AUTOMATIC COMPLEXITY CONTROL FOR LVCSR SYSTEMS

Xunying Liu

June 16, 2003

Cambridge University Engineering Department

Why are we doing complexity control?

- Most LVCSR systems are trained on large amounts of data.
- Many techniques alter system complexity and recognition performance.
  - State clustering
  - State distributions of Gaussian mixtures
  - Adaptation transforms sharing
  - Dimensionality reduction schemes
- Aiming at optimizing complexity to minimize word error rate for unseen data.
- Infeasible to train and evaluate individual systems' performance.
- Need automatic criterion to quickly predict performance ranking.

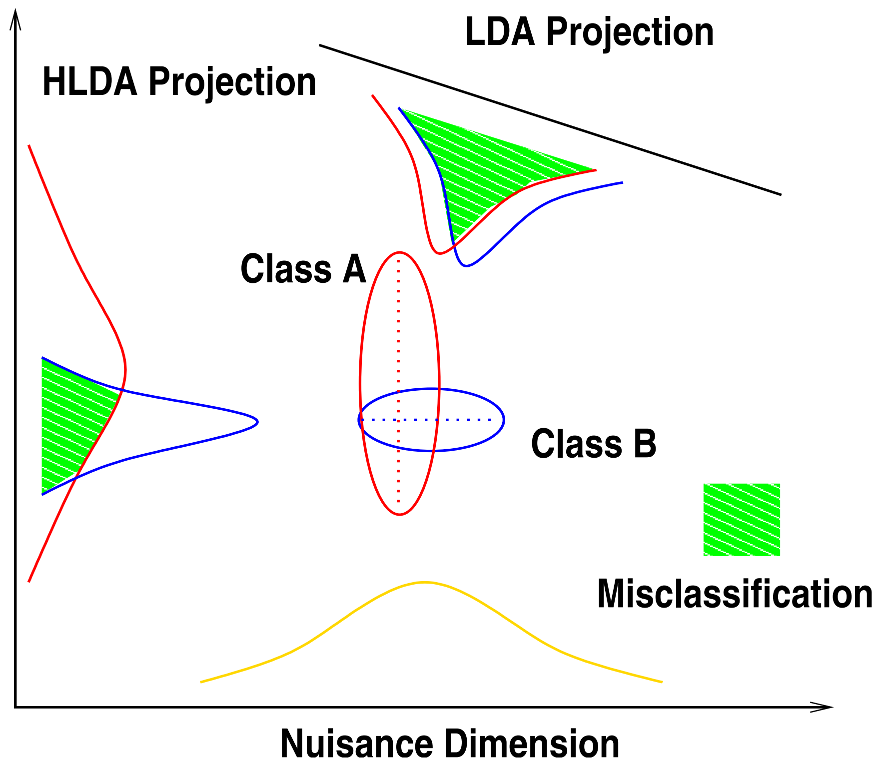
System complexity we are optimizing

- Two system complexity attributes of HLDA systems:
  - Complexity of state pdf in terms of number of Gaussians
  - Retained subspace dimensionality
- Initial aim: optimizing system complexity on global level:
  - Possible to explicitly evaluate various complexity control criteria
  - Feasible to obtain WER ranking for criterion evaluation
- Final aim: optimizing system complexity on local level:
  - Complexity of state pdf in terms of number of Gaussians
  - Infeasible to obtain WER for various systems
  - Aiming at decreasing WER given fixed system complexity

Heteroscedastic LDA (HLDA)

$$\check{\mathbf{o}} = \left[\begin{array}{c} \mathbf{A}_{[p]}\mathbf{o} \\ \mathbf{A}_{[n-p]}\mathbf{o} \end{array} \right] = \left[\begin{array}{c} \check{\mathbf{o}}_{[p]} \\ \check{\mathbf{o}}_{[n-p]} \end{array} \right]$$

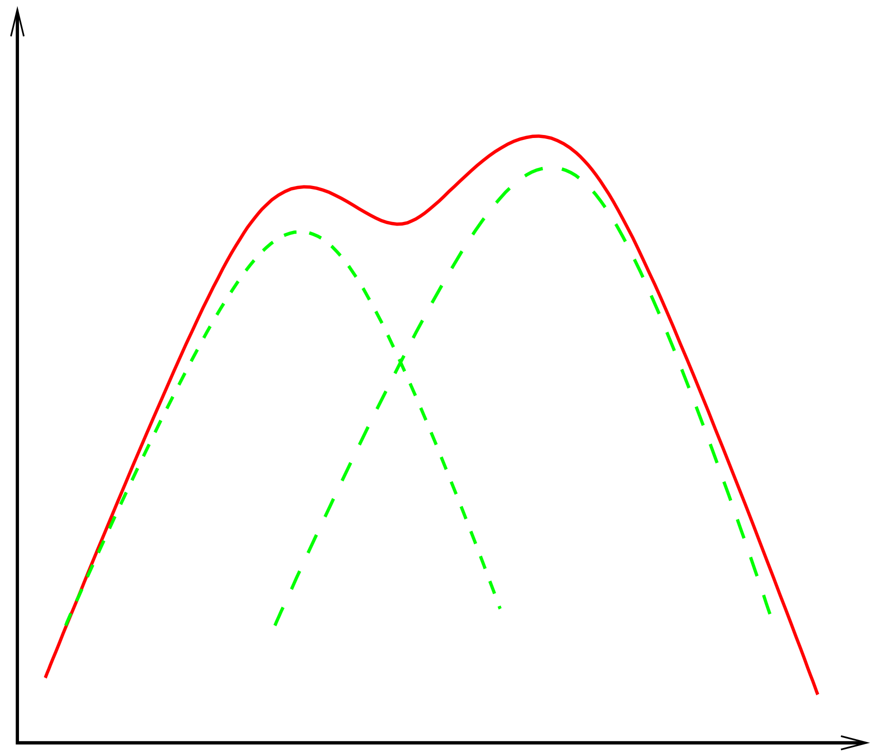
- Feature space diagonalizing and projection transform.
- Allow to incorporate higher order dynamic features.
- Iterative EM based optimization, successfully applied to LVCSR tasks.
- Need to determine optimal retained subspace dimensionality.



Mixture of Gaussians based pdf

$$b_j(\mathbf{o}) = \sum_m c_{jm} \mathcal{N}(\mathbf{o}; \boldsymbol{\mu}^{(jm)}, \boldsymbol{\Sigma}^{(jm)})$$

- Possible to approximate any form of distribution given sufficient number of Gaussian components.
- Implicitly modeling feature space correlation.
- How many components should we have then???



