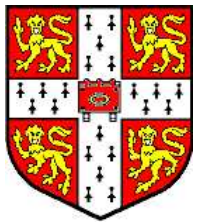
Discriminative Model Complexity Control

X. Liu & M. J. F. Gales

10th Aug 2004



Cambridge University Engineering Department

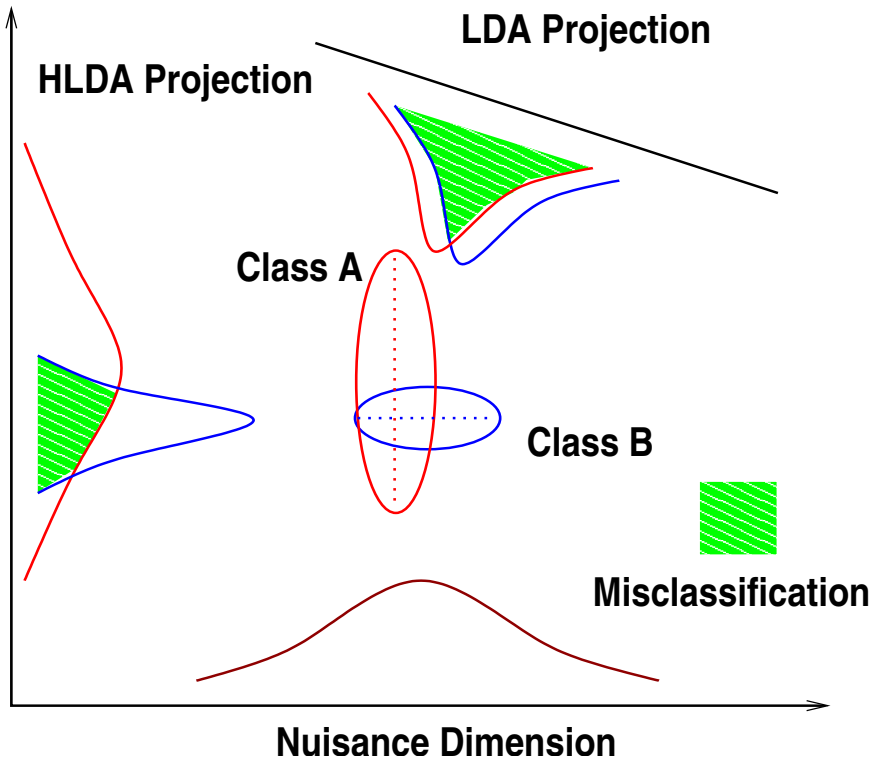
Automatic model complexity control

- Training LVCSR systems are highly complex:
  - Using large amounts of data.
  - Many techniques alter system complexity and recognition performance.
  - Infeasible to train and evaluate individual systems' performance.
- Motivations of automatic complexity control research:
  - Optimize complexity to minimize word error rate for unseen data.
  - Need automatic criterion to quickly predict performance ranking.
- Optimizing two system complexity attributes of HLDA systems:
  - *Complexity of state pdf in terms of number of Gaussians.*
  - *Retained subspace dimensionality.*

Multiple Heteroscedastic LDA (HLDA)

$$\check{\mathbf{o}}^{(r)} = \begin{bmatrix} \mathbf{A}^{(r)}_{[p]}\mathbf{o} \\ \mathbf{A}^{(r)}_{[n-p]}\mathbf{o} \end{bmatrix} = \begin{bmatrix} \check{\mathbf{o}}^{(r)}_{[p]} \\ \check{\mathbf{o}}^{(r)}_{[n-p]} \end{bmatrix}$$

- Feature space diagonalizing and transforms locally shared among groups of Gaussians.
- Allow to incorporate higher order dynamic features.
- Iterative EM based optimization, successfully applied to LVCSR tasks.
- Need to determine local retained dimensionality for multiple HLDA.

