

Adaptive Multi-Teacher Distillation for Accent-Robust Speech Recognition

Mehedi Hasan Bijoy¹ and Dejan Porjazovski² and Tamás Grósz³ and Mikko Kurimo¹

¹ Department of Information and Communications Engineering, Aalto University, Finland

² Aivot Labs, Finland

³ Walton Institute, South East Technological University, Ireland

firstname.lastname@{aalto.fi¹, aivot-ai.com², setu.ie³}

Abstract

Accented speech remains challenging for automatic speech recognition (ASR) systems due to phonetic and prosodic variation. Prior work addresses this through accent-conditioned fine-tuning, adaptation, and distillation. However, these approaches risk catastrophic forgetting and often generalize poorly to unseen accents. To address these limitations, we propose an adaptive multi-teacher knowledge distillation (MTKD) framework that unifies M heterogeneous accent-specialist teachers into a single student. We fine-tune a separate teacher model for each accent, then distill their combined knowledge into the student. A parameter-free mixer averages the teacher posteriors, while a learned adapter rebalances the transcription and distillation losses, removing manual loss tuning. Relative to conventional fine-tuning, MTKD reduces WER by $\approx 10.78\%$ on Accent20h and $\approx 10.21\%$ on EdAcc accented English ASR benchmarks, respectively. Moreover, MTKD outperforms CTC fine-tuning while improving cross-accent transfer, calibration, and zero-shot recognition across all 24 unseen accents.

1 Introduction

Over the last decade, ASR has emerged as an indispensable technology in human-computer interaction, powering hands-free conversational agents, accessibility tools, and voice-driven information retrieval. However, performance continues to degrade systematically when exposed to accented speech, thereby undermining inclusivity and user trust (Dong et al., 2024). Accents introduce systematic variation in both segmental and suprasegmental features (Strik and Cucchiaroni, 1999), consequently current adaptation strategies remain insufficiently robust across diverse speaker groups.

A broad spectrum of remedies including accent-specific modeling and mixture-of-experts (Ahmed et al., 2025; Lin et al., 2023), multi-task learning (Song et al., 2025; Zheng and Chen, 2024),

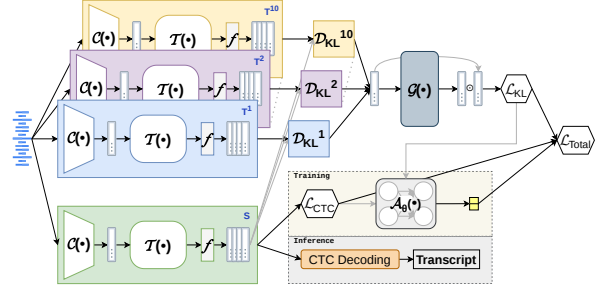


Figure 1: Overview of the proposed adaptive MTKD framework. An input waveform is processed by $M=10$ accent-specialist teachers $\{T_i\}$ and one student S , each comprising a temporal compression module $\mathcal{C}(\cdot)$, a context encoder $\mathcal{T}(\cdot)$, and a projection f . The per-teacher KL divergences $\mathcal{D}_{KL}^{(i)}$ are aggregated by the parameter-free mixer $\mathcal{G}(\cdot)$ into the distillation loss $\mathcal{L}_{KL}$. Simultaneously, the student is supervised by $\mathcal{L}_{CTC}$ on the ground-truth transcripts. The meta-learned adapter $\mathcal{A}_\theta(\cdot)$ maps both losses to bounded weights and yields $\mathcal{L}_{total}$.

embedding-based and domain-adversarial adaptation (Li et al., 2021a; Na and Park, 2021), pronunciation modification (Abbasi et al., 2025; Emara and Shaker, 2024), ensemble methods (Ghorbani and Hansen, 2024), and pretraining or self-supervised learning (SSL) followed by fine-tuning (Wang et al., 2025), has yielded notable gains but with persistent trade-offs. For instance, data augmentation via Text-to-Speech and accent conversion improves diversity yet may introduce artifacts and oversimplified prosody (Inoue et al., 2025; Do et al., 2024; Yang et al., 2025). Likewise, meta-learning, few-shot learning, and active learning accelerate adaptation but can be sensitive to sampling and computationally demanding (Eledath et al., 2024; Cheng and Nguyen, 2025; Broś et al., 2025). Moreover, empirical experiments expose demographic and geopolitical biases and reinforce the need for datasets that cover larger accent varieties and fair evaluation (Kim et al., 2024; Graham and Roll, 2024). In general, current prac-

tice remains fragmented: datasets omit socioethnic speech, fine-tuning induces catastrophic forgetting despite adapters or heuristics, and evaluation focuses narrowly on WER (Mehmood et al., 2025; Broś et al., 2025).