Two Component Mixture Model

Existing complexity control criteria

- Explicitly train up individual systems and access WER.

- Validation test using held-out data likelihood.
  - Sufficiently large and representative enough.
  - Further reducing the amount of training data available.
  - Infeasible to build individual systems for criterion evaluation.

- Bayesian evidence integration, assuming its strong correlation with held-out data likelihood.

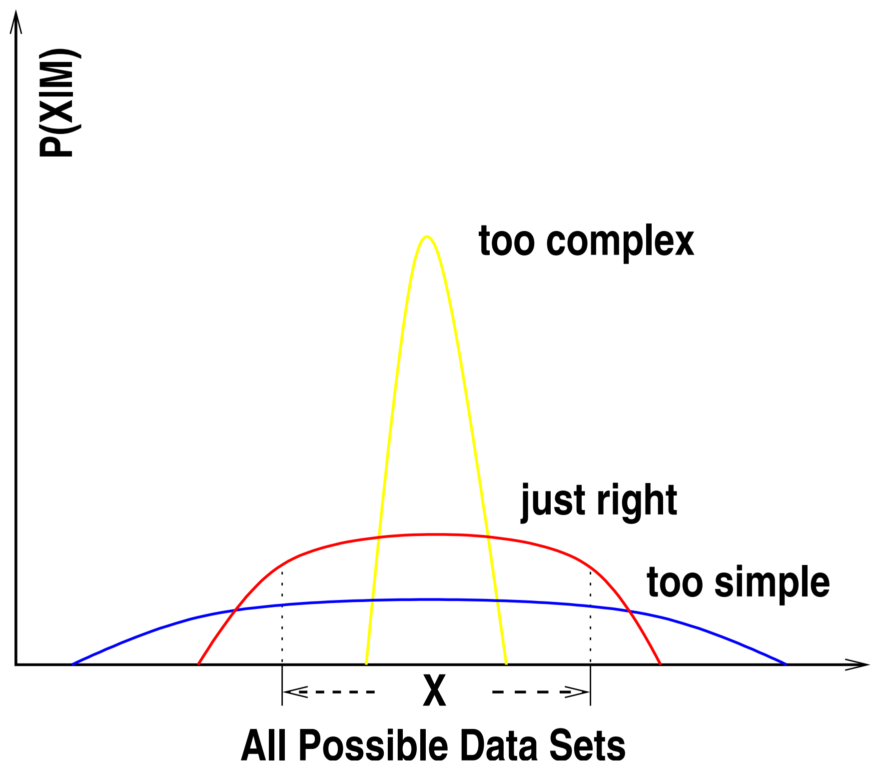
$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} \left\{ P(\mathcal{M}) \int p(\mathcal{O}|\boldsymbol{\Theta}, \mathcal{M}) p(\boldsymbol{\Theta}|\mathcal{M}) \mathrm{d}\boldsymbol{\Theta} \right\}$$

- Information theory approaches.

- Fitting complexity proportional to amount of training data, eg. VarMix

Ockham's Razor

- Important property of Bayesian evidence integral.
- Penalizes over complex model structures with bad generalization.
- Model structures with optimal complexity only model a certain range of interesting data sets.
- Over simple model structures are not powerful enough.



Approximation schemes for evidence integration

- Bayesian Information Criterion (BIC):

$$\log p(\mathcal{O}|\mathcal{M}) \approx \log p(\mathcal{O}|\hat{\boldsymbol{\Theta}}, \mathcal{M}) - \frac{k}{2} \log \mathcal{T}$$

- Laplace approximation:

$$\log p(\mathcal{O}|\mathcal{M}) \approx \log p(\mathcal{O}|\hat{\boldsymbol{\Theta}}, \mathcal{M}) - \frac{1}{2} \log \left| -\nabla^2 \log p(\mathcal{O}|\hat{\boldsymbol{\Theta}}, \mathcal{M}) \right| + \frac{k}{2} \log 2\pi$$

- Variational Approximation:

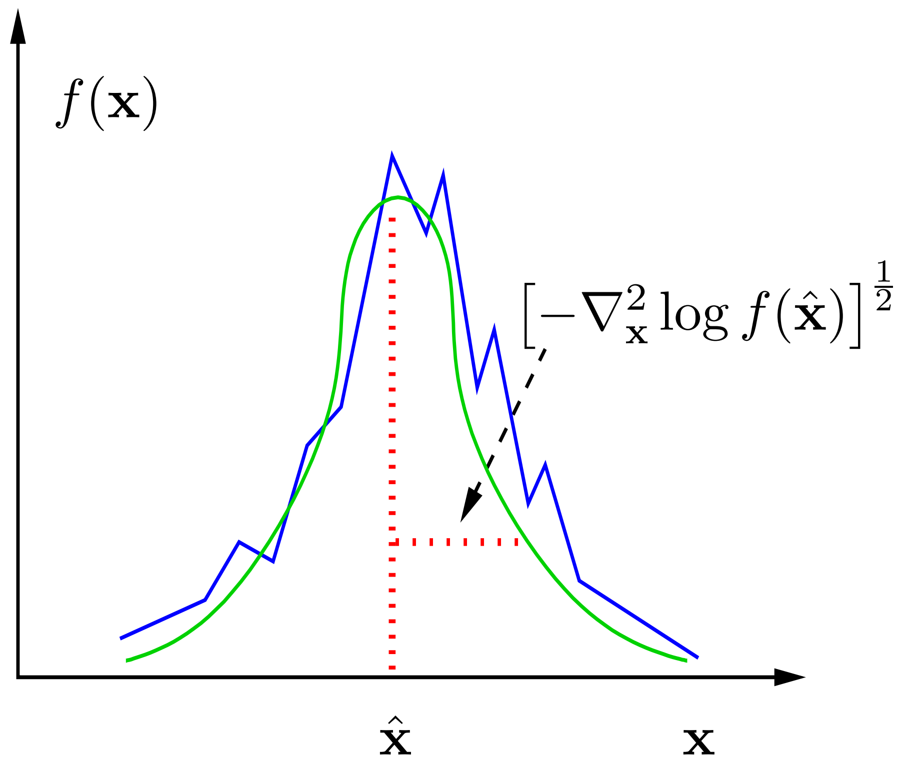
$$\log p(\mathcal{O}|\mathcal{M}) \geq \int \sum_j \mathcal{G}(\mathcal{S}_j, \boldsymbol{\Theta}) \log \frac{p(\mathcal{O}, \mathcal{S}_j, \boldsymbol{\Theta}|\mathcal{M})}{\mathcal{G}(\mathcal{S}_j, \boldsymbol{\Theta})} \mathrm{d}\boldsymbol{\Theta}$$

- Markov Chain Monte Carlo (MCMC) sampling schemes.

