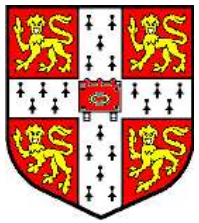
Unsupervised Bayesian Adaptation on Adaptively Trained Systems

Kai Yu & Mark Gales

May. 2006



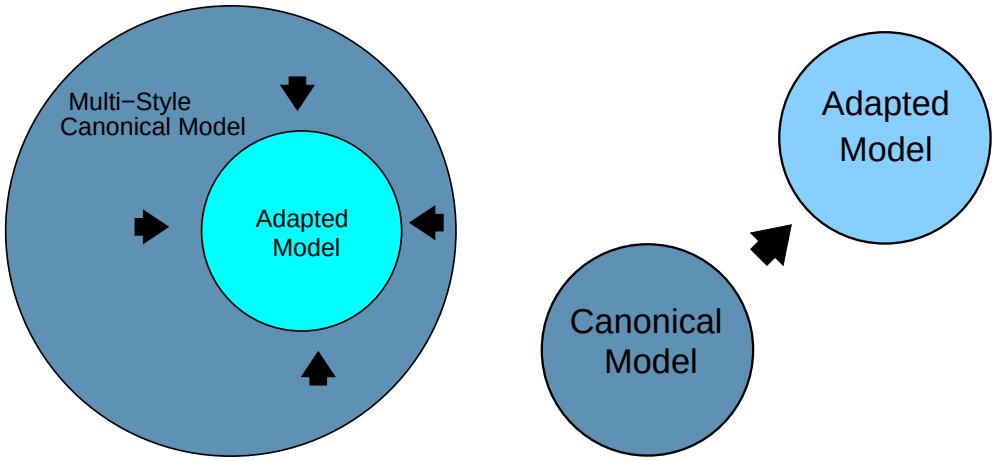
Cambridge University Engineering Department

Overview

- Adaptive training and adaptation
  - Multi-style training and adaptive training - Build model for adaptation
  - Unsupervised adaptation
- Bayesian adaptation on adaptively trained systems
  - Unsupervised adaptation from Bayesian perspective
  - Lower bound approximations
    - Point estimate and Variational Bayes
    - Lower bound inference and N-Best supervision
  - Direct approximations - Frame-independent assumption
  - Incremental Bayesian adaptation based on lower bound approx.
- Experiments on Conversational Telephone Speech recognition
  - Batch adaptation with very limited data
  - Incremental adaptation

Multi-style Training and Adaptive Training

- Adaptation requires a well built HMM system
- Training data often has various acoustic conditions - non-homogeneous

