MODEL COMPLEXITY CONTROL AND COMPRESSION USING DISCRIMINATIVE GROWTH FUNCTIONS

X. Liu and M.J.F. Gales

Cambridge University Engineering Dept,
Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {xl207,mjfg}@eng.cam.ac.uk

ABSTRACT

State-of-the-art large vocabulary speech recognition systems are highly complex. Many techniques affect both system complexity and recognition performance. The need to determine the appropriate complexity without having to build each possible system has lead to the development of automatic complexity control criteria. In this paper further experiments are carried out using a recently proposed criterion based on marginalizing an MMI growth function. The use of this criterion is much detailed for determining the appropriate dimensionality in a multiple HLDA system and the number of components per state. A scheme for also using this criterion for model compression is described. Experimental results on a spontaneous telephone speech recognition task are described. Initial system compression experiments are inconclusive. However, comparing a standard state-of-the-art system with one generated using complexity control shows a reduction in word error rate.

1. INTRODUCTION

State-of-the-art large vocabulary speech recognition system are highly complex. A wide range of techniques for refining such systems exist. These techniques will affect both system complexity and recognition performance. Automatic criteria are needed to select the appropriate complexity of system, i.e. one that generalizes well to unseen data, without having to build all possible systems. Most complexity control schemes can be classified into two types. In *Bayesian techniques* model parameters are treated as random variables and integrated out in the parametric space. In the *information theory* approaches the complexity control problem is viewed as finding an appropriate code length [2]. These two approaches are closely related to each other, and both asymptotically tend to the Bayesian Information Criterion (BIC) approximation [1]. There is an inherent assumption in these schemes that increasing the likelihood on held-out data decreases the WER. In previous work [8] this correlation has been shown to be quite weak. It would be preferable to use a cost function that is more closely related to WER, for example discriminative criteria. One form of this is to consider model confusions on the training data [5].

Recently a complexity control criterion based on discriminative *growth functions* has been proposed [9]. This is the form considered in this paper. In particular the criterion based on maximum mutual information (MMI) will be used. The MMI criterion can not be directly used given its known sensitivity to outliers. Instead a related MMI *growth function* is considered. In previous work this criterion was found to predict a good ranking of a variety of systems, and small gains in performance, using a 76 hour training set. The forms of complexity considered were the dimensionality of the nuisance subspace for multiple HLDA projection schemes [7] and varying the number of components. This paper describes further experiments using this form of complexity control criterion. First various standard complexity control schemes are described. The MMI growth function theory is then given along with implementation details and how it may be used for system compression. Results examining how the components are assigned over states and initial compression results are given on a small training set. Finally performance using a standard large training set, SwitchBoard, for both maximum likelihood (ML) and minimum phone error [4] (MPE) training are given.

This work was supported by DARPA grant MDA972-02-1-0013. The paper does not necessarily reflect the position or the policy of the US Government and no offi cial endorsement should be inferred.

2. BAYESIAN MODEL COMPLEXITY CONTROL

A standard problem in machine learning is how to find a model structure with appropriate complexity that generalizes well to unseen data. The system is normally selected from a set of candidate model structures $\{\mathcal{M}\}$, given a $\mathcal{T}$ length training data set $\mathcal{O} = \{\mathbf{o}_1, ..., \mathbf{o}_\mathcal{T}\}$ and the reference transcription $\mathcal{W}$. For speech recognition this generalization should relate to the WER performance. Standard ML estimation assumes that HMMs are genuine speech generators. Hence increasing the likelihood on unseen data decreases the WER. Bayesian complexity control techniques use this assumption, and that the *evidence* integral is strongly correlated with the held out data likelihood, to determine the appropriate system. Thus

$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} P(\mathcal{M}) \int \mathcal{F}_{\mathrm{ML}}(\Theta, \mathcal{M}) p(\Theta|\mathcal{M}) \mathrm{d}\Theta \quad (1)$$

Here Θ denotes a parameterization of $\mathcal{M}$, and

$$\mathcal{F}_{\mathrm{ML}}(\Theta, \mathcal{M}) = p(\mathcal{O}, \mathcal{W}|\Theta, \mathcal{M}) \quad (2)$$

As only the acoustic model structural optimization is considered in this paper, the complexity control of the language model $P(\mathcal{W})$ is not discussed. Furthermore, in this work the model structure prior $P(\mathcal{M})$ and parameter prior $p(\Theta|\mathcal{M})$ will be assumed to be uninformative. It is normally computationally intractable to directly compute the evidence integral in equation 1. This has led to a variety of approximation schemes, including BIC, a first order approximation, and the Laplace approximation [1, 8, 9], a second order technique.

