Automatic Model Complexity Control Using Marginalized Discriminative Growth Functions

X. Liu & M. J. F. Gales

September 3, 2003

Cambridge University Engineering Department

Automatic model 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.

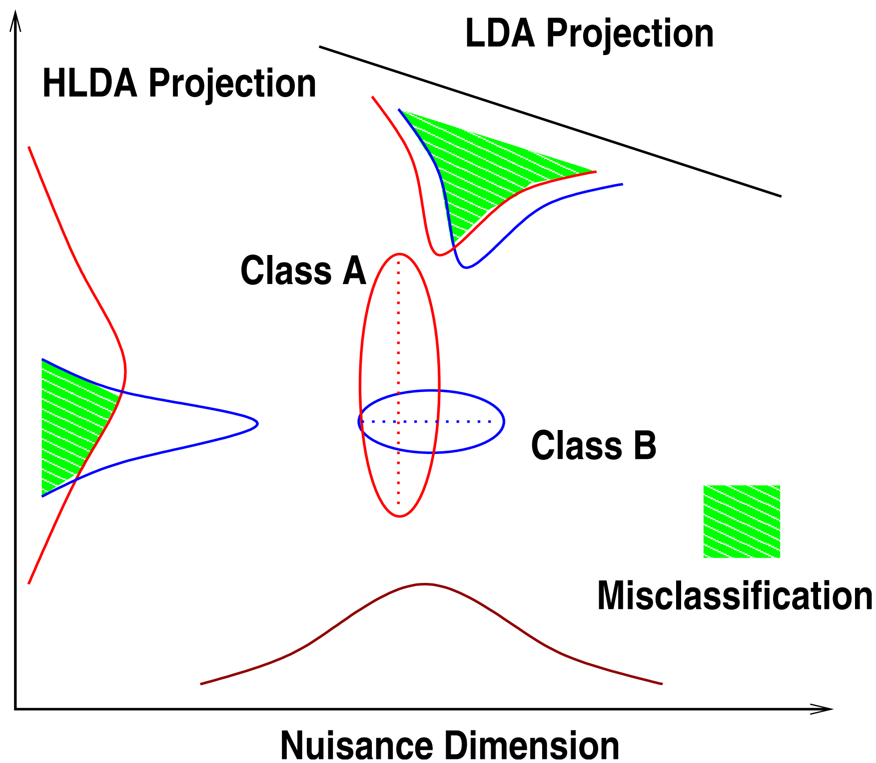
System complexity we are optimizing

- 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.
  - Transform class specific retained subspace dimensionality.
- Candidate structures to be trained using ML only during complexity control.

Multiple Heteroscedastic LDA (HLDA)

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

- Feature space diagonalizing and locally tied projection transforms.
- 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.



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.

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

- Information theory approaches.

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

Approximation schemes for evidence integration

- Bayesian Information Criterion (BIC):

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

  $\rho > 1.0$ for penalized BIC. Not suitable for optimizing multiple complexity attributes (see ICASSP03 Liu, Gales & Woodland).

- Laplace approximation:

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

- Variational Approximation:

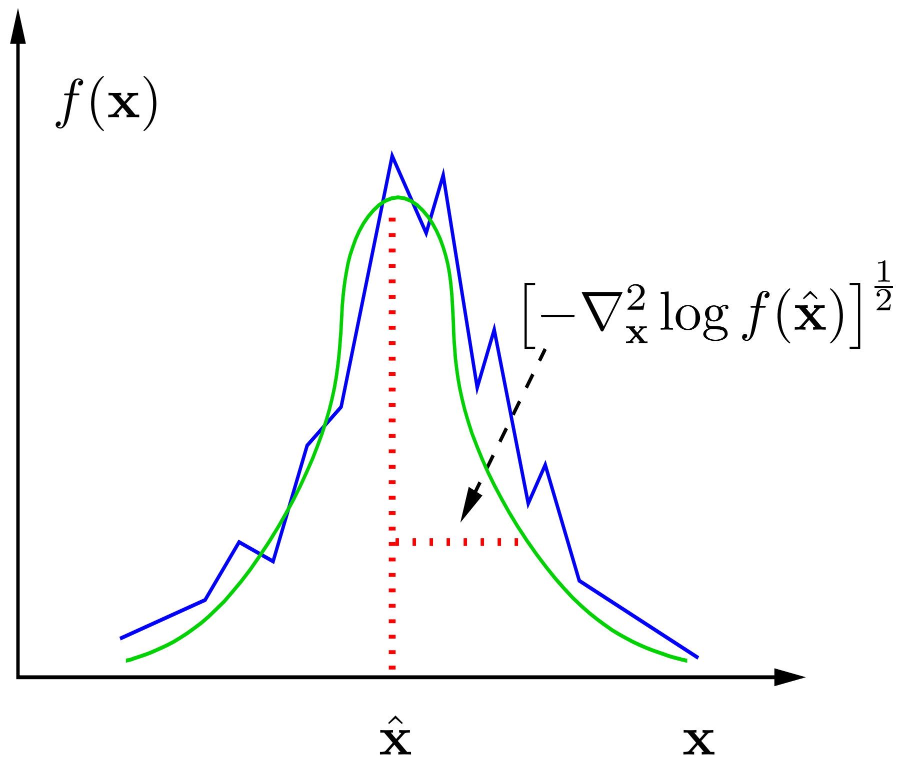
$$\log p(\mathcal{O}|\mathcal{M}) \quad \geq \quad \int \sum_{\Psi \in \{\Psi\}} \mathcal{P}(\Psi, \Theta) \log \frac{p(\mathcal{O}, \Psi, \Theta|\mathcal{M})}{\mathcal{P}(\Psi, \Theta)} \mathrm{d}\Theta$$

