types; this direction captures the systematic biases induced by pseudo-labels in that domain. Adding a scaled version of it to a target-domain pseudo-label-trained model shifts the parameters along the pseudo-to-real axis identified in the source. Moreover, since each such direction is defined between endpoints fine-tuned from the same initialization, task arithmetic shows that multiple directions can be linearly combined to compose their effects (Ilharco et al., 2023), motivating the aggregation of correction vectors across speaker subgroups. These properties motivate the two variants of our method introduced below.

Pseudo2Real: Single correction Vector. Starting from the same pre-trained backbone θ^{pre}, we fine-tune two student ASR models on the source domain: θ_s^{real}, trained on (S_s, T_s), and θ_s^{pseudo}, trained on $(S_s, \hat{T}_s)$ where $\hat{T}_s$ are pseudo-labels generated by a teacher for S_s. The difference between these two student models defines a correction vector:

$$\tau = \theta_s^{\text{real}} - \theta_s^{\text{pseudo}}. \quad (1)$$

To adapt to the target domain, we fine-tune a student model on $(S_t, \hat{T}_t)$ to obtain θ_t^{pseudo}, and then apply the correction vector:

$$\theta_t^{\text{corrected}} = \theta_t^{\text{pseudo}} + \lambda\tau, \quad (2)$$

where λ is a scaling factor controlling the correction strength. This method applies a single correction vector derived from the source domain directly to the target model.

Pseudo2Real-SC: Subgroup Correction Vectors. The second variant extends this idea by recognizing that systematic pseudo-label errors may not be homogeneous across all speakers in the source domain. In practice, pseudo-labeling quality can vary substantially due to accent, pronunciation style, or recording conditions. For example, a teacher model may systematically substitute certain consonants for speakers with a specific accent, while producing relatively accurate transcriptions for speakers from another subgroup. If all speakers are pooled together when constructing the correction vector, these fine-grained biases may not be taken into account, weakening the correction signal.

To address this, we propose partitioning the source domain into more coherent speaker subgroups and computing subgroup-specific correction vectors. Inspired by Lin et al. (2025b), we refine the correction by exploiting speaker diversity within the source domain. We use ECAPA-TDNN[1] embeddings (Desplanques et al., 2020) to extract speaker representations for each utterance, and apply k-means clustering (MacQueen, 1967) to partition the source data into speaker subgroups. For each subgroup c, we fine-tune two models from θ^{pre}: $\theta_{s,c}^{\text{real}}$, trained on real transcriptions of the subgroup, and $\theta_{s,c}^{\text{pseudo}}$, trained on pseudo-labels of the subgroup. We then compute a subgroup-specific correction vector:

$$\tau_c = \theta_{s,c}^{\text{real}} - \theta_{s,c}^{\text{pseudo}}. \quad (3)$$

The final correction is obtained by averaging across all C clusters and applying it to the target model:

$$\theta_t^{\text{corrected}} = \theta_t^{\text{pseudo}} + \frac{\lambda}{C}\sum_{c=1}^{C}\tau_c. \quad (4)$$

This aggregated vector captures systematic pseudo-label biases shared across speakers while preserving accent- or subgroup-specific corrections.

[1] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

3.3 Baselines

We compare Pseudo2Real against six baselines and a topline. As *no-adaptation references*, $\theta^{\mathbf{pre}}$ applies the pretrained Whisper directly to the target domain, and $\theta_s^{\mathbf{real}}$ fine-tunes the student on source-domain ground-truth labels; together they show how far a model can go without target-domain supervision. As the *pseudo-labeling baseline*, $\theta_t^{\mathbf{pseudo}}$ fine-tunes the student on target-domain pseudo-labels; this is the direct baseline that Pseudo2Real aims to improve upon. Prior-art methods for pseudo-label noise fall into two groups. *Label-space* methods act on the pseudo-labels: **Confidence filtering (Conf.)** retains pseudo-labels whose average word-level log-probability exceeds a development-set threshold, and **Error Correction (EC)** uses a T5 model trained on source-domain (hypothesis, ground-truth) pairs to rewrite target-domain pseudo-labels before student fine-tuning. *Weight-space* methods modify the student's parameters: **Exponential Moving Average (EMA)** maintains a running average of the student's weights during training and uses it at evaluation. Pseudo2Real is also a weight-space method, but uses a source-derived correction vector rather than a training-trajectory average. Finally, the **topline** fine-tunes the student on target-domain ground-truth labels as an upper-bound reference. EC and EMA implementation details are in Appendix F.

4 Experimental Setup

Dataset. We evaluate our method on the Afrocentric benchmark AFRISPEECH-200[2] (Olatunji et al., 2023), a 200-hour corpus of transcribed English speech from speakers representing 120 African accents, with explicit accent annotations. Accented speech remains a persistent challenge for ASR systems because strong accent variations often fall outside the distribution of large-scale pretraining corpora. AFRISPEECH-200 therefore provides a rigorous testbed for domain adaptation methods. For our experiments, we filter the corpus by accent and select the ten accents with the largest number of samples. Not all accents include complete train, development, and test splits, so we restrict our selection to accents where all three splits are available and preserve the official split to avoid data leakage. All utterances in these subsets

[2] https://huggingface.co/datasets/intronhealth/afrispeech-200

are English speech; each accent label refers to the native language background of the speakers rather than the language of the transcript (e.g., "Igbo" denotes Igbo-accented English spoken by Igbo native speakers, not transcripts in the Igbo language).

Cross-Fold Validation. To evaluate the generalization ability of our method across diverse accents, we construct a cross-fold validation setting based on the ten most represented accents in AFRISPEECH-200. We group these accents into two folds with comparable numbers of utterances and speakers, so that neither adaptation direction benefits from substantially more training data: *fold 1* consists of {Igbo, Swahili, Hausa, Zulu, Twi}, and *fold 2* consists of {Yoruba, Ijaw, Afrikaans, Idoma, Setswana}. Although all ten accents originate from African regions, they span multiple and largely unrelated language families (Niger–Congo, Afro-Asiatic, and Indo-European), and differ markedly in tonal systems, syllable structures, consonant inventories, and morphology; cross-accent adaptation across these families is therefore highly non-trivial. We list the linguistic family of each accent in Appendix D. In each experiment, one fold serves as the source domain and the other as the target. The source fold provides paired speech and transcripts that are used to derive correction vectors, while the target fold provides only speech for pseudo-labeling. We then swap the roles of the folds to form the second validation round. This design ensures that the evaluation covers accents with different phonological and prosodic characteristics.

