AUTOMATIC MODEL COMPLEXITY CONTROL USING MARGINALIZED 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

Designing a large vocabulary speech recognition system is a highly complex problem. Many techniques affect both the system complexity and recognition performance. Automatic complexity control criteria are needed to quickly predict the recognition performance ranking of systems with varying complexity, in order to select an optimal model structure with the minimum word error. In this paper a novel complexity control technique is proposed by using the marginalization of discriminative growth functions. A two stage approach is adopted to make the marginalization efficient. First a lower bound, related to the auxiliary function, is used to remove the dependence on the latent variables. Second a Laplace approximation is used for the integration. Experimental results on a spontaneous speech recognition task show that marginalized MMI growth function outperforms using held out data likelihood and standard Bayesian schemes in terms of both recognition performance ranking error and word error.

1. INTRODUCTION

How to choose an optimal model structure with the "appropriate" complexity is a standard problem for large vocabulary continuous speech recognition (LVCSR) training. State-of-the-art LVCSR systems are highly complex. A variety of techniques are used which alter the system complexity, such as state clustering, mixture of Gaussians and dimensionality reduction schemes. It is not possible to explicitly train all possible systems and obtain their word error rates (WER) on held out data for all possible structural configurations. It is therefore useful to find a criterion that predicts the WER ranking order without requiring all the systems to be built.

Most existing 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 [3]. These two approaches are closely related to each other, and both asymptotically tend to the Bayesian Information Criterion (BIC) approximation [2]. There is an inherent assumption in these schemes that increasing the likelihood on held-out data decreases the WER. In previous work [15] 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. Along these lines a discriminative measuring of model confusion over the training data has previously been used [11].

This paper presents a novel technique using the marginalization of a discriminative *growth function*, rather than the likelihood in standard Bayesian approaches. The discriminative growth function presented in this paper is based on the maximum mutual information (MMI) criterion. The MMI criterion can not be directly used as it is sensitive to outliers, instead related MMI *growth function* is used. A two stage approach is used to make the marginalization of the growth function efficient. First a lower bound, related to the auxiliary function, is used to remove the dependence on the latent variables. Second a Laplace approximation is used for the integration.

In this work the type of HMM systems being investigated use mixture of Gaussians as state output distributions and multiple Heteroscedastic LDA (HLDA) transforms shared locally among different parts of the model as a feature space diagonalizing and dimensionality reduction scheme [13]. An HLDA transform partitions the entire feature space into a *retained* subspace where all Gaussian means and variances are kept distinct, and a *nuisance* subspace where Gaussian means and variances are globally tied. Two forms of system attribute are examined. The first is the number of components associated with the states of a model. The second is the number of useful dimensions of a multiple HLDA system. The problem of examining multiple attributes makes some commonly used schemes such as BIC inappropriate for complexity control [15].

The next section describes the general framework of Bayesian complexity control. Section 3 details the derivation of marginalizing an MMI criterion growth function for automatic complexity control. Some implementation issues are discussed in section 4. Experimental results on a standard LVCSR task are presented in section 5.

2. BAYESIAN MODEL COMPLEXITY CONTROL

A standard problem in LVCSR training, and machine learning in general, is how to obtain a model structure that generalizes well to unseen data with appropriate complexity 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 directly relates to the WER performance. The standard approach is to assume that the model is "close" to the correct model, so that increasing the likelihood on unseen data decreases the WER. Bayesian complexity control

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 official endorsement should be inferred.

techniques make use of the training data, assuming the following *evidence* integral is strongly correlated with the held out data likelihood.

