AUTOMATIC COMPLEXITY CONTROL FOR HLDA SYSTEMS

Xunying Liu, Mark Gales & Phil Woodland

April 7, 2003

Cambridge University Engineering Department

Why are we doing complexity control?

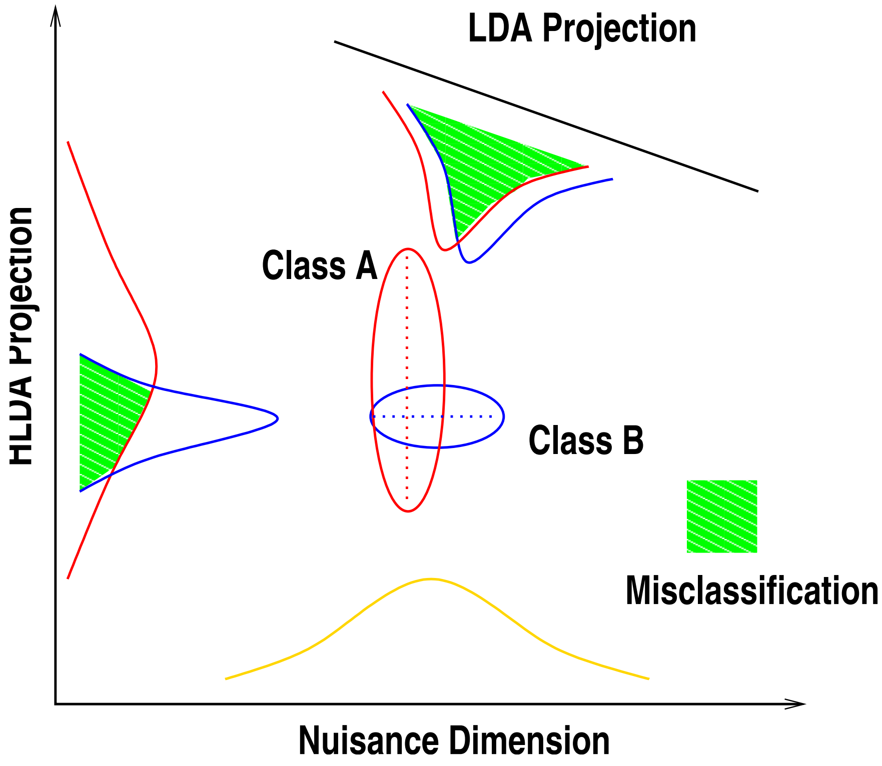
- Most of LVCSR systems are trained on large amounts of data.
- Many techniques alter system complexity and recognition performance.
  - State clustering
  - State distributions of Gaussian mixtures
  - Adaptation transforms sharing
  - Dimensionality reduction schemes
- Aiming at optimizing complexity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

HLDA Complexity

- Consider two complexity attributes of HLDA systems:
  - Complexity of state distributions in terms of number of Gaussians
  - Retained subspace dimensionality
- Initially optimizing system complexity on global level:
  - Same number of Gaussians and subspace dimensionality for all states
  - Possible to explicitly evaluate various complexity control metrics
  - Feasible to obtain WER for various systems

Heteroscedastic LDA (HLDA)

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



Existing likelihood based complexity control criteria

- 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.

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

- Information theory approaches

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:

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

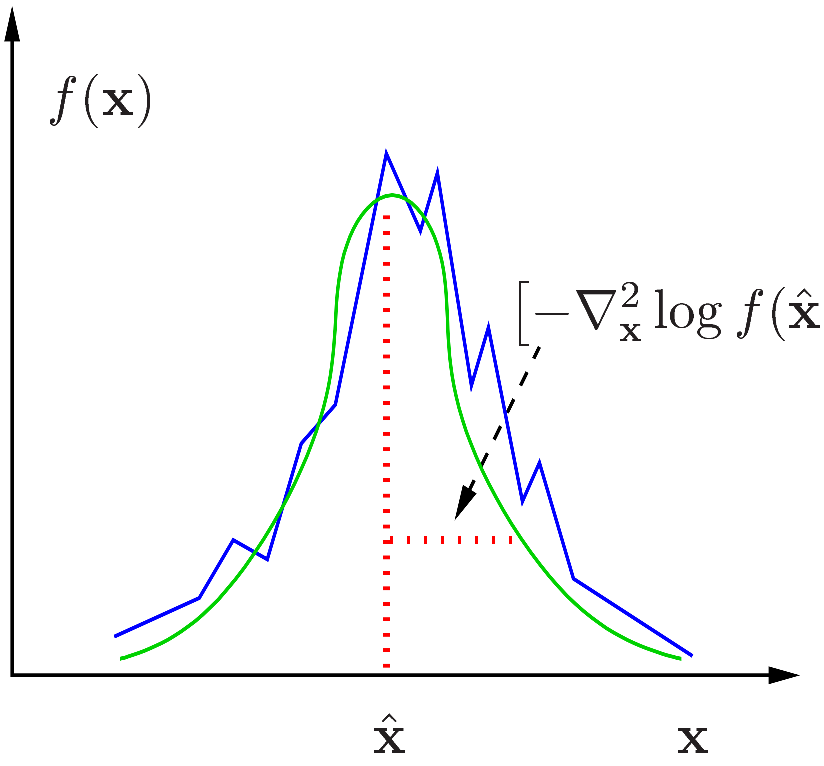
- Variational Approximation:

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

- Markov Chain Monte Carlo (MCMC) sampling schemes

Laplace approximated Bayesian evidence

- Gaussian approximation of likelihood local curvature.
- Using block diagonal Hessian matrix to reduce computation.
- Using EM auxiliary function to approximate true log likelihood.

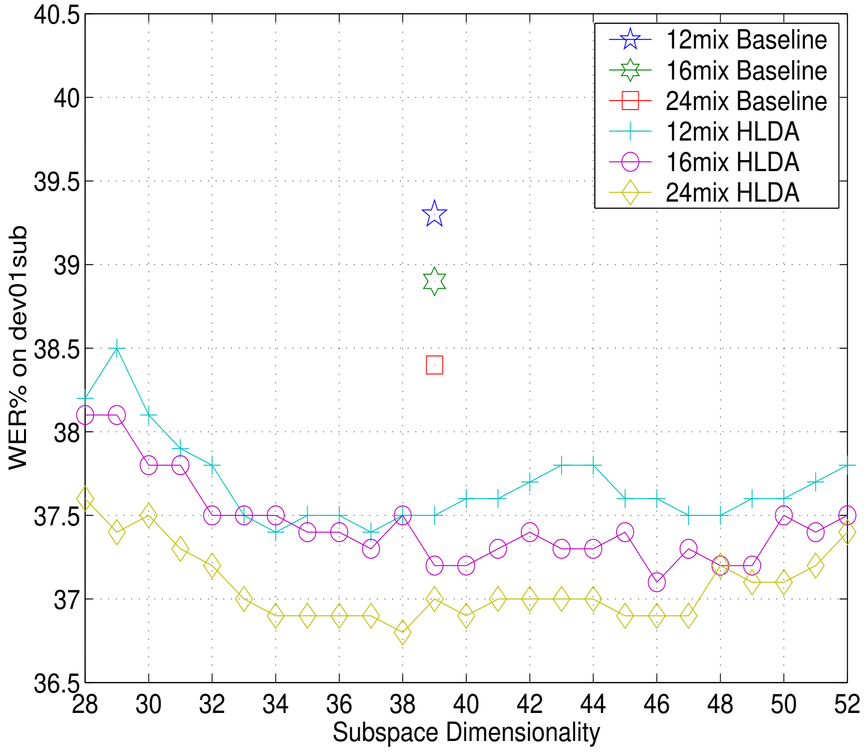


Laplace Approximation

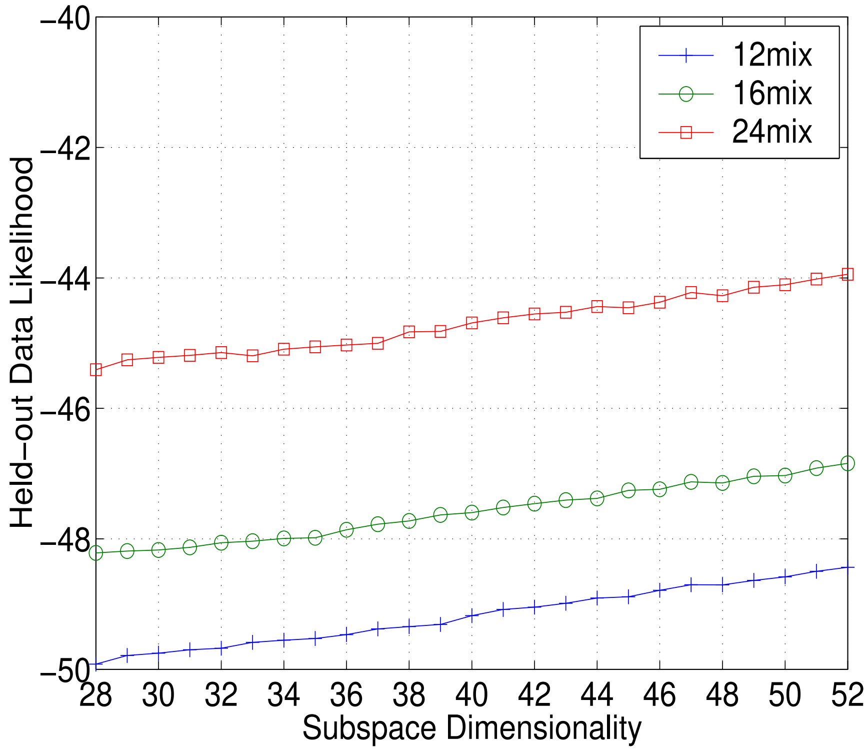
Switchboard Hub5 training setup