Knowledge distillation has also been explored in ASR to transfer complementary hypotheses from teachers to student (Chebotar and Waters, 2016; Fukuda et al., 2017; Gao et al., 2021; Fukuda et al., 2025). However, earlier efforts treat the teacher pool as a generic ensemble, rather than a structured bank of accent-specialist teachers that target systematic phonetic shifts across accent groups. In this work, we re-frame ensemble distillation specifically for accent robustness by constructing teachers specialized by accent partitions, distilling a single student without requiring accent labels at inference, and stabilizing the CTC–KD trade-off via a meta-learned objective adapter.

Against all backdrops, MTKD provides a scalable alternative by aggregating soft posteriors from accent-specialist teachers, mitigating single-teacher bias while preserving efficiency through a compact student model. We emphasize that MTKD is designed for the realistic setting where multiple accent-specialized ASR models exist or can be trained offline, but deployment requires a single model. The main goal of MTKD is to show that, unlike fine-tuning, it preserves richer accent-relevant knowledge. MTKD’s effectiveness underscores its integration of ideas from data augmentation, adversarial training, and meta-learning, which help mitigate negative transfer, improve calibration, and enhance zero-shot generalization beyond teacher models or adapter stacks (Inoue et al., 2025; Do et al., 2024). In addition, compressing multiple accent-specific models into a single student offers a practical route toward accent-robust ASR and related paralinguistic tasks such as speech emotion recognition (Bijoy et al., 2025).

The contributions of this work are as follows.

- We formalize accent robustness as accent agnostic inference through multi-teacher distillation, compressing $M(=10)$ accent specialist CTC-supervised teachers into a single deployable student without accent labels or routing.
- We introduce a dual adaptive MTKD framework that couples parameter-free teacher aggregation with a bounded meta learned objective adapter that stabilizes joint CTC and

distillation optimization and eliminates manual loss weighting.

- We provide a large scale evaluation on AESRC and EdAcc accented English ASR benchmarks covering 29 accents across seen and out of distribution unseen accent conditions including L1 and L2 speakers, and we demonstrate consistent gains in both in-domain and zero-shot generalization.
- We deliver detailed analyses of teacher heterogeneity, aggregation policies, and error decomposition, identifying the mechanisms by which MTKD improves cross accent transfer and reduces substitution errors.

Collectively, these results establish MTKD as a practical and scalable approach for inclusive accent-robust speech recognition across seen and unseen accents. Notably, these gains are realized by a single accent-agnostic student at inference, requiring neither accent labels nor explicit routing, which renders MTKD directly deployable where maintaining multiple accent-specialist models is impractical.

2 Method

We aim to develop a compact student model for accent-agnostic ASR directly from raw waveforms. Knowledge is therefore distilled from M accent-specialized teacher models $\{T_i(\cdot)\}_{i=1}^M$ to a unified student $S(\cdot)$, where each teacher is fine-tuned on a single accent partition ($M = 10$ for the AESRC dataset used in this work, corresponding to its 10 accent categories). All models share the same architecture, comprising a temporal compression module $\mathcal{C}(\cdot)$, a context-modeling encoder $\mathcal{T}(\cdot)$, and a linear projection layer. Furthermore, for an input $\mathbf{x} \in \mathbb{R}^L$, the student and teachers produce frame-level logits $\mathbf{S}$ and $\mathbf{Z}_i$, respectively, where frames are learned through adaptive down-sampling. Subsequently, a teacher mixer $\mathcal{G}(\cdot)$ assigns relevance weights α to teachers based on KL divergences between $\mathbf{S}$ and $\mathbf{Z}_i$, yielding a weighted KL loss $\mathcal{L}_{\text{KL}}$. In addition, supervision is integrated with a CTC loss $\mathcal{L}_{\text{CTC}}$, and a meta-learned adaptor $\mathcal{A}_\theta(\cdot)$ dynamically balances the two objectives. Hence, the student is supervised by a combined transcription and distillation objective whose two terms are balanced adaptively, such that $\mathcal{L}_{\text{total}} = \beta_{\text{CTC}}\mathcal{L}_{\text{CTC}} + \beta_{\text{KL}}\mathcal{L}_{\text{KL}}$, with $\beta_{\text{CTC}}, \beta_{\text{KL}} \in [0.1, 0.9]$. Finally, the student and

adapter are optimized jointly, while the parameter-free teacher mixer is applied deterministically, to achieve robust accent-generalized ASR.

