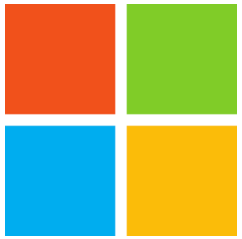
UNIFIED ACOUSTIC MODELING USING NEURAL MIXTURE MODELS

Amit Das, **Jinyu Li**, Changliang Lu, Yifan Gong



Microsoft Speech and Language

Outline of Unified Acoustic Modeling (UAM)

- Motivation

- Baseline:
  - Recurrent Adaptive Mixture Model (RADMM)

- Proposed models:
  - Improvements to RADMM
  - Linear Hidden Interpolation (LHI)
  - Hybrid Attention Mixture Model (HATMM)

- Experiments and Results

Motivation

Motivation: Why do we need unified models?

- Usually, AMs (acoustic models) are domain dependent
  - E.g. Meeting, Car, Dictation etc.

- Problem with domain dependent (expert) AMs
  - Perform well in scenarios with in-domain/seen data
  - Degrade in scenarios with out-of-domain/unseen data

- This makes AM deployment difficult for production.

- Is it possible to build a single model that can work well across many scenarios, especially new unseen scenarios?

- Yes! Unified Acoustic Modeling.

Formulation

- UAM is a Mixture-of-Experts model (similar to GMM).
- UAM = Ensemble of Experts + Switch

$$\mathbf{h} = \sum_{k=1}^{K} p(k|\mathbf{x})\ \mathbf{h}^{(k)} \quad K = \text{Total Experts}$$

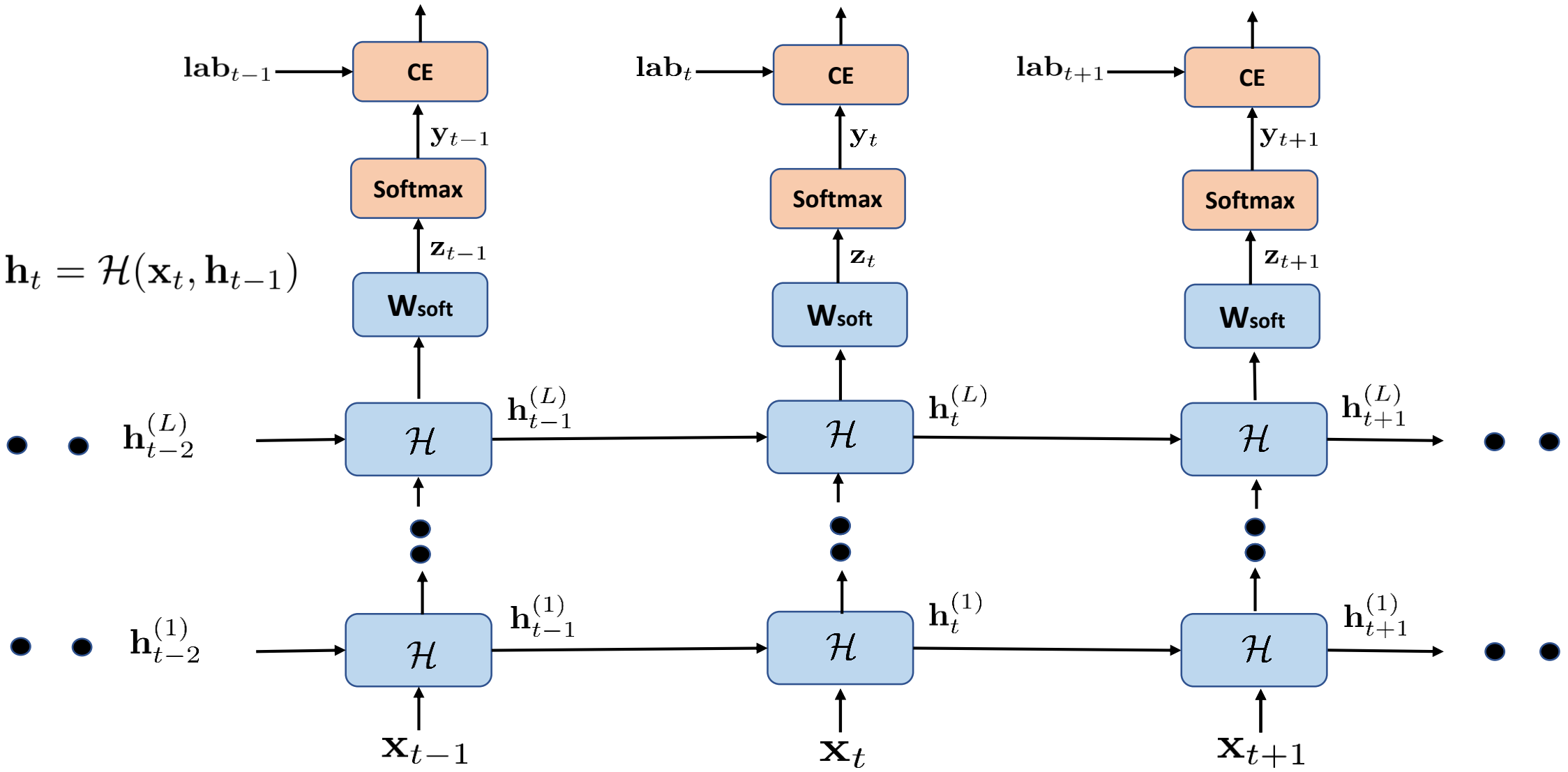
$$p(\text{class}|\mathbf{x}) = \text{softmax}(\mathbf{W}\mathbf{h} + \mathbf{b}) = \text{softmax}(\mathbf{W}\textstyle\sum_{k=1}^{K} p(k|\mathbf{x})\mathbf{h}^{(k)} + \mathbf{b})$$

- $\mathbf{h}^{(k)}$: Hidden vector output of the k[th] expert.
- $p(k|\mathbf{x})$: Probability of selecting expert 'k' given feature **x** where $\sum_{k=1}^{K} p(k|\mathbf{x}) = 1$
- $p(\text{class}|\mathbf{x})$: Posterior probability of class
- UAM Training: Learn the distribution $p(k|\mathbf{x})$ using a neural network.

