Hypothesis Posterior Student-Teacher Training

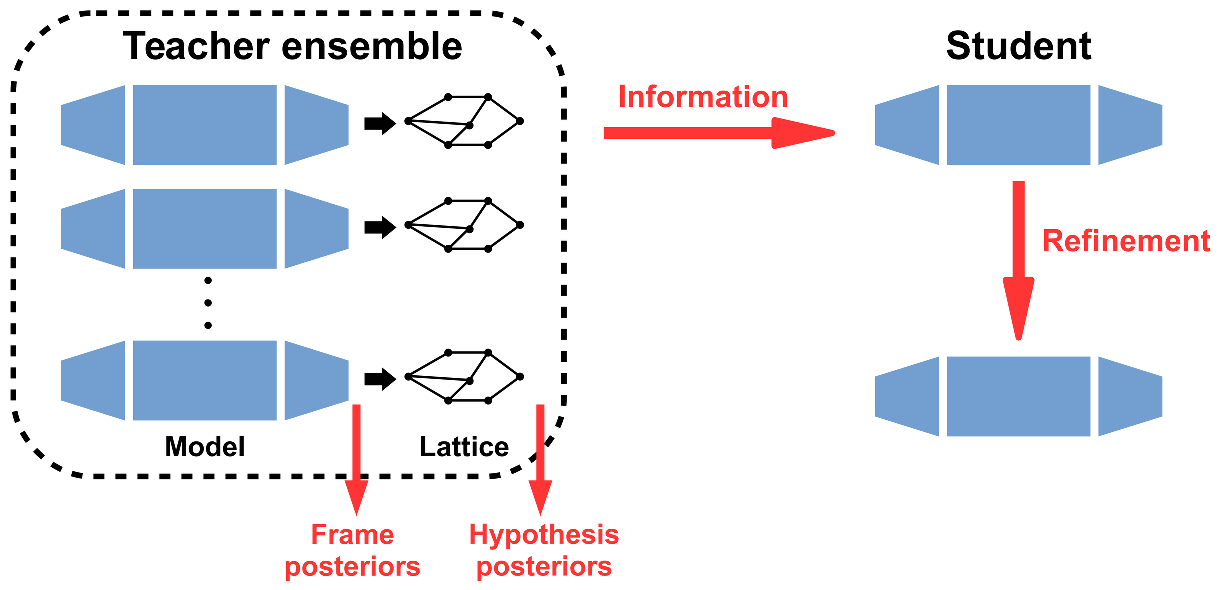
Jeremy H. M. Wong and Mark J. F. Gales

Department of Engineering, University of Cambridge

jhmw2@cam.ac.uk, mjfg@eng.cam.ac.uk

1 INTRODUCTION

- Ensemble methods
  - improve ASR performance
  - are computationally expensive to decode.
- Student-Teacher (S-T) training
  - trains single student model to emulate teacher ensemble.
  - Existing methods only transfer frame posterior information.
- This work incorporates sequence discriminative criteria into S-T training by:
  - sequence discriminative training of the teacher ensemble
  - further sequence discriminative training of the student model after frame-level S-T training
  - a proposed hypothesis-level S-T criterion.



2 TEACHER ENSEMBLE

- Diversity obtained by
  - different DNN random initialisations.
- Teachers can be trained using the following criteria:
  - Cross-Entropy (CE)

$$\mathcal{F}_{\text{CE}} = -\sum_r \sum_t \sum_{s_{rt}} \delta\left(s_{rt}, s^*_{rt}\right) \log P\left(s_{rt} | \boldsymbol{o}_{rt}, \boldsymbol{\Phi}_m\right)$$