Laplace approximated Bayesian evidence

$$\int f(\mathbf{x})\mathrm{d}\mathbf{x} \approx \frac{(2\pi)^{\frac{d}{2}} f(\hat{\mathbf{x}})}{\left|-\nabla^2_{\mathbf{x}} \log f(\hat{\mathbf{x}})\right|^{\frac{1}{2}}}$$

- Gaussian approximation of likelihood local curvature in the parametric space.
- Computationally tractable lower bound needed to approximate true log likelihood.
- Using block diagonal Hessian matrix to reduce computation.



Laplace Approximation

Variational approximated Bayesian evidence

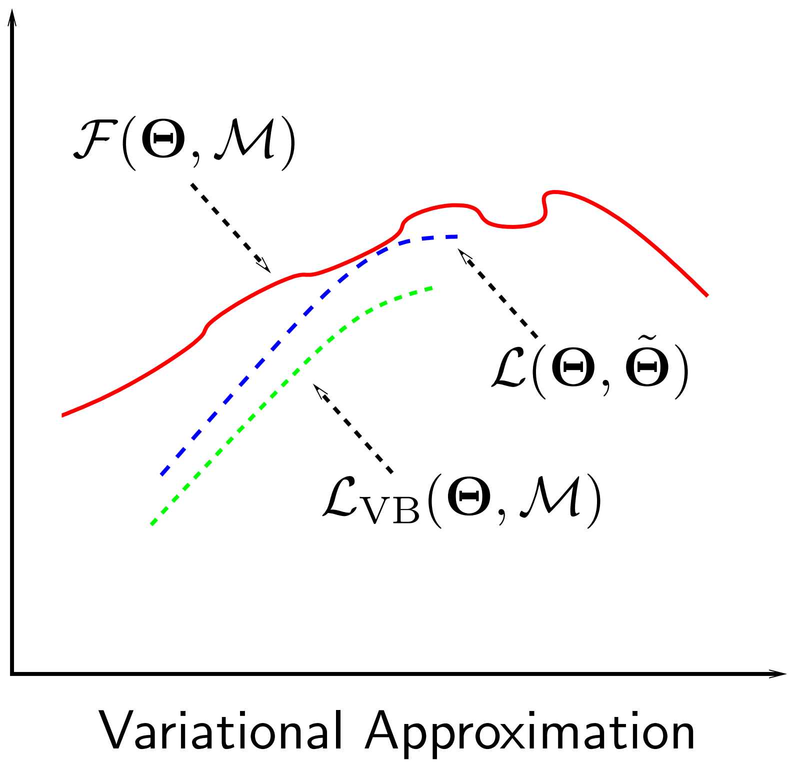
Lower bounding ML criterion marginalization

$$\log p(\mathcal{O}|\mathcal{M}) \geq \int \sum_j \mathcal{G}(\mathcal{S}_j, \boldsymbol{\Theta}) \log \frac{p(\mathcal{O}, \mathcal{S}_j, \boldsymbol{\Theta}|\mathcal{M})}{\mathcal{G}(\mathcal{S}_j, \boldsymbol{\Theta})} \mathrm{d}\boldsymbol{\Theta}$$

- Impossible to use EM strong sense auxiliary function based lower bound if joint posterior $P(\mathcal{S}_j, \boldsymbol{\Theta}|\mathcal{O}, \mathcal{M})$ is intractable.

- Using tractable approximation to $P(\mathcal{S}_j, \boldsymbol{\Theta}|\mathcal{O}, \mathcal{M})$.

- Variational lower bound may not equal to ML criterion during E step for each model instance.



Variational approximated Bayesian evidence

- Various forms of $\mathcal{G}(\mathcal{S}_j, \boldsymbol{\Theta})$ may tighten the bound.
- One choice of variational distribution:

$$\mathcal{G}(\mathcal{S}_j, \boldsymbol{\Theta}) = P(\mathcal{S}_j|\mathcal{O}, \tilde{\boldsymbol{\Theta}}, \mathcal{M})p(\boldsymbol{\Theta}|\mathcal{M})$$

- Bayesian evidence integral is then lower bounded as

$$\int p(\mathcal{O}|\boldsymbol{\Theta}, \mathcal{M})p(\boldsymbol{\Theta}|\mathcal{M})\mathrm{d}\boldsymbol{\Theta} \quad \geq \quad \mathcal{R}(\tilde{\boldsymbol{\Theta}}, \mathcal{M}) \int \exp\{\mathcal{Q}_{\mathrm{ML}}(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}})\}p(\boldsymbol{\Theta}|\mathcal{M})\mathrm{d}\boldsymbol{\Theta}$$

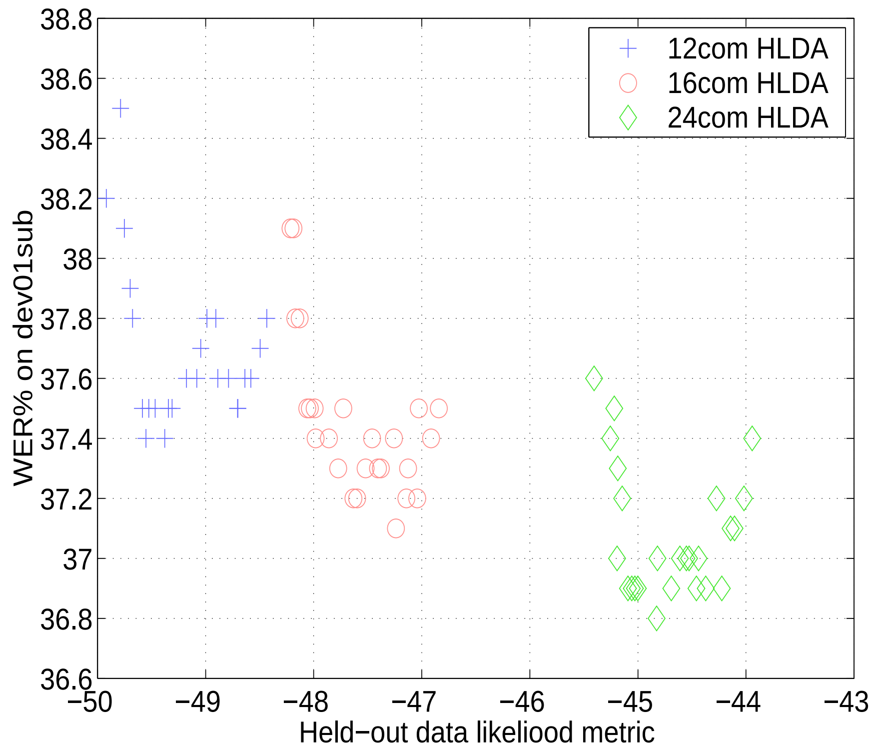
- $\mathcal{R}(\tilde{\boldsymbol{\Theta}}, \mathcal{M})$ is related to entropy of hidden variable posteriors.

