AUTOMATIC COMPLEXITY CONTROL FOR HLDA SYSTEMS

X. Liu, M. J. F. Gales and P. C. Woodland

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

ABSTRACT

Designing a state-of-the-art large vocabulary speech recognition systems is a highly complex problem. A wide range of techniques are available that affect the performance and number of free parameters. Selecting the appropriate complexity of system is both time-consuming and only a limited number of possible systems can be examined. This paper presents initial results on automatic system selection when both the number of dimensions and the number of components vary. Various complexity control schemes are discussed and evaluated. Limitations of schemes based on predicting held-out data log-likelihoods are described. In addition, problems of standard approximations for this task are detailed.

1. INTRODUCTION

A standard problem for both data modelling and classification tasks is how to determine the appropriate complexity of model given a limited amount of training data. The simplest approach, and the one adopted in the majority of speech recognition system design, is to use a held-out data set to evaluate each model's classification performance. Provided the held-out dataset is sufficiently large and representative of the data that the system is required to handle, this yields an excellent estimate of performance. However there are two major issues with this approach. First, given limited data it is undesirable to reserve a large held out data set for system evaluation. One approach to handling this problem is to use cross validation. For speech recognition this is not usually feasible due to the number of systems that would have to be built. The second problem, and the one considered in this paper, is that as the number of possible models increases it is not feasible to build and evaluate the performance of each model. This has led to the development of model selection schemes that only rely on attributes of the model and the training data. These techniques are normally split into two groups. The first is based on Bayesian techniques, where the model parameters are treated as random variables to be integrated out. One standard such approach is the Bayesian Information Criterion (BIC) [1]. The second category is information theory approaches. The complexity control problem is treated as finding a minimum code length, for example minimum description length (MDL) principle [2]. The two approaches are related, asymptotically both tend to the BIC approximation.

This paper examines automatic complexity control schemes for speech recognition using mixture of Gaussian HMM-based systems. In particular, the task is to select the appropriate complexity of system when both the dimensionality of the data and the number of Gaussian components vary. In contrast, most other work for ASR has only addressed complexity control for a single form of parameter, for example, the number of components, optimal HMM state clustering or adaptation transforms sharing [4, 5]. The form of projection scheme examined is linear, heteroscedastic linear discriminant analysis (HLDA) [6]. As this is a maximum likelihood based approach the complete feature vector is modelled. This allows valid comparison of log-likelihoods between systems of different projection dimensions. In this initial series of experiments only global decisions about the nature of the system are considered. This restricts the number of possible models. The overall aim is to be able to make *local* decisions such as using varying dimensions for different parts of speech [7]. However, using global decisions allow explicit evaluation of the word error rate (WER) and associated measures for all possible models. The next section describes the HLDA transform and optimisation scheme. Section 3 details the general area of complexity control and the schemes used in this paper. The results on a standard large vocabulary speech recognition task are given in section 4.

This work was supported by a grant from DARPA (MDA972-02-1-0013). The content does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred. Extensive use was made of equipment donated by IBM under an SUR award.

2. HLDA OPTIMIZATION

HLDA [6] is a linear projection scheme and may be viewed as a generalisation of LDA. It removes the restriction that all the within class covariance matrices are the same. The HLDA projection matrix, $\mathbf{A}^{\top}$, for a d-dimensional feature space, $\mathbf{o}$, may be written as

$$\hat{\mathbf{o}} = \mathbf{A}^{\top}\mathbf{o} = \begin{bmatrix} \mathbf{A}_{[p]}^{\top}\mathbf{o} \\ \mathbf{A}_{[d-p]}^{\top}\mathbf{o} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{o}}_{[p]} \\ \hat{\mathbf{o}}_{[d-p]} \end{bmatrix} \tag{1}$$