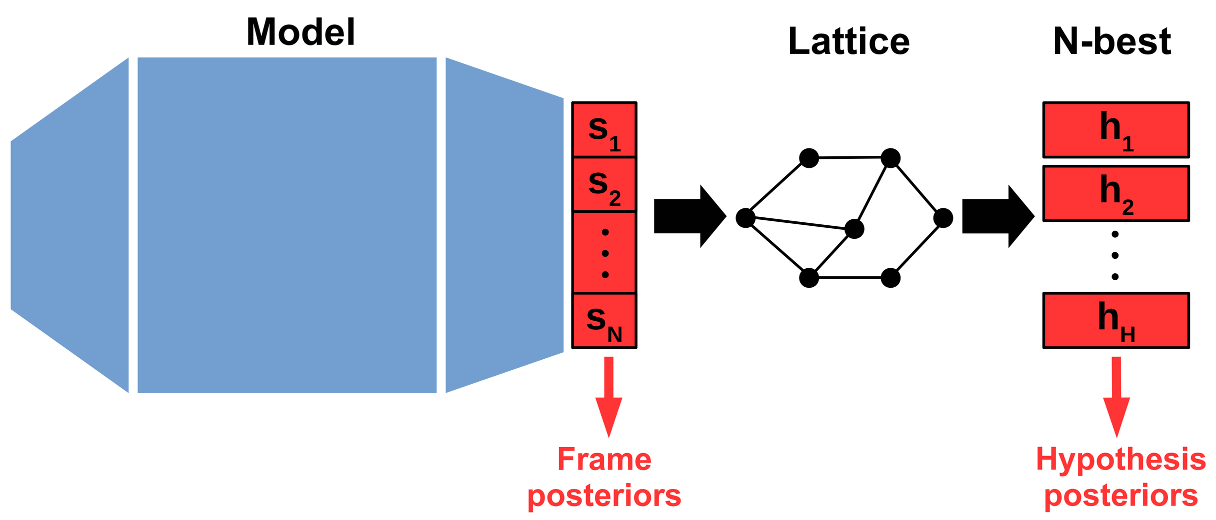
  - Maximum Mutual Information (MMI)

$$\mathcal{F}_{\text{MMI}} = -\sum_r \sum_{h_r} \delta\left(h_r, h^*_r\right) \log P\left(h_r | \mathbf{O}_r, \boldsymbol{\Phi}_m\right)$$

  - state-level Minimum Bayes Risk (sMBR)

$$\mathcal{F}_{\text{sMBR}} = \sum_r \sum_{h_r} L\left(h_r, h^*_r\right) P\left(h_r | \mathbf{O}_r, \boldsymbol{\Phi}_m\right).$$

3 INFORMATION PROPAGATION



- Frame posteriors
  - Existing method.
  - Minimise KL-divergence between frame posteriors.
  - Interpolate with hard alignments.

$$\mathcal{C}_{\text{CE}} = -\sum_r \sum_t \sum_{s_{rt}} \left[(1-\lambda)\, \delta\left(s_{rt}, s^*_{rt}\right) + \lambda \sum_m \alpha_m P\left(s_{rt} | \boldsymbol{o}_{rt}, \boldsymbol{\Phi}_m\right) \right] \log P\left(s_{rt} | \boldsymbol{o}_{rt}, \boldsymbol{\Theta}\right).$$

  - Setting $\lambda = 0$ reduces to CE.
- Hypothesis posteriors
  - Novel approach.
  - Minimise KL-divergence between hypothesis posteriors.
  - Interpolate with manual transcriptions.

$$\mathcal{C}_{\text{MMI}} = -\sum_r \sum_{h_r} \left[(1-\eta)\, \delta\left(h_r, h^*_r\right) + \eta \sum_m \beta_m P\left(h_r | \mathbf{O}_r, \boldsymbol{\Phi}_m\right) \right] \log P\left(h_r | \mathbf{O}_r, \boldsymbol{\Theta}\right).$$

  - Setting $\eta = 0$ reduces to the MMI criterion.