$$\mathcal{R}(\tilde{\boldsymbol{\Theta}}, \mathcal{M}) = \exp\left\{-\sum_j P(\mathcal{S}_j|\mathcal{O}, \tilde{\boldsymbol{\Theta}}, \mathcal{M}) \log P(\mathcal{S}_j|\mathcal{O}, \tilde{\boldsymbol{\Theta}}, \mathcal{M})\right\}$$

- Using Laplace approximation to compute the evidence integral lower bound.

Issues with ML paradigm

- No strong correlation between criteria and WER.
- Considerable prediction error.
- Making assumption about model correctness.
- Why not use criteria directly related to recognition error???



Held-out data likelihood vs. WER%

Using discriminative training criteria

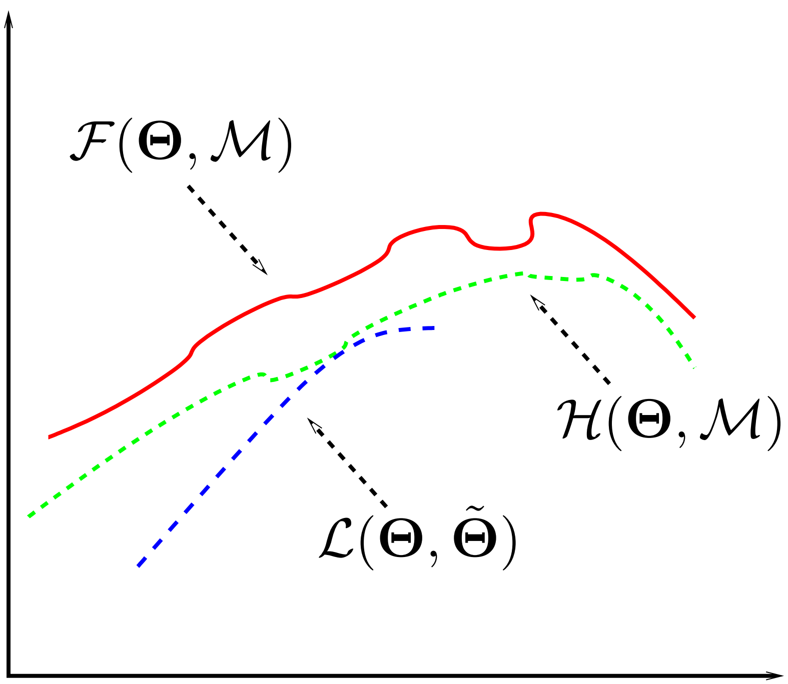
- More directly related to recognition error.
- Successfully applied for training LVCSR systems.
- Efficient lattice based implementation available.
- Criteria we will investigate:
  - Maximum Mutual Information (MMI) criterion
  - Minimum Word Error (MWE) criterion
  - Minimum Phone Error (MPE) criterion
- Can't we marginalize these criteria instead of ML criterion???

Marginalizing discriminative training criteria

Marginalizing a criterion lower bound derived using generalized EM algorithm,

$$\mathcal{L}(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}}) = \sum_j \mathcal{G}(\mathcal{S}_j, \tilde{\boldsymbol{\Theta}}) \log \frac{\mathcal{H}(\mathcal{S}_j, \boldsymbol{\Theta}, \mathcal{M})}{\mathcal{G}(\mathcal{S}_j, \tilde{\boldsymbol{\Theta}})}$$

- Initially find a criterion lower bound $\mathcal{H}(\boldsymbol{\Theta}, \mathcal{M})$ with similar curvature.
- Further lower bounding $\mathcal{H}(\boldsymbol{\Theta}, \mathcal{M})$ using generalized EM algorithm to $\mathcal{L}(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}})$.
- $\mathcal{L}(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}})$ is a strong sense auxiliary function for $\mathcal{H}(\boldsymbol{\Theta}, \mathcal{M})$ but not for $\mathcal{F}(\boldsymbol{\Theta}, \mathcal{M})$.



Marginalizing discriminative training criteria

- $\mathcal{L}(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}})$ can be related to discriminative training auxiliary functions.
  - Strong correlation between criteria and bounds in training.
  - Possible to use one set of statistics to rank multiple systems.

- This affects how to select $\mathcal{H}(\mathcal{S}_j, \boldsymbol{\Theta}, \mathcal{M})$ and $\mathcal{G}(\mathcal{S}_j, \tilde{\boldsymbol{\Theta}})$:
  - $\mathcal{H}(\mathcal{S}_j, \boldsymbol{\Theta}, \mathcal{M})$ should be related to emission probability.

$$\mathcal{H}(\mathcal{S}_j, \boldsymbol{\Theta}, \mathcal{M}) \propto p(\mathcal{O}, \mathcal{S}_j | \boldsymbol{\Theta}, \mathcal{M})$$

  - $\mathcal{H}(\mathcal{S}_j, \boldsymbol{\Theta}, \mathcal{M})$ should be related to criterion curve curvature.

$$\sum_j \mathcal{H}(\mathcal{S}_j, \boldsymbol{\Theta}, \mathcal{M}) \propto \mathcal{F}(\boldsymbol{\Theta}, \mathcal{M}) - \mathcal{F}(\tilde{\boldsymbol{\Theta}}, \mathcal{M})$$

  - $\mathcal{G}(\mathcal{S}_j, \tilde{\boldsymbol{\Theta}})$ has positive and sum to one constraint.