Model. We employ the Whisper family of models (Radford et al., 2023), which cover a wide range of capacities. Specifically, we experiment with Whisper TINY, BASE, SMALL, MEDIUM, and LARGE-V2. These models share the same encoder–decoder transformer architecture but differ in scale, ranging from 39M to 1.55B parameters, detailed in Table 6. All models are pre-trained on approximately 680k hours of weakly supervised speech and are widely adopted in both research and real-world applications (Yang et al., 2024; Wu et al., 2024; Luo et al., 2025; Lin et al., 2024). Despite its scale and multilingual coverage, prior work has shown that its performance still degrades substantially when facing strong accent variation or domain-specific shifts (Graham and Roll, 2024). Evaluating across the full model series allows us to assess the effectiveness of our correction method under both low-capacity and high-capacity regimes.

Training. In our experiments, we fully fine-tune the Whisper SMALL and TINY models as student models using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 3×10^{-5} and a weight decay of 0.1. Training is conducted for up to 40K update steps with a linear warmup of 500 steps. Each model is trained with a batch size of 16, using mixed-precision (FP16) to reduce memory consumption. Evaluation is performed every 50 steps using the word error rate (WER) metric with greedy decoding. For *task arithmetic*, we use the entire model parameter to compute the correction vector. The scaling factor λ is selected using the source-domain development sets: we perform a simple grid search over $\lambda \in \{0.1, 0.2, 0.3, ..., 1.0\}$ and choose the value that minimizes WER on the held-out source development set. All experiments are conducted on a single NVIDIA V100 GPU, totaling approximately 500 GPU-hours.

5 Result

5.1 Can Pseudo2Real improve ASR performance?

We begin by examining whether the application of Pseudo2Real can improve ASR performance in the cross-validation setting, compared to the baselines introduced in §3.3.

In our main experiments, we evaluate the case where the teacher model and the student model are identical. The results in Table 1 show that both baseline methods underperform significantly across all accent domains. The pretrained models θ^{pre} perform poorly, especially on low-resource accents such as Hausa, Ijaw, and Idoma, with average WERs of 106.5 and 54.9 for Whisper TINY and SMALL, respectively. Fine-tuning on pseudo-labeled data θ_t^{pseudo} provides noticeable gains, reducing the average WER by 17.2 points for TINY and 7.7 points for SMALL, but large error rates persist due to systematic biases in the pseudo-labels.

In contrast, **Pseudo2Real** achieves substantial improvements across all accents, demonstrating the effectiveness of parameter-space correction. For Whisper TINY, Pseudo2Real lowers the average WER from 89.3 to 57.7, representing a 35% relative improvement over pseudo-label fine-tuning. For Whisper SMALL, the average WER decreases from 47.2 to 45.0, bringing performance much closer to the topline trained on labeled target data.

Model	Variant	Igbo	Swahili	Hausa	Zulu	Twi	Yoruba	Ijaw	Afrikaans	Idoma	Setswana	Avg.
Tiny	θ^{pre}	93.2	77.7	142.8	79.0	73.3	94.8	191.2	54.4	182.9	75.6	106.5
	θ_s^{real}	**60.1**	67.3	135.3	70.8	53.8	105.0	155.8	51.8	128.2	54.4	88.2
	θ_t^{pseudo}	61.8	70.5	119.6	75.2	59.4	112.3	157.7	52.5	129.0	55.0	89.3
	conf.	73.0	58.7	147.7	**56.0**	57.7	97.8	194.6	52.0	107.2	**41.9**	88.7
	EC	81.7	64.5	93.0	79.4	75.6	122.2	137.2	47.8	83.7	53.1	83.8
	EMA	88.2	62.4	131.8	79.4	74.7	122.0	167.1	57.4	140.1	63.8	98.7
	ours	60.3	**51.9**	**78.8**	67.0	**45.5**	**61.3**	**61.3**	**45.4**	**60.7**	44.8	**57.7**
	topline	65.7	45.8	54.1	59.7	88.0	56.1	58.1	41.4	112.4	44.8	62.6
Small	θ^{pre}	56.4	49.6	60.6	49.7	49.7	62.8	63.2	41.3	66.0	49.7	54.9
	θ_s^{real}	52.9	**39.7**	73.8	40.7	37.5	56.9	52.3	36.0	55.6	40.3	48.6
	θ_t^{pseudo}	53.1	39.9	**58.2**	41.1	37.5	57.2	52.4	36.4	55.8	40.7	47.2
	conf.	52.1	43.0	78.3	42.5	38.2	50.2	49.9	36.2	55.8	41.9	48.8
	EC	50.6	40.9	63.1	44.3	43.8	50.8	137.0	**30.6**	55.5	39.8	55.6
	EMA	62.9	47.3	66.8	49.4	52.9	62.0	61.8	41.0	63.7	48.6	55.6
	ours	**46.5**	42.3	78.0	**38.5**	**35.0**	**44.1**	**47.1**	32.4	**48.9**	**37.1**	**45.0**
	topline	49.6	35.7	52.5	36.8	35.0	41.3	45.4	29.5	47.7	36.2	41.0

Table 1: **WER (%) on ten accented English target domains.** Results are shown for Whisper TINY and Whisper SMALL under eight adaptation settings. Lower is better. The best performance is **bolded**.

Pseudo2Real shows particularly strong gains on challenging accents such as **Ijaw**, **Idoma**, and **Yoruba**, where WER is reduced by up to 50 points compared with pseudo-label fine-tuning. These improvements indicate that the correction vector effectively captures accent-specific pronunciation patterns and mitigates systematic pseudo-labeling errors that standard fine-tuning fails to address.

Interestingly, Pseudo2Real occasionally outperforms the topline trained on labeled target data. For instance, Whisper TINY outperforms the topline on **Igbo**, **Twi**, and **Idoma**, while Whisper SMALL does so on **Igbo**. This behavior suggests that parameter-space correction not only compensates for pseudo-label noise but also transfers beneficial regularities learned from other domains, such as better acoustic normalization and pronunciation consistency. In these cases, the correction vector serves as a form of cross-domain regularization, enabling the adapted model to generalize more effectively than models trained solely on limited labeled target data.

