Aalto University
School of Electrical
Engineering

Multi-Teacher Knowledge Distillation for Accented English Speech Recognition

Mehedi Hasan Bijoy

Advised by Dr. Tamás Grósz and Dr. Dejan Porjazovski

Supervised by Professor Mikko Kurimo

Degree Programme : Computer, Communication and Information Sciences
Major : Speech and Language Technology
Minor : Machine Learning, Data Science and Artificial Intelligence

Outline

- Introduction
- Research Gaps & Questions
- Proposed MTKD Method
- Results
- Error Analysis
- Beyond Accented ASR: MTKD for SER
- Findings & Insights
- Challenges, Open Questions & Future Directions
- Conclusion

MTKD = Multi-Teacher Knowledge Distillation
SER = Speech Emotion Recognition

Why Do Accents Matter in Speech Recognition?

- English is spoken globally, often as a second language (L2) [1].
  - ~1.14 billion L2 English speakers vs. ~390 million native speakers.
  - Accents affect pronunciation, rhythm, and prosody.

- Automatic Speech Recognition (ASR) systems fail on non-canonical accents.
  - Most ASR models are built on “mainstream” accents → limited exposure
  - Poor accent-invariant feature learning → weak robustness to OOD speech.
  - Catastrophic forgetting during fine-tuning → limiting generalization to new speech.

[1] “English-Speaking World.” Wikipedia. Accessed June 17, 2025. https://en.wikipedia.org/wiki/English-speaking_world.

Limitations & Research Gap

Limitation	Accent-Specific FT	Adversarial / Domain Adaptation	Augmentation	Ensemble Techniques	Meta / Few - Shot / MTL	MTKD (Proposed)
High computational & storage overhead	✅	✅	✅	✅	❌	❌
Poor zero-shot generalization	✅	✅	✅	✅	❌	❌
Catastrophic forgetting in sequential adaptation	✅	❌	➖	❌	✅	❌
Static supervision	❌	❌	➖	✅	➖	❌
Rigid or manually tuned loss balancing	❌	❌	➖	✅	✅	❌
No unified model for seen & unseen accents	✅	✅	✅	✅	✅	❌

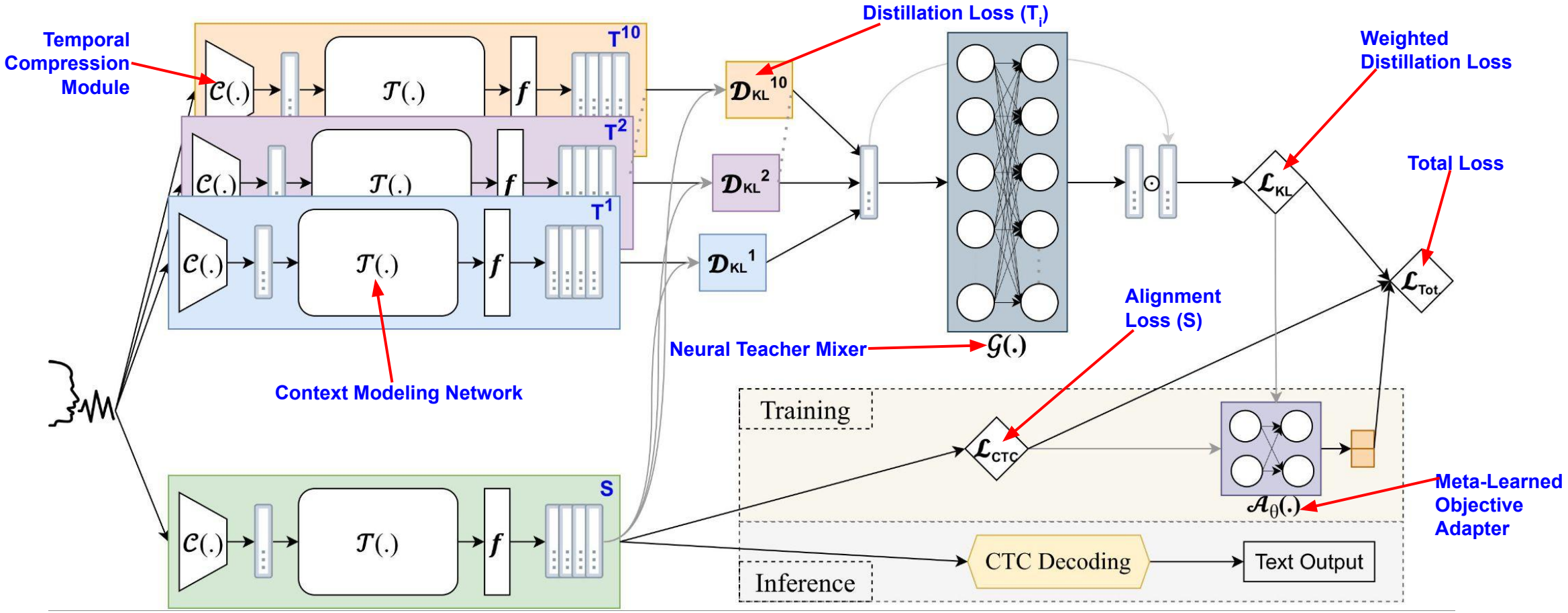
✅ = Limitation present
❌ = Limitation addressed / not present
➖ = Not applicable to this approach

Research Questions

- **RQ1**:
  Does MTKD facilitate the learning of more robust and accent-invariant acoustic representations than accent aware FT?
- **RQ2**:
  How does teacher ensemble heterogeneity influence the robustness of multi-teacher distilled student in zero-shot scenarios?
- **RQ3**:
  To what extent does MTKD mitigate catastrophic forgetting compared to FT when progressively adding new accents?