($p(k|\mathbf{x})$: K-ary switching model)

Expert Acoustic Model (Domain Dependent)

Expert AM: Stacked RNN



Trainable Parameters

Operation

Unified Acoustic Model (UAM)

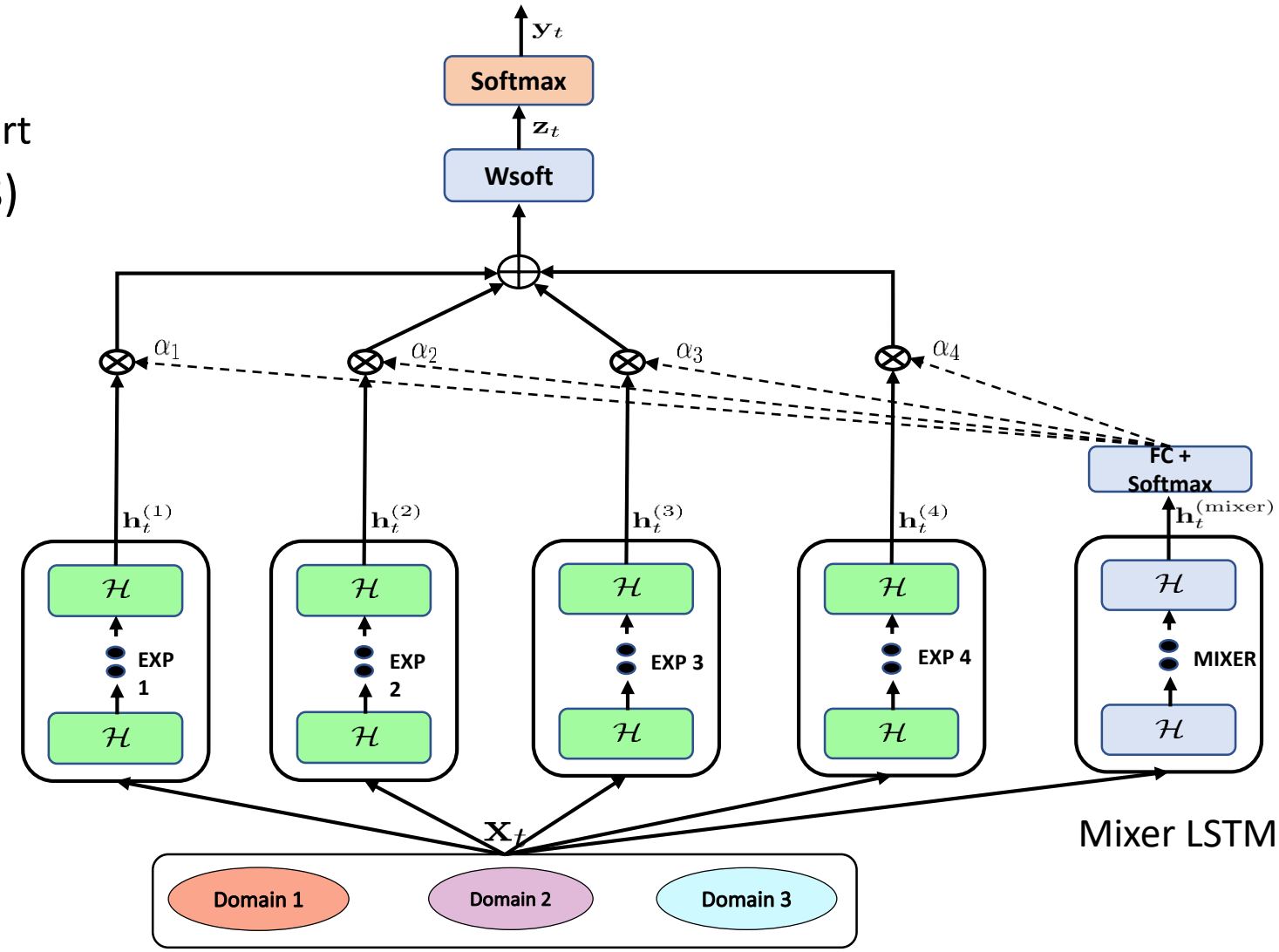
Unified Acoustic Model (UAM)

- Recurrent Adaptive Mixture Model (RADMM)
- Learned Hidden Interpolation (LHI)
- Hybrid Attention Mixture Model (HATMM)

Baseline UAM

UAM: Recurrent Adaptive Mixture Model (RADMM)

- Mixer LSTM
  - Input: Raw Features
  - Output: Weight for each expert
- Our baseline model (Irie 2018)



Constant Parameters

Trainable Parameters

Operation

[Irie 2018] K. Irie, S. Kumar, M. Nirschl, and H. Liao, "RADMM: Recurrent Adaptive Mixture Model With Applications to Domain Robust Language Modeling," ICASSP 2018.