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

Here Θ denotes a parameterization of $\mathcal{M}$. and the maximum likelihood (ML) criterion is given by

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

The language model probability $P(\mathcal{W})$ is normally optimized on additional text data, so can be ignored in terms of the complexity control considered here. Furthermore, in this work the model structure priors $P(\mathcal{M})$ and parameters priors $p(\Theta|\mathcal{M})$ 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, among which BIC is the most widely used [2]. This complexity control criterion is simply expressed in terms of penalized log likelihood evaluated at the ML estimate of model parameters $\hat{\Theta}$

$$\log p(\mathcal{O}|\mathcal{M}) \quad \approx \quad \log p(\mathcal{O}|\hat{\Theta},\mathcal{M}) - \rho \cdot \frac{k}{2} \log \mathcal{T} \quad (3)$$

where k denotes the number of free parameters in $\mathcal{M}$ and ρ is a penalization coefficient which may be tuned to the specific task [14]. Schwartz proved that when $\rho = 1$, BIC is a first order asymptotic expansion of the evidence integral. Recent research has found a major limitation of BIC when optimizing the multiple complexity attributes considered in the paper [15].

Laplace's approximation provides a second order asymptotic expansion of the evidence integral [1]. The basic idea is to make a local Gaussian approximation of likelihood curvature in the parametric space. The volume under that Gaussian is computed as an approximation.

$$\begin{aligned} \log p(\mathcal{O}|\mathcal{M}) \quad \approx \quad & \log p(\mathcal{O}|\hat{\Theta},\mathcal{M}) + \frac{k}{2}\log 2\pi \\ & -\frac{1}{2}\log \left| -\nabla^2_{\Theta} \log p(\mathcal{O}|\hat{\Theta},\mathcal{M}) \right| \quad (4) \end{aligned}$$

One issue with both of the above schemes is that the log-likelihood for each configuration is required. One method to avoid this is to derive a lower bound for the ML criterion in a tractable form that may be assumed to be applicable for multiple configurations. Using a standard EM approach this may be expressed as

$$\begin{aligned} \log p(\mathcal{O}|\Theta,\mathcal{M}) \quad \geq \quad & \log p(\mathcal{O}|\tilde{\Theta},\mathcal{M}) \\ & + \mathcal{Q}_{\rm ML}(\Theta,\tilde{\Theta}) - \mathcal{Q}_{\rm ML}(\tilde{\Theta},\tilde{\Theta}) \\ = \quad & \mathcal{L}_{\rm ML}(\Theta,\tilde{\Theta}) \quad (5) \end{aligned}$$

where the standard EM auxiliary function for HMMs is given by

$$\mathcal{Q}_{\rm ML}(\Theta,\tilde{\Theta}) \quad = \quad \sum_{j,\tau} \gamma_j(\tau) \log p(\mathbf{o}_\tau|\mathcal{S}_j,\Theta,\mathcal{M}) \quad (6)$$

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 $\mathcal{S}_j$. Using this form of bound all configurations with the same set of latent variables and statistics can be efficiently computed. This now yields a lower bound for the evidence

$$p(\mathcal{O}|\mathcal{M}) \geq \int \exp\left(\mathcal{L}_{\rm ML}(\Theta,\tilde{\Theta})\right) p(\Theta|\mathcal{M}) {\rm d}\Theta \quad (7)$$

The right hand side of inequality 7 can be efficiently integrated out using the Laplace approximation.

A special case of this is to simply use the standard ML auxiliary function [15], while ignoring the other two terms which are independent of Θ in equation 5. However when using multiple sets of statistics they may no longer be ignored. It can be shown that these two terms are equivalent to an entropy of hidden variable sequence posteriors. This is related to another popular approximation scheme, variational approximation [4], which can yield a tighter bound for the evidence integral. Markov Chain Monte Carlo sampling schemes may also be used to approximate the evidence integral, although in practice this approach is infeasible for LVCSR tasks given the high dimensionality of the sampling space.

3. DISCRIMINATIVE GROWTH FUNCTIONS

This section initially describes the standard maximum mutual information training criterion and the issue with using it for model selection. A suitable *growth* function is then described and a strict lower bound with efficient approximation is presented.