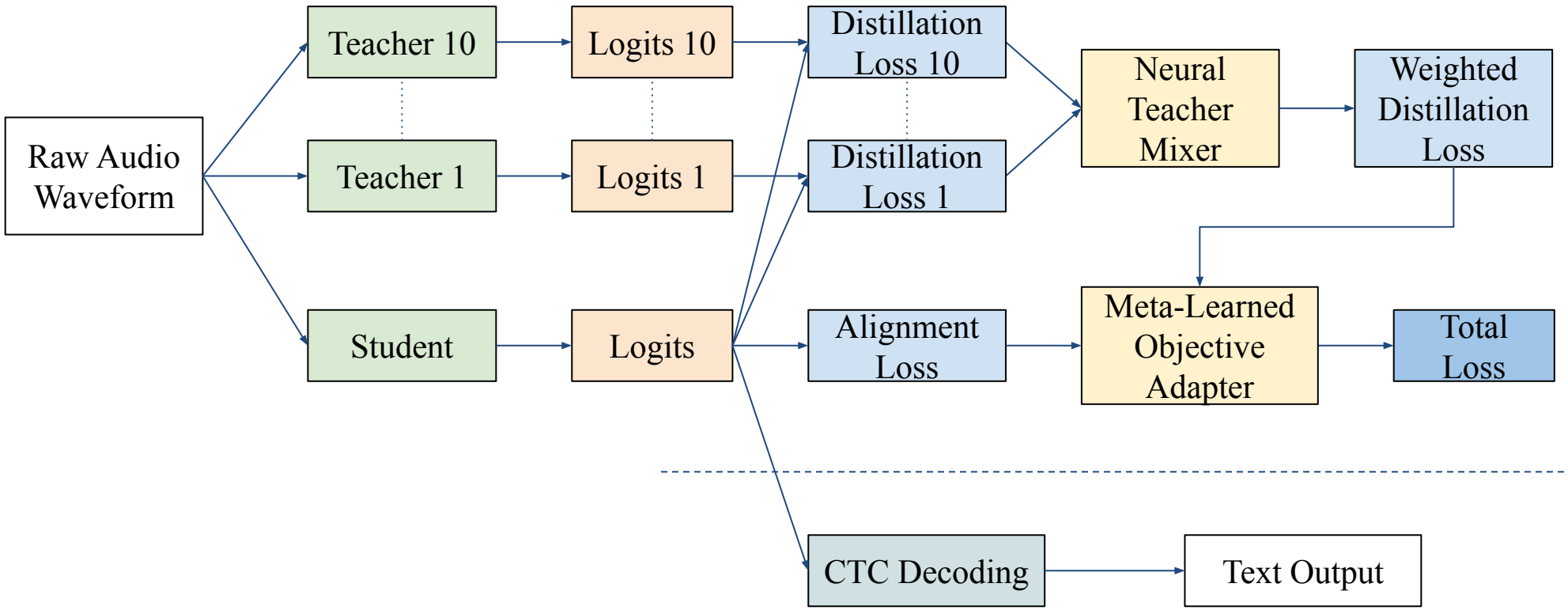
MTKD = Multi-Teacher Knowledge Distillation
FT = Fine-Tuning

Proposed Dual-Adaptive MTKD Framework



Temporal Compression Module : $\mathscr{C}(.)$ → Context Modeling Network : $\mathscr{T}(.)$
Neural Teacher Mixer : $\mathscr{G}(.)$
Meta-Learned Objective Adapter : $A\theta(\cdot)$
CTC Decoding
Multi-Objective Optimization

How Does MTKD Works?

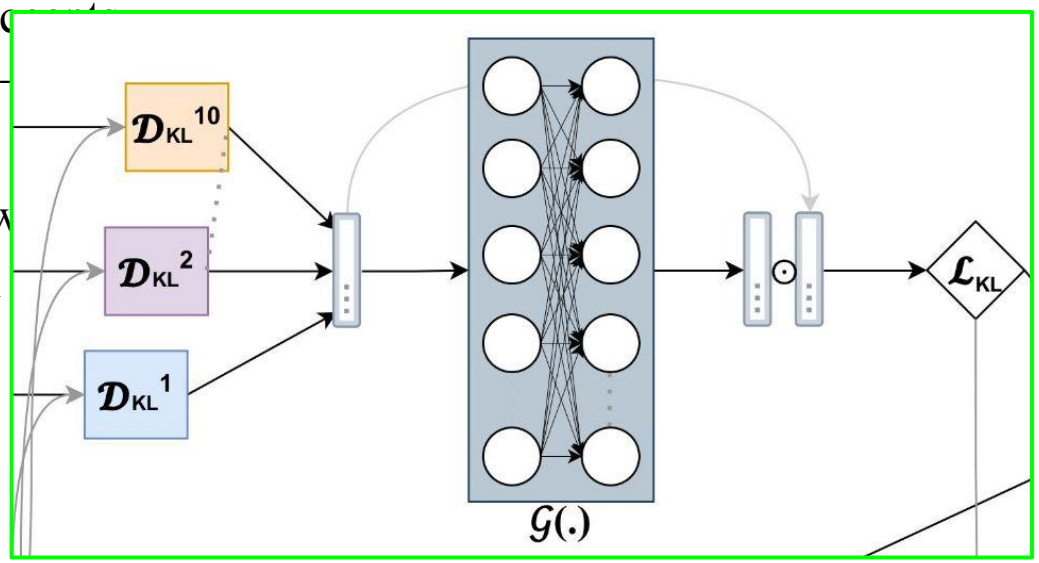


Training Configuration

- Audio resampled to 16 kHz
- Vocab size: 32
- Temporal Compression: 7-layer convolutional encoder
- Linear Projection: 512 → 768
- Context Modeling: 12 transformer encoder layers
  - 12 attention heads
  - Feed-forward dimension of 3072
- Optimizer: AdamW
- Trained for 20 epochs with batch size of 16 and learning rate of 1e-4
- Learning Rate Scheduler: Linear