2.1 Overview of the proposed method

As illustrated in Figure 1, the proposed framework integrates expert knowledge, frame-level guidance, and data-driven optimization within a unified training pipeline. First, ten accent-specialized teachers generate complementary frame-level hypotheses, which a teacher mixer consolidates through deterministic KL divergence-based weighting to emphasize reliable experts and suppress noisy ones. Meanwhile, a compact student processes the same waveform, with its outputs guided toward the mixer consensus via KL divergence loss and aligned with transcripts via CTC loss. Furthermore, a meta-learning adaptor balances these objectives by adjusting their relative weights on each mini-batch, avoiding manual tuning. Finally, joint optimization of student $\mathcal{S}(\cdot)$ and adaptor $\mathcal{A}_\theta(\cdot)$ ensures broad accent coverage and robust supervision.

2.2 Temporal compression and context modeling

The input audio is first mapped by the temporal compression module $\mathcal{C}(\cdot)$ from $\mathbf{x} \in \mathbb{R}^L$ to position-enriched frames $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1, \dots, \tilde{\mathbf{h}}_N] \in \mathbb{R}^{d \times N}$. A convolutional encoder performs learned downsampling, a linear projection with normalization sets the hidden size d , and a depth-wise positional convolution injects local temporal order. Then, the context module $\mathcal{T}(\cdot)$ applies a Transformer encoder to capture long-range dependencies, producing contextual embeddings $\mathbf{H}^{(C)} = \mathcal{T}(\tilde{\mathbf{H}}) \in \mathbb{R}^{d \times N}$. Finally, a shared linear layer projects $\mathbf{H}^{(C)}$ to vocabulary logits, yielding $\mathbf{S}$ for the student and $\{\mathbf{Z}_i\}_{i=1}^{10}$ for the teachers. Moreover, the CTC loss $\mathcal{L}_{\text{CTC}} = \text{CTC}(\text{softmax}(\mathbf{S}), \mathbf{y})$ is computed using $\mathbf{S}$ and the ground-truth transcript $\mathbf{y}$.

2.3 Teacher aggregation policies

The teacher mixer $\mathcal{G}(\cdot)$ forms a consensus supervision signal from accent-specialized teachers by measuring teacher–student agreement and then aggregating teachers accordingly. We use the frame-averaged KL divergence between teacher and student posteriors, a standard choice for matching soft targets in distillation (Li et al., 2021b). Specifically, for each teacher i , we compute:

$$\mathcal{D}_{\text{KL}}^{(i)} = \frac{1}{N} \sum_{n=1}^N \sum_{v=1}^{|\mathcal{V}|} \sigma(\mathbf{Z}_{n,v}^{(i)}) \log \left(\frac{\sigma(\mathbf{Z}_{n,v}^{(i)})}{\sigma(\mathbf{S}_{n,v})} \right) \quad (1)$$

Using the set $\{\mathcal{D}_{\text{KL}}^{(i)}\}_{i=1}^M$, we define three teacher mixing approaches that are evaluated in this work. First, in MTKD_Top1, we identify the most compatible teacher by selecting $i^* = \arg \min_i \mathcal{D}_{\text{KL}}^{(i)}$ and then distill exclusively from teacher i^* . Next, in MTKD_RankW, we rank all teachers in ascending order of $\mathcal{D}_{\text{KL}}^{(i)}$ and assign mixture weights according to $\alpha_i \propto 1/\text{rank}(i)$, ensuring $\sum_i \alpha_i = 1$. Finally, in MTKD_Avg, we assign equal weights to all teachers, such that $\alpha_i = 1/M$. We emphasize that the mixer $\mathcal{G}(\cdot)$ is parameter-free. In all three policies, the weights α are computed deterministically from the per-teacher divergences $\{\mathcal{D}_{\text{KL}}^{(i)}\}_{i=1}^M$ at each mini-batch.