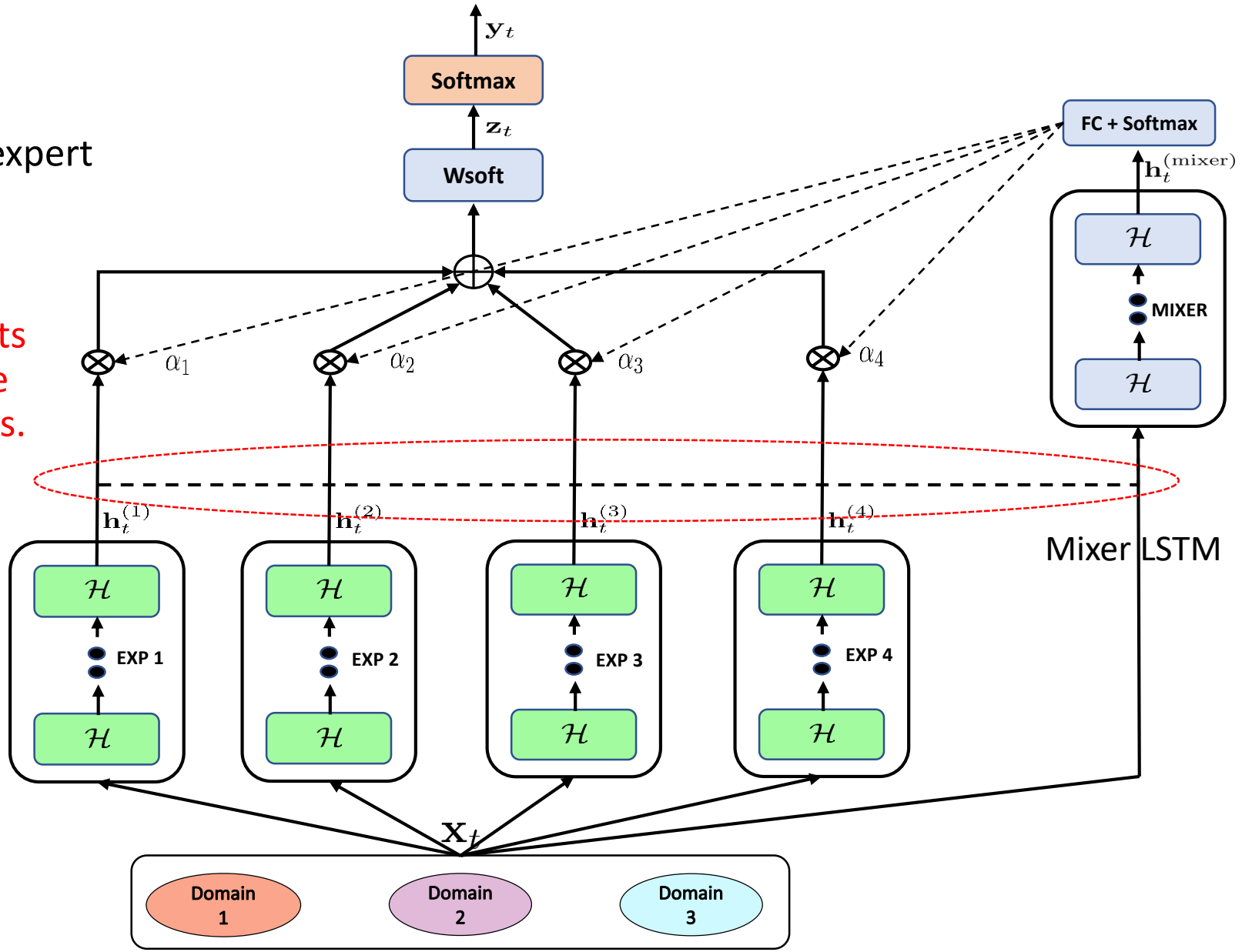
Proposed UAMs

UAM: Improvement 1 of RADMM

- Mixer LSTM
  - Input: Hidden Features
  - Output: Weight for each expert

- Why is it useful?

Hidden layer outputs from experts have better knowledge about the state of experts than raw features.