Why does Pseudo2Real sometimes beat the topline? We compare error types on the two accents where Pseudo2Real beats the topline in the tiny→tiny setting, **Twi** and **Igbo** (Table 3). On both accents, Pseudo2Real makes fewer *insertion* errors than the topline (most strikingly, 262 vs. 670 on Twi), while substitutions and deletions are comparable. Since insertions are common hallucination patterns that small target-domain fine-tuning rarely fully suppresses, this reduction suggests that the correction vector transfers additional anti-hallucination regularization from the source domain.

5.2 How does Pseudo2Real perform across different teacher model sizes?

An important question is whether Pseudo2Real can generalize across different ASR model sizes rather than be tied to a specific backbone. To examine this, we consider settings where the student and teacher may differ in capacity. In Table 2, we report results on five accents in fold 1.

For the Whisper TINY student, the strongest average gains occur with BASE and LARGE teachers, yielding +21.7% and +21.6% relative improvements, respectively. The BASE teacher produces consistent reductions across accents, including **Igbo** (+29.0%) and **Twi** (+31.3%), while the LARGE teacher achieves the largest single-accent gain on **Hausa** (+45.8%). Using SMALL or MEDIUM teachers leads to smaller average gains (+8.4% and +4.4%), with mixed outcomes such as a degradation on **Igbo** for and MEDIUM teacher (-11.5%) but a strong improvement on **Zulu** for SMALL teacher (+26.3%).

For the Whisper SMALL student, improvements are more modest overall. Pairing with a MEDIUM teacher yields a +5.2% average relative improvement, with steady gains on nearly all accents. The LARGE teacher yields nearly identical performance to the baseline on average (around +0.1%), showing mixed results across accents. It improves on **Igbo** (+14.0%) and **Setswana** (+5.6%) but degrades notably on **Hausa** (-18.8%).

Overall, Pseudo2Real is effective across a range of teacher sizes, but the magnitude of improvement

Student	Teacher	Variant	Igbo	Swahili	Hausa	Zulu	Twi	Avg.
Tiny	–	θ^{pre}	93.2	77.7	142.8	79.0	73.3	93.20
	Base	θ_t^{pseudo}	73.9	55.6	94.9	82.8	64.6	74.36
		Pseudo2Real	52.5	50.6	76.6	66.1	44.4	58.04
		Improvement (%)	**+29.0**	**+9.0**	**+19.2**	**+20.2**	**+31.3**	**+21.7**
	Small	θ_t^{pseudo}	59.88	50.44	66.59	65.58	49.66	58.43
		Pseudo2Real	52.58	49.15	73.25	48.33	44.24	53.51
		Improvement (%)	**+12.2**	**+2.6**	**-10.0**	**+26.3**	**+10.9**	**+8.4**
	Medium	θ_t^{pseudo}	58.44	51.48	72.54	49.81	46.39	55.73
		Pseudo2Real	65.19	46.78	64.66	47.42	42.33	53.28
		Improvement (%)	**-11.5**	**+9.1**	**+10.9**	**+4.8**	**+8.8**	**+4.4**
	Large	θ_t^{pseudo}	68.19	54.45	103.39	50.09	50.00	65.22
		Pseudo2Real	51.77	43.26	56.06	48.44	42.89	48.48
		Improvement (%)	**+24.1**	**+20.5**	**+45.8**	**+3.3**	**+14.2**	**+21.6**
Small	–	θ^{pre}	56.4	49.6	60.6	49.7	49.7	53.20
	Medium	θ_t^{pseudo}	46.79	39.35	50.86	40.01	36.68	42.74
		Pseudo2Real	40.31	37.25	50.86	39.27	34.76	40.49
		Improvement (%)	**+13.8**	**+5.3**	**+0.0**	**+1.8**	**+5.2**	**+5.2**
	Large	θ_t^{pseudo}	46.68	43.85	55.02	40.96	38.37	44.98
		Pseudo2Real	40.14	45.32	65.33	39.76	36.23	45.36
		Improvement (%)	**+14.0**	**-3.4**	**-18.8**	**+2.9**	**+5.6**	**+0.1**

Table 2: **WER (%) on five accented English target domains from AFRISPEECH-200, under different teacher model sizes.** Each *Improvement* row reports the relative improvement (%) of our Pseudo2Real method over pseudo-labeled fine-tuning. **Green** indicates gains (lower WER); **red** indicates degradation.

Accent	Model	Sub.	Ins.	Del.
Twi	Pseudo2Real	238	**262**	11
	Topline	251	670	4
Igbo	Pseudo2Real	1933	**3577**	205
	Topline	2004	3748	189

Table 3: Error breakdown (substitutions, insertions, deletions) for the tiny→tiny setting on two accents where Pseudo2Real surpasses the topline. Pseudo2Real consistently reduces insertion errors, while substitution and deletion counts remain comparable.

depends on the teacher–student pairing and the target accent.

5.3 What is the impact of the scaling factor λ?

We next examine how the scaling factor λ affects the magnitude of the applied correction vector. As described in § 3.2, λ controls how strongly the Pseudo2Real vector influences the target model parameters. Figure 3 presents the relationship between WER and λ for five transfer settings involving different teacher–student combinations, with λ ranging from 0.0 (no correction) to 0.5 (strong correction). The reported WER values are averaged over the same five accents used in Table 2.

Across all transfer settings, we observe a U-shaped trend. As λ increases from 0, WER first decreases, reaching its minimum around λ = 0.2–0.3, and then rises sharply for larger λ values. This pattern suggests that small scaling factors yields the best balance between correction strength and stability, whereas excessively large λ values can lead to over-correction and degraded accuracy. For instance, the TINY→TINY and TINY→SMALL perform best at $\lambda = 0.3$, while excessive scaling beyond this point causes performance degradation. Similar behavior is observed in the BASE→TINY and MEDIUM→TINY settings. However, the LARGE→TINY case shows greater instability at high λ, likely because the strong correction signal from a large teacher is difficult for a small student to absorb.

Importantly, even small scaling factors (λ = 0.1–0.2) consistently improve performance compared to the uncorrected case ($\lambda = 0.0$). This demonstrates that applying a mild parameter-space correction is generally beneficial and robust across different model configurations. Overall, the results confirm that the scaling factor λ plays a critical role in balancing correction strength and model stability.