To avoid having to estimate the likelihood for each possible system, a lower bound of the ML criterion may be derived using a standard EM approach. Multiple model structures can then share the same set of sufficient statistics. Hence

$$\begin{aligned}\log \mathcal{F}_{\mathrm{ML}}(\Theta,\mathcal{M}) &\geq \mathcal{L}_{\mathrm{ML}}(\Theta,\tilde{\Theta}) \\ &= \log \mathcal{F}_{\mathrm{ML}}(\tilde{\Theta},\mathcal{M}) + \mathcal{Q}_{\mathrm{ML}}(\Theta,\tilde{\Theta}) - \mathcal{Q}_{\mathrm{ML}}(\tilde{\Theta},\tilde{\Theta}) \quad (3)\end{aligned}$$

where the standard EM auxiliary function for HMMs is given by

$$\mathcal{Q}_{\mathrm{ML}}(\Theta,\tilde{\Theta}) = \sum_{j,\tau} \gamma_j(\tau) \log p(\mathbf{o}_\tau|\mathcal{S}_j,\Theta,\mathcal{M}) \quad (4)$$

and $\{\mathcal{S}_j\}$ is the set of discrete hidden variables allowed by the reference, $\gamma_j(\tau) = P(\mathcal{S}_{j,\tau}|\mathcal{O},\mathcal{W},\tilde{\Theta},\mathcal{M})$, $\tilde{\Theta}$ is the *current* parameterization for $\mathcal{M}$ and $\mathcal{S}_{j,\tau}$ indicates that $\mathbf{o}_\tau$ was generated by state $\mathcal{S}_j$. A tractable lower bound of the evidence integral may be approximately computed using the Laplace approximation [9].

$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} \int \exp\left(\mathcal{L}_{\mathrm{ML}}(\Theta,\tilde{\Theta})\right) p(\Theta|\mathcal{M})\mathrm{d}\Theta \quad (5)$$

In practical implementations of BIC auxiliary functions also have to be considered due again to speed considerations. Thus the main difference between equation 5 and BIC is the use of Laplace approximation. The use of the Laplace approximation is important when multiple forms of parameter are to be optimized [8]. The inherent model correctness assumption of standard Bayesian schemes make them inappropriate for speech recognition systems as the correlation between likelihood on unseen data and WER is weak [8].

3. MMI GROWTH FUNCTION

Discriminative training criteria, which are more closely related to WER that ML training, have been successfully applied to LVCSR [3, 4]. One of the most widely used criteria is the maximum mutual information (MMI), which may be expressed as

$$\mathcal{F}_{\mathrm{MMI}}(\Theta,\mathcal{M}) = \frac{p(\mathcal{O},\mathcal{W}|\Theta,\mathcal{M})}{p(\mathcal{O}|\Theta,\mathcal{M})} \quad (6)$$

Empirical results on LVCSR discriminative training tasks have shown that the Extended Baum-Welch (EBW) re-estimation formula can efficiently optimize the MMI criterion [3, 4]. The auxiliary function for EBW is

$$\mathcal{Q}_{\mathrm{MMI}}(\Theta,\tilde{\Theta}) = \sum_{j,\tau} \gamma_j^{\mathrm{MMI}}(\tau) \log p(\mathbf{o}_\tau|\mathcal{S}_j,\Theta,\mathcal{M}) \quad (7)$$

where

$$\gamma_j^{\mathrm{MMI}}(\tau) = \gamma_j^{\mathrm{num}}(\tau) - \gamma_j^{\mathrm{den}}(\tau) + D_j p(\mathcal{O}|\mathcal{S}_{j,\tau},\tilde{\Theta},\mathcal{M}) \quad (8)$$