3.1. Maximum Mutual Information Criterion

Recently discriminative training criteria, which are more closely related to WER, have been successfully applied to LVCSR training [8, 9]. One of the most widely used criteria is the maximum mutual information (MMI) criterion. This is equivalent to maximizing the posterior probability of training data over the correct transcription $\mathcal{W}$.

$$\mathcal{F}_{\rm MMI}(\Theta,\mathcal{M}) \quad = \quad \frac{p(\mathcal{O},\mathcal{W}|\Theta,\mathcal{M})}{p(\mathcal{O}|\Theta,\mathcal{M})} \quad (8)$$

Empirical results on various LVCSR discriminative training tasks have shown that the Extended Baum-Welch (EBW) reestimation formula can efficiently optimize the MMI criterion [5, 6, 8, 9]. The auxiliary function for EBW is

$$\mathcal{Q}_{\rm MMI}(\Theta,\tilde{\Theta}) \quad = \quad \sum_{j,\tau} \gamma_j^{\rm MMI}(\tau) \log p(\mathbf{o}_\tau|\mathcal{S}_j,\Theta,\mathcal{M}) \quad (9)$$

where

$$\gamma_j^{\rm MMI}(\tau) = \gamma_j^{\rm num}(\tau) - \gamma_j^{\rm den}(\tau) + D_j p(\mathcal{O}|\mathcal{S}_{j,\tau},\tilde{\Theta},\mathcal{M}) \quad (10)$$

$\gamma_j^{\rm num}(\tau)$ is the numerator posterior $P(\mathcal{S}_{j,\tau}|\mathcal{O},\mathcal{W},\tilde{\Theta},\mathcal{M})$, $\gamma_j^{\rm den}(\tau)$ 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 the convergence.

One obvious form of complexity control is to marginalize the MMI criterion over the parametric space, similar to the Bayesian evidence integral. This yields

$$\hat{\mathcal{M}} \quad = \quad \arg\max_{\mathcal{M}} \int \mathcal{F}_{\rm MMI}(\Theta,\mathcal{M}) p(\Theta|\mathcal{M}) {\rm d}\Theta \quad (11)$$

However, directly marginalizing the MMI criterion will suffer from an inherent defect - the MMI criterion computation tends to give undue weight to outliers utterances with very low posteriors. In such a case the recognition performance ranking prediction can be considerably distorted due to the presence of these outliers. One method to overcome this problem is to explicitly deweight the posteriors of outliers utterances through a criterion smoothing function [7]. An alternative method is to transform the MMI criterion into a *growth function*, $\mathcal{G}(\Theta, \mathcal{M})$. This is the approach adopted in this paper.

3.2. MMI Growth Function

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 in this paper[1]

$$\mathcal{G}(\Theta) \;=\; p(\mathcal{O}|\Theta)\left(C\mathcal{F}_{\rm ML}(\tilde{\Theta}) + \mathcal{F}_{\rm MMI}(\Theta) - \mathcal{F}_{\rm MMI}(\tilde{\Theta})\right) \quad (12)$$

The first term acts to remove the sensitivity to outliers, in this case highly unlikely sequences. The term in the bracket contains 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 $\tilde{\Theta}$, will be in the same of direction as the true criterion gradient when C approaches zero, and $p(\mathcal{O}|\Theta) > 0$.

$$\lim_{C\to 0} \left.\frac{\partial \mathcal{G}(\Theta)}{\partial \Theta}\right|_{\Theta=\tilde{\Theta}} \propto \left.\frac{\partial \mathcal{F}_{\rm MMI}(\Theta)}{\partial \Theta}\right|_{\Theta=\tilde{\Theta}} \quad (13)$$

The aim is to obtain a computationally efficient lower bound for the marginalized growth function. The growth function given in equation 12 may be re-written as

$$\begin{aligned}\mathcal{G}(\Theta) \;&=\; p(\mathcal{O},\mathcal{W}|\Theta) - P(\mathcal{W}|\mathcal{O},\tilde{\Theta})p(\mathcal{O}|\Theta)\\ &\quad +Cp(\mathcal{O},\mathcal{W}|\tilde{\Theta})p(\mathcal{O}|\Theta) \quad (14)\end{aligned}$$