Teacher	Variant	Igbo	Swahili	Hausa	Zulu	Twi	Avg.
Medium	Pseudo2Real	40.3	37.3	63.4	39.3	35.2	43.10
	Pseudo2Real-SC	41.1	37.3	49.6	38.9	35.4	40.46
	Improvement (%)	**-2.0**	**0.0**	**+21.8**	**+1.0**	**-0.6**	**+4.0**
Large	Pseudo2Real	40.1	45.3	65.3	39.8	36.2	45.34
	Pseudo2Real-SC	40.3	40.8	53.9	39.2	35.4	41.92
	Improvement (%)	**-0.5**	**+9.9**	**+17.5**	**+1.5**	**+2.2**	**+6.1**

Table 4: **Comparison between Pseudo2Real and its subgroup clustering variant (SC).** The **Improvement** row reports the relative change (%) of +SC over Pseudo2Real. **Green** indicates gains (lower WER), **red** indicates degradation. Lower is better.

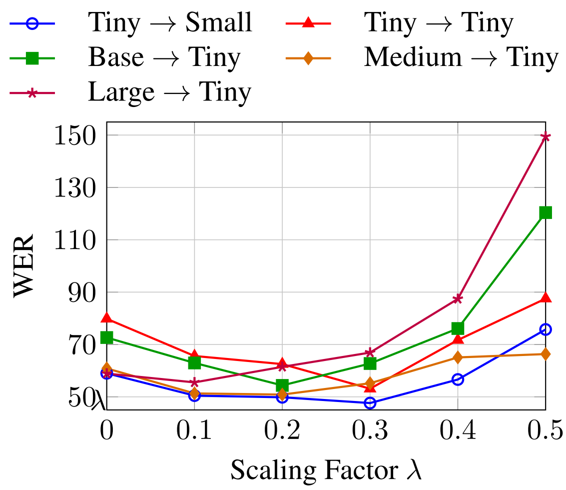

Figure 3: **WER vs. scaling factor (λ).** Each curve corresponds to a different teacher–student pairing. Here, the arrow ($\rightarrow$) denotes that pseudo-labels are generated by the teacher ASR model on the left and used to fine-tune the student model on the right (e.g., LARGE→TINY means pseudo-labels are produced by the LARGE teacher, and the TINY student's parameters are then adjusted using the Pseudo2Real correction vector). WER values are averaged over the five fold-1 accents (Igbo, Swahili, Hausa, Zulu, Twi). Lower WER indicates better performance.

6 Pseudo2Real-SC

6.1 How is the Pseudo2Real-SC compared with the simple correction vector?

We now investigate whether ensembling multiple correction vectors yields further improvements over a single correction vector. We evaluate two Whisper teacher sizes (LARGE, MEDIUM) paired with the Whisper SMALL student, comparing Pseudo2Real with its ensemble variant in Table 4. We focus on these pairings because the student benefits only modestly from a single correction vector when paired with high-capacity teachers (+5.2% for MEDIUM and +0.1% for LARGE on average; Table 2), with pronounced per-accent variation. This suggests that pseudo-label biases in these settings are more heterogeneous across speakers, making them the most informative cases for testing whether subgroup clustering adds further benefit. We use 8 clusters for k-means.

The results show that Pseudo2Real-SC generally maintains or improves performance relative to the single correction vector, with the magnitude of gains depending on the teacher size and target accent. For the MEDIUM teacher, Pseudo2Real-SC achieves an average 4.0% relative improvement, driven primarily by large gains on **Hausa** (+21.8%), while other accents remain stable. With the LARGE teacher, the ensemble variant produces a stronger average gain of 6.1%, including notable improvements on **Hausa** (+17.5%) and **Swahili** (+9.9%). These results indicate that averaging subgroup-specific correction vectors can enhance robustness by capturing complementary correction patterns from diverse speaker groups.

However, improvements are not universal. **Igbo** shows slight degradation with both the LARGE and MEDIUM teachers, indicating that excessive averaging can weaken accent-specific corrections. Overall, Pseudo2Real-SC generally improves on the single correction vector, especially when the teacher has sufficient capacity to model heterogeneous speaker variations.

6.2 Ablation on the number of clusters

We further analyze how the number of k-means clusters used in Pseudo2Real-SC affects adaptation performance. Figure 4 presents the results for the LARGE→SMALL transfer setting, where the number of clusters k varies from 1 (no clustering) to 8. The WER values are averaged across the same five accents used in previous experiments.

As the number of clusters increases, WER decreases. Using a single cluster corresponds to the standard Pseudo2Real setting, yielding a WER of

Example and Description
Training set Ground Truth (Ijaw): The codes used for the four-needle telegraph are not known, and none of the equipment has survived.
Teacher Pseudo-label: because used for the 4FN2 telegraph, I'm not known command, and I'm not the equipment as a vif
Error Type: Acoustic confusion—the teacher misinterprets the phonetic pattern of the word "survived" as the acoustically similar but meaningless phrase "as a vif".
Testing set Ground Truth: If the child survives, he or she should be monitored for the later appearance of colonic polyps.
Pretrained Tiny: if the child survives he or she should be monitored for the data appearance of colonical
Tiny (Pseudolabel): if the child is a vice he or she should be monitored for the later appearance of colonic politics
Tiny (Pseudo2Real): if the child survives he or she should be monitored for the later appearance of colonic politics
Error Mitigation: Pseudo2Real restores the correct lexical meaning "survives," correcting the acoustic corruption inherited from the teacher.

Table 5: **Qualitative examples** showing how Pseudo2Real corrects systematic pseudo-label errors. Teacher errors (top) propagate to the student trained on pseudo-labels, while Pseudo2Real effectively suppresses these patterns and restores the intended meaning. Red = error; green = corrected token.

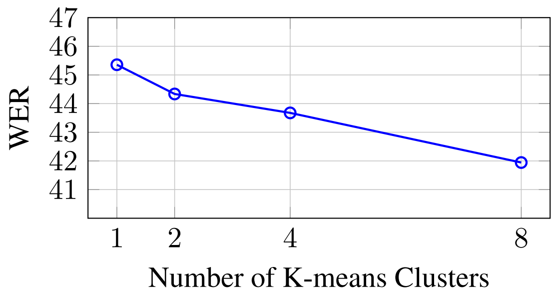

Figure 4: **WER vs. number of K-means clusters** for the LARGE→SMALL setting. WER values are averaged over the five fold-1 accents (Igbo, Swahili, Hausa, Zulu, Twi). Increasing the number of clusters improves adaptation quality (lower WER).