$\gamma_j^{\mathrm{num}}(\tau)$ is the numerator posterior $P(\mathcal{S}_{j,\tau}|\mathcal{O},\mathcal{W},\tilde{\Theta},\mathcal{M})$, $\gamma_j^{\mathrm{den}}(\tau)$ is the denominator posterior, $P(\mathcal{S}_{j,\tau}|\mathcal{O},\tilde{\Theta},\mathcal{M})$ and D_j is a positive constant regularization term to ensure stability. However, the MMI criterion is sensitive to outliers, so is not appropriate for complexity control [9].

A growth function is required that is related to the standard MMI criterion, but is not as sensitive to outliers. The form of the growth function considered is [9][1]

$$\mathcal{G}(\Theta) = p(\mathcal{O}|\Theta)\left(C\mathcal{F}_{\mathrm{ML}}(\tilde{\Theta}) + \mathcal{F}_{\mathrm{MMI}}(\Theta) - \mathcal{F}_{\mathrm{MMI}}(\tilde{\Theta})\right) \quad (9)$$

[1] In the following equations $\mathcal{M}$ is omitted for a particular model structure being considered.

The first term can remove the sensitivity to outliers, in this case highly unlikely word sequences. The term in the bracket contains gradient information about the MMI criterion and will thus give information about the curvature of the criterion surface in the parametric space. $C > 0$ is a constant regularization term. The gradient of $\mathcal{G}(\Theta)$, when evaluated at the current parameter estimate $\tilde{\Theta}$, is in the same direction as the true criterion gradient when C approaches zero, and $p(\mathcal{O}|\Theta) > 0$.

$$\lim_{C\rightarrow 0} \frac{\partial \mathcal{G}(\Theta)}{\partial \Theta}\bigg|_{\Theta=\tilde{\Theta}} \propto \frac{\partial \mathcal{F}_{\mathrm{MMI}}(\Theta)}{\partial \Theta}\bigg|_{\Theta=\tilde{\Theta}} \quad (10)$$

Similar to the ML criterion, a lower bound may be expressed using a generalized EM approach [9].

$$\begin{aligned}\log \mathcal{G}(\Theta) &\geq \mathcal{L}_{\mathrm{MMI}}(\Theta,\tilde{\Theta}) \\ &= \log \mathcal{G}(\tilde{\Theta}) + \frac{\mathcal{Q}_{\mathrm{MMI}}(\Theta,\tilde{\Theta}) - \mathcal{Q}_{\mathrm{MMI}}(\tilde{\Theta},\tilde{\Theta})}{\sum_{j,\tau}\gamma_j^{\mathrm{MMI}}(\tau)} \quad (11)\end{aligned}$$

where the auxiliary function is defined as in equation 7, however the MMI hidden variable occupancy $\gamma_j^{\mathrm{MMI}}(\tau)$ is now set as

$$\gamma_j^{\mathrm{MMI}}(\tau) = \gamma_j^{\mathrm{num}}(\tau) - \gamma_j^{\mathrm{den}}(\tau) + Cp(\mathcal{O},\mathcal{S}_{j,\tau}|\tilde{\Theta}) \quad (12)$$

The above equation is closely related to the form given in equation 8, when $D_j = Cp(\mathcal{S}_j|\tilde{\Theta})$. The choice of D_j affects the criterion convergence and test set generalization [3, 4]. A commonly used form of D_j is associated with the *denominator* occupancy $D_j = E\sum_\tau \gamma_j^{\mathrm{den}}(\tau)$, where $E > 0$. The same form of smoothing function is adopted for the growth function in this paper. In contrast to the ML lower bound, increasing the growth function lower bound only guarantees an increase in the growth function given the strict convergence of generalized EM, but not the MMI criterion. The two are only constrained to have the same gradient from equation 10. The following marginalization of the MMI growth function lower bound is used in this paper for complexity control, and computed via the Laplace approximation.

$$\begin{aligned}\hat{\mathcal{M}} &= \arg\max_{\mathcal{M}} \int \exp\left(\mathcal{L}_{\mathrm{MMI}}(\Theta,\tilde{\Theta})\right) p(\Theta|\mathcal{M})\mathrm{d}\Theta \\ &= \arg\max_{\mathcal{M}} \hat{\mathcal{G}}(\mathcal{M}) \quad (13)\end{aligned}$$

where $\hat{\mathcal{G}}(\mathcal{M})$ is the marginalized growth function value for model $\mathcal{M}$.