Marginalizing MMI criterion

- MMI criterion equivalent to posterior over the correct sentence $\mathcal{W}$.

$$\mathcal{F}_{\mathrm{MMI}}(\boldsymbol{\Theta}, \mathcal{M}) = \frac{p(\mathcal{O}, \mathcal{W}|\boldsymbol{\Theta}, \mathcal{M})}{p(\mathcal{O}|\boldsymbol{\Theta}, \mathcal{M})}$$

- Under certain constraints imposed on the parametric space we select:

$$\begin{aligned}\mathcal{H}(\mathcal{S}_j, \boldsymbol{\Theta}, \mathcal{M}) &= p(\mathcal{O}, \mathcal{S}_j|\boldsymbol{\Theta}, \mathcal{M}) \Big[\mathcal{F}_{\mathrm{MMI}}(\boldsymbol{\Theta}, \mathcal{M}) - \mathcal{F}_{\mathrm{MMI}}(\tilde{\boldsymbol{\Theta}}, \mathcal{M}) \\ &\quad + D_j \cdot p(\mathcal{O}, \mathcal{W}|\tilde{\boldsymbol{\Theta}}, \mathcal{M})\Big] \\ \mathcal{G}(\mathcal{S}_j, \tilde{\boldsymbol{\Theta}}) &= \frac{\mathcal{H}(\mathcal{S}_j, \tilde{\boldsymbol{\Theta}}, \mathcal{M})}{\sum_j \mathcal{H}(\mathcal{S}_j, \tilde{\boldsymbol{\Theta}}, \mathcal{M})}\end{aligned}$$

- $\tilde{\boldsymbol{\Theta}}$ is the "current" model parameters such that $\mathcal{F}_{\mathrm{MMI}}(\tilde{\boldsymbol{\Theta}}, \mathcal{M}) \leq \mathcal{F}_{\mathrm{MMI}}(\boldsymbol{\Theta}, \mathcal{M})$.

Marginalizing MMI criterion

The criterion lower bound $\mathcal{L}(\mathbf{\Theta}, \tilde{\mathbf{\Theta}})$ is tractable given sufficient statistics:

- MMI hidden variable occupancy.

$$\gamma_j^{\mathrm{MMI}}(\mathcal{O}) = P(\mathcal{S}_j|\mathcal{O}, \mathcal{W}, \tilde{\mathbf{\Theta}}, \mathcal{M}) - P(\mathcal{S}_j|\mathcal{O}, \tilde{\mathbf{\Theta}}, \mathcal{M}) + D_j \cdot p(\mathcal{O}, \mathcal{S}_j|\tilde{\mathbf{\Theta}}, \mathcal{M})$$

- MMI auxiliary function.

$$\mathcal{Q}_{\mathrm{MMI}}(\mathbf{\Theta}, \tilde{\mathbf{\Theta}}) = \sum_j \gamma_j^{\mathrm{MMI}}(\mathcal{O}) \log p(\mathcal{O}, \mathcal{S}_j|\mathbf{\Theta}, \mathcal{M})$$

- Hidden variable specific convergence factor D_j.

Marginalizing MWE/MPE criterion

- MWE criterion equivalent to average word error.

$$\mathcal{F}_{\mathrm{MWE}}(\mathbf{\Theta},\mathcal{M}) = \frac{\sum_{\tilde{\mathcal{W}}} p(\mathcal{O},\tilde{\mathcal{W}}|\mathbf{\Theta},\mathcal{M})\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W})}{p(\mathcal{O}|\mathbf{\Theta},\mathcal{M})}$$

- $\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W})$ is word or phone level accuracy for some path $\tilde{\mathcal{W}}$.

- Under certain constraints imposed on the parametric space we select:

$$\begin{aligned}
\mathcal{H}(\mathcal{S}_j,\mathbf{\Theta},\mathcal{M}) &= p(\mathcal{O},\mathcal{S}_j|\mathbf{\Theta},\mathcal{M})\Big[\mathcal{F}_{\mathrm{MWE}}(\mathbf{\Theta},\mathcal{M}) - \mathcal{F}_{\mathrm{MWE}}(\tilde{\mathbf{\Theta}},\mathcal{M}) \\
&\quad + D_j \cdot p(\mathcal{O}|\tilde{\mathbf{\Theta}},\mathcal{M})\Big] \\
\mathcal{G}(\mathcal{S}_j,\tilde{\mathbf{\Theta}}) &= \frac{\mathcal{H}(\mathcal{S}_j,\tilde{\mathbf{\Theta}},\mathcal{M})}{\sum_j \mathcal{H}(\mathcal{S}_j,\tilde{\mathbf{\Theta}},\mathcal{M})}
\end{aligned}$$

- $\tilde{\mathbf{\Theta}}$ satisfies that $\mathcal{F}_{\mathrm{MWE}}(\tilde{\mathbf{\Theta}},\mathcal{M}) \leq \mathcal{F}_{\mathrm{MWE}}(\mathbf{\Theta},\mathcal{M})$.

Marginalizing MWE/MPE criterion

The criterion lower bound $\mathcal{L}(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}})$ is tractable given sufficient statistics:

- MWE hidden variable occupancy.

$$\begin{aligned} \gamma_j^{\mathrm{MWE}}(\mathcal{O}) \;\; = \;\; & \sum_{\tilde{\mathcal{W}}} P(\tilde{\mathcal{W}}|\mathcal{O}, \tilde{\boldsymbol{\Theta}}, \mathcal{M}) \mathcal{A}(\tilde{\mathcal{W}}, \mathcal{W}) \Big[P(\mathcal{S}_j|\mathcal{O}, \tilde{\mathcal{W}}, \tilde{\boldsymbol{\Theta}}, \mathcal{M}) \\ & - P(\mathcal{S}_j|\mathcal{O}, \tilde{\boldsymbol{\Theta}}, \mathcal{M}) \Big] + D_j \cdot p(\mathcal{O}, \mathcal{S}_j|\tilde{\boldsymbol{\Theta}}, \mathcal{M}) \end{aligned}$$