45.36. As k increases to 2 and 4, the WER gradually decreases to 44.34 and 43.68, respectively, and reaches the lowest value of 41.94 at $k = 8$. This trend suggests that finer clustering enables the model to capture more localized speaker- or accent-specific correction patterns, resulting in improved generalization to target-domain speech.

However, increasing the number of clusters raises computational cost, since two ASR models (real and pseudo) must be trained per cluster. In practice, moderate values such as $k = 4$–8 provide a good balance between performance and efficiency. Overall, the ablation shows that leveraging speaker diversity via clustering enhances the effectiveness of parameter-space correction in Pseudo2Real.

7 Case study

To better understand how Pseudo2Real corrects systematic pseudo-labeling errors, we present a qualitative example in Table 5 that compares transcriptions from the teacher, student, and corrected models against the ground truth. More case studies can be found in Table 8.

The teacher produces nonsensical output ("as a vif"), and the fine-tuned student inherits part of this lexical confusion ("is a vice"). Pseudo2Real replaces the incorrect token with the correct verb "survives," recovering the intended meaning while keeping the rest of the sentence intact. This indicates that the correction vector adjusts the model's internal representations to reduce systematic substitution errors commonly found in pseudo-labeling.

8 Conclusion

This work proposed **Pseudo2Real**, a parameter-space correction method that mitigates systematic pseudo-label errors in ASR domain adaptation without requiring target-domain labels. Experiments on AFRISPEECH-200 across ten African accents and multiple Whisper sizes show consistent gains, achieving up to 35% relative WER reduction on Whisper TINY and occasionally surpassing topline models trained with true labels. We also introduced **Pseudo2Real-SC**, which yields additional improvements in several teacher–student settings. Future work includes extending Pseudo2Real to multilingual and spontaneous speech settings, exploring the dynamic scaling of correction strength, analyzing the interpretability of learned correction vectors, developing cluster-conditional correction vectors that can be selected or mixed adaptively at inference time based on speaker characteristics, and scaling Pseudo2Real-SC to larger numbers of clusters ($k > 8$) to study whether finer-grained subgroup partitioning yields further gains.

9 Limitation

Source domain supervision Our approach assumes access to at least one source domain with paired speech and ground-truth transcriptions in order to construct the correction vector(s). If the available source supervision is too small, unrepresentative, or collected under markedly different conditions, the estimated vector may underfit or encode mismatched biases, which can limit transfer to the target accents.

Pseudo-label assumption The method relies on an implicit stationarity assumption: systematic biases that appear in pseudo-labels for the source domain are assumed to recur in the target domain. When teacher errors are highly accent-specific or driven by channel and recording conditions that do not overlap with the source, the correction may be weak or even counterproductive. Relatedly, we observed that the scaling factor λ must be tuned, and excessive scaling can degrade WER. Although we tune λ only on held-out source development data, this still introduces a hyperparameter that may not transfer perfectly to new deployments.

Source composition. Our experiments rely on a fixed two-fold split, so robustness to alternative source-domain compositions (e.g., more folds, different mixes of accents or language families) is not directly characterized. This is partly bounded by data availability: most public ASR corpora (e.g., Common Voice) lack consistent or reliable accent annotations, making controlled cross-accent experiments difficult. AFRISPEECH-200 is one of the few datasets that provides curated accent labels with complete train/dev/test splits, which guided our selection. Investigating the sensitivity of the correction vector to source-fold composition, and extending the protocol to more source partitions, is an important direction for future work.

Language Our experiments focus on English accents within AFRISPEECH-200. Generalization to other languages and domains beyond reading or conversational speech is not validated here and remains future work.

Accent Representation Our experiments focus on English accents within AFRISPEECH-200. We filtered to accents that provide complete train, development, and test splits to ensure a fair protocol, but this choice may bias the evaluation toward better-represented accents and does not cover underrepresented varieties or code-switching scenarios. Generalization to other languages, domains beyond read or conversational speech, or far-field conditions is not validated here and remains future work.

10 Ethical considerations

This work focuses on improving the robustness of ASR through parameter-space correction of pseudo-labeling errors. The research is primarily methodological and does not involve the collection of new speech data or the deployment of real-world systems. Nevertheless, several potential risks and ethical considerations merit discussion.

Bias and fairness. ASR systems often exhibit disparities in accuracy across accents, dialects, and demographic groups (Jahan et al., 2025; Ngueajio and Washington, 2022; Fuckner et al., 2023). While our method aims to mitigate such disparities by improving adaptation to underrepresented accents, it may also amplify biases present in the teacher models or source-domain data. We encourage practitioners to evaluate model fairness carefully across linguistic and demographic subgroups when applying this technique, and to accompany adaptation with representative validation datasets.

Privacy and data use. Our experiments rely on publicly released corpora with consented speech recordings. No personally identifiable information or private data is used. However, adaptation methods in general could be misapplied to voice data collected without consent. Researchers and practitioners should ensure compliance with data protection regulations and obtain appropriate permissions before applying domain adaptation to sensitive speech.

Dual use and misuse. The proposed parameter-space correction could, in principle, be used to enhance ASR systems deployed in surveillance or monitoring settings. Our intention is to support low-resource and accessibility-oriented speech technologies, rather than enabling intrusive applications.

Acknowledgement

This work was supported by the Ministry of Education (MOE) of Taiwan under the Taiwan Centers of Excellence in Artificial Intelligence project, through the NTU Artificial Intelligence Center of Research Excellence (NTU AI-CoRE). We also thank the National Center for High-performance Computing (NCHC) of the National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources. We sincerely thank Dianna Yee, Colin Lea, and Ting-Yao Hu for their valuable feedback and insightful suggestions on this work.

References

Rehan Ahmad, Muhammad Umar Farooq, and Thomas Hain. 2024. Progressive unsupervised domain adaptation for asr using ensemble models and multi-stage training. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11466–11470.

Tom Bäckström. 2025. Privacy in speech technology. *Preprint*, arXiv:2305.05227.

Yu Chen, Wen Ding, and Junjie Lai. 2023. Improving noisy student training on non-target domain data for automatic speech recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Dimitrios Damianos, Georgios Paraskevopoulos, and Alexandros Potamianos. 2025. MSDA: Combining Pseudo-labeling and Self-Supervision for Unsupervised Domain Adaptation in ASR. In *Interspeech 2025*, pages 3863–3867.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, pages 3830–3834.