In the same fashion as the bound for Bayesian evidence this must be rewritten in terms of the hidden state sequences $\{\Psi\}$. A lower bound may be expressed using a generalized EM approach [10].

$$\begin{aligned}\log \mathcal{G}(\Theta) \;&=\; \log \sum_{\Psi} \mathcal{G}(\Psi,\Theta)\\ &\geq\; \sum_{\Psi} \mathcal{P}(\Psi,\tilde{\Theta}) \log \frac{\mathcal{G}(\Psi,\Theta)}{\mathcal{P}(\Psi,\tilde{\Theta})}\\ &=\; \mathcal{L}_{\rm MMI}(\Theta,\tilde{\Theta}) \quad (15)\end{aligned}$$

where the hidden variable sequence version of the growth function is defined as[2]

$$\begin{aligned}\mathcal{G}(\Psi,\Theta) \;&=\; p(\mathcal{O},\Psi,\mathcal{W}|\Theta) - P(\mathcal{W}|\mathcal{O},\tilde{\Theta})p(\mathcal{O},\Psi|\Theta)\\ &\quad +Cp(\mathcal{O},\mathcal{W}|\tilde{\Theta})p(\mathcal{O},\Psi|\Theta) \quad (16)\end{aligned}$$

For this inequality to be valid using Jensen's inequality the sequence posterior distribution $\mathcal{P}(\Psi,\tilde{\Theta})$ must satisfy the positive and sum to one constraint. The form of the hidden variable sequence "posterior" considered in this paper is

$$\mathcal{P}(\Psi,\tilde{\Theta}) \;=\; \frac{\mathcal{G}(\Psi,\tilde{\Theta})}{\sum_{\Psi}\mathcal{G}(\Psi,\tilde{\Theta})} \;=\; \frac{\gamma^{\rm MMI}_{\Psi}(\mathcal{O})}{\sum_{\Psi}\gamma^{\rm MMI}_{\Psi}(\mathcal{O})} \quad (17)$$

When C is big enough, such $\mathcal{P}(\Psi,\tilde{\Theta})$ is guaranteed to be positive and satisfy the sum to one constraint required by Jensen's inequality. This form of posterior will be shown to yield bounds closely related to the standard MMI auxiliary function.

Various forms of this growth function will be used in this work. First at the "current" model parameter value

$$\mathcal{G}(\Psi,\tilde{\Theta}) \;=\; \gamma^{\rm MMI}_{\Psi}(\mathcal{O})p(\mathcal{O},\mathcal{W}|\tilde{\Theta}) \quad (18)$$

where the MMI hidden variable sequence occupancy $\gamma^{\rm MMI}_{\Psi}(\mathcal{O})$ is defined as

$$\gamma^{\rm MMI}_{\Psi}(\mathcal{O}) = P(\Psi|\mathcal{O},\mathcal{W},\tilde{\Theta}) - P(\Psi|\mathcal{O},\tilde{\Theta}) + Cp(\mathcal{O},\Psi|\tilde{\Theta}) \quad (19)$$

This gives the simple form of posterior in equation 17. It is also possible to write

$$\mathcal{G}(\Psi,\Theta) = \left(C\mathcal{F}_{\rm ML}(\tilde{\Theta}) + P(\mathcal{W}|\Psi) - \mathcal{F}_{\rm MMI}(\tilde{\Theta})\right) p(\mathcal{O},\Psi|\Theta) \quad (20)$$

since the observations are conditionally independent of the word sequence given the hidden state sequence, Ψ.