- 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: $\{12, 16, 24\}$

Held-out data likelihood



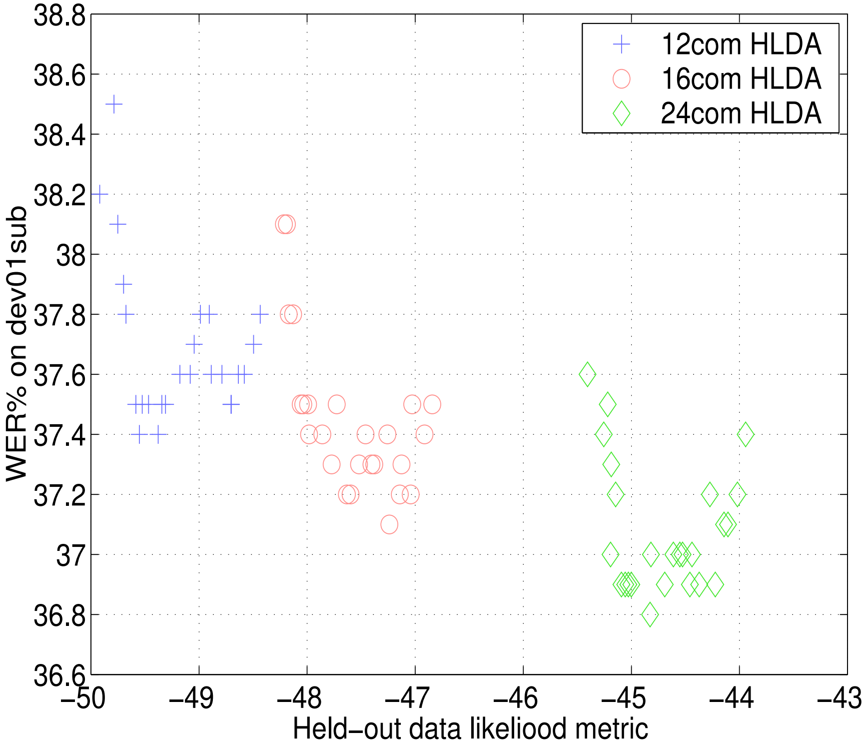
Word error on dev01sub



Held-out data likelihood metric

Held-out data likelihood predicting WER

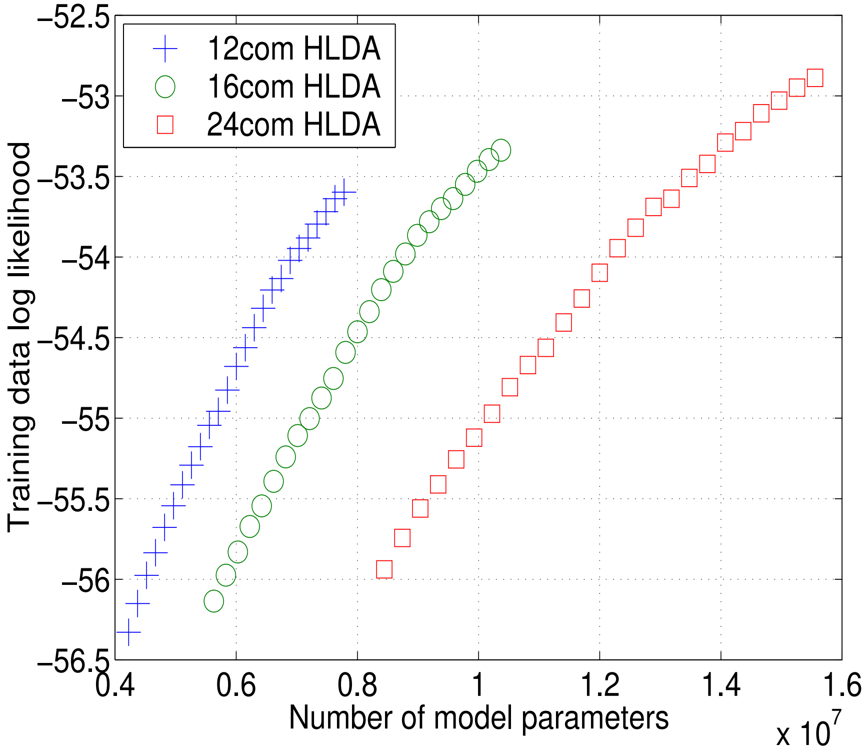
- Some correlation with WER.
- Considerable prediction error.
- Computationally expensive.
- Favoring high dimensional systems.



Held-out data likelihood vs. WER