- 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

EM lower bound of Bayesian Evidence

- EM lower bound of ML criterion can be expressed as

$$\begin{aligned}\log \mathcal{F}_{\mathrm{ML}}(\Theta, \mathcal{M}) &\geq \log p(\mathcal{O}|\tilde{\Theta}, \mathcal{M}) + \mathcal{Q}_{\mathrm{ML}}(\Theta, \tilde{\Theta}) - \mathcal{Q}_{\mathrm{ML}}(\tilde{\Theta}, \tilde{\Theta}) \\ &= \mathcal{L}_{\mathrm{ML}}(\Theta, \tilde{\Theta})\end{aligned}$$

- Evidence may then be approximated as.

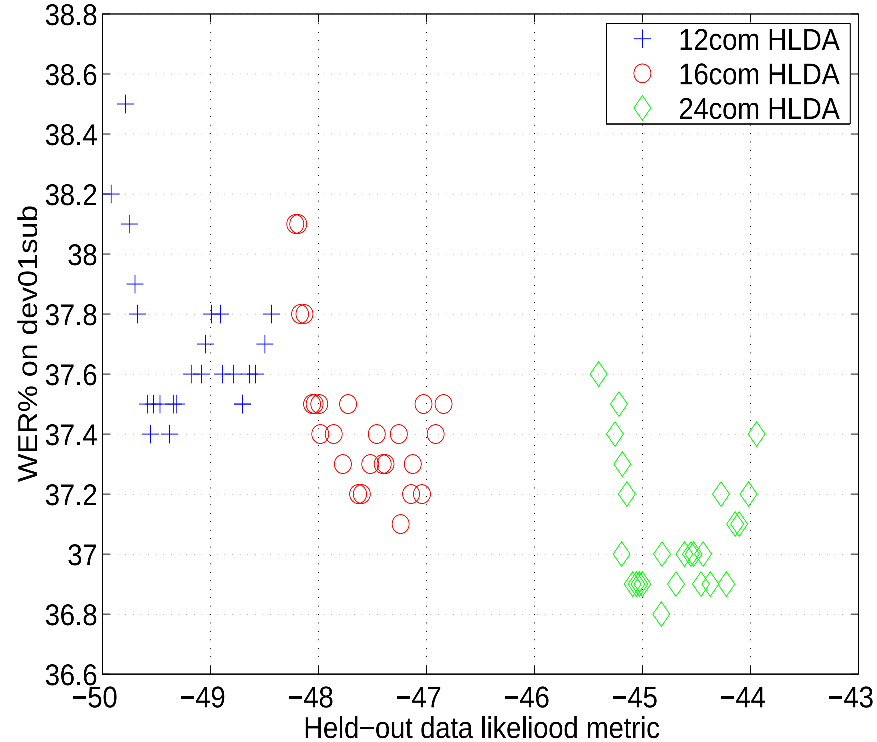
$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} \int \exp\left(\mathcal{L}_{\mathrm{ML}}(\Theta, \tilde{\Theta})\right) p(\Theta|\mathcal{M})\mathrm{d}\Theta$$

- Laplace approximation can be used to approximate the integral.
- Multiple model structures may share the same set of statistics.
- Related to variational approximation.

Issues with ML paradigm

75 global HLDA systems built with varying retained dimensionality $\{28, ..., 52\}$ and number of Gaussians $\{12, 16, 24\}$ on a 68 hour CTS **h5etrain00sub** and 54k LM2002 trigram full decoding on 3 hour development set **dev01sub**.