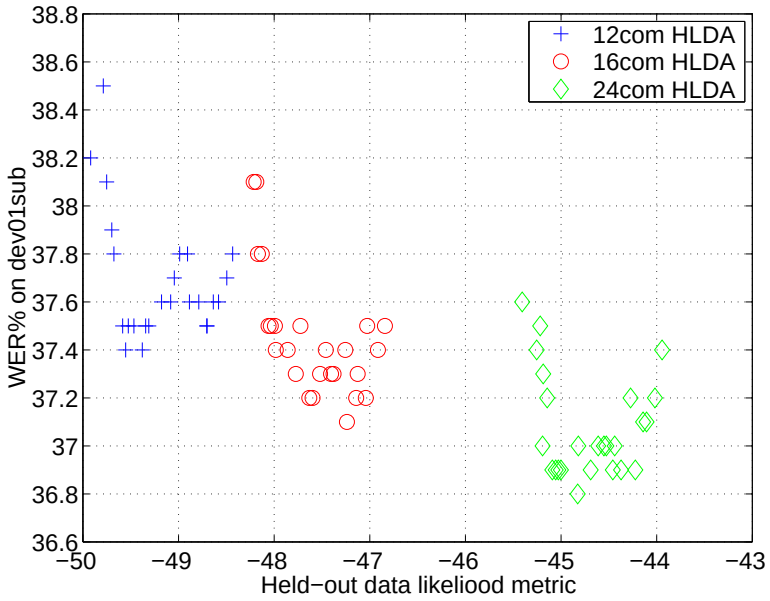


Standard Bayesian complexity control

Marginalization of standard ML criterion to optimize likelihood on held-out data.

$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} P(\mathcal{M}) \int \mathcal{F}_{\mathrm{ML}}(\lambda, \mathcal{M}) p(\lambda|\mathcal{M}) \mathrm{d}\lambda$$

- Assuming strong correlation with held-out data likelihood and WER.
- Making assumption about model correctness.
- Poor performance ranking prediction.
- Why not marginalize criteria directly related to recognition error???



Held-out data likelihood vs. WER

Marginalizing discriminative criteria

- Sensitive to outliers utterances, inappropriate to directly marginalize.
- A generic discriminative growth function,

$$\mathcal{G}(\lambda, \mathcal{M}) \quad = \quad p(\mathcal{O}|\lambda, \mathcal{M}) \left[\mathcal{F}(\lambda, \mathcal{M}) - \mathcal{F}(\tilde{\lambda}, \mathcal{M}) + C\mathcal{F}_{\mathrm{sm}}(\tilde{\lambda}, \mathcal{M}) \right]$$

  - Reduced sensitivity to outliers utterances.
  - Retaining gradient of original criteria at *current* parameterization $\tilde{\lambda}$.

- Integrated out in the parametric space for complexity control.

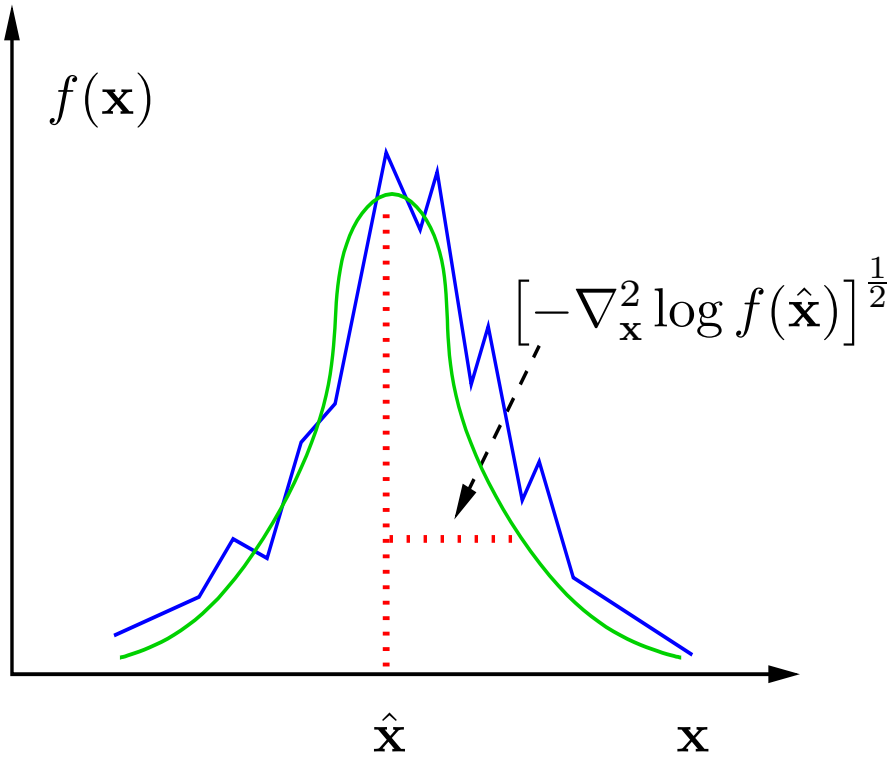
$$\hat{\mathcal{M}} \quad = \quad \arg\max_{\mathcal{M}} P(\mathcal{M}) \int \mathcal{G}(\lambda, \mathcal{M}) p(\lambda|\mathcal{M}) \mathrm{d}\lambda$$

- Marginalization computed via Laplace approximation.