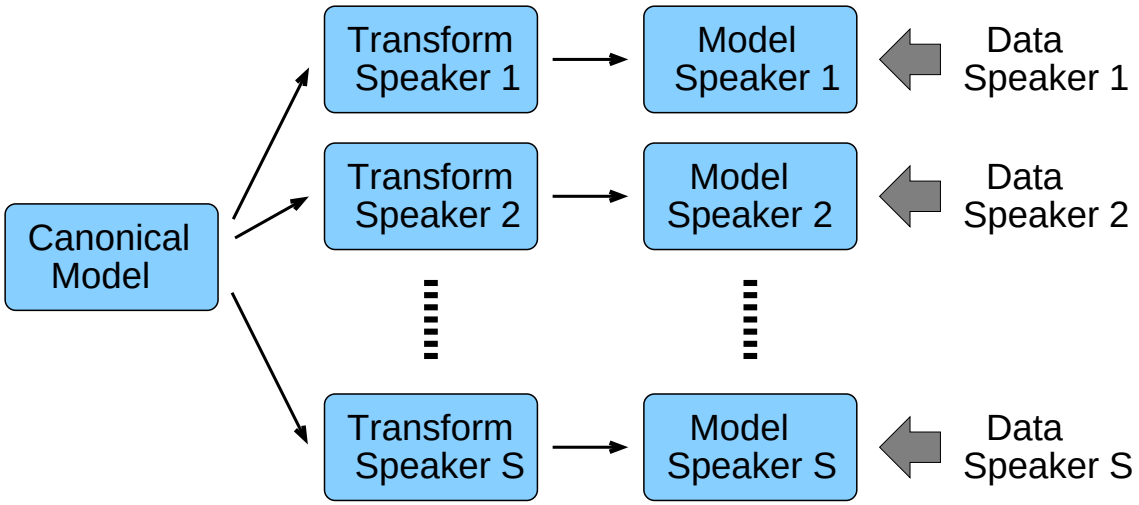


(a) Multi-Style System (b) Adaptive System

- Two schemes to build systems on non-homogeneous training data
  - Multi-style training: generic model - all kinds of variabilities
  - Adaptive training: canonical model - pure speech variability

Adaptive Training

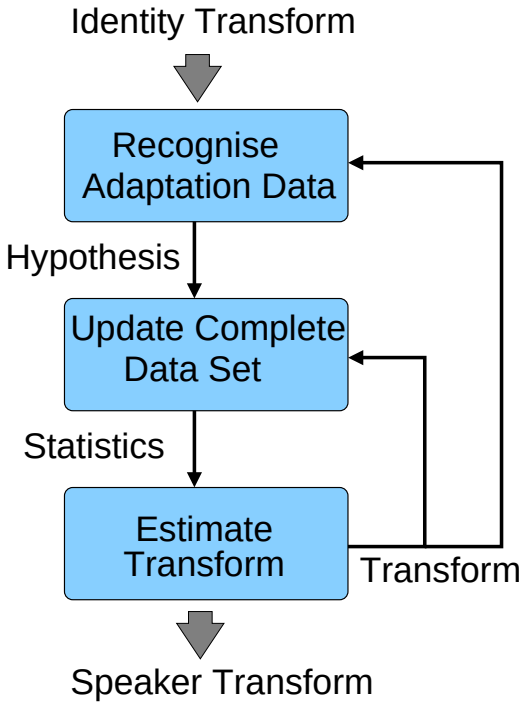
- Separate modelling of speech/non-speech variabilities



  - Canonical model - pure speech variability
  - A set of transforms
    - Represent unwanted non-speech variabilities
    - Each transform associated with one homogeneous block
  - Interleave canonical model and transform estimation

Unsupervised Adaptation

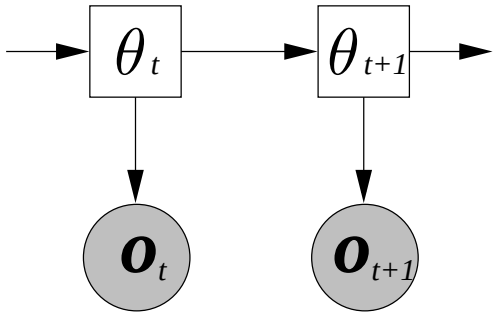
- No transcriptions available for test data
- Multi-style trained system - Directly used for decoding, adaptation optional



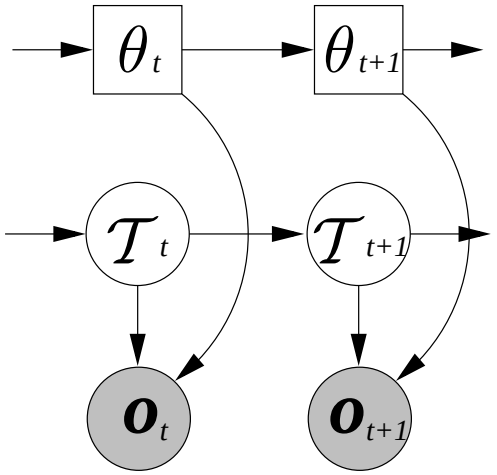
  - Unsupervised adaptation process:
    1. Initial 1-Best supervision generation:
       Decode with the multi-style model
    2. Adaptation:
       Estimate transforms using hypothesis
    3. Recognition:
       Adapt model and re-decode all data

- Adaptively trained system - Not suited for direct decoding, adaptation required
  - **How to directly use canonical model for unsupervised adaptation**