Evaluation Setup

- Datasets:
  - Accent200h [1]: ~200 hours from 10 seen English ac
  - EdAcc_Dev [2]: ~16 hours from 23 unseen accents
  - Standard Test Sets for Benchmarking:
    - Accent20h [1]: ~20 hours from 10 seen accents (w
    - EdAcc_Test [2]: ~17 hours from the same unseen



- Metrics:
  - Word Error Rate (WER)
  - Character Error Rate (CER)
- Baselines:
  - FT: Standard accent-specific fine-tuning
  - MTKD_Top1 → *Only the most aligned teacher selected by the Neural Teacher Mixer is used.*
  - MTKD_RankWeighted → *Teachers are ranked by their distillation losses and weighted accordingly.*

[1] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, Y. Qian, and L. Xie, "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). IEEE, 2021, pp. 6918–6922.

[2] R. Sanabria, N. Bogoychev, N. Markl, A. Carmantini, O. Klejch, and P. Bell, "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). IEEE, 2023, pp. 1–5.

Pretrained Model Selection for Cross-Accent ASR

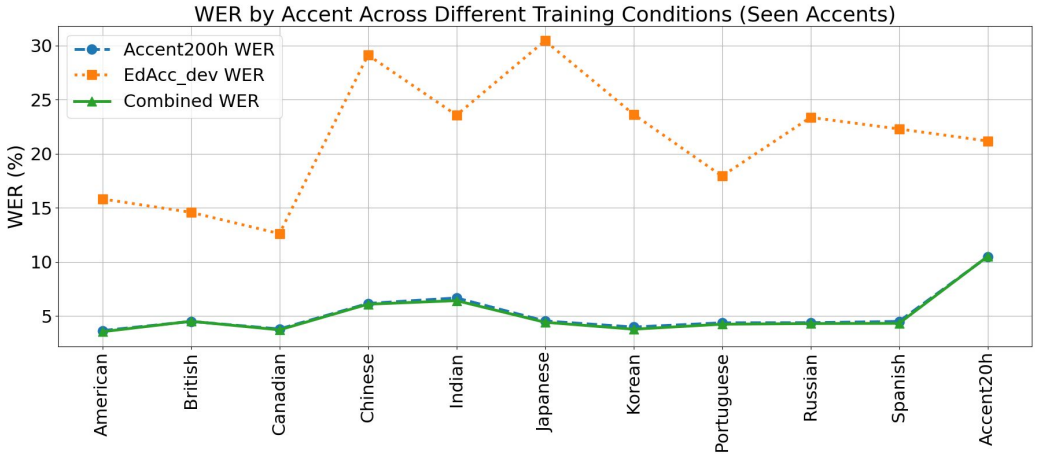
Accent	w2v2-base		w2v2-base-timit		w2v2-base-xlsr		w2v2-base-960h	
	WER	CER	WER	CER	WER	CER	WER	CER
American	**0.49**	**0.19**	3.09	0.87	1.45	0.46	1.24	0.42
British	17.49	7.0	23.11	9.5	17.52	7.14	**12.04**	**4.97**
Canadian	9.72	3.34	13.35	4.69	8.61	2.96	**6.74**	**2.41**
Chinese	29.99	13.31	35.8	16.43	28.45	12.64	**24.96**	**11.37**
Indian	26.12	10.97	36.923	17.12	20.78	8.31	**19.81**	**8.26**
Japanese	26.33	12.05	32.43	15.28	24.79	11.34	**22.58**	**10.57**
Korean	18.39	7.95	23.08	10.15	17.03	07.34	**15.39**	**6.89**
Portuguese	14.29	5.29	18.73	7.07	12.44	4.54	**11.03**	**4.18**
Russian	19.23	7.94	24.98	10.71	17.28	7.09	**15.74**	**6.66**
Spanish	18.02	7.25	23.04	9.57	16.19	6.39	**14.68**	**6.02**
Accent20h	20.01	8.06	25.51	10.8	17.83	7.07	**16.16**	**6.62**

