CAMBRIDGE UNIVERSITY

ENGINEERING DEPARTMENT

AUTOMATIC MODEL COMPLEXITY CONTROL USING MARGINALIZED DISCRIMINATIVE GROWTH FUNCTIONS

X. Liu & M. J. F. Gales

CUED/F-INFENG/TR.490

Aug 2004

Cambridge University Engineering Department
Trumpington Street
Cambridge. CB2 1PZ
England

E-mail: {xl207,mjfg}@eng.cam.ac.uk
http://mi.eng.cam.ac.uk/~{xl207,mjfg}

Abstract

How to select the optimal model structure with the "appropriate" complexity, is a standard problem for training large vocabulary continuous speech recognition (LVCSR) systems. State-of-the-art LVCSR systems are highly complex. A wide variety of techniques are currently used which alter the system complexity and word error rate (WER). Explicitly evaluating systems with varying structural configurations is computationally infeasible. Most existing complexity control schemes can be classified into two types, Bayesian learning techniques and information theory approaches. An inherent assumption is made in these schemes, such that increasing the likelihood on held-out data can decrease the WER. However this correlation has been found quite weak for speech recognition systems. It would be preferable to use a complexity control scheme that is more closely related to WER. This paper presents a novel discriminative complexity control technique, the marginalization of a discriminative growth function. A two stage approach is used to make the marginalization of growth functions 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 discriminative growth functions outperforms using held out data likelihood and standard Bayesian schemes, in terms of both recognition performance ranking error and word error.

Table of Notations

General expressions for probability and matrix algebra:

$p(\cdot)$	probability density function
$P(\cdot)$	probability mass or a posterior distribution
$\{\cdot\}^\top$	transpose of a matrix
$\vert \cdot \vert$	determinant of a square matrix
$\{\cdot\}^{-1}$	inverse of a square matrix
$\nabla\{\cdot\}$	gradient of a matrix function

Expressions for model complexity control:

$\mathcal{M}$	model structural configuration
λ	valid parameterization of some model structure
$\tilde{\lambda}$	current parameter estimate
$\hat{\lambda}$	optimal parameter estimate
$\mathcal{O}$	utterance with finite length
$\mathcal{W}$	reference word sequence
$\tilde{\mathcal{W}}$	arbitrary word sequence
$\mathcal{S}$	hidden variable
ψ	hidden variable sequence
$\mathcal{F}$	training criterion
$\mathcal{A}(\tilde{\mathcal{W}}, \mathcal{W})$	accuracy measuring of a word sequence $\tilde{\mathcal{W}}$ against a reference $\mathcal{W}$
$\mathcal{G}$	growth function of a discriminative criterion
$\mathcal{L}$	strict lower bound of ML criterion or a discriminative growth function
$\mathcal{Q}$	auxiliary function derived for ML or a discriminative criterion
$\mathcal{P}$	variational or pseudo posterior distribution
$\mathbf{o}_\tau$	n dimensional acoustic observation at a time instance τ
$\gamma_j(\tau)$	Gaussian posterior occupancy
C	global scalar regularization term in discriminative growth functions
D	smoothing constant used in EBW
E	smoothing constant used in MMI and MPE training

Expressions for HLDA subspace projection:

$\mathbf{A}$	$n \times n$ square HLDA full transform
$\mathbf{A}_{[p]}$	$p \times n$ non-square useful HLDA projection
$\mathbf{A}_{[n-p]}$	$(n-p) \times n$ non-square nuisance HLDA projection
$\boldsymbol{\mu}^{(g)}$	global mean vector
$\boldsymbol{\Sigma}^{(g)}$	global covariance matrix
$\boldsymbol{\mu}^{(j)}$	mean vector of a Gaussian density
$\boldsymbol{\Sigma}^{(j)}$	covariance matrix of a Gaussian density
$\check{\boldsymbol{\mu}}^{(j)}$	transformed Gaussian mean
$\check{\boldsymbol{\mu}}^{(j)}_{[p]}$	transformed Gaussian mean in the useful subspace
$\check{\boldsymbol{\Sigma}}^{(j)}$	transformed Gaussian covariance
$\check{\boldsymbol{\Sigma}}^{(j)}_{[p]}$	transformed Gaussian covariance in the useful subspace
$\mathbf{a}_i$	ith row of an HLDA transform
$\mathbf{c}_i$	cofactor vector of the ith row of an HLDA transform
$\mathbf{G}$	second order statistics for HLDA optimization
r	index of HLDA transformation classes