Adaptive Training From Bayesian Perspective



Standard HMM



Adaptive HMM

- Observation dependent on state/component θ and transform $\mathcal{T}$
- Transform is constant for each testing acoustic condition $\mathcal{T}_t = \mathcal{T}_{t+1}$
- Output of adaptive training with sufficient data
  - Point estimate of canonical model
  - Prior distribution of transform parameters - $p(\mathcal{T})$
    - Form - conjugate prior (e.g. single Gaussian for MLLR)
    - Hyper-parameters - estimated using empirical Bayes

Adaptation Using Bayesian Inference

- Acoustic score - marginal likelihood of the whole sequence

$$p(\mathbf{O}|\mathcal{H}) = \int_{\mathcal{T}} p(\mathbf{O}|\mathcal{H},\mathcal{T})p(\mathcal{T})\,d\mathcal{T}$$

- Two modes of Bayesian adaptation
  - Supervised mode: $p(\mathcal{T})$ updated to posterior distribution
  - Unsupervised mode: $p(\mathcal{T})$ directly used as above

- Acoustic score calculated for *every* possible hypothesis sequence
  - Observations not conditionally independent due to constant transform
  - Viterbi algorithm is not applicable
  - N-Best rescoring used

Bayesian Inference Approximations

- Bayesian integral in $p(\mathbf{O}|\mathcal{H})$ calculation intractable - approx. required
  - Real inference evidence:

$$\hat{\mathcal{H}} = \arg\max_{\mathcal{H}} p(\mathbf{O}|\mathcal{H})P(\mathcal{H})$$

  - Practical inference evidence: approx. value used instead of real likelihood

$$\hat{\mathcal{H}} = \arg\max_{\mathcal{H}} \mathcal{A}(\mathbf{O}|\mathcal{H})P(\mathcal{H})$$

- Approximation approaches
  - Lower bound approximations: $p(\mathbf{O}|\mathcal{H}) \geq \mathcal{A}(\mathbf{O}|\mathcal{H})$
    - Point estimate (MAP/ML)
    - Variational Bayes
  - Direct approximations: $p(\mathbf{O}|\mathcal{H}) \approx \mathcal{A}(\mathbf{O}|\mathcal{H})$
    - Sampling
    - Frame-independent assumption

Lower Bound Approximations

- Lower bound to log marginal likelihood - Jensen's inequality
- Joint variational distribution of state/component sequence $\boldsymbol{\theta}$ and transform parameters $\mathcal{T}$ is introduced

$$\begin{aligned}\log p(\mathbf{O}|\mathcal{H}) &= \log \int_{\mathcal{T}} p(\mathbf{O}|\mathcal{H},\mathcal{T})p(\mathcal{T})\,d\mathcal{T} \\ &\geq \int_{\mathcal{T}} q(\boldsymbol{\theta},\mathcal{T}) \log \frac{p(\mathbf{O},\boldsymbol{\theta}|\mathcal{T},\mathcal{H})p(\mathcal{T})}{q(\boldsymbol{\theta},\mathcal{T})}\,d\mathcal{T}\end{aligned}$$

- Equality condition

$$q(\boldsymbol{\theta},\mathcal{T}) = P(\boldsymbol{\theta}|\mathbf{O},\mathcal{H},\mathcal{T})p(\mathcal{T}|\mathbf{O},\mathcal{H})$$

- Iterative algorithm (EM-like) employed to update $q(\boldsymbol{\theta},\mathcal{T})$

Form of Lower Bound Approximations

- Various forms of $q(\boldsymbol{\theta}, \mathcal{T})$ may be used in lower bound approx.
  - **Point estimate (MAP/ML)** - Sufficient data assumption

$$q(\boldsymbol{\theta}, \mathcal{T}) = P(\boldsymbol{\theta}|\mathbf{O}, \mathcal{H}, \hat{\mathcal{T}})\delta(\mathcal{T} - \hat{\mathcal{T}})$$