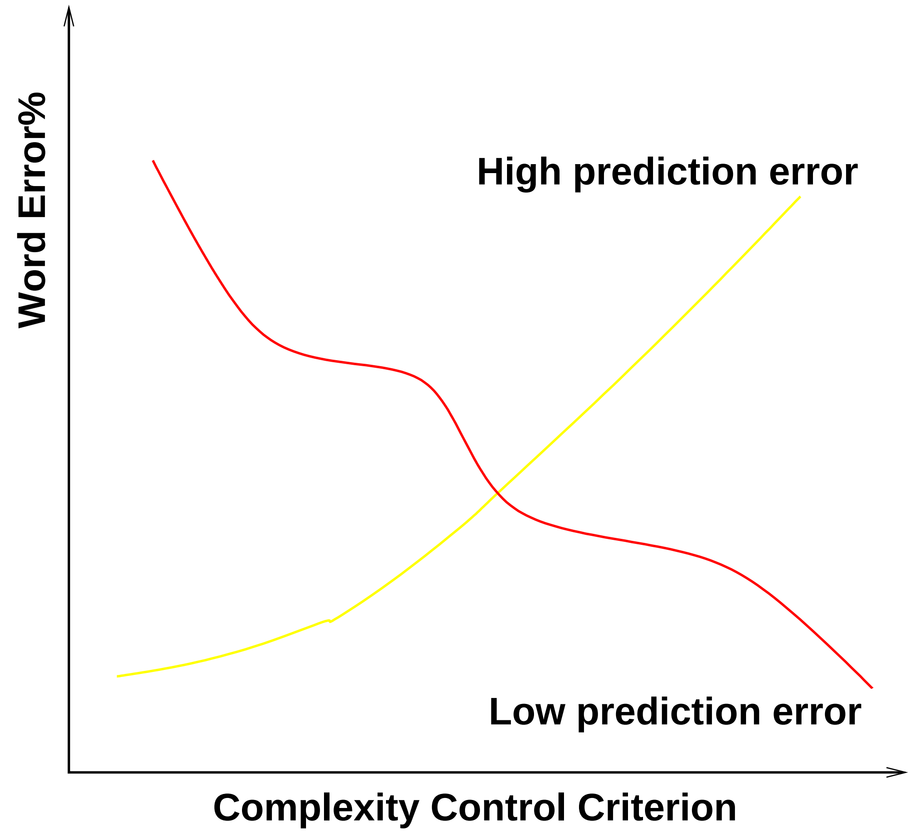
- MWE auxiliary function.

$$\mathcal{Q}_{\mathrm{MWE}}(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}}) = \sum_j \gamma_j^{\mathrm{MWE}}(\mathcal{O}) \log p(\mathcal{O}, \mathcal{S}_j|\boldsymbol{\Theta}, \mathcal{M})$$

- Hidden variable specific convergence factor D_j.

Evaluation of complexity control criteria

- Expecting strong correlation between criterion and WER.
- Increasing a good criterion should never deteriorate WER.
- Increasing a bad criterion leads to high error in WER ranking prediction.
- Intuitive and efficient to compare various criteria.



Quantizing criteria ranking prediction error

- Average ranking prediction error is computed using:
  - Amount of position shifts due to mis-ranking.
  - Pairwise WER difference between the mis-ranked systems.
  - Normalization by maximum WER difference and position shifts.

- Simple example: criterion $\mathcal{F}_2$ outperforms $\mathcal{F}_1$ in ranking prediction.
  - Correct ranking: 38.5 38.2 38.1 38.0
  - $\mathcal{F}_1$: 38.1 38.2 38.5 38.0

$$\Longrightarrow \frac{(38.5-38.1)\times(3-1)}{4\times(38.5-38.0)\times 3} = 13.3\% \quad \times$$

  - $\mathcal{F}_2$: 38.5 38.0 38.2 38.1

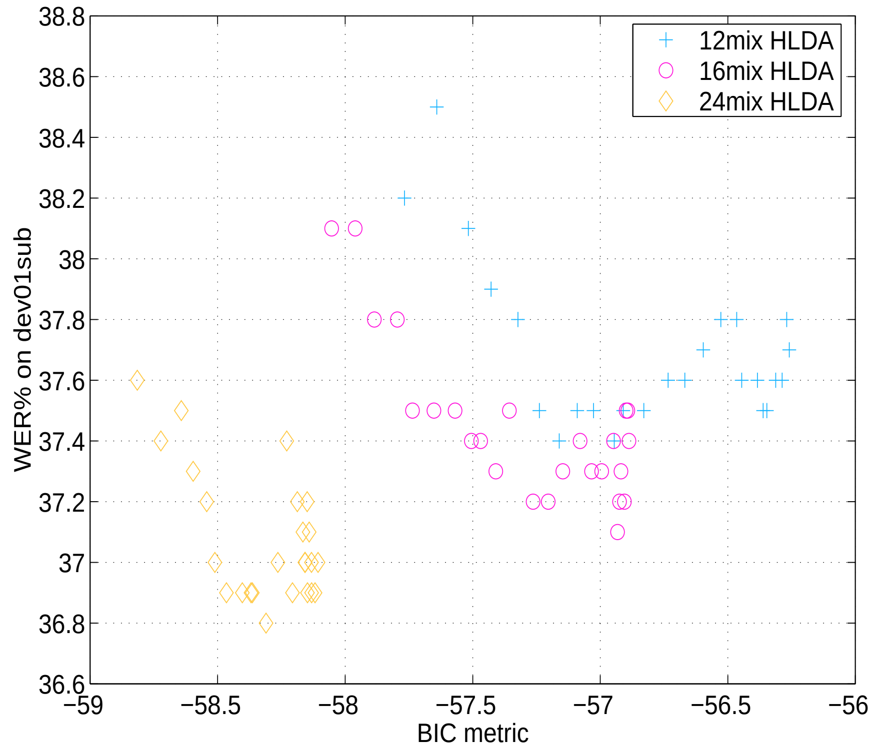
$$\Longrightarrow \frac{(38.2-38.1)\times(3-2)+(38.1-38.0)\times(4-3)}{4\times(38.5-38.0)\times 3} = 3.3\% \quad \surd$$

Switchboard Hub5 training setup

- 68 hours switchboard corpus h5train00sub
  - PLP features with differentials up to third order
  - VTLN with side based cepstral mean and variance normalization
  - Decision tree based cross word triphone
  - trigram language model for decoding
- 3 hours of test and held-out data set dev01sub
- System complexity attributes to optimize on global level:
  - Retained subspace dimensionality: $\{28, ..., 52\}$
  - Number of Gaussians per state: $\{12, 16, 24\}$
- System complexity attributes to optimize on local level:
  - Variable number of mixture components per state
  - Fixed total number of components in the system: 74k

Bayesian Information Criterion (BIC)