1 Introduction

How to select the optimal model structure with the "appropriate" complexity is a standard problem for training large vocabulary continuous speech recognition (LVCSR) systems. State-of-the-art LVCSR systems are highly complex. A wide variety of techniques are currently used which alter the system complexity and word error rate (WER). These techniques include decision tree based state clustering, using a mixture of Gaussians as a state density distribution and various dimensionality reduction schemes. Explicitly evaluating word error rates (WER), for all possible model structural configurations, is infeasible. It is therefore useful to find a criterion that predicts the WER ranking order, without requiring a variety of systems to be built.

Most existing complexity control schemes can be classified into two types. In *Bayesian learning* techniques model parameters are treated as random variables. The maximum likelihood (ML) criterion is integrated out in the parametric space. Normally such a marginalization is computationally intractable, which leads to various approximation schemes. These include Bayesian Information Criterion (BIC), Laplace approximation, variational approximation and Markov chain Monte Carlo (MCMC) sampling schemes. In the *information theory* approaches a complexity control problem is viewed as finding an appropriate code length, for a underlying data generation process [4]. These two approaches are closely related to each other. Both asymptotically tend to either a BIC [2] based first order, or Laplace approximation based second order expansion. These schemes have been previously studied for various complexity control problems for speech recognition systems. For instance, how to optimally determine the number of states in a decision tree based clustering, or the number of linear transforms for speaker and environment adaptation [24, 25]. An inherent assumption is made in these schemes, such that increasing the likelihood on held-out data can decrease the WER. However this correlation has been found quite weak for speech recognition systems [27]. It would be preferable to employ a complexity control scheme that is more closely related to WER. Along these lines a discriminative measuring of model over the data has previously been used [20], as a method of incrementally splitting Gaussian mixture components in a hidden Markov model (HMM) based speech recognition system. State alignment is fixed for both the correct and incorrect word sequences. For each state a splitting operation is considered, if it increases the posterior probability of the data over the correct state labeling. Unfortunately such a scheme does not consider the information of the underlying model structure size, during the splitting process. No penalization is given to over-complex structures. Hence such a Gaussian components splitting process can not be regarded as a control of model complexity, because no stopping criterion is available.

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) and minimum phone error (MPE) criterion. Due to sensitivity to outliers, for discriminative training criteria such as MMI, these can not be directly integrated over. Instead a related discriminative growth function is marginalized. The growth function of a discriminative criterion retains the criterion's curvature in the parametric space, and largely removes the underlying criterion's sensitivity to outliers. Using a generalized EM approach, a strict lower bound of the growth function can derived . It may simply be expressed in a computationally more tractable form for marginalization, using appropriate sufficient discriminative statistics. In this paper a two stage approach is used to make the marginalization of the growth function efficient. First a generalized EM based lower bound, related to a discriminative auxiliary function, is used to remove the dependence on the hidden variables. In the second stage 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. These transforms are shared locally among different parts of the system, as a feature space diagonalizing and dimensionality reduction scheme [23]. 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 each HMM state distribution. The second is the number of useful dimensions of for each HLDA transformation. The problem of simultaneously examining multiple complexity attributes, makes commonly used schemes such as BIC inappropriate for complexity control [27].

This paper is organized as follows. The next section reviews various standard complexity control techniques and their limitations. Section 3 discussed various discriminative training criteria, and issues with the directly marginalization of them for complexity control. Then two forms of discriminative growth functions are introduced, with detailed derivation of marginalization for automatic complexity control. Some implementation issues are discussed in section 4. Experimental results on a standard LVCSR tasks are presented in section 5.