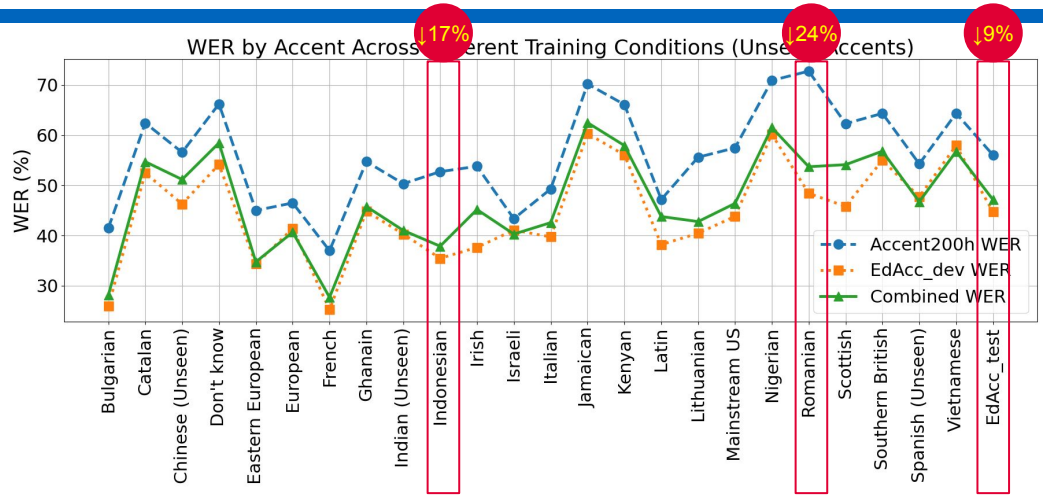
Table 1: WER and CER across ten English accents and the Accent20h test set after fine-tuning four pretrained Wav2Vec2 models on American-accented speech.

- **Task**: FT on American subset → evaluate on 10 seen accents + Accent20h set
- Models Compared: 4
- w2v2-base-960h:
  - Best performer!
  - Avg WER: ~14.4% (12% relative improvement compared to next best)
  - Consistently lower CER reflects subword modeling, not memorization.
  - Balanced trade-off between specialization and generalization.
  - Behaves like a general linguist → neutralizing accent variation while retaining performance.

facebook/wav2vec2-base
elgeish/wav2vec2-base-timit-asr
facebook/wav2vec2-large-xlsr-53
facebook/wav2vec2-base-960h

Balancing Generalization and Retention in Fine-Tuning





Training Scenarios:

- **Accent200h** → performs well on seen accents but fails to generalize
- **EdAcc_dev** → adapts well but forgets previously seen accents (catastrophic forgetting)
- **Combined** → optimal trade-off: strong seen performance while adapting well to OOD accents

EdAcc_test WER: 47.13% (Combined) from 56.03% (Accent200h)

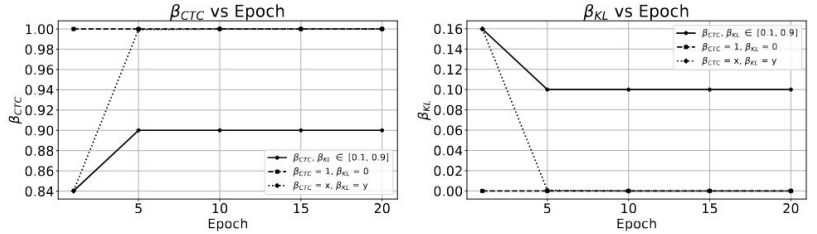
Accent-Specific Insights:

- Romanian (↓24%)
- Indonesian (↓17%)
- Same label ≠ same phonetics (e.g. Chinese)

OOD = Out-of-Distribution

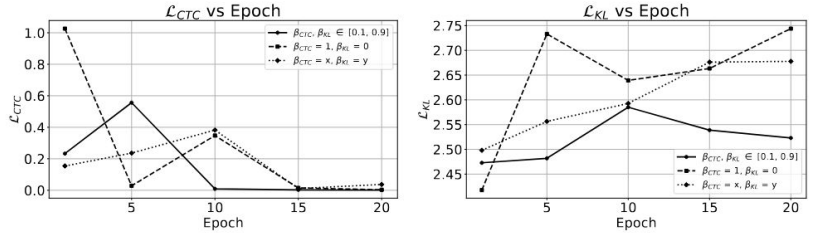
Meta-Learned Weighting Strategies in MTKD

- Four β-weighting configurations were tested:
  - **Regularized**: Meta-learned and softly constrained.
  - **CTC-only**: Ignores teacher guidance.
  - **Unregularized Dynamic**: Adaptive but unconstrained.
  - **KL-only**: Training failed!
- β_{CTC} preserves acoustic supervision *(f)*
  - CTC-only setup hinders generalization.
- β_{KL} collapse to 0 in unregularized setups → leads to weak teacher signal. *(b)*
  - KL-only fails completely.
- Regularized setup ensures balanced reduction in L_{CTC}, L_{KL}, and L_{Total}. *(c-e)*



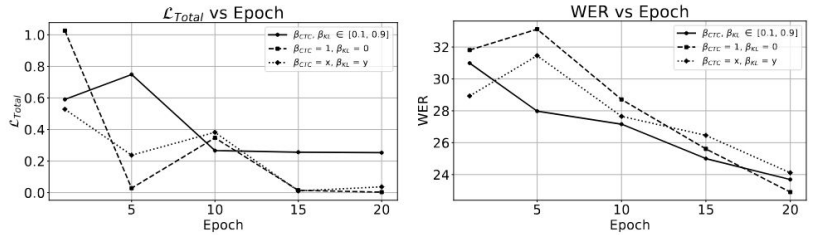
(a) Variations in β_{CTC} over epochs, showing stability in the regularized case.

(b) Tracked evolution of β_{KL} across training epochs in the three examined setups.



(c) Progression of $\mathcal{L}_{CTC}$ over epochs under different weighting schemes.

(d) Comparison of $\mathcal{L}_{KL}$ dynamics across training setups with varying β_{KL} strategies.



(e) Evolution of the combined loss $\mathcal{L}_{Total}$ over training epochs for different β configurations.

(f) WER trends across epochs for each loss weighting configuration.