2.4 Meta-learned objective adapter

Next, the adapter $\mathcal{A}_\theta(\cdot)$ takes the transcription and distillation losses $\mathbf{l} = [\mathcal{L}_{\text{CTC}}, \mathcal{L}_{\text{KL}}]^\top$ and outputs bounded weights: $[\beta_{\text{CTC}}, \beta_{\text{KL}}] = 0.8 \cdot \sigma(\mathbf{W}_\theta \mathbf{l} + \mathbf{b}_\theta) + 0.1$. The adapter is a single linear layer with parameters, which is trained jointly with the student by standard backpropagation through $\mathcal{L}_{\text{total}}$ and updated on every mini-batch, so there is no separate inner and outer optimization loop. In effect, it learns a data-driven mapping from the current magnitudes of $\mathcal{L}_{\text{CTC}}$ and $\mathcal{L}_{\text{KL}}$ to bounded mixing coefficients, normalizing the scale mismatch between the two losses. In turn, objectives are automatically rebalanced at each mini-batch, preventing loss domination and avoiding manual tuning. The bounds $[0.1, 0.9]$ were selected based on preliminary experiments. In particular, allowing the weights to reach 0 or 1 often caused one objective to dominate while the other vanished. Conversely, using wider ranges diluted the effect of dynamic adaptation. Therefore, constraining the coefficients within $[0.1, 0.9]$ consistently yielded stable training and reliable performance across accents.

2.5 Multi-objective optimization

The student is trained with a multi-objective loss that combines distillation and transcription supervision. The KL divergence encourages the student to mimic teacher distributions, while the CTC loss enforces alignment with ground-truth transcripts. These are merged into a single objective using adaptive weights from the meta-learned

Table 1: WER and CER across accents for 3 MTKD mixing policies and FT, along with test set statistics (utterance count, word count) and 95% confidence intervals for MTKD_Avg WER. Trained only on the American subset, MTKD_Avg yields lower errors, indicating stronger cross-accent knowledge transfer and improved generalization.

Accent	#utt	#words	MTKD_Avg			MTKD_RankW		MTKD_Top1		FT	
			WER	95% CI	CER	WER	CER	WER	CER	WER	CER
American	18311	169250	0.90	[0.83, 0.97]	0.34	1.39	0.48	0.92	0.36	1.24	0.43
British	19523	206205	11.55	[11.36, 11.74]	4.72	14.43	5.99	13.57	5.75	13.04	5.97
Canadian	18403	172158	6.39	[6.24, 6.55]	2.26	7.80	2.80	6.99	2.52	7.74	3.40
Chinese	15073	150546	23.18	[22.87, 23.50]	10.41	28.07	12.96	25.38	11.69	25.96	12.37
Indian	14664	164598	18.81	[18.55, 19.07]	7.69	19.66	8.19	21.41	9.19	20.81	9.25
Japanese	16785	152360	20.98	[20.70, 21.28]	9.66	25.35	11.96	22.80	10.82	23.58	11.57
Korean	17652	160817	14.59	[14.33, 14.85]	6.41	17.51	7.86	15.33	6.93	16.39	7.89
Portuguese	17959	165775	10.36	[10.17, 10.55]	3.85	12.37	4.67	11.30	4.32	12.03	5.18
Russian	16561	151513	14.74	[14.51, 14.98]	6.15	18.28	7.88	16.09	6.86	16.74	7.66
Spanish	17675	162661	13.67	[13.45, 13.89]	5.53	16.38	6.71	14.75	6.15	15.68	7.02
AESRC20H	18194	163787	15.31	[15.05, 15.58]	6.15	17.97	7.40	16.43	6.83	17.16	7.62

adapter: $\mathcal{L}_{\text{total}} = \beta_{\text{CTC}}\mathcal{L}_{\text{CTC}} + \beta_{\text{KL}}\mathcal{L}_{\text{KL}}$, where $\beta_{\text{CTC}}, \beta_{\text{KL}} \in [0.1, 0.9]$. As a result, the model benefits from both hard labels and soft targets, while dynamic weighting prevents loss domination and stabilizes training. Finally, $\mathcal{S}(\cdot)$ and $\mathcal{A}_{\theta}(\cdot)$ are optimized jointly, yielding an accent-robust and generalizable ASR system.

3 Experimental setup

Two datasets were used to validate MTKD: AESRC (read speech) with standard splits (Shi et al., 2021), and EdAcc (spontaneous) with development and test sets (Sanabria et al., 2023). The EdAcc test set serves as out-of-distribution, while its development set is combined with AESRC200h data to assess adaptation and generalization.