Section 6 is the conclusion and possible extension of future research.

2 Complexity Control

A standard problem in LVCSR training, and machine learning in general, is how to select a model structure that generalizes well to unseen data with appropriate complexity. Model structures which are too simple lack of sufficient power to learn the observed data. On the other hand, over-fitting structures are too complex and will lead to bad generalization performance. A complexity control problem considers the selection of an optimal model structural configuration $\hat{\mathcal{M}}$ 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 a speech recognition task, the generalization to unsee test data $\mathcal{O}^\mathcal{H}$ is commonly measured by word error rate (WER). The ultimate aim of controlling the complexity of a speech recognition system is to achieve minimum WER on test data, by optimizing the underlying model structure. In this sense WER is a "golden" complexity control criterion with zero risk. However explicitly accessing WER requires systems with various structural configurations to be built and evaluated. For highly complex speech recognition systems this is infeasible. Instead of directly measuring the WER for each candidate model structure, conventional complexity control schemes assume that all candidate structures are "close" to the genuine generator of human speech patterns, so that increasing the likelihood on unseen data will decrease the WER.

$$\hat{\mathcal{M}} \quad = \quad \arg\max_{\mathcal{M}} p(\mathcal{O}^\mathcal{H}, \mathcal{W}|\hat{\lambda}, \mathcal{M})p(\hat{\lambda}|\mathcal{M})P(\mathcal{M}) \tag{1}$$

This inherent assumption over the correlation between WER and likelihood of unseen data, motivates the use of validation test as an alternative to directly accessing WER. A set of appropriately selected unseen data can be used to perform likelihood test, over various model structures instantiated with their optimal parameters $\hat{\lambda}$. These parameters are estimated either in a maximum likelihood (ML) or maximum *a posteriori* (MAP) sense. It however still requires training of individual systems. Most of existing complexity control techniques make use of the training data, as there is a strong correlation between unseen data likelihood, and the *marginal* likelihood of the training data given a particular model structure. These schemes can classified into two categories. In *Bayesian learning* 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 [4]. These two approaches are closely related to each other. Both can be explicitly expressed as the training data marginal likelihood given a model structure, and asymptotically tend to the Bayesian Information Criterion (BIC) approximation [2]. In this section these two categories of complexity control schemes will be reviewed, and their limitations discussed.

2.1 Bayesian Complexity Control

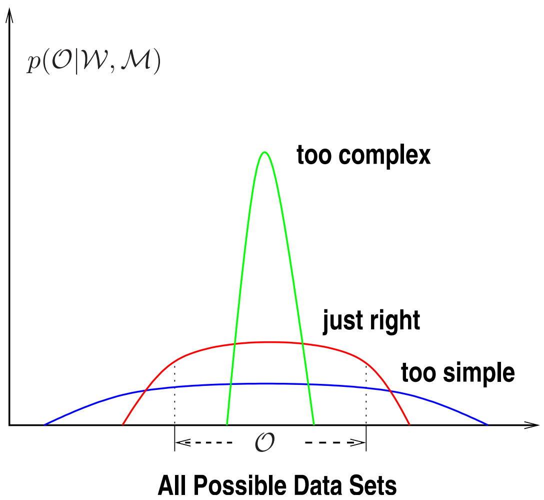

Figure 1: *Three model structures with different complexity*

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

$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} P(\mathcal{M}) \int \mathcal{F}_{\mathrm{ML}}(\lambda, \mathcal{M}) p(\lambda|\mathcal{M}) \mathrm{d}\lambda \tag{2}$$

Here λ denotes a parameterization of $\mathcal{M}$, and the ML criterion is given by

$$\mathcal{F}_{\mathrm{ML}}(\lambda, \mathcal{M}) = p(\mathcal{O}|\lambda, \mathcal{W}, \mathcal{M}) P(\mathcal{W}). \tag{3}$$