Constant Parameters

Trainable Parameters

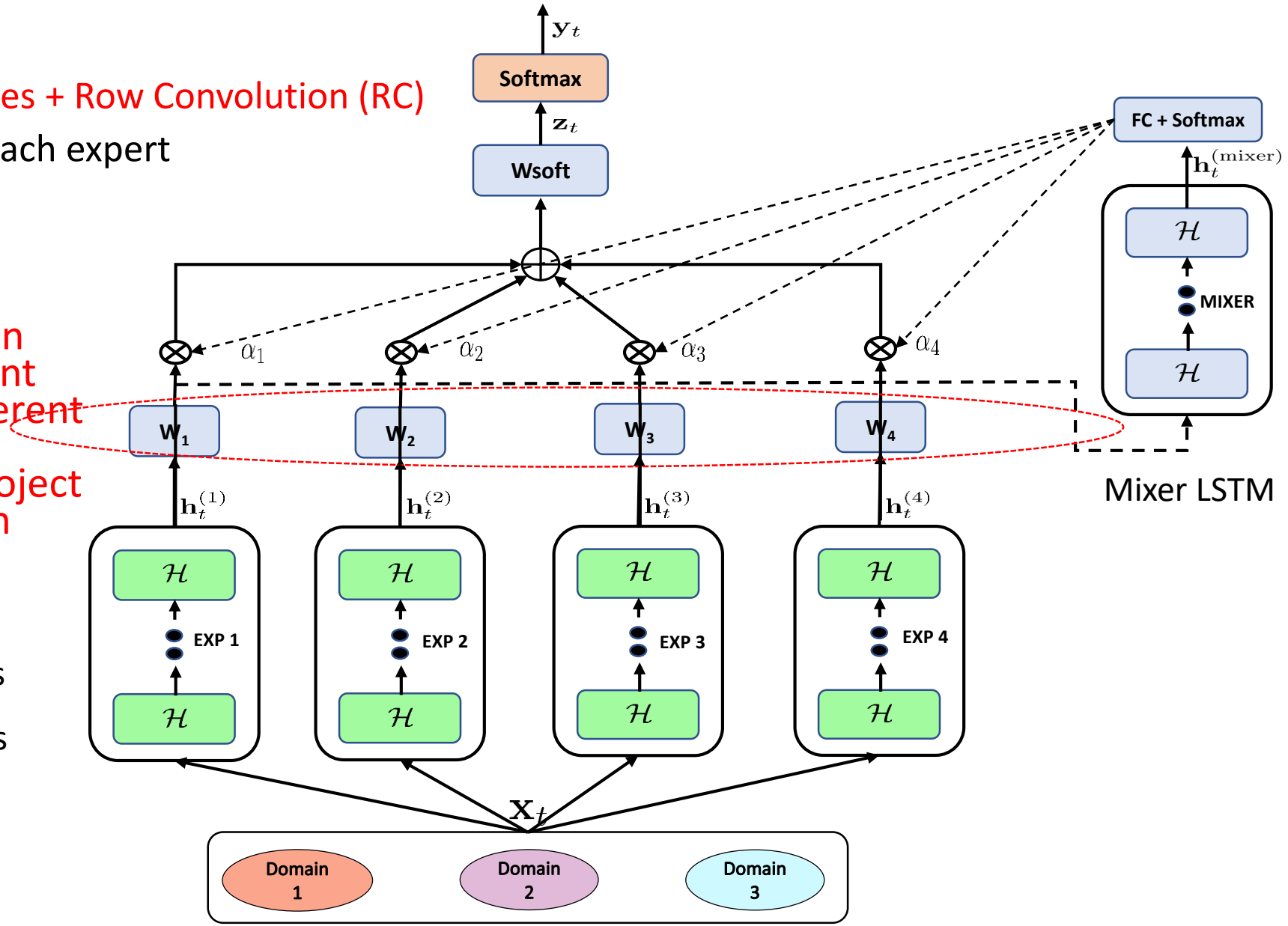
Operation

UAM: Improvement 2 of RADMM

- Mixer LSTM
  - Input: Hidden Features + Row Convolution (RC)
  - Output: Weight for each expert

- Why is it useful?

We believe that the hidden layer outputs from different experts may reside in different sub-spaces. Therefore, a transform is needed to project all of them into a common sub-space



Constant Parameters

Trainable Parameters

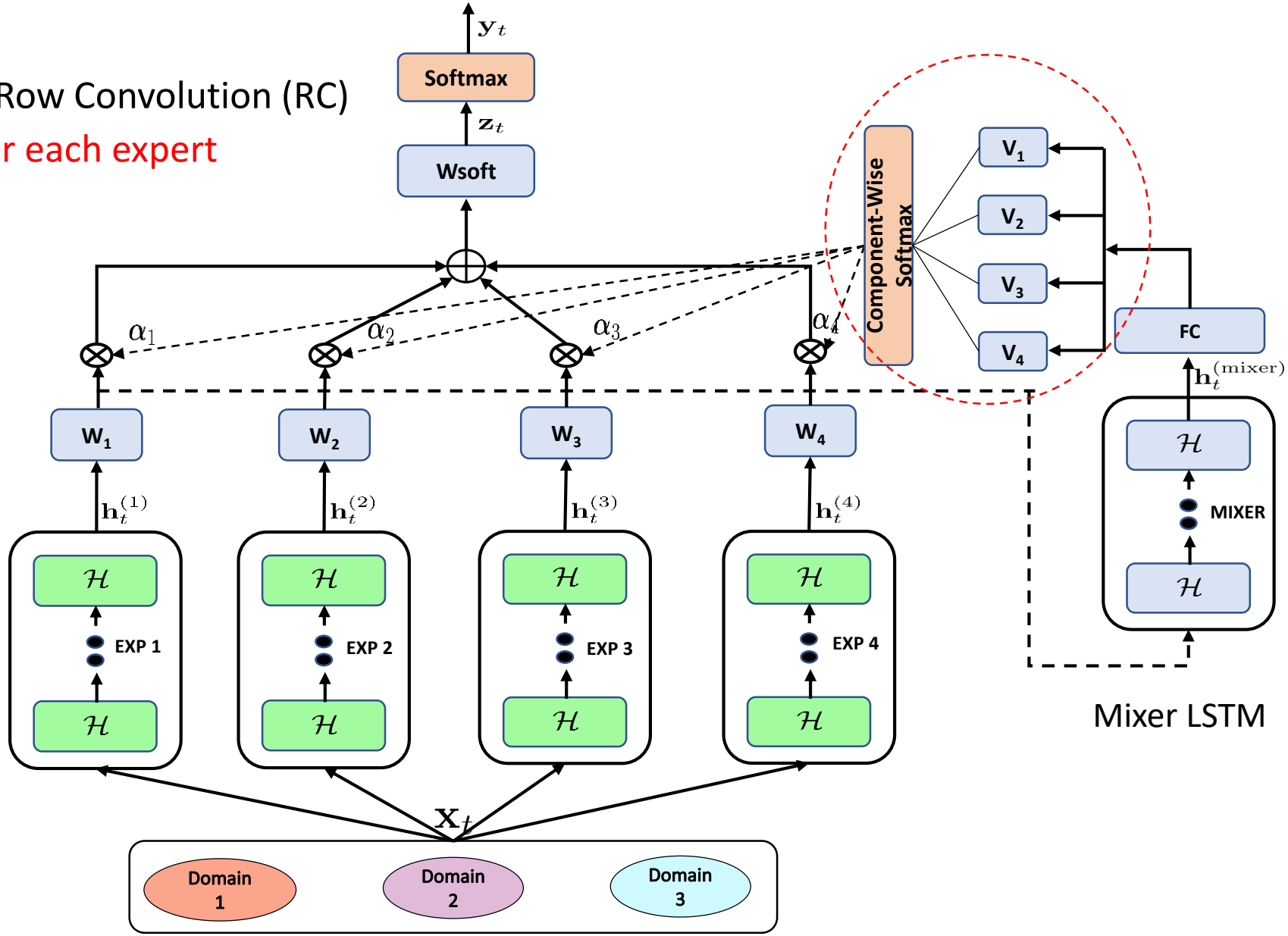
Operation

UAM: Improvement 3 of RADMM

- Mixer LSTM
  - Input: Hidden Features + Row Convolution (RC)
  - Output: Vector weights for each expert

- Why is it useful?

Applying component-wise weighting gives us greater control in cherry-picking the individual components of the hidden features.