Accent200h (Shi et al., 2021). It contains ≈ 200 hours of labeled speech from 10 accents, with ≈ 20 hours per accent. Unlike the original setup, we use the full dataset for training, and also conduct per-accent training to analyze cross-accent generalization.

EdAcc_dev (Sanabria et al., 2023). It provides 9,848 utterances (≈ 15.8 h) across 23 accents plus a “Don’t know” class. This pool is leveraged for training augmentation and adaptation studies.

Benchmark Test Sets. Experiments are benchmarked on the following two publicly available standard test sets:

Accent20h (Shi et al., 2021): ≈ 20 h of speech (2h/accent) from the same 10 AESRC accents, without explicit accent labels.

EdAcc_test (Sanabria et al., 2023): ≈ 16.9 h spanning 23 accents plus a “Don’t know” category,

mirroring accent categories from EdAcc_dev set.

3.1 Models

We compare the empirical outcomes of the following four models, all sharing the same student architecture and training setup:

FT. Conventional supervised fine-tuning of the *wav2vec2-base-960h*¹ (Baevski et al., 2020) model using CTC loss.

MTKD_Top1. A MTKD variant where, for each mini-batch, the teacher with the lowest frame-level KL divergence is chosen, and the student is trained only on its soft targets.

MTKD_RankWeighted. Teachers are ranked by KL divergence, and reciprocal-rank weights α_{rank} yield $\mathcal{L}_{\text{KL}} = \sum_i \alpha_i \times \mathcal{D}_{\text{KL}}^{(i)}$, emphasizing the most relevant teacher while blending others without extra parameters.

MTKD_Avg. A MTKD variant where, for each frame, teacher posteriors are uniformly averaged and the student is trained to match this consensus via $\mathcal{L}_{\text{KL}} = \frac{1}{N} \sum_{i=1}^N \mathcal{D}_{\text{KL}}^i$.

On the input side, all audio was resampled to 16kHz and transcripts were normalized. The model was initialized from checkpoint¹ and fine-tuned for 20 epochs with batch size 16 and learning rate 10^{-4} . The architecture includes a 7-layer convolutional feature encoder, a 12-layer Transformer encoder with 12 attention heads and 3072 hidden units, and a linear projection to a 32-token vocabulary. For reproducibility, we open-source our codebase at [this GitHub repository](https://github.com/aalto-speech/MTKD4AccentedASR)².

¹facebook/wav2vec2-base-960h

²<https://github.com/aalto-speech/MTKD4AccentedASR>

4 Results

Four pretrained models were fine-tuned on American-accented subset of Accent200h and evaluated on Accent20h. Among them, *w2v2-base-960h*¹ yielded the best performance (WER 16.16, CER 6.62), followed by *w2v2-base-xlsr*³ (WER 17.83, CER 7.07), whereas *w2v2-base*⁴ (WER 20.01, CER 8.06) and *w2v2-base-timit*⁵ (WER 25.51, CER 10.8) lagged behind. Therefore, *w2v2-base-960h* is selected as the backbone for further experiments because it provides the strongest overall cross-accent performance, consistently outperforming the alternatives on non-American accents with a macro WER of 14.4%.

4.1 Empirical performance of FT and MTKD

The empirical outcomes confirm that performance depends on both the loss weighting scheme and teacher aggregation strategy. While fixed or unregularized β_{CTC} , β_{KL} lead to instability in teacher supervision, a lightly regularized mixture ensures stable optimization, lower losses, and superior WER. Building on this, Table 1 shows that *MTKD_Avg* consistently outperforms *MTKD_RankWeighted*, *MTKD_Top1*, and *FT* across all accents and benchmark test set. The advantage is most pronounced for high-variability accents such as Chinese and Japanese. This superiority arises from two factors: implicit variance reduction, which aligns gradients with CTC supervision, and balanced accent coverage, which prevents bias amplification. Thereby, uniform logit averaging with constrained loss weights emerges as the most reliable strategy for cross-accent generalization.