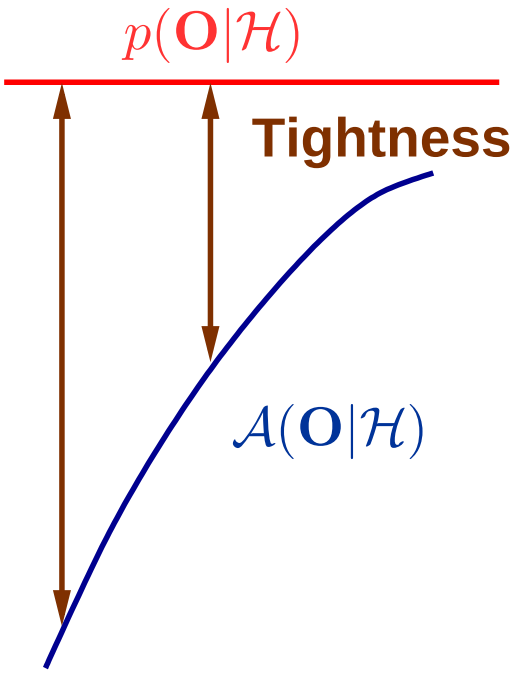
    * Transform posterior becomes Dirac delta function $\Rightarrow$ point estimate
    * Non-informative prior: MAP $\Rightarrow$ ML estimate
    * Point estimate for transform $\hat{\mathcal{T}}$ used to calculate $\mathcal{A}(\mathbf{O}|\mathcal{H})$
  - **Variational Bayes (VB)** - state/component $\boldsymbol{\theta}$ and transform $\mathcal{T}$ conditionally independent

$$q(\boldsymbol{\theta}, \mathcal{T}) = P(\boldsymbol{\theta}|\mathbf{O}, \mathcal{H})p(\mathcal{T}|\mathbf{O}, \mathcal{H})$$

    * Decoupling of $\boldsymbol{\theta}$ and $\mathcal{T}$ $\Rightarrow$ integral tractable
    * More robust due to use of real distributions
    * Non-point distribution $p(\mathcal{T}|\mathbf{O}, \mathcal{H})$ used to calculate $\mathcal{A}(\mathbf{O}|\mathcal{H})$

Tightness of Lower Bound

- Tightness of lower bound greatly affects inference



- Forms of lower bound
  - Point estimate - loose
  - Variational Bayes - tighter
- Iteration number in EM-like algorithm
  - More iterations - tighter bounds
  - Tightness controllable

1-Best vs. N-Best Supervision

- 1-Best supervision - standard adaptation concept
  - Choose 1-Best hypothesis as "supervision"
  - One transform (dist.) for all possible hypothesis
  - All lower bounds optimised using the same transform (dist.)

- N-Best supervision - obtain tight lower bound
  - Each hypothesis as "self supervision"
  - Distinct transform (dist.) for every possible hypothesis
  - Lower bound optimised using specific transform (dist.)

- Effects on Bayesian inference
  - 1-Best supv. - generally looser lower bound $\Rightarrow$ poor infer. performance
  - N-Best supv. - tighter lower bound $\Rightarrow$ better infer. performance

Lower Bound Inference

Exact Evidence	Exact Value	Value of lower bound evidence	
		1-Best (Loose)	N-Best (Tight)
$p(\mathbf{O}\|\texttt{bat})P(\texttt{bat})$	0.88	0.66	0.80
$p(\mathbf{O}\|\texttt{fat})P(\texttt{fat})$	0.84	0.78	0.78
$p(\mathbf{O}\|\texttt{mat})P(\texttt{mat})$	0.80	0.68	0.74

- Decoding using lower bound instead of exact marginal likelihood
- Assumption: lower bound approx. yields consistent rank ordering
  - 1-Best supervision is `fat`
  - Lower bound with 1-Best supervision is looser than N-Best one
  - Looser bound yields inconsistent rank ordering: $\texttt{fat} > \texttt{mat} > \texttt{bat}$
  - Tighter bound yields consistent rank ordering: $\texttt{bat} > \texttt{fat} > \texttt{mat}$
- Lower bound as tight as possible $\Rightarrow$ N-Best supervision