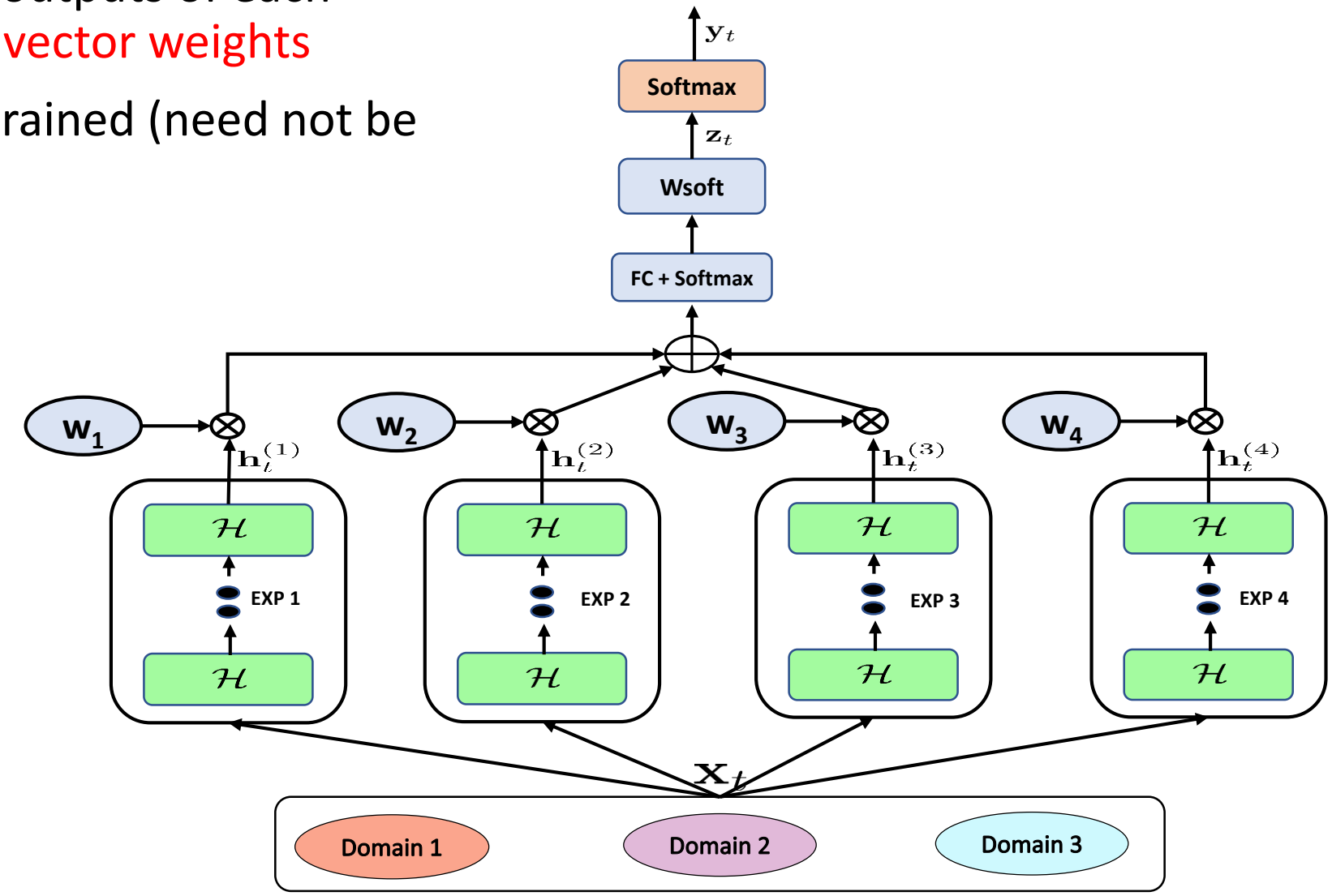
Constant Parameters

Trainable Parameters

Operation

UAM: Learned Hidden Interpolation (LHI)

- Combine the hidden outputs of each expert using learned vector weights
- Weights are unconstrained (need not be in [0,1])

