JHU CLSP Workshop 2009

An Implementation of Sub-Space Model Based ASR

OUTLINE

- Implementation of joint subspace based model (SGMM)
- Training and evaluation on the Wall Street Journal Task

$$p(x \mid s_j) = \sum_{m=1}^{M} c_{j,m} \sum_{i=1}^{I} w_{j,m,i}\, p(x; \mu_{j,m,i}^{(s)}, \Sigma_i)$$

HMM State

Substates

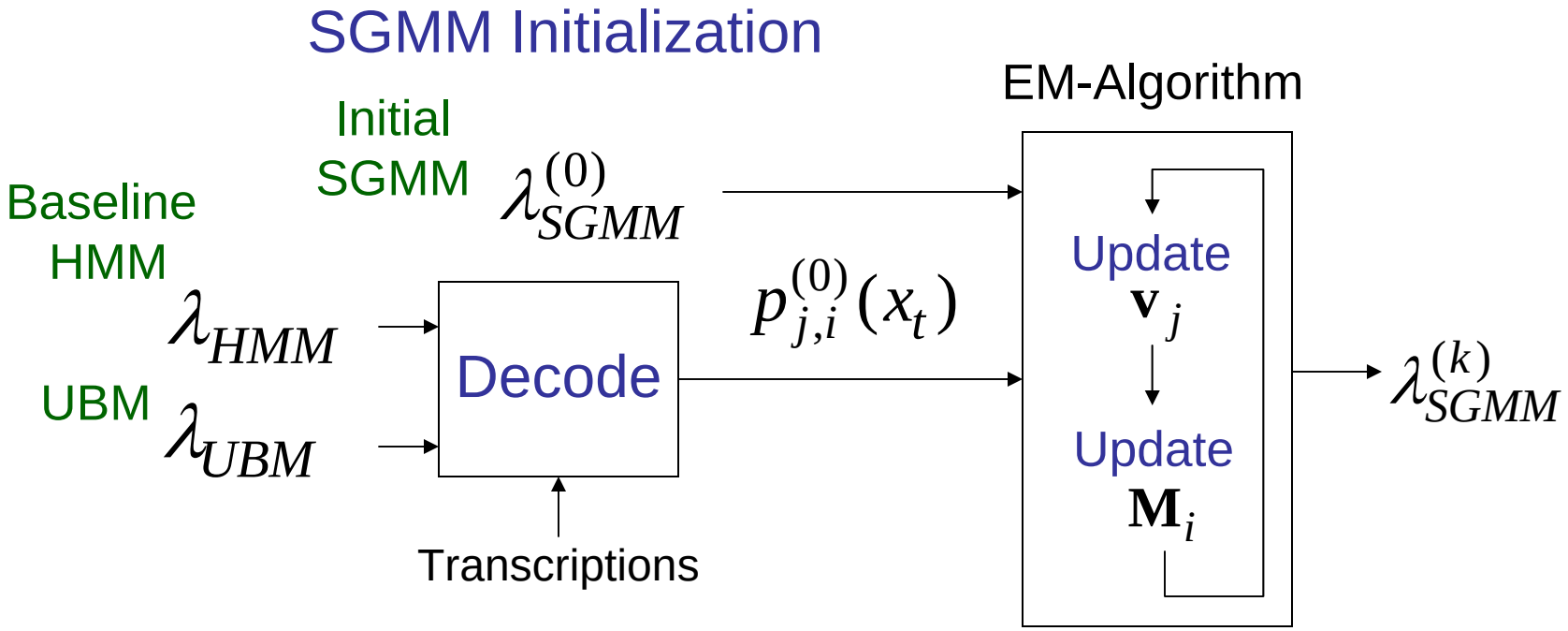
Shared Gaussians/ Subspaces

$$\mu_{j,m,i}^{(s)} = \mathbf{M}_i \mathbf{v}_{j,m} + \mathbf{N}_i v^{(s)}$$

Model Space Model (phonetic variability)

Speaker Space Model (speaker variability)