Using equation 17 and 20, equation 15 can be re-written as

$$\begin{aligned}\mathcal{L}_{\rm MMI}(\Theta,\tilde{\Theta}) \;&=\; \log \mathcal{G}(\tilde{\Theta}) + \sum_{\Psi}\mathcal{P}(\Psi,\tilde{\Theta})\log p(\mathcal{O},\Psi|\Theta)\\ &\quad - \sum_{\Psi}\mathcal{P}(\Psi,\tilde{\Theta})\log p(\mathcal{O},\Psi|\tilde{\Theta}) \quad (21)\end{aligned}$$

This lower bound can be re-expressed as

$$\mathcal{L}_{\rm MMI}(\Theta,\tilde{\Theta}) = \log \mathcal{G}(\tilde{\Theta}) + \frac{\mathcal{Q}_{\rm MMI}(\Theta,\tilde{\Theta}) - \mathcal{Q}_{\rm MMI}(\tilde{\Theta},\tilde{\Theta})}{\sum_{\Psi}\gamma^{\rm MMI}_{\Psi}(\mathcal{O})} \quad (22)$$

where the auxiliary function is defined as[3]

$$\begin{aligned}\mathcal{Q}_{\rm MMI}(\Theta,\tilde{\Theta}) \;&=\; \sum_{\Psi}\gamma^{\rm MMI}_{\Psi}(\mathcal{O})\log p(\mathcal{O}|\Psi,\Theta)\\ &=\; \sum_{j,\tau}\gamma^{\rm MMI}_{j}(\tau)\log p(\mathbf{o}_{\tau}|\mathcal{S}_j,\Theta) \quad (23)\end{aligned}$$

and the MMI hidden variable occupancy $\gamma^{\rm MMI}_{j}(\tau)$ is,

$$\gamma^{\rm MMI}_{j}(\tau) = \gamma^{\rm num}_{j}(\tau) - \gamma^{\rm den}_{j}(\tau) + Cp(\mathcal{O},\mathcal{S}_{j,\tau}|\tilde{\Theta}) \quad (24)$$

The above equation is closely related to the form given in equation 10, when $D_j = Cp(\mathcal{S}_j|\tilde{\Theta})$. Previous research on LVCSR MMI training shows that the selection of D_j considerably affects the criterion convergence and test set generalization [8, 9]. A commonly used form of D_j is associated with the *denominator* occupancy $D_j = E\sum_{\tau}\gamma^{\rm den}_{j}(\tau)$, where $E > 0$. The same form of smoothing function may be used for the growth function. Now the value of C in equation 16 will vary depending on the hidden variable sequence. However this means that the marginalization of discriminative growth functions is not a parameter free scheme. For complexity control tasks the appropriate setting of D_j can also be important. This paper uses the form of D_j as described above.

The form of lower bound defined in equation 22 and the ML case given in equation 5 have similar forms. Given the current value of the criterion, the new value is estimated as the old value

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

[2] Here the hidden variable sequence consists of both the state sequence and the component sequence.

[3] For this definition the likelihood of the hidden variable sequence $P(\Psi|\tilde{\Theta})$, has been ignored. The influence of the component priors and transition matrices have thus been removed.

plus the change in auxiliary function. However the ML case has a simple closed form for maximizing the auxiliary function, whereas for the MMI growth function the value selected is sensitive to C. Note in both cases an increase in the auxiliary function guarantees an increase in the related ML, or MMI growth function. However increasing the MMI growth function does not guarantee an increase in the MMI criterion, as they are simply constrained to have the same gradient from equation 13.

The following marginalization of the MMI growth function lower bound is used in this paper for complexity control.

$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} \int \exp\left(\mathcal{L}_{\mathrm{MMI}}(\Theta, \tilde{\Theta})\right) p(\Theta|\mathcal{M}) \mathrm{d}\Theta \qquad (25)$$

In common with the marginalized lower bound for ML this expression is approximated by using Laplace's approximation.

4. IMPLEMENTATION ISSUES

There are three main implementation issues associated with the marginalized growth function complexity control described in the previous section: obtaining the "complete" dataset; using Laplace's approximation; and choice of the value of C (or the hidden variable specific D_j).