Direct approximations

- Direct approximate marginal likelihood - $p(\mathbf{O}|\mathcal{H}) \approx \mathcal{A}(\mathbf{O}|\mathcal{H})$
- Form of direct approximations
  - Sampling approach
    - Only applicable to systems with small number of parameters
  - Frame-independent assumption
    - Transform not constant within a homogeneous block
- Closeness of approx. value $\mathcal{A}(\mathbf{O}|\mathcal{H})$ to $p(\mathbf{O}|\mathcal{H})$ is not well controllable

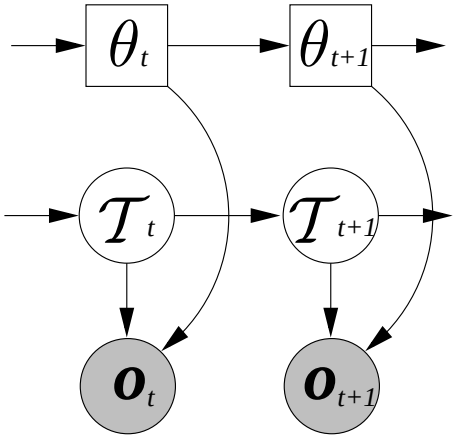
Sampling Approximation

$$p(\mathbf{O}|\mathcal{H}) \approx \frac{1}{N}\sum_{n=1}^{N} p(\mathbf{O}|\mathcal{H}, \hat{\mathcal{T}}_n)$$

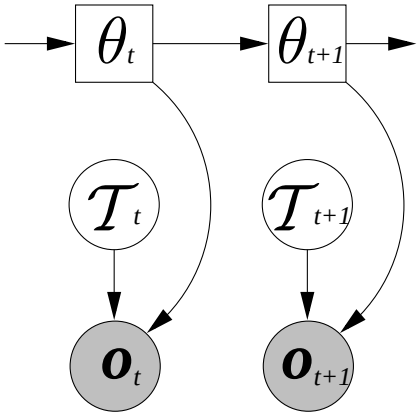
- Converge to true evidence when $N \rightarrow \infty$
- N likelihood calculations for each possible hypothesis - high cost
- Only applicable to systems with small number of parameters
  - CAT (2-clusters): 2
  - MLLR-SAT (39-dim feature): $39 \times 39 + 39$

Frame-independent (FI) Assumption

- Transform can swap at each frame



Strict Adaptation



FI Assumption

- Integral performed at each frame - Bayesian predictive distribution

$$p(\mathbf{O}|\mathcal{H}) = \int_{\mathcal{T}} \sum_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathcal{M}) \prod_t b(\mathbf{o}_t|\mathcal{T},\theta_t) p(\mathcal{T})\, d\mathcal{T}$$

$$\approx \sum_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathcal{M}) \prod_t \int_{\mathcal{T}} b(\mathbf{o}_t|\mathcal{T},\theta_t) p(\mathcal{T})\, d\mathcal{T}$$

Incremental Bayesian Adaptation

- Many tasks require results causally - e.g. dictation
- Adaptation data comes causally (on-line adaptation)
- Basic Process of incremental adaptation
  1. **Initialisation**. Use canonical model and transform prior distribution to decode the 1^{st} utterance
  2. **Propagation**. Using adaptation information from previous utterances to adapt the canonical model
  3. **Inference**. Find the best hypothesis sequence upto the current utterance
  4. Next utterance comes. Go to the propagation step 2.
- Lower bound inference used in step 3
- Key issue - **what information to propagate?**

Information Propagation Strategy

- Information propagation affects efficiency of adaptation
- **No information propagated**
  - Redo inference on all utterances
  - High computational cost
- **All information propagated**
  - Inferred hypothesis sequence - No need to redo inference on previous utterances
  - Estimated transform (distribution) propagated - align current utterance
  - Accumulated statistics propagated - No need to re-cal. statistics of the previous utterances
- Efficient recursive formulae obtained with the above propagation