where the top p dimensions, $\hat{\mathbf{o}}_{[p]}$, are deemed to be those dimensions that contain discriminatory information, the *useful* dimensions, and the final (d-p)-dimensions, $\hat{\mathbf{o}}_{[d-p]}$, contain no useful information, the *nuisance* dimensions. HLDA transforms are trained using maximum likelihood (ML) estimation and the EM algorithm. The auxiliary function to be optimised is

$$\begin{aligned} \mathcal{Q}_{\mathrm{HLDA}}(\mathcal{M},\hat{\mathcal{M}}) &= \frac{\mathcal{T}}{2}\log\left|\mathbf{A}^{\top}\right|^{2} - \frac{\mathcal{T}}{2}\log\left|\check{\boldsymbol{\Sigma}}_{[d-p]}^{(g)}\right| \\ &\quad -\frac{1}{2}\sum_{m,\tau}\gamma_m(\tau)\log\left|\check{\boldsymbol{\Sigma}}_{[p]}^{(m)}\right| \end{aligned} \tag{2}$$

where

$$\check{\boldsymbol{\Sigma}}_{[p]}^{(m)} = \mathrm{diag}\left(\mathbf{A}_{[p]}^{\top}\boldsymbol{\Sigma}^{(m)}\mathbf{A}_{[p]}\right) \tag{3}$$

$$\check{\boldsymbol{\Sigma}}_{[d-p]}^{(g)} = \mathrm{diag}\left(\mathbf{A}_{[d-p]}^{\top}\boldsymbol{\Sigma}^{(g)}\mathbf{A}_{[d-p]}\right) \tag{4}$$

$\mathbf{\Sigma}^{(m)}$ and $\mathbf{\Sigma}^{(g)}$ are the covariance of component m and the global covariance in the original feature space, $\mathbf{o}$, $\mathcal{T}$ is the number of training data frames and $\gamma_m(\tau)$ is the posterior probability of being in component m at time τ.

Directly maximizing equation 2 via numerical methods is computationally very expensive. An alternative iterative optimisation scheme is given in [7] where equation 2 is expressed as

$$\mathcal{Q}_{\mathrm{HLDA}}(\mathcal{M},\hat{\mathcal{M}}) = \frac{\mathcal{T}}{2}\log\left|\mathbf{A}^{\top}\right|^2 - \frac{1}{2}\sum_i \mathbf{a}_i^{\top}\mathbf{G}^{(i)}\mathbf{a}_i - \frac{\mathcal{T}}{2}\log\left|\check{\mathbf{\Sigma}}^{(g)}_{[d-p]}\right| - \frac{1}{2}\sum_{m,\tau}\gamma_m(\tau)\log\left|\check{\mathbf{\Sigma}}^{(m)}_{[p]}\right| \qquad (5)$$

where $\mathbf{a}_i^{\top}$ is the i^{th} row of $\mathbf{A}^{\top}$ and

$$\mathbf{G}^{(i)} = \begin{cases} \sum_m \frac{\gamma^{(m)}}{\check{\sigma}^{(m)2}_{[p]i}}\mathbf{\Sigma}^{(m)} & i \notin \nu \\ \frac{\mathcal{T}}{\check{\sigma}^{(g)2}_{[d-p]i}}\mathbf{\Sigma}^{(g)} & i \in \nu \end{cases} \qquad (6)$$

$\gamma^{(m)} = \sum_\tau \gamma_m(\tau)$ is the mixture occupancy counts, ν is the set of indices of the nuisance dimensions. An iterative scheme is then used, alternating between updating the estimate of the covariance matrices, given the current transform, and the columns of the transformation. The transformation matrix columns are given by

$$\mathbf{a}_i^{\top} = \mathbf{c}_i\mathbf{G}^{(i)^{-1}}\sqrt{\frac{\mathcal{T}}{\mathbf{c}_i\mathbf{G}^{(i)^{-1}}\mathbf{c}_i^{\top}}} \qquad (7)$$

where $\mathbf{c}_i$ is the cofactor vector of row i of $\mathbf{A}^{\top}$. For any iterative estimation scheme it is necessary to initialise the model parameters, Here the projection matrix is initialised by examining the Fisher ratio values and selecting those dimensions with the largest Fisher ratios. In this form of system model parameters are estimated for *all* the dimensions, including the nuisance dimensions.