Regularized: β_{CTC}, $\beta_{KL} \in [0.1, 0.9]$
CTC-only: $\beta_{CTC}=1$, $\beta_{KL}=0$
Unregularized Dynamic: $\beta_{CTC}=x$, $\beta_{KL}=y$
KL-only: $\beta_{CTC}=0$, $\beta_{KL}=1$

MTKD Strategy Comparison Across Accents

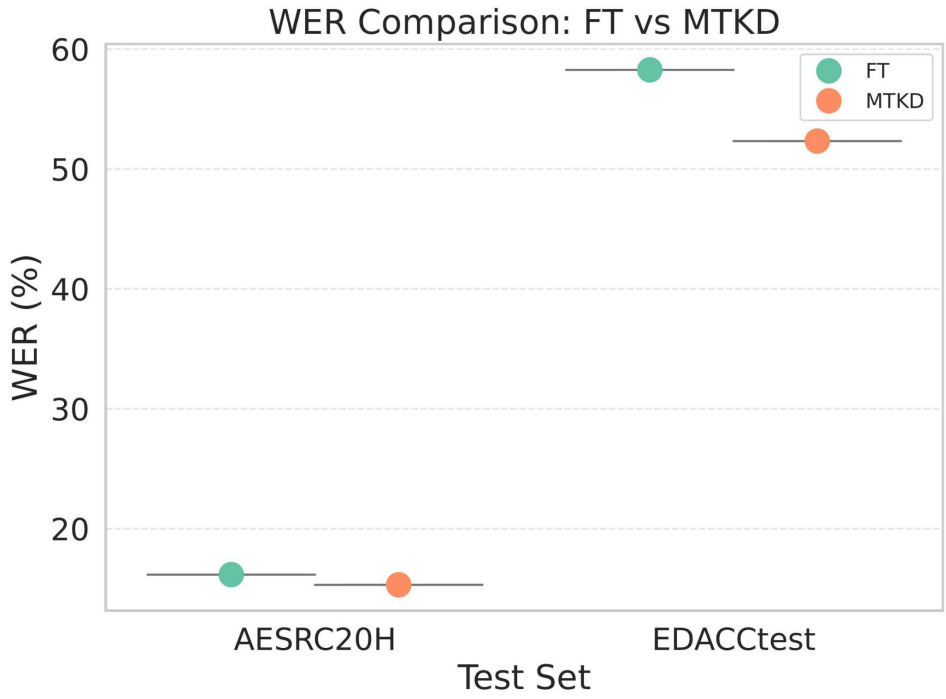
- 3 teacher aggregation strategies are benchmarked:
  - MTKD_Avg: Uniform logit averaging
  - MTKD_RankWeighted: Rank-based weighting
  - MTKD_Top1: Selects the top-aligned teacher
- MTKD_Avg consistently outperforms others
- Reduces systematic biases via ensemble smoothing
- Ensures balanced accent coverage, preserving minority-accent cues
- Works well with the regularized β-weight schedule
  - MTKD_Avg + Regularized β
    - Most robust setup for cross-accent generalization
    - Generalizes better under phonetic variability and accent imbalance

Accent	MTKD_Avg		MTKD_RankWeighted		MTKD_Top1		
	WER	CER	WER	CER	WER	CER	
American	**0.90**	**0.34**	1.39	0.48	0.92	0.36	
British	**11.55**	**4.72**	14.43	5.99	13.57	5.75	
Canadian	**6.39**	**2.26**	7.80	2.80	6.99	2.52	
Chinese	**23.18**	**10.41**	28.07	12.96	25.38	11.69	↓4.9
Indian	**18.81**	**7.69**	19.66	8.19	21.41	9.19	
Japanese	**20.98**	**9.66**	25.35	11.96	22.80	10.82	↓4.4
Korean	**14.59**	**6.41**	17.51	7.86	15.33	6.93	
Portuguese	**10.36**	**3.85**	12.37	4.67	11.30	4.32	
Russian	**14.74**	**6.15**	18.28	7.88	16.09	6.86	↓3.5
Spanish	**13.67**	**5.53**	16.38	6.71	14.75	6.15	
AESRC20H	**15.31**	**6.15**	17.97	7.40	16.43	6.83	

Table 3: WER and CER comparison across three MTKD strategies for different accents.

Standard Test Set Verdict: FT vs MTKD

- WER Reduction with MTKD:
  - Accent20h: from 16.16% → 15.31% (**RI: ~5.26%**)
  - EdAcc_test: from 58.27% → 52.33% (**RI: ~10.19%**)
- MTKD offers better generalization across accent variations.
- MTKD reduces accent overfitting by exposing the student to diverse expert views.
- FT suffers from memorization, MTKD encourages phonetic interpolation.