Bayesian approaches perform model selection, by evaluating the evidence integral for each model structure, and treating the model parameters as unknown random variables to be integrated out in the parametric space. Over-simple model structures are not powerful enough to model the observed data. On the other hand, over-complex model structures are penalized for having too much freedom in their parametric space. They are over-fitting to a small amount of observed data, with bad generalization performance. This is shown in figure 2.1. Hence the over-fitting problem may be avoided using a Bayesian approach. This is because having more parameters certainly provides more power in modeling observed data. Furthermore a system with appropriate complexity is expected to be capable of modeling a certain range of "interesting" observed or unsee data sets, giving a high evidence for those but little for others. This property of Bayesian evidence is often referred to as *Ockham's Razor* [1].

The language model probability $P(\mathcal{W})$ is normally optimized on additional text data. The optimization of its complexity, such as the number of distinct tokens in a standard N-gram, or the number of word clusters in a class based language model, is not considered in this paper. Only the structural optimization of acoustic models is studied. If one does not have any knowledge about the model prior distribution $P(\mathcal{M})$, or a reason to favor one model structure over another, then a uniform model prior may be used. Similarly the parameter prior distribution $p(\lambda|\mathcal{M})$ may also be treated as uninformative, due to lack of knowledge about the underlying distribution. In this work it is assumed to be a uniform distribution. Alternatively various assumptions may be made about this distribution's structure, which typically constrain it to be a conjugate prior distribution for $p(\mathcal{O}|\lambda, \mathcal{W}, \mathcal{M})$, so that the evidence integration problem may be relatively simplified. Nevertheless how to select a form of the parameter prior distribution is always subjective.

In practice it is however often computationally intractable to directly integrate out the marginal likelihood in equation 2. Therefore an appropriate approximation scheme is required. In the following sections four approximation schemes are discussed. These are a first order asymptotic expansion using Bayesian Information Criterion (BIC), a Laplace approximation based second order asymptotic expansion, an expectation maximization (EM) and variational approximation based approximation, and Markov chain Monte Carlo (MCMC) style sampling schemes.

2.1.1 Bayesian Information Criterion (BIC)

Bayesian Information Criterion (BIC) is the most widely used [2] approximation scheme for evidence integral. This complexity control criterion is simply expressed in terms of penalized log likelihood evaluated at the ML or MAP estimate of model parameters $\hat{\lambda}$,

$$\log p(\mathcal{O}|\mathcal{W}, \mathcal{M}) \approx \log p(\mathcal{O}|\hat{\lambda}, \mathcal{W}, \mathcal{M}) - \rho \cdot \frac{k}{2} \log \mathcal{T} \tag{4}$$

where k denotes the number of free parameters in $\mathcal{M}$. ρ is a penalization coefficient which may be tuned to the specific task [24]. Schwartz proved that when $\rho = 1$, BIC is a first order asymptotic expansion of the evidence integral, under certain regular assumptions upon the probability density $p(\mathcal{O}|\lambda, \mathcal{M})$. The Bayesian evidence integration in equation 2, may be asymptotically approximated via Taylor series expansion around the optimum of model parameters $\hat{\lambda}$, when the number of training samples $\mathcal{T}$ is infinitely large. Recent research has found a major limitation of BIC when simultaneously optimizing the multiple complexity attributes considered in [27].

2.1.2 Laplace Approximation

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 the likelihood curvature, in the parametric space. The Gaussian mean is at the optimum of the model parameters. Its covariance is the Hessian, or often referred to as *fisher information* matrix in various statistical inference literature, is also evaluated at the optimum of model parameters. The volume under that Gaussian is computed as an approximation, as is shown in figure 2.1.2.

$$\log p(\mathcal{O}|\mathcal{W},\mathcal{M}) \quad \approx \quad \log p(\mathcal{O}|\hat{\lambda},\mathcal{W},\mathcal{M}) - \frac{1}{2}\log\left|-\nabla^2_{\lambda=\hat{\lambda}}\log p(\mathcal{O}|\lambda,\mathcal{W},\mathcal{M})\right| + \frac{k}{2}\log 2\pi \tag{5}$$