4. IMPLEMENTATION ISSUES

The previous sections has described how a marginalized MMI growth function may be calculated and used to control the complexity of the system. However, there are a number of implementation issues that affect the performance of this form of complexity control. These are briefly described below.

1) **Sufficient statistics**. For LVCSR training the majority of the time is spent accumulating the sufficient statistics to estimate the model parameters.Since EM, or the extended Baum-Welch algorithm, are used to train the models they rely on the use of a *current* model set to obtain the alignments to gather the statistics. For the marginalized growth function this yields the form in equation 11. When varying the number of Gaussian components in a state it is impractical to obtain new statistics for each number of components, even if the state alignments are fixed. To overcome

this problem the sufficient statistics are obtained using the current model. From this model it is only possible to reduce the number of components associated with each state. This allows the use of *component merging* to obtain sufficient statistics. In model merging a pair of Gaussians are combined together to form a single component. Thus to join components j and k to yield l,

$$\gamma_l^{\text{MMI}}(\tau) \quad = \quad \gamma_j^{\text{MMI}}(\tau) + \gamma_k^{\text{MMI}}(\tau) \tag{14}$$

similarly for the first and second order moments. All possible pairs of component merging are considered and the one with the largest marginalized value selected. In the case of determining the number of retained HLDA dimensions no component merging is required as the sufficient statistics may be assumed to be the same for all numbers of retained dimensions.

2) **Hessian calculation**. The Laplace approximation requires the storage of a Hessian matrix. The number of model parameters in an LVCSR system can be in the millions. Computing the Fisher Information matrix for this is impractical. The solution used in this work is the same as that in [8], where each component is assumed independent of all other components, and the mean vector and variance vector are additionally assumed independent.

3) **Maximum structure change**. When using the MMI growth function for controlling complexity the value of the growth function depends on all the competing models. However the sufficient statistics are accumulated given the current model. If the structure of the competing models vary dramatically from this then the estimate of the growth function will be poor. To overcome this problem a maximum change in the model complexity can be imposed. For the case of altering the number of components, the maximum number that may be removed from any state is set to 2 in these experiments. This was found to yield a reasonable compromise between accuracy and speed.

4) **Smoothing constant**. For the MMI growth function there is a smoothing constant D_j. This affects both the stability of the optimization and the rate of convergence. It will therefore change the marginalized growth function and rate of model structure change. In practice similar values to those used in MMI training were found to be a reasonable compromise.

5) **Model compression**. Equation 11 requires that in order to alter the system structure there must be an increase in the value of the marginalized growth function. For the case of determining the number of components where it is only possible to reduce the number of model parameters, it is possible to use the same approach for generating discriminatively determined compact systems. To allow a reduction in the complexity of the system it is required that

$$\hat{\mathcal{G}}(\mathcal{M}) - \hat{\mathcal{G}}(\hat{\mathcal{M}}) \quad \geq \quad -\alpha\hat{\mathcal{G}}(\hat{\mathcal{M}}_0) \tag{15}$$

where α is the compression factor and $\hat{\mathcal{M}}_0$ is the a current model structure to accumulate sufficient statistics from. $\alpha = 0$ corresponds to the standard structural optimization problem. $\alpha = \infty$ is a standard mix-down approach, where the number of components are reduced to the value specified in the maximum structure change.