RI = Relative Improvement

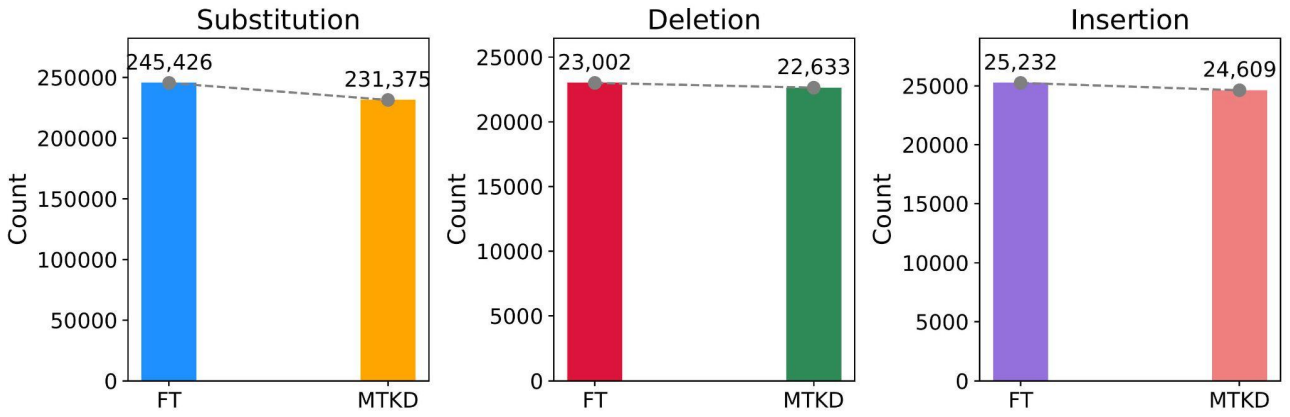
Zero-Shot Adaptation: FT vs MTKD

- Setup: All models trained on American English; evaluated on 24 diverse accents in EdAcc test set. → Tests generalization.
- MTKD advantages:
  - Avg. WER: 58.27% → **52.33%** (≈*10.18% relative improvement*)
  - Avg. CER: 33.89% → **29.42%** (≈*13.18% relative improvement*)
  - **Every accent benefits from MTKD**
- Top Gains: Italian, Kenyan, Israeli, Ghanaian (↓10–14% WER)
- Why MTKD works?
  - Multi-teacher setup produces avg logits → acts like a soft consensus — makes the learning more resilient to outliers.
  - Balanced optimization through meta-learned loss adaptation → helps the model retain generaliable phonetic patterns → avoids overfitting to dominant ones.

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↓

Understanding Errors: Not All Mistakes are Equal

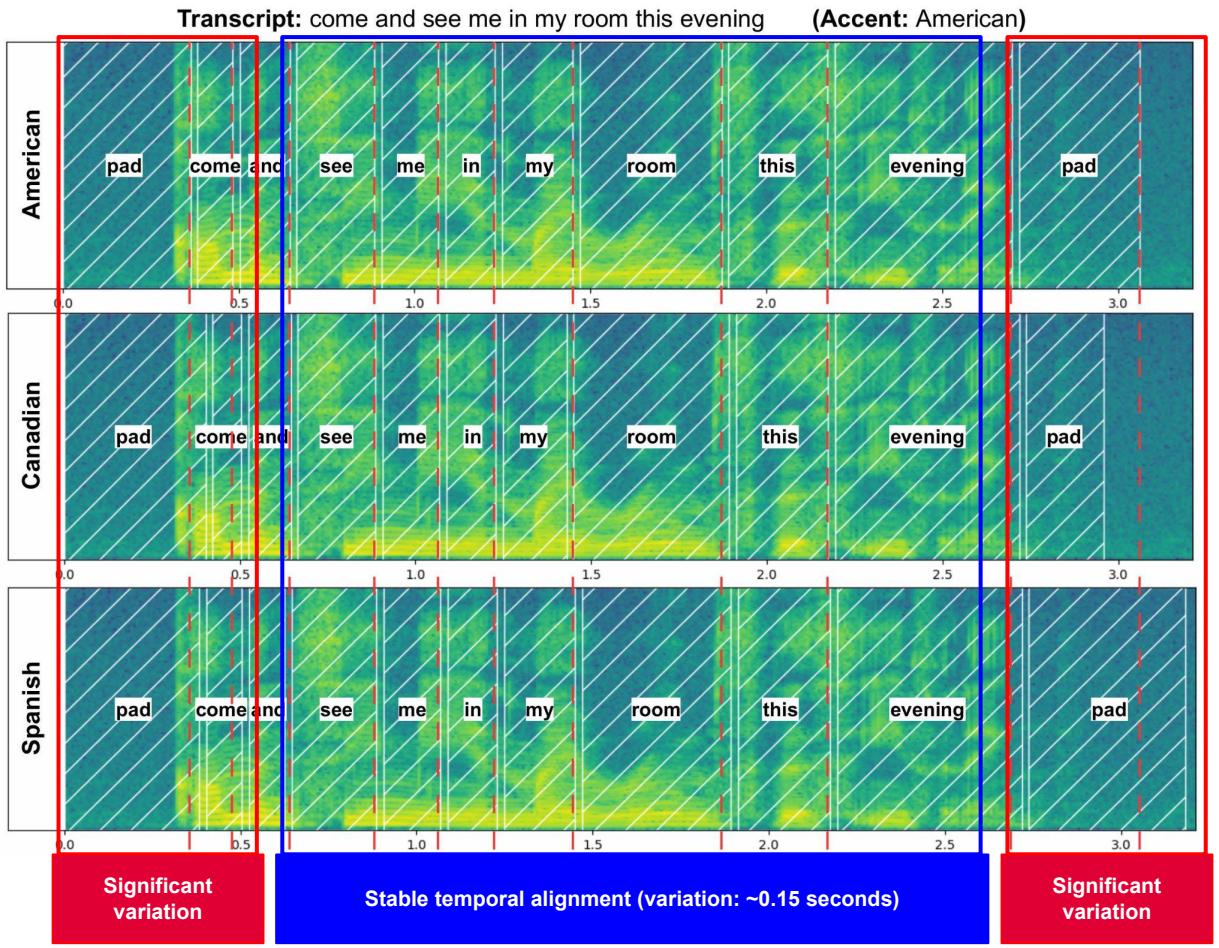
- Substitution Errors: FT: 245,426 → MTKD: 231,375 = ↓14,051 errors
  - Substitution errors dominate total count ← significant improvement.
    - reflects improved phoneme-level disambiguation → better phonetic understanding.
- Deletion & insertion drop modestly → suggesting general fluency gains.
- MTKD captures more distinct acoustic-phonetic mappings.