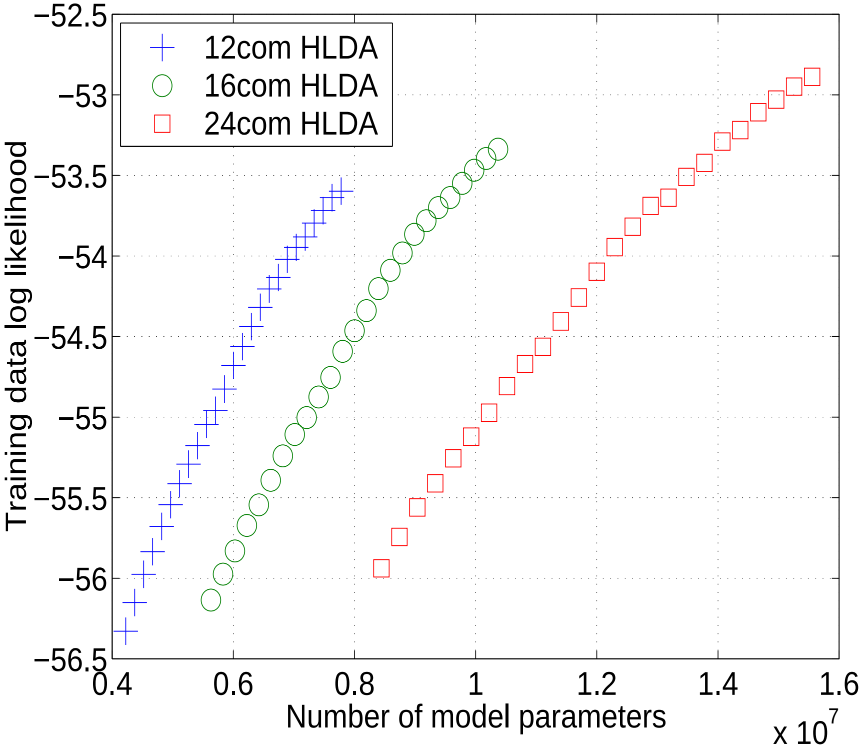
- High ranking prediction error.
- Wrong prediction of optimal number of Gaussian components.
- Favoring higher dimensional systems.
- Computationally expensive.



BIC predicting WER%

Bayesian Information Criterion (BIC)

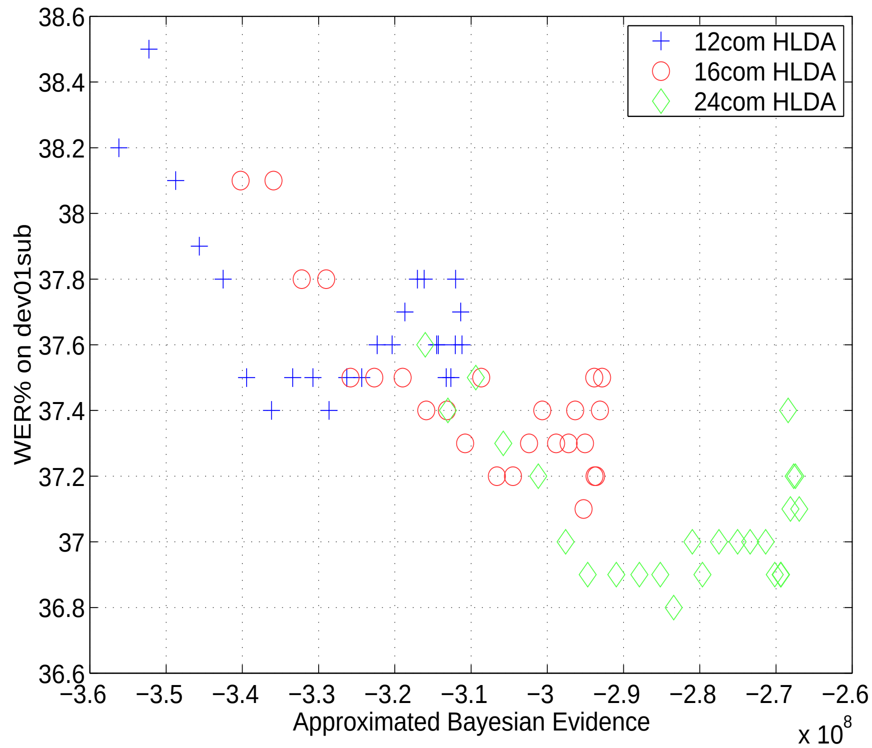
- Criterion ambiguity: non-monotonic increment of training data log-likelihood against the number of free parameters.
- Limitation for optimizing multiple system complexity attributes.
- Unsuitable for LVCSR complexity control tasks.



Training data likelihood vs. #Parm

Variational approximated Bayesian evidence

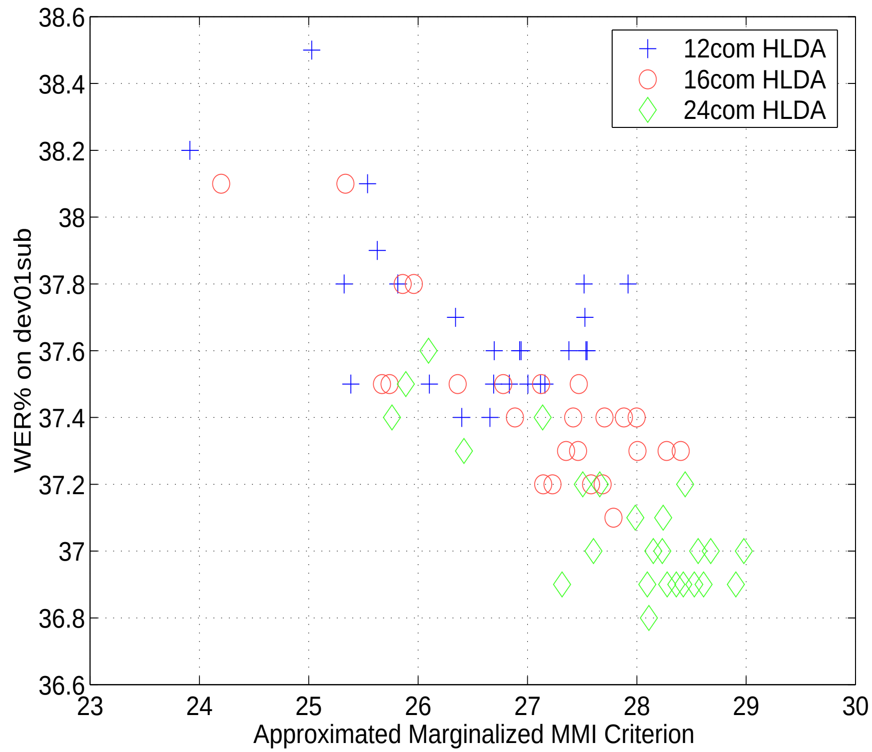
- General trend of reduced WER vs. increased criterion.
- Robust prediction for optimizing multiple system complexity attributes.
- Low prediction error given the assumptions made.
- Computationally cheaper.