Laplace approximation

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

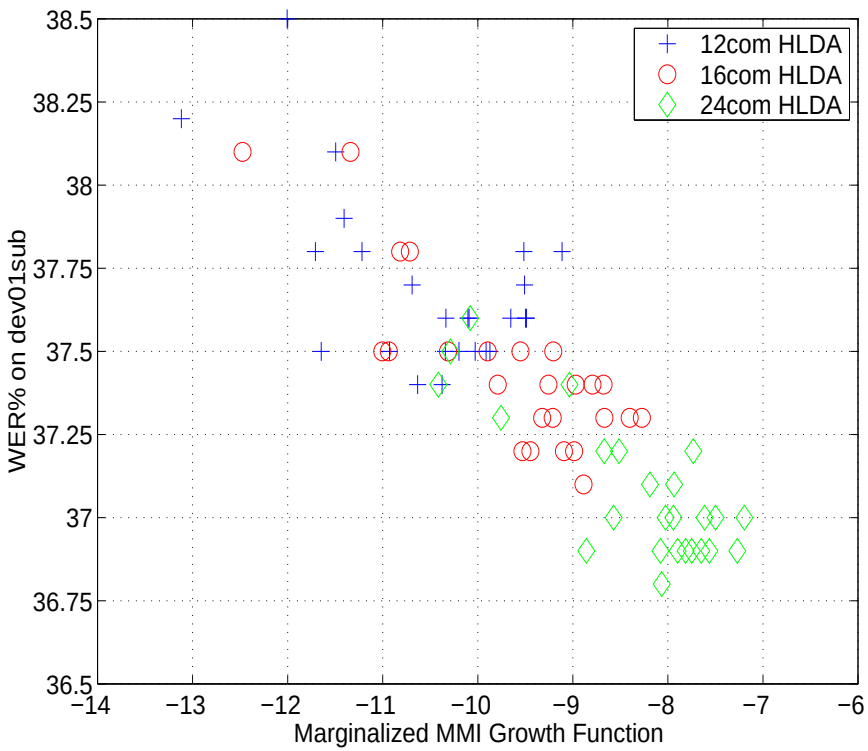
- Gaussian approximation of growth function local curvature in the parametric space.
- Computationally tractable lower bound needed to approximate true log likelihood.
- Using block diagonal Hessian matrix to reduce computation.



Laplace Approximation

Marginalizing discriminative growth functions

- Very strong correlation between criterion and WER.
- Robust in optimizing multiple system complexity attributes.
- Computationally cheaper by sharing same set of statistics among multiple model structures.
- Predicted best system only 0.2% worse than the actual best.