- Interpretation:
  - MTKD retains rare phonetic patterns without overfitting.
  - FT suffers from catastrophic forgetting and memorization.
- Analogy:
  - FT: It talks fast and listens less.
  - MTKD: It listens carefully before speaking.

Temporal Variability & Accent-Specific Decoding

- Mel-spectrogram → how different models align and attend to the input features ← accent variations.
- Same words but different temporal segmentation
  - different way of attending acoustic cues based on accent condition.
  - non-universality of phoneme timing.
- **"The models agree on what is said—but not when it is said."**
- **MTKD_Avg**:
  - preserved temporal diversity across models → better generalization.



Beyond Accent ASR: MTKD for Speech Emotion Recognition

Why MTKD for Speech Emotion?

- SER focuses on paralinguistic cues (prosody & timbre) → how well MTKD captures raw acoustic reps.?
- Emotions tend to be universal across languages → reveals MTKD's language-agnostic power
- Success in SER ⇒ transferable features for other non-lexical tasks (speaker traits, health)

- Proposed MTKD variants for both monolingual and multilingual speech emotion recognition.
- Baselines: FT-Mono., FT-Multi., KD-Mono.
- Languages: English, Finnish, French
- Datasets: IEMOCAP, FESC, CaFE

Mehedi Hasan Bijoy, et al. "Multi-Teacher Language-Aware Knowledge Distillation for Multilingual Speech Emotion Recognition." arXiv preprint arXiv:2506.08717 (2025).

MTKD for Speech Emotion Recognition

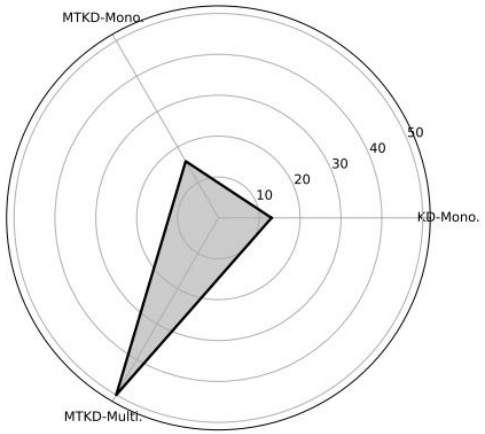
Split	FT-Mono.		FT-Multi.		KD-Mono.		**MTKD-Mono.**		**MTKD-Multi.**	
	UR	WR	UR	WR	UR	WR	UR	WR	UR	WR
Split 1	69.8 $\frac{74.5}{64.4}$	65.8 $\frac{70.6}{65.8}$	70.6 $\frac{74.1}{66.7}$	70.7 $\frac{74.0}{67.2}$	72.5 $\frac{77.4}{66.9}$	70.3 $\frac{75.1}{64.9}$	72.4 $\frac{76.8}{67.1}$	68.5 $\frac{73.1}{63.2}$	**72.5** $\frac{76.0}{68.7}$	**73.4** $\frac{76.6}{69.9}$
Split 2	**79.3** $\frac{83.9}{73.3}$	**78.1** $\frac{82.6}{72.8}$	70.9 $\frac{74.4}{67.3}$	74.9 $\frac{77.9}{71.6}$	76.7 $\frac{81.5}{71.0}$	74.8 $\frac{79.5}{69.3}$	77.8 $\frac{82.4}{72.1}$	76.0 $\frac{80.5}{70.6}$	71.2 $\frac{74.8}{67.3}$	71.0 $\frac{74.4}{67.3}$
Split 3	66.2 $\frac{71.3}{60.7}$	65.9 $\frac{71.0}{60.4}$	67.7 $\frac{71.2}{64.0}$	69.5 $\frac{72.8}{66.0}$	69.5 $\frac{74.5}{64.0}$	69.1 $\frac{74.1}{63.6}$	68.9 $\frac{73.9}{63.5}$	69.0 $\frac{74.0}{63.6}$	**71.4** $\frac{74.9}{67.8}$	**71.5** $\frac{74.8}{67.9}$
Split 4	68.5 $\frac{73.9}{62.7}$	70.2 $\frac{75.1}{64.8}$	64.2 $\frac{68.0}{60.3}$	69.7 $\frac{73.0}{66.1}$	70.2 $\frac{75.3}{64.3}$	70.1 $\frac{75.0}{64.7}$	**72.3** $\frac{77.6}{66.3}$	72.8 $\frac{77.8}{67.2}$	71.6 $\frac{75.3}{67.7}$	**73.8** $\frac{77.0}{70.3}$
Split 5	72.2 $\frac{76.9}{66.8}$	70.5 $\frac{75.1}{65.4}$	67.9 $\frac{71.3}{64.2}$	70.2 $\frac{73.4}{66.8}$	72.2 $\frac{77.0}{66.7}$	71.2 $\frac{75.8}{66.0}$	71.0 $\frac{75.9}{65.5}$	69.7 $\frac{74.4}{64.5}$	**73.0** $\frac{76.3}{69.3}$	**74.7** $\frac{77.8}{71.3}$
Mean	71.2 $\frac{76.1}{65.6}$	70.1 $\frac{74.9}{65.8}$	68.2 $\frac{71.8}{64.5}$	71.0 $\frac{74.2}{67.6}$	72.2 $\frac{77.1}{66.6}$	71.1 $\frac{76.1}{65.8}$	**72.5** $\frac{77.3}{66.9}$	71.2 $\frac{76.0}{65.8}$	71.9 $\frac{75.5}{68.2}$	**72.9** $\frac{76.1}{69.3}$

Table 5: Performance of the proposed MTKD methods alongside other baselines on the IEMOCAP dataset, where 'Mono.' refers to monolingual, and 'Multi.' stands for multilingual configurations. The upper (numerator) and lower (denominator) bounds of the confidence interval are reported for both UR and WR.