Robert Flynn and Anton Ragni. 2024. Self-Train Before You Transcribe. In *Interspeech 2024*, pages 2840–2844.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. 2020. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Marcio Fuckner, Sophie Horsman, Pascal Wiggers, and Iskaj Janssen. 2023. Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 146–151.

Calbert Graham and Nathan Roll. 2024. Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2):025206.

Yosuke Higuchi, Niko Moritz, Jonathan Le Roux, and Takaaki Hori. 2022. Momentum pseudo-labeling: Semi-supervised asr with continuously improving pseudo-labels. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1424–1438.

Brandon Huang, Chancharik Mitra, Leonid Karlinsky, Assaf Arbelle, Trevor Darrell, and Roei Herzig. 2024. Multimodal task vectors enable many-shot multimodal in-context learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

Maliha Jahan, Priyam Mazumdar, Thomas Thebaud, Mark Hasegawa-Johnson, Jesús Villalba, Najim Dehak, and Laureano Moro-Velazquez. 2025. Unveiling performance bias in asr systems: A study on gender, age, accent, and more. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Ji-Hun Kang, Jae-Hong Lee, Mun-Hak Lee, and Joon-Hyuk Chang. 2024. Whisper Multilingual Downstream Task Tuning Using Task Vectors. In *Interspeech 2024*, pages 2385–2389.

Eungbeom Kim and Kyogu Lee. 2025. Uncertainty-aware self-training for ctc-based automatic speech recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24330–24338.

Hongkang Li, Yihua Zhang, Shuai Zhang, Pin-Yu Chen, Sijia Liu, and Meng Wang. 2025. When is task vector provably effective for model editing? a generalization analysis of nonlinear transformers. In *The Thirteenth International Conference on Learning Representations*.

Tatiana Likhomanenko, Ronan Collobert, Navdeep Jaitly, and Samy Bengio. 2023. Continuous soft pseudo-labeling in asr. In *Proceedings on "I Can't Believe It's Not Better! - Understanding Deep Learning Through Empirical Falsification" at NeurIPS 2022 Workshops*, volume 187 of *Proceedings of Machine Learning Research*, pages 66–84. PMLR.

Tzu-Quan Lin, Wei-Ping Huang, Hao Tang, and Hung yi Lee. 2025a. Speech-ft: Merging pre-trained and fine-tuned speech representation models for cross-task generalization. *Preprint*, arXiv:2502.12672.

Yi-Cheng Lin, Huang-Cheng Chou, and Hung yi Lee. 2025b. Mitigating Subgroup Disparities in Multi-Label Speech Emotion Recognition: A Pseudo-Labeling and Unsupervised Learning Approach. In *Interspeech 2025*, pages 2053–2057.

Yi-Cheng Lin, Tzu-Quan Lin, Chih-Kai Yang, Ke-Han Lu, Wei-Chih Chen, Chun-Yi Kuan, and Hung-Yi Lee. 2024. Listen and speak fairly: a study on semantic gender bias in speech integrated large language models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 439–446.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yu-Xiang Luo, Yi-Cheng Lin, Ming-To Chuang, Jia-Hung Chen, I-Ning Tsai, Pei Xing Kiew, Yueh-Hsuan Huang, Chien-Feng Liu, Yu-Chen Chen, Bo-Han Feng, Wenze Ren, and Hung yi Lee. 2025. ToxicTone: A Mandarin Audio Dataset Annotated for Toxicity and Toxic Utterance Tonality. In *Interspeech 2025*, pages 4008–4012.

J MacQueen. 1967. Multivariate observations. In *Proceedings ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability*, volume 1, pages 281–297.

Long Mai and Julie Carson-Berndsen. 2022. Unsupervised domain adaptation for speech recognition with unsupervised error correction. In *Interspeech 2022*, pages 5120–5124.

Vimal Manohar, Tatiana Likhomanenko, Qiantong Xu, Wei-Ning Hsu, Ronan Collobert, Yatharth Saraf, Geoffrey Zweig, and Abdelrahman Mohamed. 2021. Kaizen: Continuously improving teacher using exponential moving average for semi-supervised speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 518–525.

Niko Moritz, Takaaki Hori, and Jonathan Le Roux. 2021. Semi-supervised speech recognition via graph-based temporal classification. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6548–6552.

Haruki Nagasawa, Shinta Otake, and Shinji Iwata. 2025. Task vector arithmetic for low-resource asr. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. 2020. What is being transferred in transfer learning? In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Mikel K. Ngueajio and Gloria Washington. 2022. Hey asr system! why aren't you more inclusive? In *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, pages 421–440, Cham. Springer Nature Switzerland.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. 2023. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Daniel S. Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V. Le. 2020. Improved noisy student training for automatic speech recognition. In *Interspeech 2020*, pages 2817–2821.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830.

Jeena Prakash, Blessingh Kumar, Kadri Hacioglu, Bidisha Sharma, Sindhuja Gopalan, Malolan Chetlur, Shankar Venkatesan, and Andreas Stolcke. 2025. Better pseudo-labeling with multi-asr fusion and error correction by speechllm. *Preprint*, arXiv:2506.11089.

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.

Gowtham Ramesh, Kartik Audhkhasi, and Bhuvana Ramabhadran. 2024. Task vector algebra for asr models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12256–12260.

Milind Rao, Gopinath Chennupati, Gautam Tiwari, Anit Kumar Sahu, Anirudh Raju, Ariya Rastrow, and Jasha Droppo. 2023. Federated self-learning with weak supervision for speech recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Fabian Ritter-Gutierrez, Yi-Cheng Lin, Jui-Chiang Wei, Jeremy H.M Wong, Eng Siong Chng, Nancy F. Chen, and Hung yi Lee. 2025a. Distilling a speech and music encoder with task arithmetic. In *Interspeech 2025*, pages 3858–3862.

Fabian Ritter-Gutierrez, Yi-Cheng Lin, Jeremy H. M Wong, Hung yi Lee, Eng Siong Chng, and Nancy F. Chen. 2025b. A correlation-permutation approach for speech-music encoders model merging. *Preprint*, arXiv:2506.11403.

Michael Shoemate, Kevin Jett, Ethan Cowan, Sean Colbath, James Honaker, and Prasanna Muthukumar. 2022. Sotto voce: Federated speech recognition with differential privacy guarantees. *arXiv preprint arXiv:2207.07816*.

Satwinder Singh, Feng Hou, and Ruili Wang. 2023. A novel self-training approach for low-resource speech recognition. In *Interspeech 2023*, pages 1588–1592.