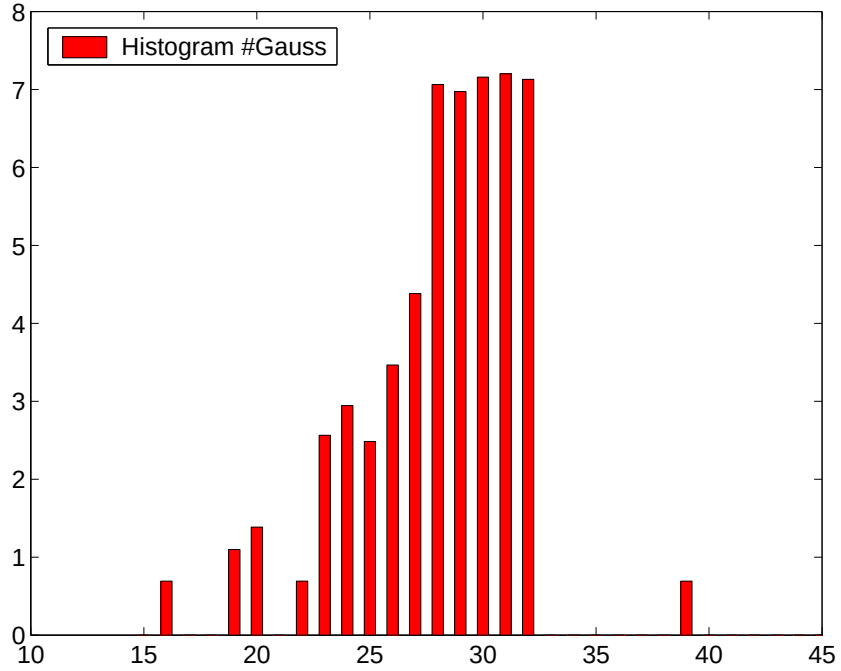
Marginalized growth function vs. WER

Implementation Issues

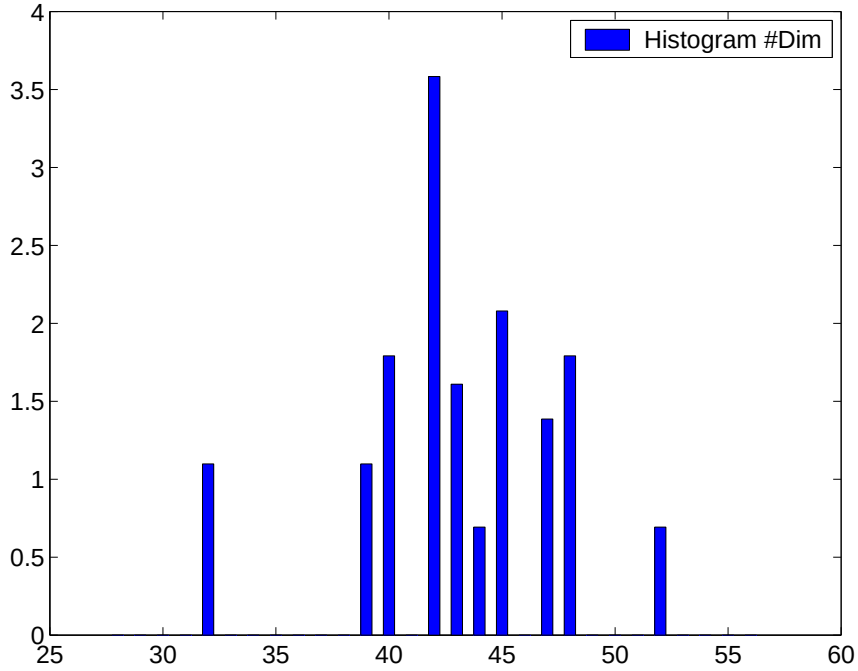
- Determining model structure:
  - Sharing same set of statistics among multiple structures.
  - Only merge pairs of Gaussians giving criterion increment.
- Using Laplace approximation:
  - Means and variances of different Gaussians assumed independent.
  - Block diagonal Hessian matrix structure.
- Adapting complexity controlled system:
  - 52x53 full matrix transforms estimated in original feature space.
  - Diagonal variance approximation based MLLR mean adaptation.
  - Full variance adaptation unavailable.

Controlling model complexity on CTS fsh2004sub

- Standard systems use 28 component *VarMix*, one global 39 dim HLDA.
- **COMCTRL** system has 65 HLDA transforms.
- Avg. 29.9 Gaussians per state and 42.6 dimensions per Gaussian.



Determining #Gauss via MMI GFunc



Determining #Dim via MPE GFunc

Development Results

System		eval03			dev04
		s25	fsh	Avg	
P3a	SAT	25.2	17.3	21.4	17.6
P3b	HLDA	25.3	18.0	21.8	18.0
P3c	SPron	25.3	18.0	21.8	17.9
P3d	**COMCTRL**	**24.7**	**17.7**	**21.3**	**17.7**
P3a-cn	SAT	24.5	17.1	20.9	17.3
P3b-cn	HLDA	24.8	17.7	21.4	17.5
P3c-cn	SPron	24.7	17.6	21.3	17.6
P3d-cn	**COMCTRL**	**24.5**	**17.5**	**21.1**	**17.6**
P3a+P3c	CNC	23.9	16.8	20.5	16.9
P3a+P3c+P3d		**23.6**	**16.5**	**20.1**	**16.7**

400 hour `fsh2004sub` 10xRT system performance

- Gains of 0.2%∼0.5% and retained after system combination.

Conclusion

- Likelihood based schemes unsuitable for LVCSR complexity control:
  - Considerable prediction error on recognition performance.
  - Poor performance when optimizing multiple complexity attributes.
  - No direct relation with recognition word error.
- Discriminative complexity control schemes:
  - Stronger relation with recognition error.
  - Low prediction error on recognition performance.
  - More compact model structures.
- Future work will be concentrated on:
  - Integrate discriminative complexity control with discriminative training.
  - Improved test set adaptation.
  - Alternative integral approximation schemes.