5. EXPERIMENTAL RESULTS

Two series of experiments were conducted to investigate the performance of the marginalized growth function for complexity control and compression. Two training sets were used. The first is a full system using a 296 hour training set `h5etrain03`, consisting of 4800 Switchboard I, 228 CHE and 418 LDC Cellular conversation sides. A 76 hour subset of this, `h5etrainsub`, was used for initial development. For evaluation a 3 hour subset of 2001 development data, `dev01sub`, was used. Cepstral features were extracted and normalized for each conversation side via side based VTLN, mean and variance normalization. A 52 dimensional acoustic feature was then generated by appending derivatives to the third order. This was projected down using one, or more, HLDA projections to a 39 dimensional feature vector, or a complexity determined dimensionality. Continuous density, mixture of Gaussian, cross-word triphone, gender independent HMM system were used. All recognition experiments used a 58k word trigram language model.

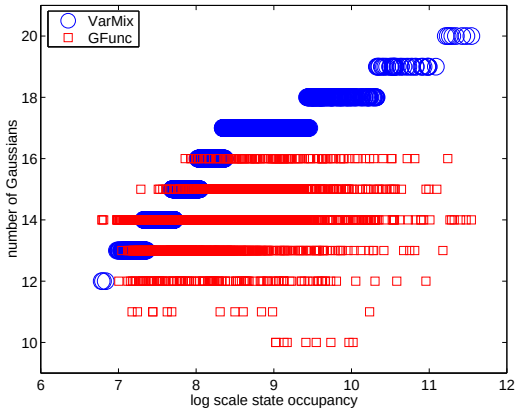

Fig. 1. Number of Gaussians vs. state occupancy

A 16 component system was generated on the `h5etrain03sub` training data. The components per state was then varied proportionally to the occupancy raised to a power of 0.2. This is the *varmix* approach used at CUED. The original 16 component was then compressed using the discriminative growth function to 13.9. Figure 1 shows a plot of number of components against state occupancy for these two systems. The use of components is dramatically different in the two systems. The marginalized growth function is not correlated with the state occupancy.

	Compression factor α ($\times 10^{-4}$)				
	—	0	5	10	∞
#Gauss	24.0	20.7	20.7	17.7	16.0
WER (%)	35.3	35.1	35.2	35.3	35.5

Table 1. WER on `dev01sub` and system size against α, MLE training on `h5etrain03sub`

The 16 component systems was then iteratively split until the number of components was 24. This was then used as the baseline model for the complexity control experiments. Table 1 shows the performance of these systems. The baseline error rate for the 24 component system on `dev01sub` was 35.3%. Using the "optimal" structure, determined with $\alpha = 0$ and 4 iterations of structure optimization the error rate fell to 35.1% and the average number of components per state was 20.7. The model structure was further compressed by increasing the value of α. Using $\alpha = 1 \times 10^{-3}$, the average number of components was 17.7, a parameter reduction of 26% over the standard 24 component system with the same

error. As a contrast a system with $\alpha = \infty$ where 2 components were removed from each state at each iteration is also shown. An equivalent *varmix* 16 component system gave an error rate 35.5%. The compression results on these experiments were disappointing. This may be partly due to the small performance gain in the range considered, only 0.2% absolute was gained by going from 16 to 24 components. In contrast using a single HLDA transform with the `h5etrain03` training data the performance gain from 16 to 28 components was 1.5% absolute. This should allow a better assessment of the complexity control.

Previous complexity control experiments in [9] concentrated on either controlling the number of components, or the dimensionality of the HLDA projection, using the `h5etrain03sub` training data. In practice state-of-the-art systems use larger amounts of training data. It would also be preferable to examine the performance when both parameters are varied. For these experiments a 28 component system was used as the baseline. Note for complexity control α was set to 0.