3. MODEL COMPLEXITY CONTROL

A standard problem in speech recognition, and machine learning in general, is how to obtain the appropriate complexity of system given a limited amount, $\mathcal{T}$, of training data, $\mathcal{X}$. In the case of HMM-based speech recognition this normally requires deciding the number of states in the system and the number of components to assign to each of those states. When *global* decisions about the structure of the models are made, for example how many components *all* the states should have, then it is possible to search over the space of all models, $\{\mathcal{M}\}$, evaluating the classification performance of each model on some held out data set, $\mathcal{D}$. However, as *local* decisions about complexity are made, the space of possible models increases. Explicit evaluation of all possible models becomes impractical. Some form of measure, which does not explicitly require calculating an error rate, is required. This section examines various criteria with respect to an HLDA system. Here the number of mixture components per state may vary as well as as the useful dimension size, p.

The simplest approach to reducing the computational cost is to compute the log-likelihood of some held-out data, given the MAP or ML estimate, $\hat{\mathbf{\Theta}}$, of each of the possible models. The appropriate model is then determined by

$$\hat{\mathcal{M}} = \arg\max_{\{\mathcal{M}\}}\left\{\log p(\mathcal{D}|\mathcal{M},\hat{\mathbf{\Theta}})p(\hat{\mathbf{\Theta}}|\mathcal{M})P(\mathcal{M})\right\} \qquad (8)$$

This assumes that the set of models is "close" enough to the correct model, that an increase in held-out data log-likelihood will decrease WER. This assumption may be poor, particularly for highly complex processes such as speech recognition. Even if this assumption is reasonable, it may still be impractical to evaluate the log-likelihood when the set of possible models becomes very large. In addition, it is preferable to make use of all the available training data, rather than using a held-out data set, or cross validation.

An alternative approach is to use Bayesian model selection techniques. Rather than using the MAP or ML estimates of the model parameters, the model parameters are integrated out to get an estimate of the marginal likelihood, or *evidence*, for each model, $p(\mathcal{X}|\mathcal{M})$. This is then weighted by the prior for each model, $P(\mathcal{M})$. This selection process is normally written as

$$\hat{\mathcal{M}} = \arg\max_{\{\mathcal{M}\}}\left\{\log P(\mathcal{M})\int p(\mathcal{X}|\mathbf{\Theta},\mathcal{M})p(\mathbf{\Theta}|\mathcal{M})\mathrm{d}\mathbf{\Theta}\right\} \qquad (9)$$

This form of marginalisation automatically penalises overly complex models. However it is normally computationally intractable to directly compute the marginal log-likelihood in equation 9. This has led to the development of various approximate schemes. For this paper the model parameters priors, $p(\mathbf{\Theta}|\mathcal{M})$, and model priors, $P(\mathcal{M})$, are assumed to be uninformative.

One of the simplest and most commonly used approximations to Bayesian model selection in the Bayesian information criterion (BIC) [1]. A general form is shown in equation 10.

$$\log p(\mathcal{X}|\mathcal{M}) \approx \log p(\mathcal{X}|\hat{\mathbf{\Theta}},\mathcal{M}) - \rho\cdot\frac{k}{2}\log\mathcal{T} \qquad (10)$$

where $\hat{\mathbf{\Theta}}$ is the ML estimate of model parameters. Schwartz [1] proved that, with the restriction that $\rho = 1$, this is a first order asymptotic approximation to 9 as $\mathcal{T}\to\infty$. For a d-dimensional HLDA system with p useful dimensions, the number of free parameters, k, is given by $k = d^2 + 2Mp + 2(d-p) + M$, where M is the total number of Gaussian components and the number of transitions and states are fixed. In [4] penalized BIC, where ρ is not constrained to be one, was proposed. ρ is usually tuned to a particular task or model set. It is designed to take into account two aspects of the approximation. For real data, such as speech, the samples are seldom independent, or conditionally independent, of one another. ρ allows this dependence to be taken into account. This results in ρ typically being greater than 1. The second aspect is to compensate for the ignored higher order terms when there is finite training data. If these higher-order terms vary dramatically depending in the nature of the parameters the performance of penalised BIC can be limited.