As previously described to make the complexity control efficient, a lower bound related to the standard ML and MMI auxiliary functions, is marginalized over the model parameters. The nature of the complexity attributes to be controlled determines how complex this process is. For the multiple HLDA case complexity control is simple, where the number of useful dimensions retained in an HLDA transform is allowed to vary from transform to transform. Only the retained dimensionality is allowed to vary, there are no differences in the hidden state sequence for any of the systems to choose over[4]. In contrast if the number of components associated with a state is to be determined, then the hidden state sequence will vary. In this situation it is necessary to obtain statistics for multiple systems. The approach adopted in this work is to fix state level posteriors using the current model parameters and train systems with varying numbers of components per state given the state posteriors. The discriminative statistics are then accumulated for each of the various systems. By fixing the state level alignments, again yielding a lower bound, the influence of the state complexity on the state posterior has been removed.

The second issue is the Laplace approximation. 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 [15], where each component is assumed independent of all other components, and the mean vector and variance vector are additionally assumed independent.

In common with the EBW training it is necessary to set a value for the smoothing constant C, or component specific version D_j. The larger the value of this constant the more stable, and slower, the MMI optimization. The equivalent for the complexity control is that at each iteration the change in the system structure will be smaller.

5. EXPERIMENTAL RESULTS

Initial evaluation of the complexity control system was based on a conversational telephone speech task (referred to as SwitchBoard). The results presented in this paper consist of two distinct parts. In the first part, various techniques were used to optimize multiple model complexity attributes on a global level. This allowed all systems to be trained and evaluated. In the second part, a single model complexity attribute was optimized on a local level. All model structures considered were trained using the ML criterion after the complexity has been determined.

5.1. Complexity evaluation

When a complete set of possible systems are built and evaluated, for example in the global experiments described here, it is possible to compare the error rate and rank ordering for the complexity control criteria being used. A good complexity control criterion should yield the correct rank ordering for all the systems to be compared. A measure of the distance between the two rankings is required.

In this paper, an empirical ranking prediction error metric is defined as

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

Here $\{\mathbf{w}_1, ..., \mathbf{w}_N\}$ denotes the WER of all N possible systems according to a ranking order generated by some complexity control criterion, and the binary function $\delta(\mathbf{w}_i, \mathbf{w}_j)$ will be true only if the ranking between $\mathbf{w}_i$ and $\mathbf{w}_j$ is incorrect and the difference in WER is significant, in this case bigger than a given *WER threshold*. This has a good intuitive feel as penalizing systems that differ only slightly in error rate is felt inappropriate. Thus differences of systems whose word error rate is less than a *WER threshold* are ignored. Thus, the ranking error is related to the total amount of position shifts over a minimum threshold, weighted by WER difference between all mis-ranked pairs of systems. The normalization term guarantees the ranking error will be positive and less than one.

5.2. Optimizing Global Complexity Attributes

The global complexity control experiments used the same configuration as [15]. This used a 68 hour subset of all the available Switchboard I and Call Home English (CHE) conversation sides. A continuous density, mixture of Gaussian, cross-word triphone, gender independent HMM system was trained using the ML criterion, with 6168 physical speech states after decision tree based tying. All recognition experiments used a trigram language model. A 3 hour subset of the 2001 development data was used both as the test and held-out data. For more details see [15]. Two global complexity attributes of a single transform HLDA system were optimized: the number of Gaussian components per state from the range $\{12, 16, 24\}$; the number of useful dimensions in the range $\{28, ..., 52\}$. The permutation of these two attributes led to 75 model structures.

In previous work on this set-up the correlation between WER and held out data likelihood for the 75 systems was examined [15]. Though there was a very general trend that error rate decreased with increased held-out data likelihood, the precise ordering of

[4] As the structural change is too big the correlation between the growth function and its lower bound may be arguably weak.

systems was poor. Noticeably this scheme favors the most complex system, the best model structure predicted has 24 Gaussians per state and 52 retained dimensions, which is significantly worse than the actual best system by 0.6% absolute. Previous work also examined Laplace's approximation and showed that the Bayesian evidence (which should be closely related to the held-out data log-likelihood) may be reasonably approximated and that BIC is a poor criterion when considering multiple forms of parameters in the complexity control [15].