Model Subspace Estimation



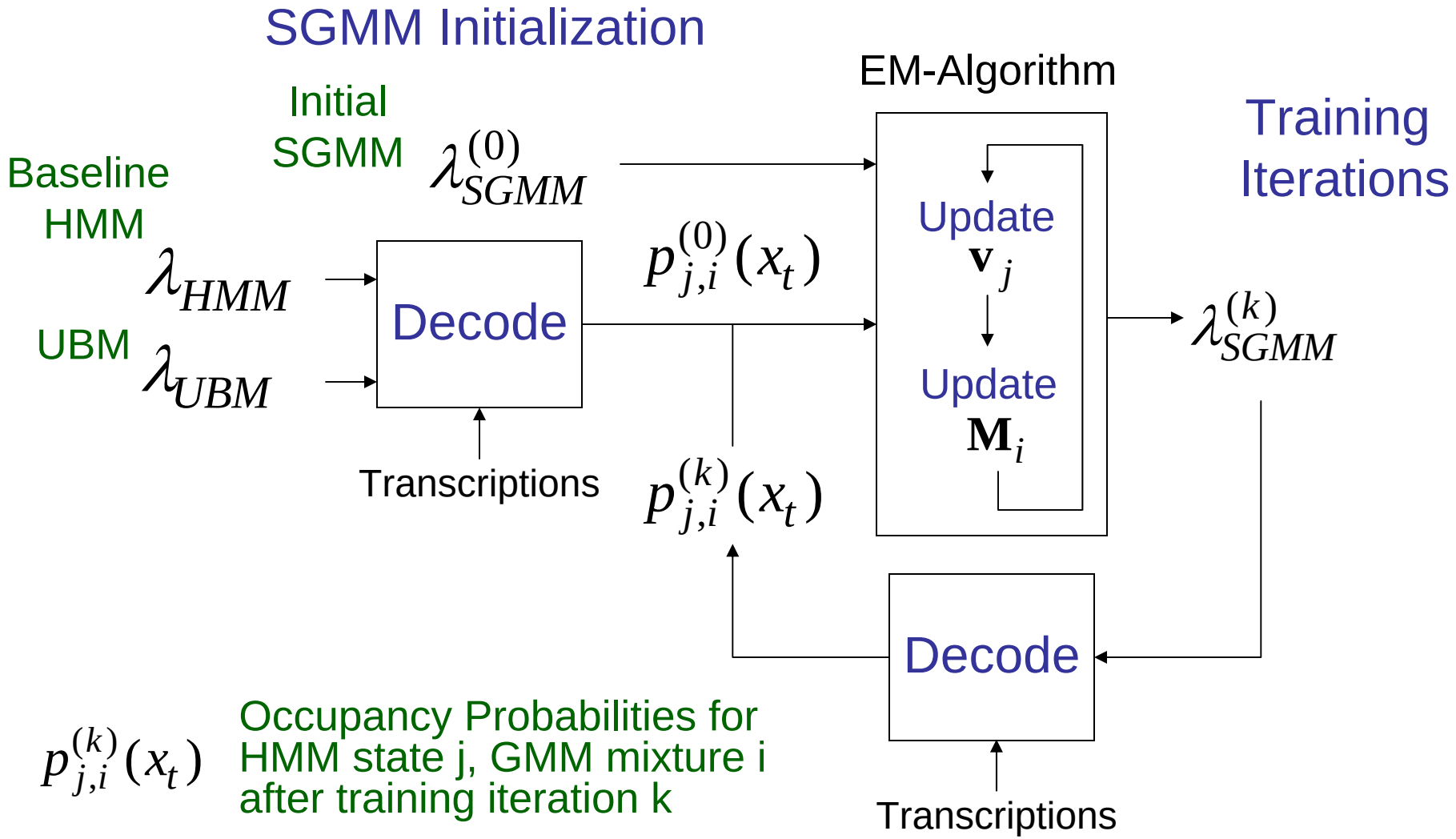
$p_{j,i}^{(k)}(x_t)$ Occupancy Probabilities for HMM state j, GMM mixture i after training iteration k

Training Joint Subspace SGMM Model

$$p(x \mid s_j) = \sum_{m=1}^{M} c_{j,m} \sum_{i=1}^{I} w_{j,m,i}\, p(x; \mu_{j,m,i}^{(s)}, \Sigma_i) \qquad \mu_{j,i}^{(s)} = \mathbf{M}_i \mathbf{v}_j + \mathbf{N}_i v^{(s)}$$

- Initialization:
  - Initial Baseline HMM model λ_{HMM}
  - Pool of full covariance Gaussian mixtures λ_{UBM}
  - Randomly initialized SGMM projection matrices $\mathbf{M}_i^{(0)}$
- Iteratively train model space model parameters
  - EM updates: $\mathbf{M}_i^{(k)}$ $\mathbf{v}_{j,m}^{(k)}$ $c_m^{(k)}$
  - Not estimating weight vectors $w_{j,m,i}$
  - Increase number of sub-states by randomly perturbing weight vectors
- Iteratively train speaker space parameters from initial model space parameters
  - EM updates: $\mathbf{N}_i^{(k)}$ $v^{(s)}$

Model Subspace Estimation



Wall Street Journal Task

- Read Speech Task
- 5K Bigram Language Model
- Close Talking Microphone Quiet Conditions
- Training Data: 988 Speakers, 80 Hours of Speech
- Baseline HMM System:
  - 6015 States
  - 96240 Gaussians
- Baseline HMM-Based ASR Performance on Nov92 WSJ test set: 5.8 % WER
- Additional model adaptation and feature normalization gives only marginal improves over this baseline performance

SGMM Model Space Results

- Baseline HMM: 5.8% WER, ~8 Million parameters
- WER with respect to SGMM Parameterizations:

Parameter Allocation			Number of parameters			Word Error Rate
UBM Gaussians	Sub-space Dim.	Sub-States	Shared	State-Specific	Total	
256	39	1	600K	235K	835K	9.5
256	39	~2	600K	470K	1.07M	8.2
256	39	~3	600K	940K	1.54M	7.6
256	39	~4	600K	1.88M	2.48M	7.3

Discussion

- Model subspace based WER is within 15% of the WSJ HMM baseline
  - Additional tuning (and bug fixes) should make up the difference
- Subspace based speaker adaptation is currently not having an effect on WER
  - Likelihoods improve, WER does not
  - Surprising because of previous success with model space adaptation on WSJ
- Post-workshop work:
  - Clean up short-cuts made in HTK based implementation (and fix bugs in the process)
  - Evaluate effects of SGMM parameterizations on other tasks – Call Home ASR and pronunciation verification task

Thanks

- Acknowledgment: Implementation based on Cambridge University's HTK
- Thanks to BUT for use of their cluster
- Thanks for the help and advice from all of the team members:
  - Mohit Agarwal, Pinar Akyazi, Lukas Burget, Arnab Ghoshal, Nagendra Goel, Dan Povey, Petr Schwarz, Samuel Thomas

  … and McGill colleagues Yun Tang and Shou-Chun Yin
- Thanks to our hosts at the CLSP