The penalised BIC approximation in equation 10 uses the log-likelihood of the training data. For each model this must be computed using the training data. An alternative is to use the log-likelihood of the complete dataset, the auxiliary function. Multiple systems are then able to share the same complete dataset, for example all the 12 component systems could use the same set of posteriors, $\gamma_m(\tau)$. This can dramatically reduce the computational cost. Even when the number of components varies the state posteriors may be fixed. There are two issues with this approximation. First, the ordering obtained from the log-likelihoods is assumed to be same as that of the auxiliary functions. However differences in auxiliary function values (for a fixed model) are underestimates of differences in the log-likelihood. Second, the approximations are normally based around the ML, or MAP, estimate of the model parameters. These are not obtained if the complete dataset is fixed.

Laplace's method approximates the marginal likelihood by fitting a Gaussian at the optimum of the model parameters and computing the volume under that Gaussian [3]. This yields

$$\log p(\mathcal{X}|\mathcal{M}) \approx \log p(\mathcal{X}|\hat{\Theta}, \mathcal{M}) + \frac{k}{2}\log 2\pi - \frac{k}{2}\log \mathcal{T} - \frac{1}{2}\log\left|\frac{1}{\mathcal{T}}\mathbf{I}(\hat{\Theta})\right| \quad (11)$$

where $\mathbf{I}(\hat{\Theta})$ is the *Fisher information matrix* defined as $\mathbf{I}(\hat{\Theta}) = \nabla^2 \log p(\mathcal{X}|\hat{\Theta}, \mathcal{M})$. Laplace's approximation is a second order approximation to the marginal likelihood. Unfortunately, computing $\mathbf{I}(\hat{\Theta})$ for a large vocabulary speech recognition systems is impractical. In this paper $\mathbf{I}(\hat{\Theta})$ is approximated using a block-diagonal structure. Each component is assumed independent of all others and the mean and variance parameters of each component are assumed independent. Furthermore, the HLDA transform parameters are assumed independent of other parameters. To further simplify the problem, the auxiliary function rather than the log-likelihood is used. The approximated log marginal auxiliary function may be expressed as (from equation 5)

$$\begin{aligned}\hat{\mathcal{M}} = \arg\max_{\{\mathcal{M}\}} \Bigg\{ & \mathcal{Q}_{\mathrm{HLDA}}(\mathcal{M}, \tilde{\mathcal{M}}) - \frac{1}{2}(p+1)\sum_m \log \gamma^{(m)} \\ & - \frac{1}{2}\sum_i \log\left|\frac{\mathcal{T}}{|\hat{\mathbf{A}}^\top|^2}\hat{\mathbf{c}}_i\hat{\mathbf{c}}_i^\top + \mathbf{G}^{(i)}\right| + \frac{3}{2}\sum_m \log\left|\check{\mathbf{\Sigma}}^{(m)}_{[p]}\right| \\ & + \frac{3}{2}\log\left|\check{\mathbf{\Sigma}}^{(g)}_{[d-p]}\right| + \frac{1}{2}[(p+1)M - p + d + 1]\log 2 \\ & - \frac{1}{2}(d-p+1)\log\mathcal{T} + \frac{k}{2}\log 2\pi \Bigg\} \quad (12)\end{aligned}$$

where $\tilde{\mathcal{M}}$ is the model used to obtain the complete dataset.

