TIED-STATE BASED DISCRIMINATIVE TRAINING OF CONTEXT-EXPANDED REGION-DEPENDENT FEATURE TRANSFORMS FOR LVCSR

Zhi-Jie Yan, Qiang Huo, Jian Xu, and Yu Zhang
Microsoft Research Asia, Beijing, China

Microsoft Research

1. Summary

- Formulate feature transform using a set of context-expanded, region-dependent linear transforms (RDLTs)
- Train RDLTs by a lattice-free, tied-state based maximum mutual information (MMI) criterion
- Leverage both long-span features and contextual weight expansion
- Achieve relative word error reduction of 10% and 6% respectively compared with conventional RDLT baselines

2. CE-RDLT

- Context-expanded region-dependent feature transform

$$\hat{\mathbf{o}}_t = \sum_{m=1}^{M} \kappa_{m,t} \cdot \mathbf{W}_m \xi_t$$

$\kappa_{m,t}$: a weight of the m^{th} transform $\mathbf{W}_m$ at time t, which is calculated by using the so-called "acoustic context expansion" in fMPE;

ξ_t: a long-span feature vector obtained by concatenating several neighboring frames of feature vectors around $\mathbf{o}_t$, i.e., $\xi_t = \left[1 \ \ \mathbf{o}_{t-L}^{\top} \ldots \mathbf{o}_t^{\top} \ldots \mathbf{o}_{t+L}^{\top}\right]^{\top}$.

- Conventional fMPE, FE-RDLT and WE-RDLT are special cases of CE-RDLT

	fMPE	FE-RDLT	WE-RDLT	CE-RDLT
Bias only / Full transform	Bias	Full	Full	Full
Long-span features		✓		✓
Contextual weight expansion	✓		✓	✓

3. Training criterion / optimization

- An MMI criterion formulated on decision-tree tied tri-phone HMM states is used

$$\mathcal{F}(\mathbf{W}) = \sum_t \log p(s_t^r|\hat{\mathbf{o}}_t) = \sum_t \log \frac{p(\hat{\mathbf{o}}_t|s_t^r)p(s_t^r)}{\sum_s p(\hat{\mathbf{o}}_t|s)p(s)}$$

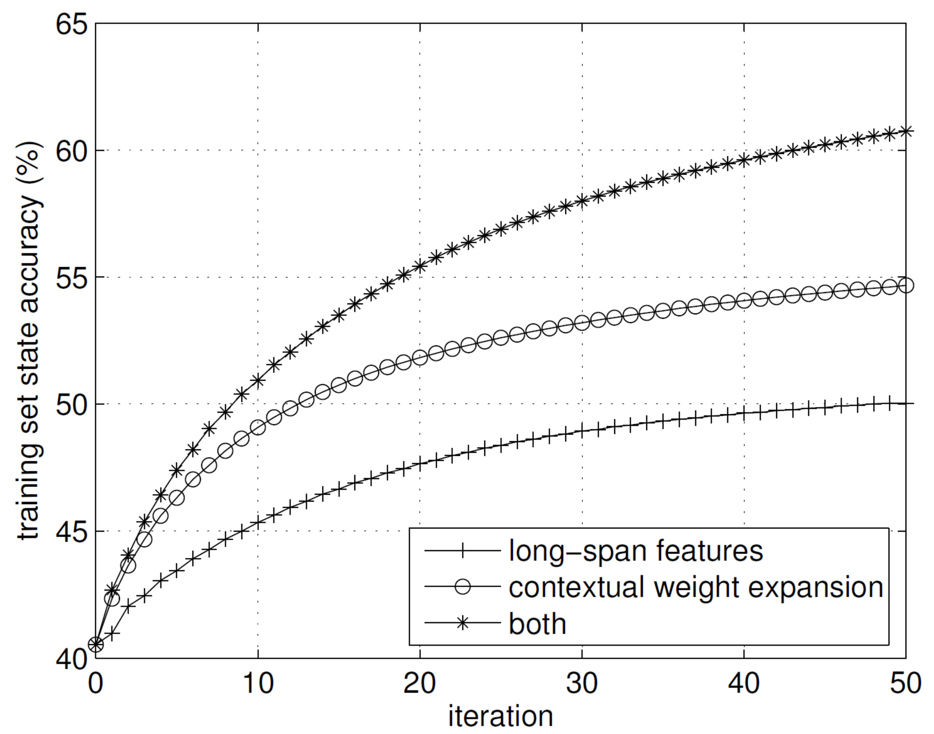
- Lattice-free – more efficient and less dependent on the language model used in training
- A batch-mode, adaptive Rprop algorithm is performed in optimization
- Per-parameter initial step sizes are set in a semi-automatic process (refer to the paper for details)

4. Experimental setups

- Training: 309hr Switchboard-1 conversational telephone speech transcription
  - fMPE: 7x50k 39-dimensional bias vectors
  - FE-RDLT: 1k 39x573 RDLTs
  - WE-RDLT: 7x1k 39x53 RDLTs
  - CE-RDLT: 7x1k 39x573 RDLTs
- Testing: NIST 2000 Hub5

5. Experimental results

- Learning curve



- Comparison of RDLTs trained w/ and w/o lattices (ML baseline WER: 26.5)
  - w/ lattices

fMPE	FE-RDLT	WE-RDLT	CE-RDLT
23.4 (11.7)	23.8 (10.2)	23.2 (12.5)	23.0 (13.2)

  - w/o lattices

FE-RDLT	WE-RDLT	CE-RDLT
23.7 (10.6)	23.2 (12.5)	22.0 (17.0)

- Comparison with deep neural network (DNN-HMM) on the same task
  - Both achieve similar training set tied-state classification accuracies (~60%)
  - DNN-HMM achieves much lower WER on testing set (17.1%)

- Combined with GMM-HMM training using lattice-free tied-state based MMI
  - Training set tied-state classification accuracy increased to 63%
  - No WER reduction on testing set
- Combined with GMM-HMM training using lattice-based BMMI

Discriminative Feature Transform	fMPE	FE-RDLT	
	Lattice	Lattice	Tied-State
+BMMI HMM Training	22.6 (14.7)	22.8 (14.0)	21.5 (18.9)

Discriminative Feature Transform	WE-RDLT		CE-RDLT	
	Lattice	Tied-State	Lattice	Tied-State
+BMMI HMM Training	21.9 (17.4)	21.3 (19.6)	21.8 (17.7)	20.6 (22.3)

6. Conclusions

- Both the long-span features and the contextual weight expansion are helpful in the proposed context-expanded RDLT (CE-RDLT) feature transform
- The best practice is to train the feature-space CE-RDLTs by using lattice-free, tied-state based discriminative training, while model-space GMM-HMMs are trained by using a conventional word-lattice based discriminative training method
- Future work: train the output (softmax) layer of a DNN by lattice-based discriminative training, while other layers were trained by lattice-free tied-state based discriminative training