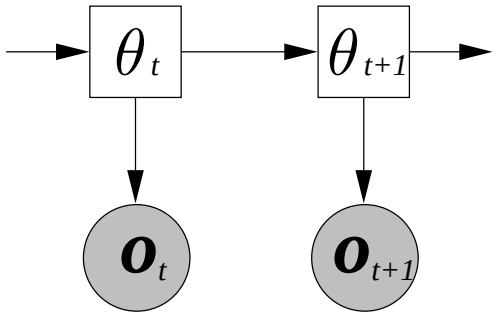
Experiments on Conversational Telephone Speech Task

- Switchboard (English): conversational telephone speech task
  - Training dataset: about 290hr, 5446spkr
  - Test dataset: 6hr, 144spkr
  - Front-end: PLP+Energy+1^{st},2^{nd},3^{rd} derivatives
  - HLDA and VTLN used
  - 150-Best list rescoring in inference
- 16 Gaussian components per state systems
  - ML and MPE speaker independent (SI) system - baseline
  - MLLR based speaker adaptive training (SAT) - ML and MPE version

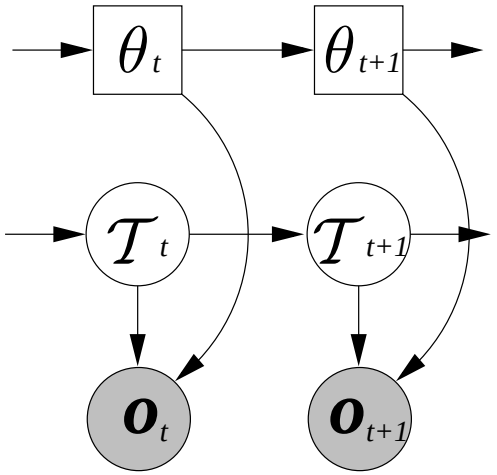
$$\hat{\boldsymbol{\mu}}^{(s)} = \mathbf{A}^{(s)}\boldsymbol{\mu} + \mathbf{b}^{(s)}$$

  - MLLR prior distribution - **Single Gaussian distribution**
  - MPE-SAT only discriminatively updated the canonical model given ML estimated transforms

Multi-style vs. Adaptive Systems



Multi-style



Adaptive

ML-SI	ML-SAT + VB
32.83	31.50

- Each utterance is a single homogeneous block - ave. length is 3.13 s
- ML-SAT+VB is the best performance for adaptively trained system
- Adaptive training significantly outperformed multi-style training by **1.3%**

Utterance Level Bayesian Adaptation - ML

Bayesian Approx	ML Train	
	SI	SAT
—	32.83	—
FI	—	32.90
ML	35.54	35.16
MAP	32.16	31.76
VB	31.77	31.50

- FI similar to SI - single Gaussian prior
- ML adaptation much worse than SI - insufficient adaptation data
- MAP improved WER - use prior information
- VB significantly better - non-point distribution, tighter bound
- SAT outperformed SI after adaptation by **0.3%** - **0.4%**

Lower Bound Tightness - N-Best Supervision

- VB trans. dist. and MAP transform updated with 1-Best and N-Best supv.
- Experiments done on ML-SAT system