Hsuan Su, Hua Farn, Fan-Yun Sun, Shang-Tse Chen, and Hung-yi Lee. 2024. Task arithmetic can mitigate synthetic-to-real gap in automatic speech recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8905–8915, Miami, Florida, USA. Association for Computational Linguistics.

Haibin Wu, Ho-Lam Chung, Yi-Cheng Lin, Yuan-Kuei Wu, Xuanjun Chen, Yu-Chi Pai, Hsiu-Hsuan Wang, Kai-Wei Chang, Alexander Liu, and Hung-yi Lee. 2024. Codec-SUPERB: An in-depth analysis of sound codec models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10330–10348, Bangkok, Thailand. Association for Computational Linguistics.

Chih-Kai Yang, Yu-Kuan Fu, Chen-An Li, Yi-Cheng Lin, Yu-Xiang Lin, Wei-Chih Chen, Ho Lam Chung, Chun-Yi Kuan, Wei-Ping Huang, Ke-Han Lu, Tzu-Quan Lin, Hsiu-Hsuan Wang, En-Pei Hu, Chan-Jan Hsu, Liang-Hsuan Tseng, I-Hsiang Chiu, Ulin Sanga, Xuanjun Chen, Po chun Hsu, and 2 others. 2024. Building a taiwanese mandarin spoken language model: A first attempt. *Preprint*, arXiv:2411.07111.

Jun Zhang, Jingyue Wu, Yiyi Qiu, Aiguo Song, Weifeng Li, Xin Li, and Yecheng Liu. 2023. Intelligent speech technologies for transcription, disease diagnosis, and medical equipment interactive control in smart hospitals: A review. *Computers in Biology and Medicine*, 153:106517.

Kevin Zhang, Luka Chkhetiani, Francis McCann Ramirez, Yash Khare, Andrea Vanzo, Michael Liang, Sergio Ramirez Martin, Gabriel Oexle, Ruben Bousbib, Taufiquzzaman Peyash, Michael Nguyen, Dillon Pulliam, and Domenic Donato. 2024. Conformer-1: Robust asr via large-scale semisupervised bootstrapping. *Preprint*, arXiv:2404.07341.

Qiuming Zhao, Guangzhi Sun, and Chao Zhang. 2025. Low-rank and sparse model merging for multilingual speech recognition and translation. *Preprint*, arXiv:2502.17380.

Han Zhu, Dongji Gao, Gaofeng Cheng, Daniel Povey, Pengyuan Zhang, and Yonghong Yan. 2023. Alternative pseudo-labeling for semi-supervised automatic speech recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:3320–3330.

A Model characteristic

Table 6 summarizes the key characteristics of the Whisper models used in our experiments. Parameter counts are approximate values reported by the official release.

Model	Parameters (M)	Encoder	Decoder
Tiny	39	4	4
Base	74	6	6
Small	244	12	12
Medium	769	24	24
Large v2	1550	32	32

Table 6: Whisper models used in our experiments. All models share the same encoder–decoder transformer architecture but differ in scale. Parameter counts are reported in millions. **Encoder** and **Decoder** are the number of transformer layers in the encoder and decoder, respectively.

B Data, Artifacts, and Licensing

Dataset statistic Table 7 summarizes the number of samples for each language across the train, development, and test splits.

Licenses and terms of use. All datasets and pre-trained models used in this work are publicly available for research purposes under open licenses. We use the AfriSpeech-200 corpus (Olatunji et al., 2023), which is distributed under a Creative Commons Attribution NonCommercial ShareAlike v4.0 (CC BY-NC-SA 4.0) license. The Whisper models (Radford et al., 2023) are released by OpenAI under the MIT license. Our use of these resources fully complies with their stated terms and intended use for non-commercial academic research. No additional data scraping or private data collection was conducted. All artifacts associated with this work, including source code, trained correction vectors, fine-tuned model checkpoints, and documentation, will be released under the CC BY-NC-SA 4.0 license after acceptance.

Intended use and compatibility. All artifacts are used within the scope of their original research purpose, which is speech recognition and domain adaptation studies. We do not deploy or fine-tune any model for commercial, surveillance, or identification applications. Derived models and results are intended solely for academic analysis and benchmarking. Any derived artifacts that we release will include clear documentation of intended use and

Split	Igbo	Swahili	Hausa	Zulu	Twi	Yoruba	Ijaw	Afrikaans	Idoma	Setswana
Train	8083	5480	5437	1306	1315	14369	2357	1911	1760	1273
Dev	216	313	116	310	186	361	48	82	50	215
Test	355	521	196	175	58	648	80	54	60	97

Table 7: Number of samples per split for each accent that we used for our experiment.

license terms to prevent misuse outside of research contexts.

Anonymization and privacy protection. The AFRISPEECH-200 dataset contains anonymized speech recordings collected with participant consent. No personally identifiable information (PII) or metadata that could reveal speaker identity is used or released. We performed a manual spot-check and confirmed that no audio files or transcripts contain sensitive, offensive, or private information. All models were trained and evaluated locally on anonymized data, with no connection to external user data or APIs.

C Use of AI assistants

This manuscript was refined with the assistance of large language models, which were used to improve clarity, grammar, and readability of the text. All conceptual development, experimental design, data analysis, and interpretation were conducted entirely by the authors. The AI assistants were not involved in generating research ideas or writing original scientific content.

D Language families of the ten accents

Although all ten accents used in our experiments originate from African regions, they belong to distinct and often unrelated linguistic families, each with different phonological and prosodic characteristics. Table 9 lists each accent together with its language family and subfamily/branch. These groups differ markedly in tonal systems, syllable structures, consonant inventories, and morphology. Therefore, while the accents are geographically African, they are not linguistically homogeneous, and adaptation across these families remains highly non-trivial.

E K-Means Implementation

For the speaker clustering procedure used in the Pseudo2Real-SC variant, we employ the standard K-means algorithm from the scikit-learn library (Pedregosa et al., 2011) with default configuration. The initialization method is set to `k-means++` to improve convergence speed and stability. The maximum number of iterations per run is fixed at 300, and the convergence tolerance is set to 10^{-4}. The standard Lloyd's algorithm is used as the clustering method.