4 EXPERIMENTS

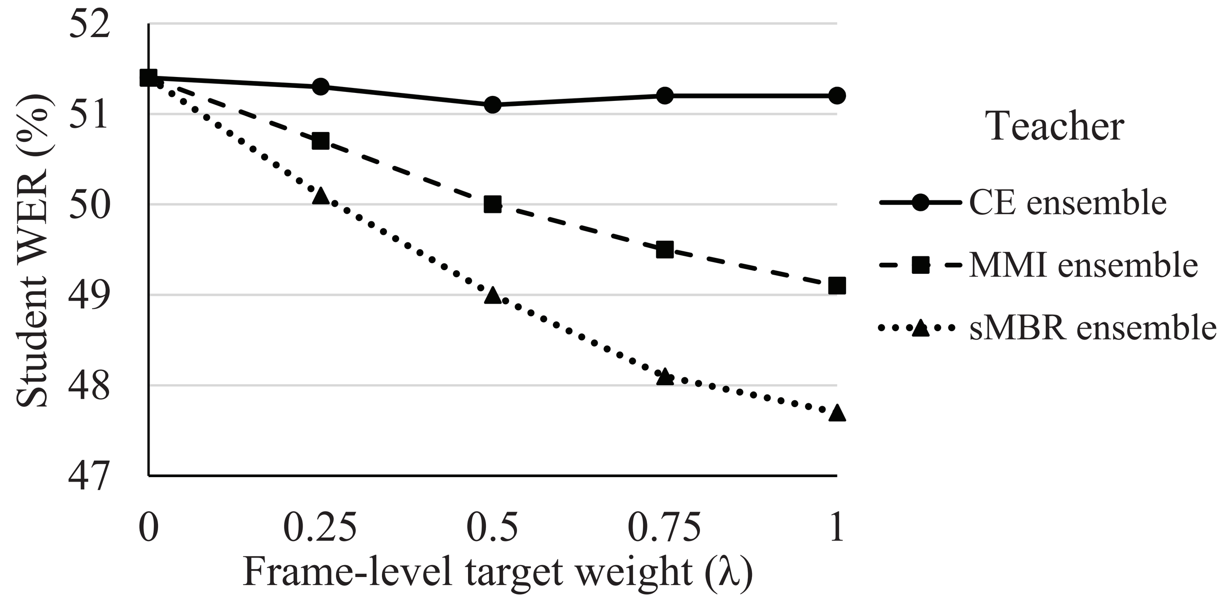
- Datasets:
  - IARPA Babel Tok Pisin (*IARPA-babel207b-v1.0e*)
    - 3 hour VLLP training set
    - 10 hour development set
  - WSJ
    - 14 hour *si-84* training set
    - 64K words open-vocabulary *eval92* test set.
- Setup:
  - **Ensemble size** = 10 (Tok Pisin), 4 (WSJ)
  - **Combination method** = MBR combination decoding
  - **Acoustic model** = DNN-HMM hybrid
    - 1000 nodes × 4 layers for Tok Pisin
    - 2000 nodes × 6 layers for WSJ.
- The student and teacher models have the same architecture.

4.1 TEACHER ENSEMBLE TRAINING CRITERION

- Training ensemble with different criteria, in Tok Pisin

Ensemble criterion	Single system WER (%) mean	best	worst	std dev	Combined WER (%)
CE	51.4	51.3	51.5	0.1	50.5
MMI	49.3	49.1	49.4	0.1	48.4
sMBR	48.2	48.1	48.4	0.1	47.0

  - Training teachers with sequence discriminative criteria improves combined ensemble performance.
- Frame-level S-T training with sequence-trained teachers, in Tok Pisin



  - Gains from sequence discriminative training of teachers are carried through to student.
  - $\lambda = 1$ produces the best student performances for sequence-trained teacher ensembles.

4.2 REFINEMENT OF THE STUDENT MODEL

Training	WER (%) Tok Pisin	WSJ
frame level S-T	47.7	5.07
frame level S-T + MMI	47.6	5.09
frame level S-T + sMBR	47.2	4.94

- Student is initialised using frame-level S-T training with the sMBR-trained teacher ensemble.
- For WSJ,
  - **mean single sMBR system WER** = 5.09 %
  - **combined ensemble WER** = 4.84 %.
- Further sMBR training of student improves performance.
- Further MMI training does not give significant gains, as the teacher ensemble has been sMBR-trained.

4.3 PROPAGATING HYPOTHESIS POSTERIOR INFORMATION

Training	η	WER (%) Tok Pisin	WSJ
frame level S-T	-	47.7	5.07
frame level S-T + MMI	0.0	47.6	5.09
hypothesis level S-T	0.5	47.0	4.91
hypothesis level S-T	1.0	47.4	4.94

- Hypothesis-level S-T training improves the student performance beyond frame-level S-T training, even with further MMI training.

5 CONCLUSIONS

- Sequence discriminative training of the teacher ensemble improves the resulting student performance.
- Further sequence discriminative training after frame-level S-T training brings additional gains.
- Proposed hypothesis-level S-T training yields gains over frame-level S-T training, even with further sequence discriminative training.

ACKNOWLEDGMENT

This work was supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD / ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government. The authors would like to thank the LORELEI team for providing the KWS infrastructure and multilingual deep neural network features.