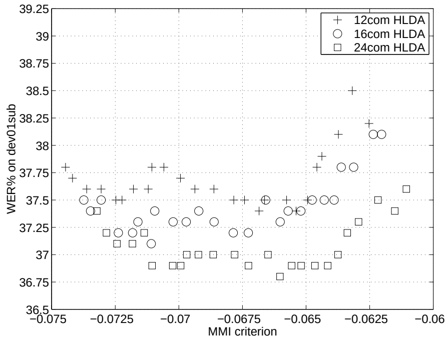

Fig. 1. Held out data MMI criterion vs. WER

Figure 1 shows the held-out data MMI criterion value against WER. The correlation between the two is quite poor. The outliers were found to heavily influence the value of the MMI criterion. This was the motivation for using a marginalized growth function, rather than the MMI criterion. In an initial plot of the held-out data, the growth function showed reasonable correlation. However in contrast to the standard MMI training the value of C was globally set.

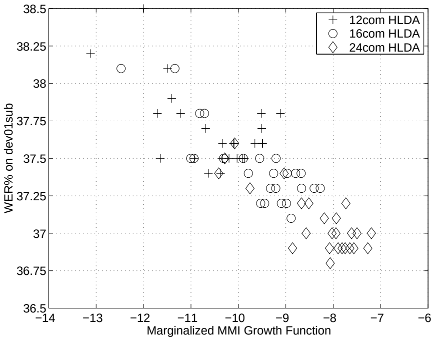

Fig. 2. Marginalized MMI growth function vs. WER

Figure 2 shows the marginalized MMI growth function against WER. For these experiments the smoothing constant was set on a per component basis in the standard MMI fashion. A strong correlation is observed with the WER. The best system selected is only 0.2% absolute worse than the actual best one.

A good complexity control scheme should correctly rank all the systems. Table 1 shows the recognition performance ranking prediction error computed using the method in section 5.1. The table consists of three sections. The first is the baseline number by ranking the systems according to the training data likelihood. This will simply yield an ordering on system complexity with no penalization. The first section also shows the ranking performance on using the held-out data scores. The poor performance of the MMI criterion is clearly shown. The likelihood score and growth function score are quite close. The second section of the table shows the ranking errors for evidence based approximations. As described in [15] there are issues with BIC when controlling multiple attributes. Using both standard BIC and penalized BIC ($\rho = 2$), the ranking scores were poor. The ML bound, using equation 7, yielded good performance. It is interesting that the approximations for the ML bound gave a slightly lower ranking error that the held-out likelihood. The best performance was obtained using the marginalized growth function (*GFunc Integral*) and is the score related to figure 2. Part of the performance gain over the held-out growth function score that it approximates, is because the global setting of C in equation 12 is not as powerful in terms of test set generalization and criterion convergence. The local setting of D_j described in section 3.2, was used in the marginalization of MMI function over the training data. As expected if the WER threshold is increased then the ranking error decreases, though the general ranking of all complexity control schemes remain about the same.

	WER threshold		
	0.0	0.1	0.2
Training Like	22.08	22.08	21.59
Held-out Like	8.94	8.89	8.19
Held-out MMI	37.40	37.40	35.91
Held-out GFunc	9.03	8.99	8.14
BIC (ρ = 1)	48.43	48.36	47.35
BIC (ρ = 2)	55.68	55.68	55.42
ML Bound	7.40	7.35	6.25
GFunc Integral	4.74	4.64	3.10

Table 1. Complexity control scheme ranking error (%)

5.3. Optimizing Local Complexity Attributes

In these experiments system complexity attributes were optimized on a "local" level. The complexity attributes to be optimized were: the number of Gaussians per state for a single transform HLDA system; the transform class specific number of useful dimensions of a multiple HLDA system. A larger 76 hour training corpus `he5train03sub` was used, which subsumes the `h5train00sub` corpus and includes an additional 166 Switchboard II Cellular conversation sides. An updated trigram language model was also used in the following full decoding experiments, while the test set remained the same.