4.2 Zero-shot accent generalization: FT vs. MTKD

The comparison between FT and MTKD highlights the advantage of multi-teacher distillation. On benchmark sets, MTKD consistently reduces WER relative to FT, with large gains on EdAcc_test, which includes diverse and unseen accents. More importantly, the zero-shot results in Table 2 demonstrate that MTKD outperforms FT across all accents, lowering average WER from 58.27% to 52.33% and CER from 33.89% to 29.42%. The largest improvements occur for Italian, Kenyan, Israeli, and Ghanaian speech, though even near-

domain accents such as Southern British show modest gains. We attribute these outcomes mainly to ensemble-averaged teacher guidance, which mitigates accent-specific overfitting. We further hypothesize that the bounded loss schedule helps by keeping the CTC branch active enough to retain source-domain fidelity while the KL branch drives cross-accent transfer. Accordingly, MTKD emerges as more robust and transferable than FT in handling accent variability.

Table 2: Zero-shot results of the American baseline (FT) and multi-teacher distilled student (MTKD_Avg) in *EdAcc_test* set.

Accent	FT		MTKD_Avg	
	WER	CER	WER	CER
Italian	55.16	32.51	41.34 ^{13.8↓}	23.74 ^{8.8↓}
Kenyan	70.72	41.46	59.01 ^{11.7↓}	33.57 ^{7.9↓}
Israeli	50.73	27.19	40.11 ^{10.6↓}	21.56 ^{5.6↓}
Ghanain	58.34	35.32	47.85 ^{10.5↓}	24.58 ^{10.7↓}
Irish	51.57	29.01	41.22 ^{10.4↓}	21.98 ^{7.0↓}
European	47.68	25.81	39.66 ^{8.0↓}	19.07 ^{6.7↓}
Don't know	71.87	42.43	64.01 ^{7.9↓}	40.20 ^{2.2↓}
Romanian	69.53	44.93	62.60 ^{6.9↓}	37.53 ^{7.4↓}
French	40.29	21.74	34.64 ^{5.7↓}	17.34 ^{4.4↓}
Mainstream US	57.08	33.45	51.43 ^{5.7↓}	29.03 ^{4.4↓}
Latin	50.10	27.57	44.76 ^{5.3↓}	24.14 ^{3.4↓}
Spanish	57.98	33.67	52.67 ^{5.3↓}	29.73 ^{3.9↓}
Chinese	57.25	31.30	52.36 ^{4.9↓}	28.25 ^{3.1↓}
Scottish	59.93	34.76	55.29 ^{4.6↓}	32.44 ^{2.3↓}
Indian	51.61	29.26	47.31 ^{4.3↓}	25.37 ^{3.9↓}
Indonesian	53.68	32.27	49.56 ^{4.1↓}	27.92 ^{4.4↓}
Bulgarian	46.67	23.66	42.59 ^{4.1↓}	21.39 ^{2.3↓}
Lithuanian	56.41	32.23	52.57 ^{3.8↓}	29.16 ^{3.1↓}
Vietnamese	69.25	42.24	65.45 ^{3.8↓}	38.45 ^{3.8↓}
Nigerian	71.27	43.83	68.34 ^{2.9↓}	40.78 ^{3.1↓}
Jamaican	72.48	43.29	69.95 ^{2.5↓}	39.82 ^{3.5↓}
Catalan	66.76	41.95	64.24 ^{2.5↓}	38.76 ^{3.2↓}
Eastern European	45.94	23.55	43.84 ^{2.1↓}	21.98 ^{1.6↓}
Southern British	66.22	40.02	65.09 ^{1.1↓}	39.32 ^{0.7↓}

5 Ablations and analysis

We conduct a series of ablations to better understand the contributions of individual design choices in the MTKD, with quantitative results presented in Table 1. First, teacher aggregation strongly shapes outcomes. Uniform averaging consistently yields the lowest WER and CER, whereas rank weighting amplifies dominant teachers and harms minority accents, and top-1 discards useful diversity and destabilizes training. Hence, uniform blending proves most robust. Next, adaptive weighting is crucial. Bounding β_{CTC} , $\beta_{KL} \in [0.1, 0.9]$ stabilizes training, reduces gradient noise, and ensures monotonic loss drops. In contrast, static or unconstrained schedules that collapse the KL branch ($\beta_{KL} = 0$) yield a WER of 24.08, compared to 22.78 with

³facebook/wav2vec2-xls-r-300m

⁴facebook/wav2vec2-base

⁵elgeish/wav2vec2-base-timit-asr

bounded weights ($\beta_{KL} = 0.1$), corresponding to a 5.7% relative degradation on the Indian accent when the model is trained on the American accent.