	IEMOCAP		FESC		CaFE	
	UR	WR	UR	WR	UR	WR
FT-Mono.	71.2	70.1	59.5	62.0	78.1	78.6
KD-Mono.	72.2	71.1	62.7	67.8	**82.3**	**79.8**
MTKD-Mono.	**72.5**	**71.2**	**62.9**	**68.1**	78.1	75.0
FT-Multi.	68.2	71.0	62.7	64.3	**77.1**	**79.8**
MTKD-Multi.	**71.9**	**72.9**	**63.4**	**66.1**	73.4	72.3

Table 6: Comparison of the proposed methods with baselines for SER in English, Finnish, and French.

- **Table 5:**
  - Best overall: **MTKD-Multi.** achieves highest mean WR (72.9)
  - Generalization: **MTKD-Multi.** achieves top WR scores in 4 out of 5 splits
- **Table 6:**
  - MTKD-Mono. outperforms monolingual baselines.
  - MTKD-Multi. outperforms multilingual baseline.



(b) Reduction in misclassification counts for MTKD methods compared to the FT baseline.

Mehedi Hasan Bijoy, et al. "Multi-Teacher Language-Aware Knowledge Distillation for Multilingual Speech Emotion Recognition." arXiv preprint arXiv:2506.08717 (2025).

Findings

- MTKD showcased superior performance across benchmarks:
  - Outperformed strongest FT baseline: >↑5% WER on Accent20h and >↑10% WER on EdAcc_test
- Strong generalization to unseen / out-of-distribution accents:
  - Improved zero-shot performance across 24 accents
- Improved error robustness:
  - Major drop in substitution errors → better phoneme understanding
- Aggregation strategy matters:
  - Uniform Averaging best balances accent diversity ← enabled generalization via smooth posterior fusion
- Dynamic loss weighting is crucial:
  - Adaptive scheduling with regularization stabilized training
- Versatile across tasks ← validated through SER.

SER = Speech Emotion Recognition

Answers to the Research Questions

RQ1: Does MTKD facilitate the learning of more robust and accent-invariant acoustic representations than accent aware FT?

✅ **Yes**. MTKD outperforms FT on both seen and unseen accents.

→ Relative WER and CER reduction on both Accent20h and EdAcc_test benchmarks.

RQ2: How does teacher ensemble heterogeneity influence the robustness of multi-teacher distilled student in zero-shot scenarios?

✅ **Positively**. Greater teacher diversity → better generalization.

→ Diverse teacher ensembles significantly improve zero-shot performance.

→ Students trained on heterogeneous teachers are more accent-agnostic.

RQ3: To what extent does MTKD mitigate catastrophic forgetting compared to FT when progressively adding new accents?

✅ **Markedly**. MTKD preserves prior knowledge better than FT.

→ Maintains low and stable WER across incremental accent additions.

→ FT suffers from performance degradation, unlike MTKD.

Challenges, Open Questions, and Future Directions

- **Accent Representation without Labels**
  - **Challenge**: Accent labels are often missing
  - **Open Question**: Can accent clusters emerge without explicit labels?
  - **Future Direction**: Use unsupervised learning to incorporate accent-agnostic embeddings.
- **Temporal Misalignment**
  - **Challenge**: No time-warping between student and teacher peaks → misalign supervision.
  - **Open Question**: Can we correct misalignments without global attention?
  - **Future Direction**: Add local attention (10–20 frames) to guide frame-level distillation
- **Teacher Selection Based on KL, Not CTC**
  - **Challenge**: KL-aligned teacher ≠ best in transcription quality
  - **Open Question**: Does lowest CER lead to better learning?
  - **Future Direction**: Try hybrid selection: balance KL fit + transcription error
- **Shallow Supervision at Output Only**
  - **Challenge**: No internal layer guidance in MTKD
  - **Open Question**: Can intermediate features improve distillation depth?
  - **Future Direction**: Extend to multi-layer distillation via attention maps

Conclusion

- Proposed a novel **Dual-Adaptive MTKD** for Accented English Speech Recognition.
  - 10 accent-specialist teachers → one compact, accent-robust ASR model
  - A neural teacher mixer + meta-learned objective adapter
  - Consistently outperformed per-accent fine-tuning baselines
  - Uniform averaging proved superior to rank-based or top-one strategies
  - Dynamic loss weighting stabilized training
- Extended MTKD to SER, achieving state-of-the-art performance in 3 languages
  - Proposed a novel **Language-Aware MTKD** for Multilingual Speech Emotion Recognition. [Accepted to INTERSPEECH 2025]
- MTKD emerges as one of the principal routes toward inclusive and resource-efficient speech technology.

Mehedi Hasan Bijoy, et al. "Multi-Teacher Language-Aware Knowledge Distillation for Multilingual Speech Emotion Recognition." arXiv preprint arXiv:2506.08717 (2025).

Aalto University
School of Electrical
Engineering

Thank You!

Questions?

mehedi.bijoy@aalto.fi