Bayesian Information Criterion (BIC)

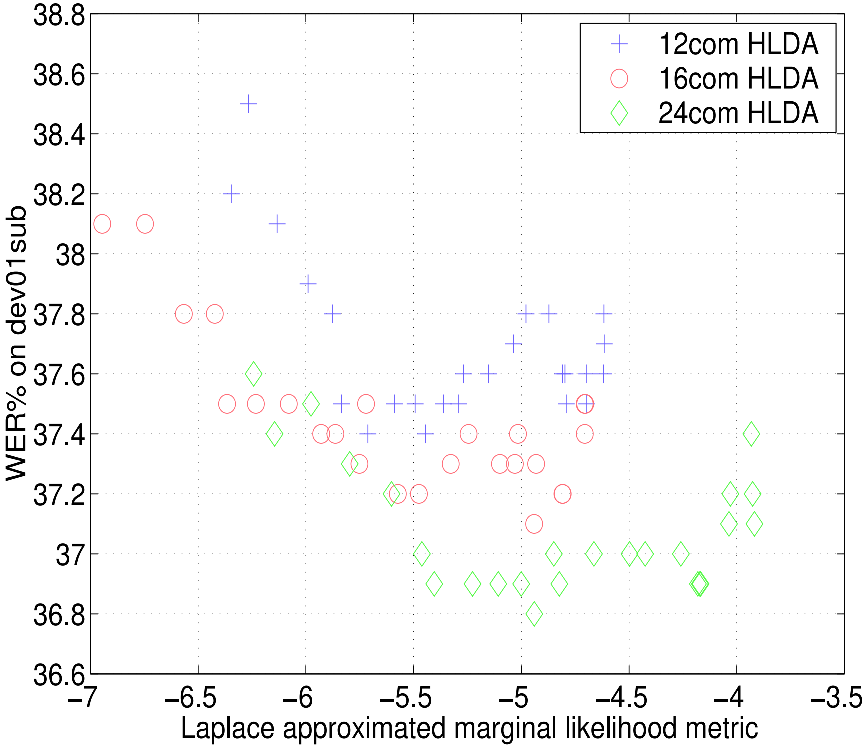
- Criterion ambiguity: non-monotonic increment of training data log-likelihood against the number of free parameters.
- Limitation for optimizing multiple system complexity attributes.
- Computationally expensive.
- Unsuitable for LVCSR complexity control tasks.



Training data likelihood vs. #Parm

Laplace 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.