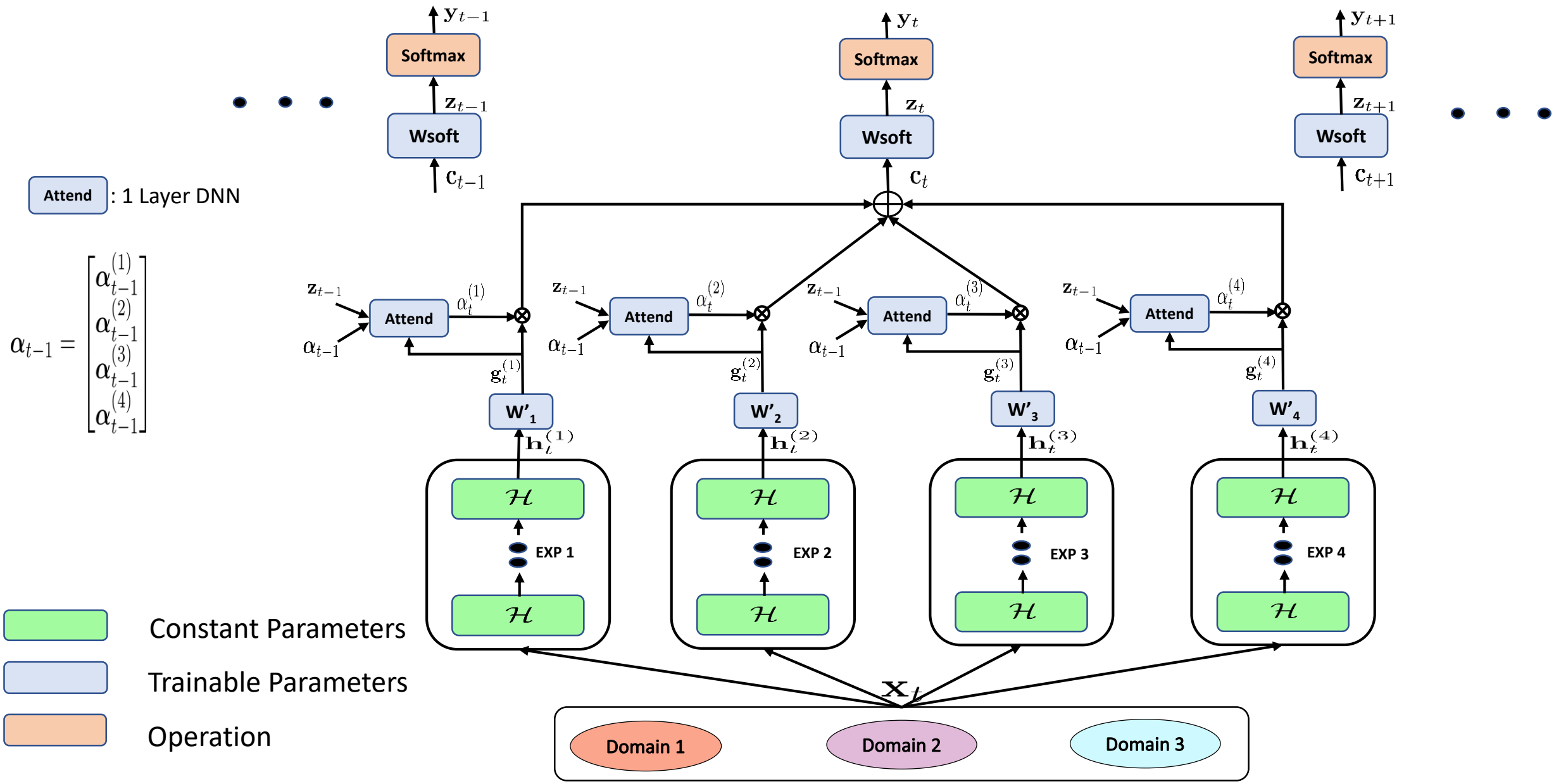


Constant Parameters

Trainable Parameters

Operation

UAM: Hybrid Attention Mixture Model (HATMM)



Experimental Set-Up

- Expert Acoustic Model and Training Data
  - Model
    - 6 layers, Uni-LSTM with 1024 memory cells with cell output linearly projected to 512 dimensions
    - CE trained
    - One expert model/domain x 4 domains = 4 expert models
  - Data
    - 4 Domains: S1, S2, S3, S4
    - 30k hours

- Unified Acoustic Model and Training Data
  - Model
    - 4 Experts (4 Uni-LSTMs) + 4-ary switch
  - Data
    - 4 Domains: S1, S2, S3, S4
    - 300 hours (1% of training data used for experts)

Experimental Set-Up

- Test Data
  - In-Domain (Seen during training): S1, S2, S3, S4
  - Out-of-Domain (Unseen during training):U1, U2, U3, U4
- Features
  - Log Mel Filterbank Energy (LMFE), 80-dimensional
- Output Labels
  - 9404 Senones
- Decoder
  - LM based decoding

WER Experts

				Expert Models				
Test Scenario	**Domain**	**Utt Count**	**Word Count**	**S1**	**S2**	**S3**	**S4**	**Best Expert Model (Oracle)**
S1	Seen	6114	34150	**11.45**	12.91	16.71	18.71	**11.45**
S2	Seen	5806	32134	13.04	**6.97**	23.83	23.14	**6.97**
S3	Seen	3878	46471	19.73	50.12	**14.1**	15.7	**14.1**
S4	Seen	1671	21690	30.87	51.66	21.51	**20.69**	**20.69**
U1	Unseen	4556	22618	**16.39**	17.36	22.64	23.39	**16.39**
U2	Unseen	5485	29480	37.9	**19.7**	33.6	25.84	**19.7**
U3	Unseen	7520	28553	14.91	32.37	14.54	**14.47**	**14.47**
U4	Unseen	1986	14715	**13.69**	18.6	15.55	19.99	**13.69**
Wtd Average (In)			134445	17.83	30.60	18.28	19.05	**12.79**
Wtd Average (Out)			95366	22.18	22.77	22.51	20.95	**16.42**
Wtd Average (All)			229811	**19.63**	**27.35**	**20.04**	**19.84**	**14.29**