- 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.
- Maximum Mutual Information (MMI) criterion has been investigated.
- MMI criterion sensitive to outliers utterances.

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

- Inappropriate to directly marginalize, high performance ranking prediction error.

Marginalizing MMI growth function

- MMI criterion equivalent to posterior of reference transcription $\mathcal{W}$.

$$\mathcal{F}_{\mathrm{MMI}}(\Theta, \mathcal{M}) \quad = \quad P(\mathcal{W}|\mathcal{O}, \Theta, \mathcal{M})$$

- MMI criterion is transformed into a growth function,

$$\mathcal{G}(\Theta, \mathcal{M}) = p(\mathcal{O}|\Theta, \mathcal{M}) \left(C\mathcal{F}_{\mathrm{ML}}(\tilde{\Theta}, \mathcal{M}) + \mathcal{F}_{\mathrm{MMI}}(\Theta, \mathcal{M}) - \mathcal{F}_{\mathrm{MMI}}(\tilde{\Theta}, \mathcal{M}) \right)$$

  – Reduced sensitivity to outliers utterances.
  – Retaining gradient of MMI criterion at *current* parameterization.

- C is a positive constant regularization term.

Marginalizing MMI growth function

- A generalized EM based growth function lower bound exists.

$$\mathcal{L}_{\text{MMI}}(\Theta,\tilde{\Theta}) \quad = \quad \log \mathcal{G}(\tilde{\Theta},\mathcal{M}) + \frac{\mathcal{Q}_{\text{MMI}}(\Theta,\tilde{\Theta}) - \mathcal{Q}_{\text{MMI}}(\tilde{\Theta},\tilde{\Theta})}{\sum_{j,\tau}\gamma_j^{\text{MMI}}(\tau)}$$

- Tractable given sufficient discriminative statistics, $\gamma_j^{\text{MMI}}(\tau)$ is the MMI hidden variable occupancy.

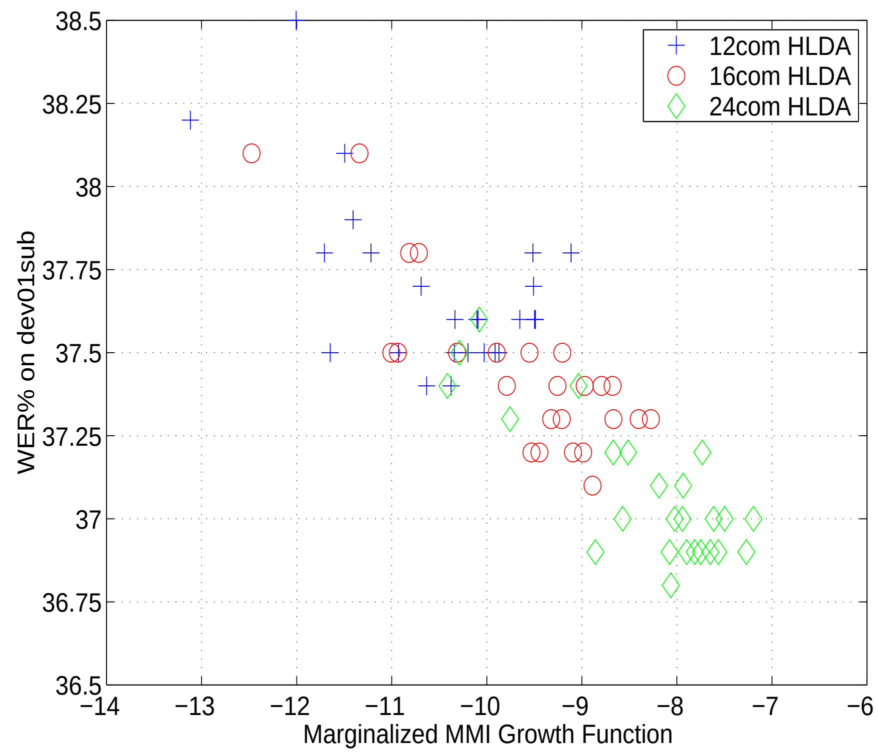
- Integrated out in the parametric space for complexity control.

$$\hat{\mathcal{M}} \quad = \quad \arg\max_{\mathcal{M}} \int \exp\left(\mathcal{L}_{\text{MMI}}(\Theta,\tilde{\Theta})\right) p(\Theta|\mathcal{M})\mathrm{d}\Theta$$

- $\mathcal{L}_{\text{MMI}}(\Theta,\tilde{\Theta})$ is a *strong* sense auxiliary function for the growth function, but a *weak* sense auxiliary function for the MMI criterion.