Complexity Control	#Gauss	#Trans	#Dim	Training	
				MLE	MPE
Std	28	—	39	34.7	—
Fixed	28	1	39	33.4	30.1
			52	33.2	—
Fixed	28	65	39	33.3	29.8
			52	32.9	—
GFunc	25.6	65	41.5	32.7	29.6

Table 2. WER on `dev01sub` against number of Gaussians and retained HLDA dimensions, training on `h5etrain03`

Table 2 shows the performance of various structural configurations on the `h5etrain03` training data using MLE and Minimum Phone Error (MPE) [4] criteria. The standard non-HLDA system used static features with first and second derivatives. Appending the third derivatives and projecting back to 39 dimensions decreased the error rate by 1.3% absolute for MLE. If instead of an HLDA projection, a global semi-tied covariance matrix was used a slight decrease in error rate was observed. The number of transforms was then increased to 65. The assignment of component to transform was determined by clustering in the acoustic space, silence components were assigned to a distinct transform. Both multiple HLDA projections and semi-tied systems were built. The use of all 52 dimensions gave better performance than the HLDA projection to 39 dimensions. For the projection to 39 dimension configuration an MPE system was built. This gave a 0.3% absolute reduction over the single transform MPE result. Structural optimization was then performed in two stages. First the number of components per state was determined for a standard non-HLDA 28 component system. The projections for each of the 65 transforms was then determined given this number of components per state. Compared to the standard MLE single HLDA projection to 39 dimensions, as used in the CUED evaluation systems, the error rate was reduced by 0.7% absolute. Using MPE training a 0.2% gain over the 65 transform system was observed. This is significantly less than the 0.6% gained using MLE training. One possible reason is that the MPE training for all systems in the table use the same set of word and phone lattices, generated by the standard MLE global HLDA system. As the underlying model structure becomes more different from it, these lattices becomes more inappropriate as the both its word level confusion and model level alignment are also turning more different. Compared to the standard evaluation set-up (single transform, 39 dim projection), the complexity control yielded a 0.5% gain. This performance gain was maintained after adaptation.

6. CONCLUSION

An efficient model structure compression technique was presented using marginalized discriminative growth functions. The discriminative growth function investigated is closely related to maximum mutual information (MMI) criterion, with a reduced sensitivity to outliers utterances with very low posteriors. Model structure compression on a typical LVCSR task was performed where the number of Gaussians per state and the useful dimensions of an HLDA system were reduced. Initial experiments indicate that this form of discriminative structural optimization may be useful in speech recognition.

As a general model complexity control framework, the discriminative growth functions are not restricted to the MMI criterion. Future work will examine using other discriminative criteria which are more closely related to WER, such as minimum word error (MWE) [6] and MPE criteria [4]. In addition optimizing the complexity of discriminatively trained model structures will also be investigated.

7. REFERENCES

[1] G. Schwartz (1978). Estimating the Dimension of a Model, *The Annals of Statistics*, pp. 461–464, Vol. 6, No. 2, February 1978.

[2] A. R. Barron, J. J. Rissanen & B. Yu (1998). The Minimum Description Length Principle in Coding and Modeling, *IEEE Transactions on Information Theory*, pp. 2743–2760, Vol. 44, No. 6, October 1998.

[3] P. C. Woodland & D. Povey (2002). Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition. *Computer Speech and Language*, Vol. 16, pp. 25-47.

[4] D. Povey (2003). *Discriminative Training for Large Vocabulary Speech Recognition*, PhD thesis, Cambridge University Engineering Department, England.

[5] M. Padmanabhan & L. R. Bahl (2000). Model Complexity Adaptation Using a Discriminant Measure, *IEEE Transactions on Speech and Audio Processing*, pp. 205–208, Vol. 8, No. 2, March 2000.

[6] J. Kaiser, B. Horvat & Z. Kacic, A Novel Loss Function for the Overall Risk-criterion Based Discriminative Training of HMM Models, *ICSLP'2000*.

[7] M. J. F. Gales (2002). Maximum Likelihood Multiple Projection Schemes for Hidden Markov Models, *IEEE Transactions on Speech and Audio Processing*, pp. 37–47, Vol. 10, 2002.

[8] X. Liu, M. J. F. Gales & P. C. Woodland (2003). Automatic Complexity Control for HLDA Systems, *Proc. ICASSP'03*, Vol. 1, Hong Kong.

[9] X. Liu & M. J. F. Gales (2003). Automatic Model Complexity Control Using Marginalized Discriminative Growth Functions, *Proc. ASRU'03*, St. Thomas, U.S. Virgin Islands.