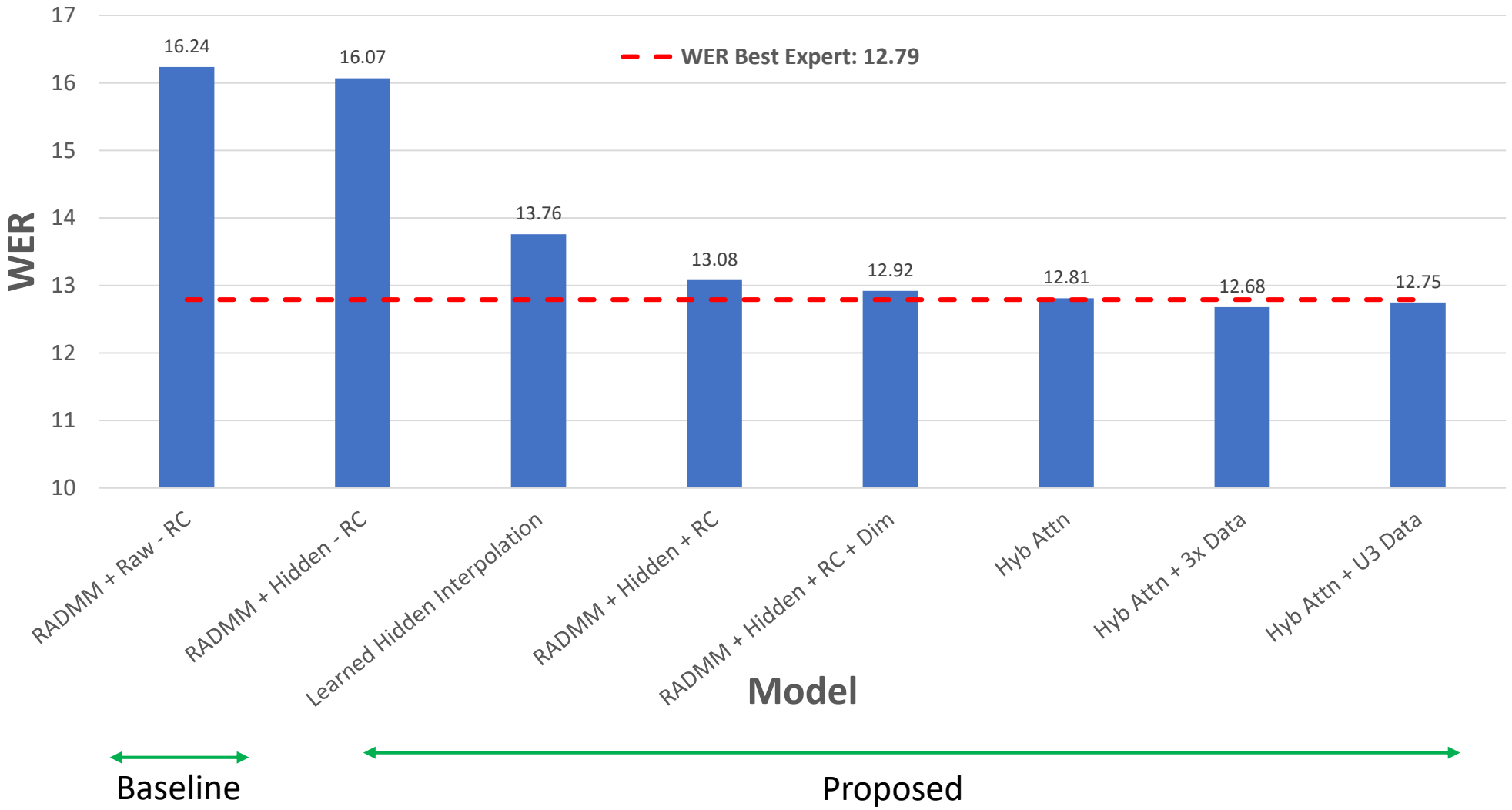
Domain = Seen: **Match** in training and testing scenarios.
Domain = Unseen: **Mismatch** in training and testing scenarios.

- Clearly, each expert model (S1-S4) performed significantly worse than the Oracle model.
- What do we want?: Can UAM pick the best expert for each test scenario and achieve the WER of the oracular “best expert model”?

WER of UAMs: Seen Data

Baseline vs Best Proposed

Model	WER	Abs %	Rel %
RADMM + Raw – RC	16.24	-	-
Hybrid Attention	12.81	3.43	21.12