Marginalizing MMI growth function

- 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

Recognition performance ranking prediction error

Ideal complexity control schemes rank all systems correctly - simple measure of ranking error is the total position shifts weighted by WER differences of mis-ranked pairs of systems.

$$\text{RankErr\%} = \frac{\sum_{i,j} \delta(\mathbf{w}_i, \mathbf{w}_j) \times |\mathbf{w}_i - \mathbf{w}_j| \times |i - j|}{N \times \max_{i,j}\{|\mathbf{w}_i - \mathbf{w}_j|\} \times \max_{i,j}\{|i - j|\}}$$

	WER threshold		
	0.0	0.1	0.2
Held-out Like	8.94	8.89	8.19
Held-out MMI	37.40	37.40	35.91
BIC ($\rho = 1$)	48.43	48.36	47.35
BIC ($\rho = 2$)	55.68	55.68	55.42
GFunc Integral	4.74	4.64	3.10

Ranking error (%) over 75 global HLDA systems

Implementation Issues

- Optimizing the number of Gaussians per state:
  - Start from canonical structure with same number of Gaussians per state.
  - Sharing same set of statistics among multiple structures.
  - Merging pairs of Gaussians giving increment in marginalized growth function.
- Optimizing retained dimensionality per transform class:
  - Start from non-HLDA canonical structure.
  - Sharing same set of Gaussian level statistics.
  - Select dimensionality giving maximum marginalized growth function.
- Using Laplace approximation:
  - Block diagonal Hessian matrix structure.
  - Means and variances of difference Gaussians assumed independent.
- Diagonal variance approximation based MLLR mean adaptation.

Experiments on CTS English

- 76 hours switchboard corpus **h5etrain03sub**, 862 Swbd1, 90 CHE and 166 LDC Cellular conversation sides, 5920 tied states, 12 Gaussians per state
- 296 hours switchboard corpus **h5etrain03**, 4800 Swbd1, 228 CHE and 418 LDC Cellular conversation sides, 6189 tied states, 16 Gaussians per state
- PLP features with VTLN and side based CMN and CVN
- 58k trigram language model LM2003 for full decoding
- 3 hours of test and held-out data set **dev01sub**, 20 sides Swbd2 (eval98), 20 sides Swbd1 (eval00), 19 sides Swbd2 cellular (for manual segments)
- System complexity attributes to optimize on local level:
  - Variable number of mixture components per state
  - Retained subspace dimensionality of a multiple HLDA system

Optimizing #Gaussians per state

System	Selection	#Gaussians	WER%
12com	-	71k	36.1
16com		95k	35.5
20com		119k	35.5
24com		142k	35.3
16com	VarMix	92k	35.6
20com		118k	35.3
24com		138k	35.3
16com	GFunc	82k	35.4
20com		105k	35.2
24com		124k	35.1

Optimizing #Gaussians on 76 hour h5etrain03sub

- Slight improvement in WER over standard schemes like *VarMix*.
- Marginalized MMI growth function leads to more compact model structures.

Optimizing HLDA retained dimensionality

System	#Trans	AvgDim	WER%		
			MLE	MPE	MLLR
std	-	39	37.5	-	-
Fixed	1	39	36.1	33.1	31.2
		52	36.3	-	-
Fixed	65	39	35.5	32.7	30.9
		52	35.5	-	-
GFunc	65	48.7	35.2	32.4	30.5

Optimizing retained dimensionality on 76 hour h5etrain03sub

- 0.3% abs reduction in WER.
- Gain additive to discriminative training and mean adaptation.
- MLLR mean adaptation possible for multiple HLDA/STC systems.

Using multiple HLDA transforms

System	#Trans	AvgDim	WER%			
			16com	28com	MPE	MLLR
std	-	39	35.9	-	-	-
Fixed	1	39	34.9	33.4	30.1	28.5
		52	-	-	-	-
Fixed	65	39	34.2	-	-	-
		52	33.4	32.9	29.7	27.8
GFunc	65	51.3	33.6	33.0	29.7	27.9

Using multiple HLDA transforms on 296 hour h5etrain03

- Gain retained after discriminative training and mean adaptation.
- Diagonal variance approximation, impossible for constrained MLLR.
- 52dim, 65 transforms and 16com structure still not over-fitting !!! Gain NOT additive after over-fitting structural change in mixing up.

Conclusion

- Likelihood based schemes unsuitable.
  - 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
  - Using other discriminative criteria, such as MWE/MPE.
  - Integrate discriminative complexity control with discriminative training.
  - Generalization to other tasks like BN.