Variational approximated Bayesian evidence predicting WER%

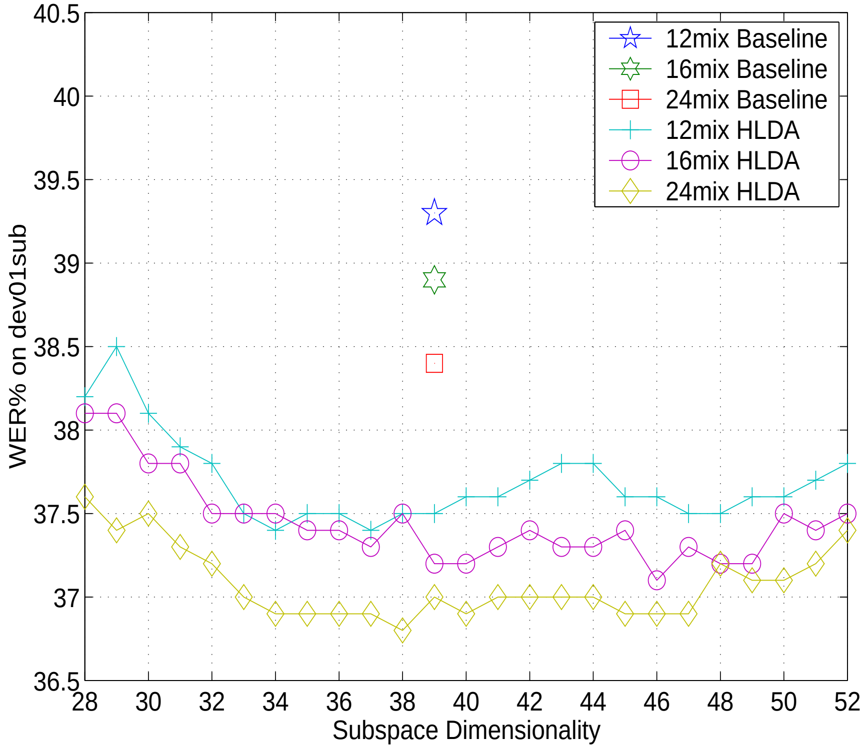
Marginalized MMI criterion

- Considerably strong correlation between criterion and WER%.
- Robust in optimizing multiple system complexity attributes.
- Low prediction error.
- Computationally cheaper.

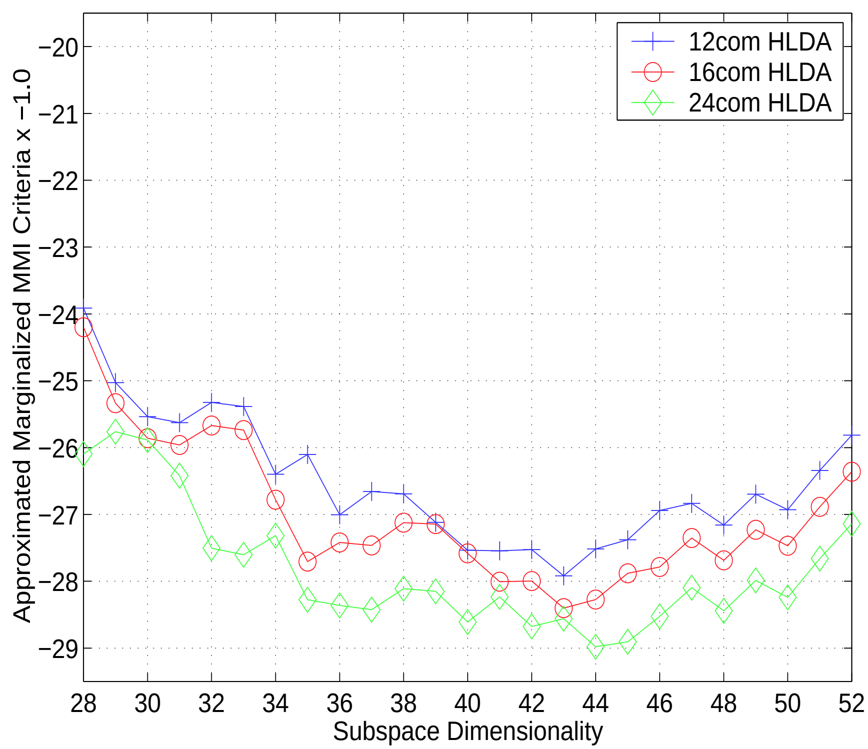


Marginalized MMI criterion predicting WER%

Marginalized MMI criterion



Word error on dev01sub



Negated marginalized MMI criterion

Closely capturing WER variation across different model structures!!!

Ranking prediction error and run time

- Average ranking prediction error and run time of various criteria across all 75 ML HLDA systems with different WER% thresholds.

	Ranking Error%			Run time
WER% threshold	0.0	0.1	0.2	($\times$RT)
BIC	48.43	48.36	47.35	1200.0
Held-out data likelihood	8.94	8.89	8.19	1237.5
Variational approximation	7.50	7.46	6.40	8.5
Marginalized MMI	7.37	7.35	5.79	29.0
WER	0.0	0.0	0.0	1575.0

- Marginalized MMI criterion outperforms all other criteria with the lowest overall ranking prediction error.

- Criterion run time of marginalized MMI criterion and variational approximated Bayesian evidence is significantly smaller.

Optimizing local complexity attributes

- Fixing the total number of Gaussians in the system using various criteria to optimize state pdf complexity on local level.

	WER% on dev01sub			
	Swbd1	Swbd2	Cellular	Total
Baseline (12com)	27.7	44.9	44.7	39.0
VarMix	27.6	45.0	44.4	38.9
Variational approximation	27.6	44.9	44.0	38.7
Marginalized MMI	27.6	44.8	44.0	38.7

- 0.3% abs gain from both variational approximated Bayesian evidence and marginalized MMI criterion.

- Most of the gain on cellular data, improvements over all three subsets.

Conclusion

- Likelihood based schemes like BIC unsuitable.
  - Considerable prediction error on recognition performance.
  - Poor performance when optimizing multiple complexity attributes.
  - No direct relations with recognition word error.

- Future work will be concentrated on
  - Marginalized discriminative training criteria.
  - Optimizing HLDA retained subspace dimensionality on local level.