System	WER%
12 component	36.1
VarMix	35.8
BIC (ρ = 1)	36.2
BIC (ρ = 2)	36.1
ML Bound	36.0
GFunc Integral	35.8

Table 2. Number of Gaussian components per state

Table 2 shows the word error rate for various complexity control systems used to determine the number of Gaussian components to train on each state. After the number of components was determined 4 iterations of Baum-Welch training were used to refine the estimates. The *VarMix* system is a simple occupancy based scheme where the number of components in a state is proportional to the state occupancy raised to a power, in this case 0.2. There are no restrictions on the number of components that may be assigned to a state. Using this system a 0.3% absolute reduction in error rate was obtained from the baseline 12 components per state system. All the other schemes for the table had the restriction that the number of components could only increase, or decrease, by one. The optional systems for all schemes were generated by fixing the state alignment and using the HTK `MU` and `MD` commands to change the number of components. These systems were then refined, given the fixed state alignment using Baum-Welch training. Comparing the various schemes the marginalized growth function gave better performance (though not significantly) than the ML-based schemes. Though the marginalized growth function performance was the same as the VarMix system, there were far more restrictions on changing the number of components in a state.

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

Table 3. Number of retained HLDA dimensions

Experiments were performed on a 12 component HLDA system with 65 HLDA transforms, with the aim of optimizing transform class specific number of retained dimensions. All the silence Gaussians were assigned to one transform class while all speech Gaussians were split into 64 distinct classes. The range of number of retained dimensions for all classes is in the set $\{28, ..., 52\}$. Table 3 shows the performance of various configurations. The use of multiple transforms shows significant gains over a single transform. Furthermore the use of marginalized growth function further reduced the error rate. More importantly the gain from this structural optimization process is found additive to MPE criterion based discriminative training [9] and MLLR based speaker adaptation after fixing the model structure. An overall gain of 0.7% absolute WER reduction is obtained over the 39 dimensional global transform system.

6. CONCLUSION

A complexity control 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. Complexity attributes optimized for a typical LVCSR task were the number of Gaussians per state and the useful dimensions of an HLDA system. Both attributes were optimized at global and local level. Initial experiments indicate that this form of discriminative complexity control 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) [12] and Minimum Phone Error (MPE) criteria [9]. In addition optimizing the complexity of discriminatively trained model structures will also be investigated, rather than the current scheme that uses ML training.

7. REFERENCES

[1] A. H. Welsh (1996). Aspects of Statistical Inference, John Wiley & Sons, Inc., 1996.

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

[3] 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.

[4] Z. Ghahramani & M. J. Beal (2000). Graphical Models and Variational Methods, *Advanced Mean Field Method—Theory and Practice*. MIT Press 2000.

[5] P.S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo (1989). Generalization of the Baum algorithm to Rational Objective Functions, *Proc. ICASSP'89*, pp. 631-634.

[6] Y. Normandin (1991). *Hidden Markov Models Maximum Mutual Information Estimation and the Speech Recognition Problem*, PhD thesis, McGill University, Canada.

[7] V. Valtchev (1995). *Discriminative Methods for HMM-based Speech Recognition*, PhD thesis, Cambridge University Engineering Department, England.

[8] 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.

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

[10] A. Gunawardana (2001). *The Information Geometry of EM Variants for Speech and Image Processing*. PhD Thesis, John Hopkins University, April 2001.

[11] 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.

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

[13] 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.

[14] W. Chou & W. Reichl (1999). Decision Tree State Tying Based on Penalized Bayesian Information Criterion, *Proc. ICASSP'99*, Vol. 1, Phoenix.

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