Beyond optimization, a detailed error decomposition revealed how MTKD modifies the error profile compared to fine-tuning. Specifically, MTKD reduced substitutions by 14,051 instances (5.7%), deletions by 369 instances (1.6%), and insertions by 623 instances (2.5%). Thus, while all three error types declined, the most substantial improvements came from substitutions, indicating that MTKD primarily strengthens phoneme discrimination. This effect can be attributed to averaging teacher posteriors, which smooths accent-specific peaks and provides accent-invariant acoustic evidence on ambiguous segments. Deletions and insertions also decline, though modestly, pointing to residual alignment challenges rather than phonetic confusion. In CTC, over-confident blank probabilities tend to increase deletions, whereas poorly localized non-blank spikes promote insertions. By pooling teacher targets, MTKD slightly redistributes mass away from over-confident blanks and sharp accent-idiosyncratic spikes. We therefore observe consistent decreases in deletions and insertions. This partially supports the hypothesis that averaging targets reduces blank probability.

Moreover, alignment analysis shows that teachers differ in boundary placement despite identical transcripts. MTKD preserves this temporal diversity while steering the student toward a consensus, improving performance on ambiguous segments (substitutions) and providing secondary benefits (deletions and insertions), although the lack of explicit temporal warping may limit frame-level transfer. Therefore, local alignment mechanisms remain a promising extension. Finally, zero-shot tests confirm generalization. MTKD outperforms FT across all 24 unseen accents, with largest gains for Italian, Kenyan, and Israeli English. Nevertheless, performance drops when teacher diversity shrinks, and top-1 selection risks accent idiosyncrasies. Overall, uniform averaging with bounded loss weights emerges as the most reliable recipe for distillation.

6 Conclusion

In this study, we addressed the persistent challenge of accent variability in ASR and proposed a dual-adaptive MTKD framework that integrates accent-specialist teachers through a mixer and a meta-learned loss adapter. Through systematic

evaluation on AESRC and EdAcc benchmarks, the method consistently outperformed fine-tuning, yielding about 10.78% relative WER reduction on Accent20h and over 10.21% on EdAcc, while also mitigating catastrophic forgetting and generalizing well to unseen accents. Moreover, ablations demonstrated that uniform logit averaging and bounded dynamic loss weighting yield stable and bias-resistant learning, while zero-shot experiments confirmed robust generalization to out-of-distribution accents. Nonetheless, the approach still expects labeled data for teacher models and leaves gaps for prosodically extreme varieties, suggesting future directions such as self-supervised accent discovery and efficient teacher mixer. Taken together, these findings position MTKD as a viable approach toward inclusive and resource-efficient speech recognition systems.

Acknowledgements

The computational resources were provided by Aalto ScienceIT. This work was funded by the Business Finland project LAREINA under Grant 7817/31/2022 and the SaameAI project by the Finnish Cultural Foundation.

References

- Abdul Malik Abbasi, Imtiaz Husain, Illahi Bakhsh, Neda Kameh Khosh, and Ahlam Khan. 2025. Investigating voice onset time in pakistani english speech. *Speech Communication*, 168:103174.
- Ghayas Ahmed, Aadil Ahmad Lawaye, Vishal Jain, Jyotir Moy Chatterjee, and Shubham Mahajan. 2025. Enhancing english accent identification in automatic speech recognition using spectral features and hybrid cnn-bilstm model. *Multimedia Tools and Applications*, pages 1–28.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.
- Mehedi Hasan Bijoy, Dejan Porjazovski, Tamás Grósz, and Mikko Kurimo. 2025. [Multi-teacher language-aware knowledge distillation for multilingual speech emotion recognition](#). In *Proc. Interspeech 2025*, pages 146–150.
- Karolina Broś, Katarzyna Czarnocka-Gołębiewska, Janina Mołczanow, and Małgorzata Szupica-Pyrzanowska. 2025. The role of expertise in coping with accents during simultaneous interpreting: A pupillometric study. *Interpreting*, 27(1):52–86.