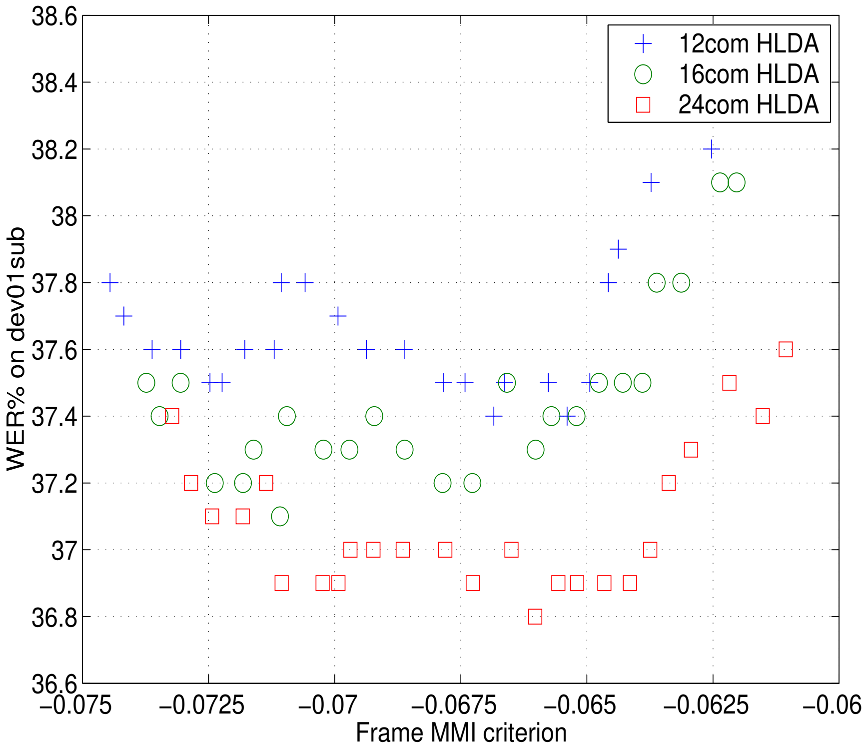
Laplace approximated Bayesian evidence

Discriminative training criteria based schemes

- More directly related to recognition error.
- Successfully applied for training LVCSR systems.
- Lattice based efficient implementation available.
- Criteria we will investigate:
  - Maximum Mutual Information (MMI) criterion
  - Minimum Word Error (MWE) criterion
  - Minimum Phone Error (MPE) criterion

MMI criterion

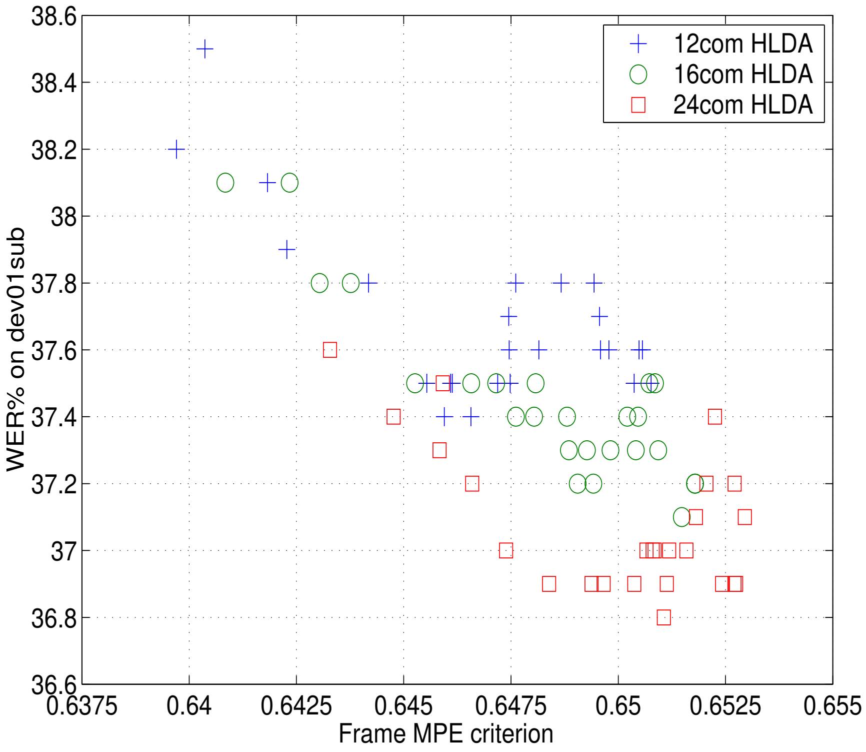
- Closely related to sentence error in a Bayesian sense
- Criterion computation dominated by outliers with low sentence posteriors
- High prediction error
- Can be approximated via a valid lower bound of a tractable form



MMI criterion vs. WER

MWE/MPE criterion

- Closely related to word error
- Low prediction error
- Feasible to implement using word or phone recognition lattices
- Can be approximated via a valid lower bound of a tractable form



MPE criterion vs. WER

Conclusion

- Likelihood based schemes 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 system complexity on local level.
  - Variational Bayesian learning techniques.