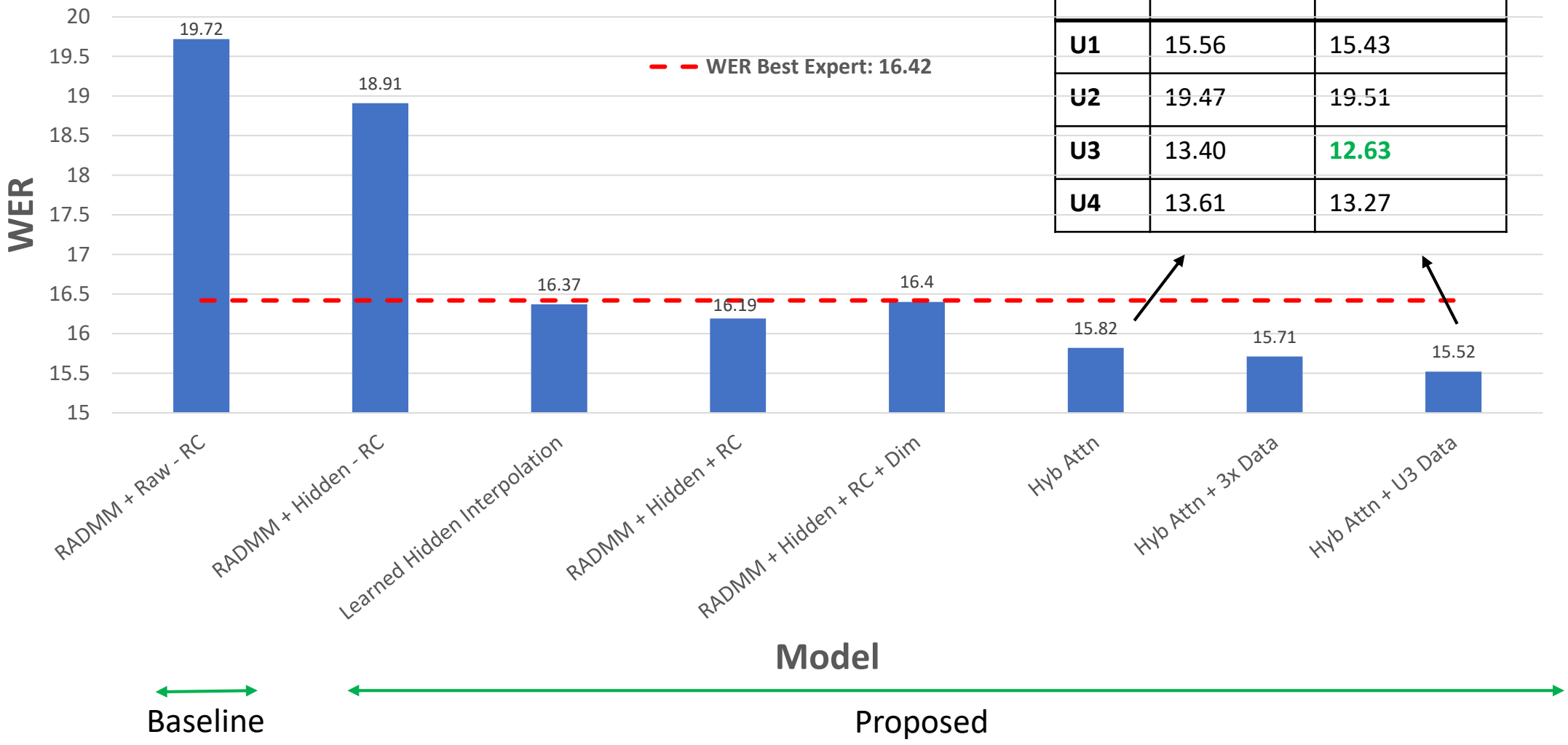


WER of UAMs: Unseen Data

Baseline vs Best Proposed

Model	WER	Abs %	Rel %
RADMM + Raw – RC	19.72	-	-
Hybrid Attention	15.82	3.90	19.78

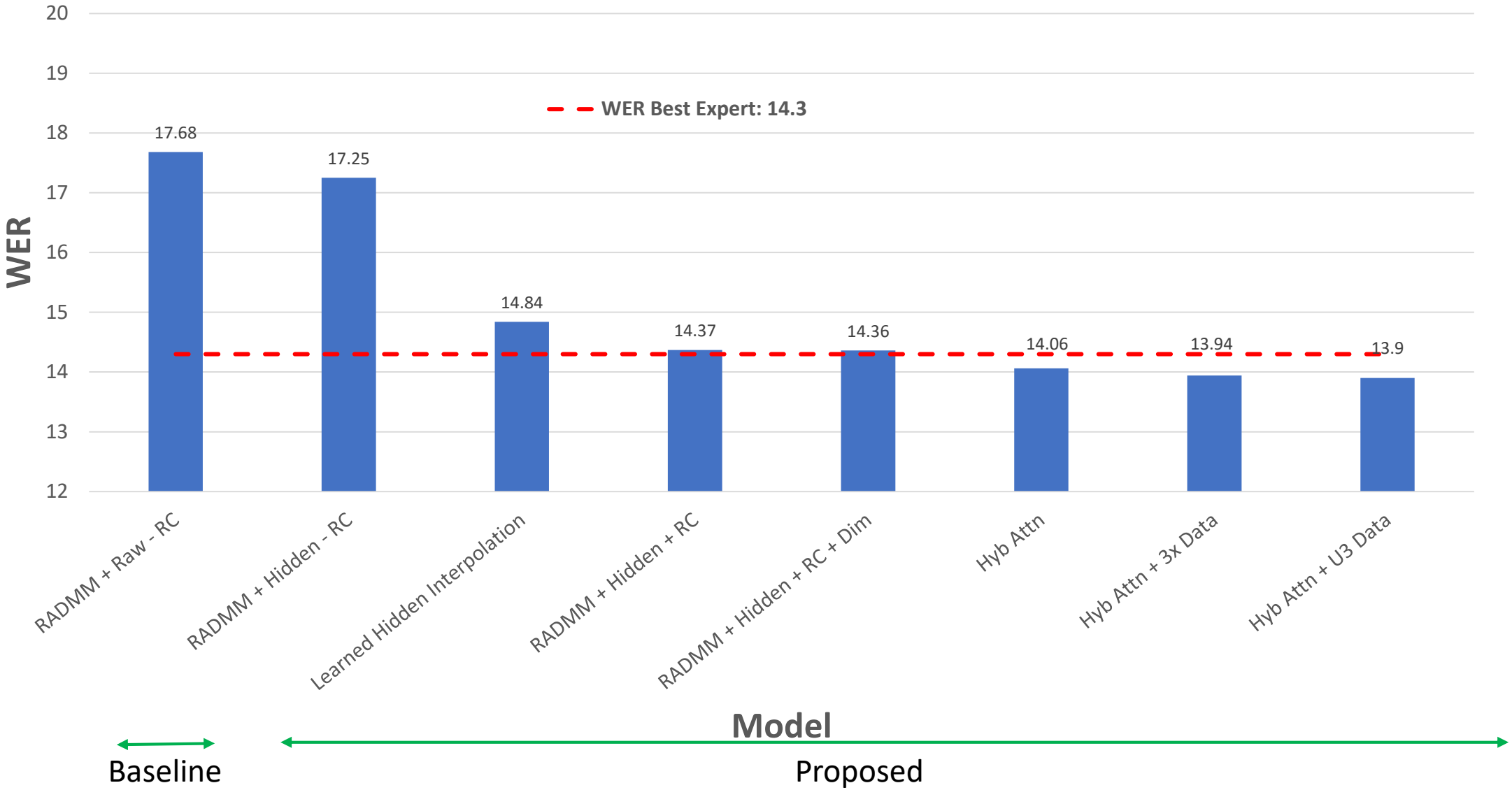
Data	Hyb Attn	Hyb Attn + U3
U1	15.56	15.43
U2	19.47	19.51
U3	13.40	12.63
U4	13.61	13.27



WER of UAMs: All Data

Baseline vs Best Proposed

Model	WER	Abs %	Rel %
RADMM + Raw – RC	17.68	-	-
Hybrid Attention	14.06	3.62	20.48



Summary and Conclusion

- Proposed UAM using Hybrid Attention (HATMM) and several improvements over RADMM.

Model	Baseline: RADMM	Best Proposed: HATMM	Absolute WERR	Relative WERR
WER (Seen Data)	16.24	12.81	**3.43**	**21.12**
WER (Unseen Data)	19.72	15.82	**3.90**	**19.78**
WER (All Data)	17.68	14.06	**3.62**	**20.48**

- Conclusion: HATMM is the best proposed UAM.
  - HATMM (best proposed) > Best expert (oracle) > RADMM (baseline)

Thank You