- Yevgen Chebotar and Austin Waters. 2016. Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech*, pages 3439–3443.
- Jian Cheng and Sam Nguyen. 2025. Speech few-shot learning for language learners’ speech recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Cong-Thanh Do, Shuhei Imai, Rama Doddipatla, and Thomas Hain. 2024. Improving accented speech recognition using data augmentation based on unsupervised text-to-speech synthesis. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 136–140. IEEE.
- Renchang Dong, Yijie Li, Dongxing Xu, and Yanhua Long. 2024. Cross-modal parallel training for improving end-to-end accented speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10396–10400. IEEE.
- Dhanva Eledath, Arun Baby, and Shatrughan Singh. 2024. Robust speech recognition using meta-learning for low-resource accents. In *2024 National Conference on Communications (NCC)*, pages 1–6. IEEE.
- Ingy Farouk Emara and Nabil Hamdy Shaker. 2024. The impact of non-native english speakers’ phonological and prosodic features on automatic speech recognition accuracy. *Speech Communication*, 157:103038.
- Takashi Fukuda, Gakuto Kurata, and George Saon. 2025. Knowledge distillation based training of unified conformer ctc models for multi-form asr. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. 2017. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, pages 3697–3701.
- Yan Gao, Titouan Parcollet, and Nicholas D Lane. 2021. Distilling knowledge from ensembles of acoustic models for joint ctc-attention end-to-end speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 138–145. IEEE.
- Shahram Ghorbani and John HL Hansen. 2024. Advanced accent/dialect identification and accentedness assessment with multi-embedding models and automatic speech recognition. *The Journal of the Acoustical Society of America*, 155(6):3848–3860.
- Calbert Graham and Nathan Roll. 2024. Evaluating openai’s whisper asr: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2).
- Sho Inoue, Shuai Wang, Wanxing Wang, Pengcheng Zhu, Mengxiao Bi, and Haizhou Li. 2025. Macst: Multi-accent speech synthesis via text transliteration for accent conversion. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Seoyoung Kim, Yeon Su Park, Dakyeom Ahn, Jin Myung Kwak, and Juho Kim. 2024. Is the same performance really the same?: Understanding how listeners perceive asr results differently according to the speaker’s accent. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–22.
- Jialu Li, Vimal Manohar, Pooja Chitkara, Andros Tjandra, Michael Picheny, Frank Zhang, Xiaohui Zhang, and Yatharth Saraf. 2021a. Accent-robust automatic speech recognition using supervised and unsupervised wav2vec embeddings. *arXiv preprint arXiv:2110.03520*.
- Zerui Li, Yue Ming, Lei Yang, and Jing-Hao Xue. 2021b. Mutual-learning sequence-level knowledge distillation for automatic speech recognition. *Neurocomputing*, 428:259–267.
- Yuqin Lin, Shiliang Zhang, Zhifu Gao, Longbiao Wang, Yanbing Yang, and Jianwu Dang. 2023. Wav2vec-moe: An unsupervised pre-training and adaptation method for multi-accent asr. *Electronics Letters*, 59(11):e12823.
- Haaris Mehmood, Karthikeyan Saravanan, Pablo Peso Parada, David Tuckey, Mete Ozay, Gil Ho Lee, Jungin Lee, and Seokyeong Jung. 2025. Valsub: Sub-sampling validation data to mitigate forgetting during asr personalization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Hyeong-Ju Na and Jeong-Sik Park. 2021. Accented speech recognition based on end-to-end domain adversarial training of neural networks. *Applied Sciences*, 11(18):8412.
- Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. The edinburgh international accents of english corpus: Towards the democratization of english asr. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Xian Shi, Fan Yu, Yizhou Lu, Yuhao Liang, Qiangze Feng, Daliang Wang, Yanmin Qian, and Lei Xie. 2021. The accented english speech recognition challenge 2020: open datasets, tracks, baselines, results and methods. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6918–6922. IEEE.
- Tianyu Song, Linh Thi Hoai Nguyen, and Ton Viet Ta. 2025. Mpsa-densenet: A novel deep learning model for english accent classification. *Computer Speech & Language*, 89:101676.

- Helmer Strik and Catia Cucchiaroni. 1999. Modeling pronunciation variation for asr: A survey of the literature. *Speech Communication*, 29(2-4):225–246.
- Jun-You Wang, Sheng Li, Li-An Lu, Sydney Chia-Chun Kao, and Jyh-Shing Roger Jang. 2025. Similarity-based accent recognition with continuous and discrete self-supervised speech representations. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Guanrou Yang, Fan Yu, Ziyang Ma, Zhihao Du, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2025. Enhancing low-resource asr through versatile tts: Bridging the data gap. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Hao-Tian Zheng and Berlin Chen. 2024. Improving speech recognition by enhancing accent discrimination. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.