4. EXPERIMENTAL RESULTS

The CU-HTK Hub5 system for Switchboard evaluation was used to evaluate the model selection schemes. A standard 68 hour subset of the Hub5 training was selected, containing training data for 862 Switchboard I conversation sides and a subset of 92 Call Home English (CHE) sides. The standard acoustic features were 13 order cepstral coefficients, including the zero order, with first order and second order derivatives, a 39 dimensional feature vector. For all HLDA projection experiments third derivatives were added to give a 52-dimensional feature vector. The cepstral coefficients are derived from a modified PLP analysis, using Mel-scale filter-bank data, with the frequency range from 125Hz to 3.8kHz. Cepstral features were normalized for each conversation side via side based cepstral mean and variance normalization, and vocal tract length normalization. Using this data, a continuous density, mixture of Gaussian, cross-word triphone, gender independent HMM system was trained using maximum likelihood estimation. State clustered decision tree tying was used to specify 6168 speech states. All recognition experiments used a tri-gram language model. A 3 hour subset of the 2001 development data was used as the test data and held-out data.

This paper presents initial experiments where global decisions are made about the complexity of the model. This allows all the measures to be evaluated, including word error rate. The aim of these experiments is to establish which measures are most closely related to word error rate. The range of model varied by: the number of components per state from the set $\{12, 16, 24\}$; the number of useful dimensions from the set $\{28, \ldots, 52\}$. This gave a total of 75 models. In addition $\{12, 16, 24\}$ component systems using the standard front-end were evaluated.

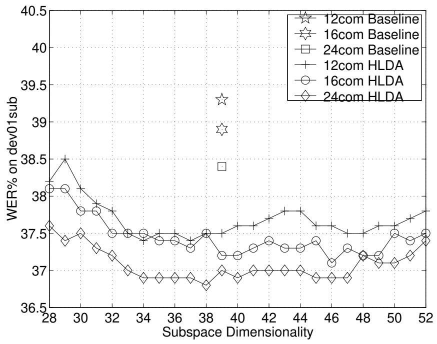

Fig. 1. Test set word error rate for all possible models, with the standard front-end 12, 16 and 24 component performance

Figure 1 shows the WER performance of the 75 HLDA systems and the three standard front-end systems on the test data. This is the "gold-standard" of model selection, classification performance. The HLDA systems significantly out performed the standard front-ends for the equivalent number of components per state. As expected there was some random variation in the performance, since we do not have an infinite held out test set. The best performance, 36.8%, was obtained using 24 components per state and an HLDA projection from 52 dimensions to 38 dimensions. However, similar performance was obtained from a range of useful dimensions, approximately 33 to 47.

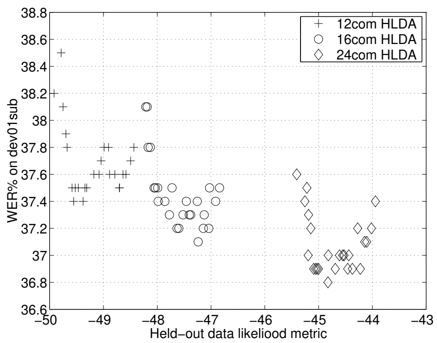

Fig. 2. Test data word error rate against held-out data log-likelihood metric

The standard complexity measures are based on producing systems that model unseen data well, i.e. the log-likelihood of held-out data. Figure 2 shows the relationship of the held-out data log-likelihood with WER. If log-likelihood truly predicted the WER, WER would decrease as log-likelihood increased. Though the figure shows this general trend, there is significant variation. Using this measure, the most complex system, 24 components per state and 52 useful dimensions, was selected. This had an error rate of 37.4% significantly worse than the best system performance of 36.8%.