Bayesian Approx.	Supervision	
	N-Best	1-Best
MAP	31.76	32.00
VB	31.50	32.04

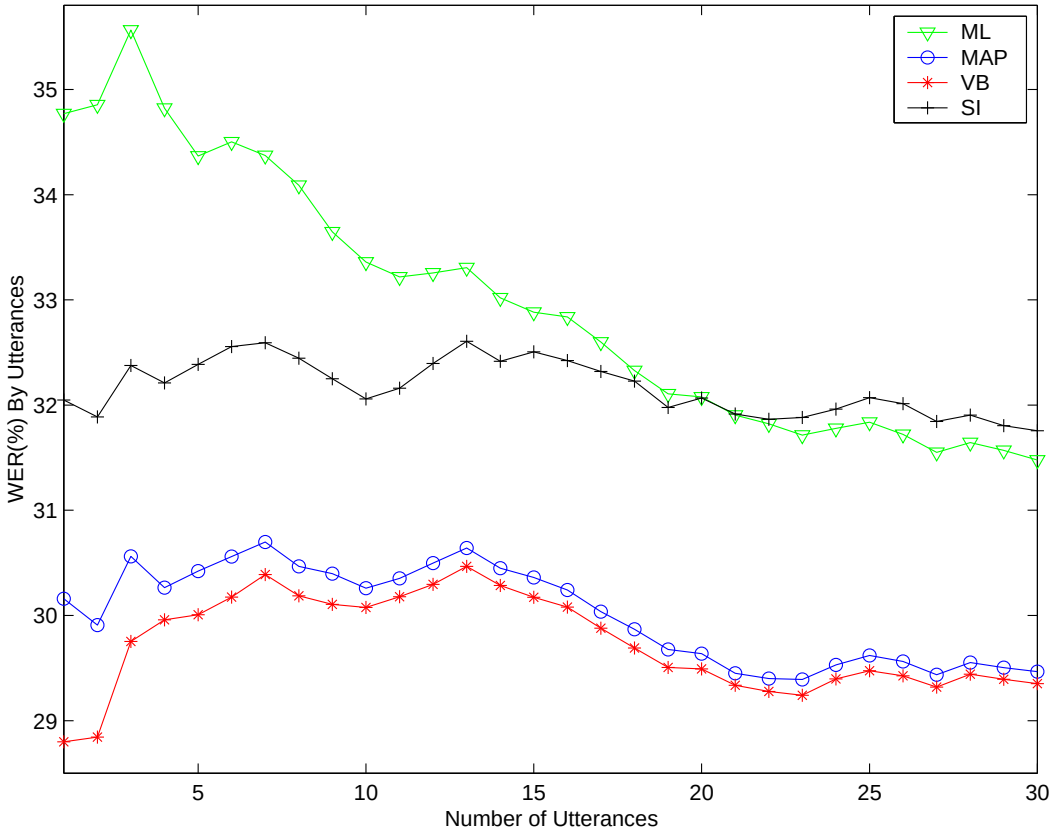
- N-Best supv. significantly better than 1-Best supv.
- VB degradation (0.5%) more than MAP degradation (0.2%)
  - VB more sensitive to supv.

Utterance Level Bayesian Adaptation - MPE

Bayesian Approx	MPE Train	
	SI	SAT
—	29.20	—
FI	—	29.74
ML	32.44	32.27
MAP	29.01	28.80
VB	28.75	28.63

- FI significantly worse than SI - **0.5%** degradation
- Similar trends for lower bound approximation as ML case
  - VB $>$ MAP $>$ SI $>$ ML
  - Corresponding gains (SAT vs. SI) reduced: **0.1%** - **0.2%**
- Reason for degradation (FI) or reduced gain (lower bound):
  - Prior distribution estimated on ML transforms
  - Prior applied in a non-discriminative way

Incremental Bayesian Adaptation



- Full Bayesian approximation significantly better at beginning
- Smaller difference between MAP and VB with more data available
- Trends for ML-SAT and MPE-SAT are similar

Incremental Bayesian Adaptation - Final Performance

Bayesian Approx	ML Train		MPE Train	
	SI	SAT	SI	SAT
ML+thresh	**31.23**	—	**27.81**	—
ML	32.23	31.84	28.86	28.72
MAP	30.92	30.40	27.65	27.47
VB	30.88	**30.31**	27.73	**27.44**

- Similar observations utterance level adaptation
  - `ML+thresh` is the baseline for incremental adaptation
  - ML not robust, VB outperformed MAP
  - Adaptive training significantly better than non-adaptive training
  - Gain of MPE adaptive training reduced
    - Poor prior and non-discriminative adaptation
- Overall performance better than utterance level adaptation

Conclusion

- Adaptive training - Deals with non-homogeneous training data
- Bayesian framework allows adaptively trained systems to be used in decoding
  - Approximations required: Lower bound/direct approximations
  - Efficient recursive formulae for incremental adaptation
- Observations in experiments
  - Adaptively trained systems significantly outperformed multi-style systems
  - Variational Bayes more robust with limited data
  - Point estimates gradually become reasonable with data amount increasing
  - Gains of discriminative adaptive systems are smaller than ML adaptive systems due to the use of ML transform prior distribution