F Baseline Implementation Details

Error Correction (EC). We fine-tune a T5-base model to perform post-hoc correction of ASR hypotheses. To preserve the unsupervised domain adaptation (UDA) setting, in which no target-domain ground-truth labels are available, all T5 supervision is drawn exclusively from the source domain. Concretely, we construct training pairs *(hypothesis, ground-truth)* by running the Whisper teacher on source-domain speech (the five source accents of each cross-fold split) and pairing the resulting hypotheses with the corresponding human transcripts. The T5 model is thus trained to map noisy, teacher-produced hypotheses back to their reference transcripts; to the extent that the teacher makes similar mistakes on source and target speech, the learned mapping transfers to cleaning target-domain pseudo-labels at inference. Input and target sequences are truncated or padded to 256 tokens. We use a per-device batch size of 4 with 4-step gradient accumulation (effective batch size 16), a learning rate of 3×10^{-4}, and 500 warmup steps. The model is trained for up to 20,000 optimization steps with evaluation every 1,000 steps using cross-entropy loss over target tokens, and the checkpoint with the lowest validation loss is selected.

At inference, the trained T5 model is applied to the teacher-generated pseudo-labels on target-domain speech, producing a corrected pseudo-label set. The Whisper student is then fine-tuned on the corrected target-domain data *(target speech, T5-corrected pseudo-label)*, following the same recipe as θ_t^{pseudo} but with the raw pseudo-labels replaced by their T5-corrected counterparts. EC therefore isolates the effect of text-level pseudo-label cleaning, and since T5 is trained only on source-domain supervision and applied to target-domain pseudo-

Example and Description
Example 1
Training set Ground Truth: Emmanuel Opuru said the suspects will face murder charge after investigation are complete.
Teacher Pseudo-label: the man on the opposite side of the suspect with face, mother, child after investigation are complete.
Error Type: Accent-induced lexical confusion—teacher model mishears "said" as "side" due to strong accent variation.
Testing set Ground Truth: It was a great pleasure, an audience member said later.
Pretrained Tiny: it was a great page and audience will see the later
Tiny (Student): it was a great pleasure and audience will be a side later
Tiny (Pseudo2Real): it was a great pleasure and audience member said later
Error Mitigation: Pseudo2Real effectively suppresses the accent-induced error by restoring the correct lexical item ("said"), aligning the transcription with the intended semantic meaning.
Example 2
Training set Ground Truth: In figure skating, sometimes women or men skate alone, or they skate in couples.
Teacher Pseudo-label: In figure skating, sometimes women or men skate alone or they skate in couples full stop
Error Type: Teacher hallucination—an extra "full stop" token is added.
Testing set Ground Truth: In Nigeria, too, the May Day celebrations also happen.
Pretrained Tiny: in nigeria 2 the media celebrations also happen 1st
Tiny (Student): in nigeria 2 the medial celebrations also happen full stop
Tiny (Pseudo2Real): in nigeria 2 the may day celebrations also happen
Error Mitigation: Pseudo2Real removes the inherited hallucinated "full stop" and restores correct lexical content ("May Day").

Table 8: Examples showing how Pseudo2Real corrects systematic pseudo-label errors, including accent-induced confusion (Example 1) and hallucinated tokens (Example 2). Red = error; green = correct token.

Language	Family	Subfamily / Branch
Igbo	Niger–Congo	Volta–Niger
Swahili	Niger–Congo	Bantu (Sabaki)
Hausa	Afro-Asiatic	Chadic
Zulu	Niger–Congo	Bantu (Nguni)
Twi	Niger–Congo	Kwa (Akan)
Yoruba	Niger–Congo	Volta–Niger
Ijaw	Niger–Congo	Ijoid
Afrikaans	Indo-European	Germanic
Idoma	Niger–Congo	Volta–Niger (Idomoid)
Setswana	Niger–Congo	Bantu (Sotho–Tswana)

Table 9: Language family and subfamily/branch for each of the ten accents used in our experiments. The ten accents span three major language families (Niger–Congo, Afro-Asiatic, Indo-European) and numerous subfamilies, highlighting that the evaluation is not linguistically homogeneous despite the shared geographic origin.

labels without access to any target ground truth, this baseline remains within the UDA setting.

Exponential Moving Average (EMA). EMA is applied on top of the θ_t^{pseudo} baseline, which fine-tunes the Whisper student on target-domain pseudo-labels. Let $\theta^{(k)}$ denote the student's parameters at training step k and $\hat{\theta}^{(k)}$ the EMA parameters, initialized as $\hat{\theta}^{(0)} = \theta^{(0)}$. After each optimizer update, the EMA weights are updated as

$$\hat{\theta}^{(k)} = \beta\,\hat{\theta}^{(k-1)} + (1-\beta)\,\theta^{(k)},$$

with decay factor $\beta = 0.999$, which yielded the best average validation performance among the values we tried. The EMA parameters are maintained separately from the training weights and do not participate in gradient computation. During evaluation and checkpoint saving, the training weights are temporarily swapped with the EMA weights and restored immediately afterward to continue optimization. All other training details are identical to the θ_t^{pseudo} baseline; in particular, no target-domain ground-truth labels are used, preserving the UDA setting.

G Additional case study

To further illustrate how Pseudo2Real mitigates systematic pseudo-label errors, we provide more qualitative examples in Table 8. These cases highlight two common types of errors in teacher-generated pseudo-labels and show how Pseudo2Real effectively addresses them.

Example 1 illustrates an accent-induced lexical confusion. Here, the teacher model mishears "said" as "side" due to strong accent variation in the training data, producing semantically inconsistent pseudo-labels. This error propagates to the student model, which reproduces the mistaken "side" in the testing example, leading to incorrect transcription. Pseudo2Real successfully corrects this accent-induced error by restoring the intended token "said," aligning the transcription with the ground truth.

Example 2 shows that the teacher model introduces an extraneous token ("full stop") which does not exist in the ground-truth transcription. Such hallucinated tokens often arise from overconfident language modeling behavior and can propagate into the student model trained on these pseudo-labels. When the same error type appears in the testing example, the student model reproduces this pattern, again appending a spurious "full stop" at the end.

By contrast, Pseudo2Real completely removes the hallucinated "full stop" pattern, demonstrating that the erroneous token sequence no longer appears in the output. Moreover, it restores the correct lexical content ("May Day") that aligns with the ground-truth transcription. This indicates that the parameter-space correction in Pseudo2Real not only suppresses inherited hallucinations but also reinforces meaningful acoustic-text alignment, leading to more faithful and semantically accurate transcriptions than direct pseudo-label fine-tuning.