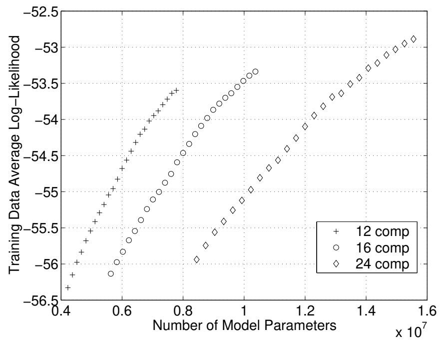

Fig. 3. Average training data log-likelihood against number of model parameters for all systems

Despite the limitations of using held-out log-likelihood to predict WER, it is useful to examine the performance of the standard approximations. For BIC and penalised BIC there is assumed to be sufficient training data that the effects of the normalised Fisher information matrix is negligible. The approximate marginal log-likelihood is only a function of the log-likelihood and number of model parameters. Hence, the *training data* log-likelihood is expected to monotonically increase as the number of model parameters increases, irrespective of the form of the parameters. Figure 3 shows this relationship for the 75 HLDA models. There are three distinct lines associated with the 12, 16 and 24 component systems. For each of these systems the training data log-likelihood increases as the number of useful dimensions increases. However, if the number of parameters is selected, for example, as approximately 6 million, the average training data log-likelihood is either -54.7 or -55.8, depending on whether parameters are used for increasing the number of components or the number of useful dimensions. This indicates a major limitation of BIC when selecting over models with different forms of parameters, even when ρ is allowed to vary. Varying ρ to minimise WER gave a performance of 37.4%. This is disappointing as ρ was tuned using the true WER. However, if only the useful dimensions, or number of components is varied, it is possible to select ρ so that the minimum WER system is selected. Similar trends are observed when penalising the auxiliary function with BIC, rather than log-likelihood.

In contrast to BIC, Laplace's approximation takes into account the second order terms. Figure 4 shows the WER against Laplace's approximation value. In a similar fashion to the average held-out data log-likelihood there is a general trend of improved performance as the Laplace's approximation value increases. Though again there is significant variation. Using this measure the best model had an error rate of 37.1%. At present no experiments have been performed where a form of ρ is introduced to help account for the dependencies in the training data.

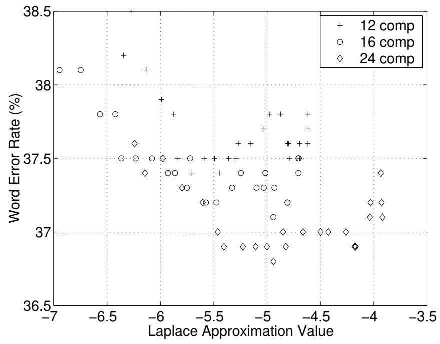

Fig. 4. Test set word error rate against Laplace's approximation measure

5. CONCLUSION

This paper has investigated how to select the appropriate complexity of a HMM-based speech recognition system when both the number of components per state and the number of dimensions retained with an HLDA projection. In this initial investigation a restricted number of possible models were examined to allow all measures, including word error, to be assessed. Current complexity controls are normally derived from either Bayesian schemes based on correctly modelling the data, or coding schemes for minimising the bit rate. When applying these measures for classification, the assumption is made that classification performance improves as the log-likelihood on held-out data increases. For speech recognition it is found that, though there is some level of correlation, for detailed selection it is not appropriate. In addition the limitations of some standard complexity control schemes were examined. None of the likelihood based schemes currently appear suitable for detailed complexity control task. Future research work will be focused on discriminative criteria that are more closely related to predicting held-out data word error rates, rather than held-out data log-likelihoods.

6. REFERENCES

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

[2] P. D. Grünwald (1998). *The Minimum Description Length Principle and Reasoning under Uncertainty*, Ph.D. Thesis, University of Amsterdam, Amsterdam, 1998.

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

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

[5] K. Shinoda & T. Watanabe (1995). Speaker Adaptation with Autonomous Model Complexity Control by MDL Principle, *IEEE Transactions on Speech and Audio Processing*, pp. 717–720, Vol. 8, 1995.

[6] N. Kumar (1997). *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*, Ph.D. Thesis, John Hopkins University, Baltimore.

[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.