In practice the computation of the Hessian is non-trivial for model structures containing vast amounts of free parameters. This means computing and storing the Hessian in a full matrix structure is infeasible. An alternatively approach is to use a block diagonal approximation. It is assumes that model parameters belonging to different part of the underlying model structure, such as individual Gaussian components, are considered largely independent against one other. This is the approach adopted in this paper, and will be addressed with other implementation issues with more details in later sections. It should also be noted that under a large number assumption, when the number of training data samples $\mathcal{T}$ is infinitely large, Laplace approximation tend to the same asymptotic expansion as BIC.

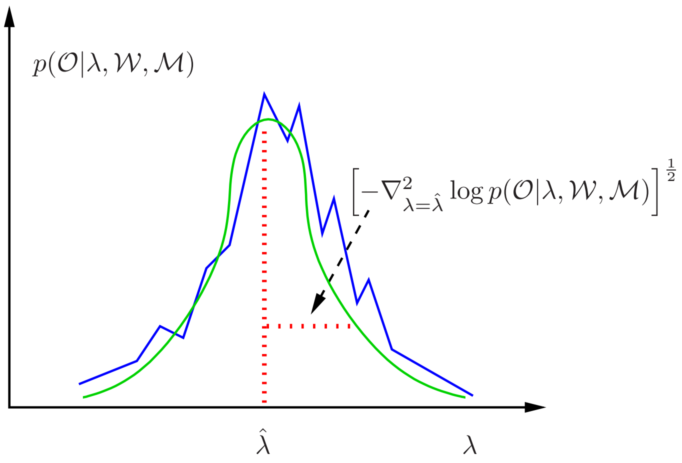

Figure 2: *Laplace Approximation*

2.1.3 EM and Variational EM

One issue with both of the above two approximation schemes is that the log-likelihood for each structural 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 expectation maximization (EM) approach this may be expressed as

$$\begin{aligned}\log p(\mathcal{O}|\lambda,\mathcal{W},\mathcal{M}) &\geq \log p(\mathcal{O}|\tilde{\lambda},\mathcal{W},\mathcal{M}) + \mathcal{Q}_{\mathrm{ML}}(\lambda,\tilde{\lambda}) - \mathcal{Q}_{\mathrm{ML}}(\tilde{\lambda},\tilde{\lambda}) \\ &= \mathcal{L}_{\mathrm{ML}}(\lambda,\tilde{\lambda})\end{aligned} \tag{6}$$

where the standard EM auxiliary function for HMMs is given by

$$\mathcal{Q}_{\mathrm{ML}}(\lambda,\tilde{\lambda}) \quad = \quad \sum_{j,\tau}\gamma_j(\tau)\log p(\mathbf{o}_\tau|\psi_\tau=\mathcal{S}_j,\lambda,\mathcal{M}) \tag{7}$$

and $\{\psi\}$ is the set of discrete hidden variables allowed by the reference transcription. $\gamma_j(\tau) = P(\psi_\tau = \mathcal{S}_j|\mathcal{O},\mathcal{W},\tilde{\lambda},\mathcal{M})$ is the frame hidden variable posterior. $\tilde{\lambda}$ is the *current* parameterization for $\mathcal{M}$. $\psi_\tau = \mathcal{S}_j$ indicates that an acoustic feature vector $\mathbf{o}_\tau$ was generated by a hidden 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 integral.

$$\log p(\mathcal{O}|\mathcal{W},\mathcal{M}) \quad \geq \quad \log\int\exp\left(\mathcal{L}_{\mathrm{ML}}(\lambda,\tilde{\lambda})\right)p(\lambda|\mathcal{M})\mathrm{d}\lambda \tag{8}$$

The right hand side of inequality 8 can be efficiently integrated out using the Laplace approximation. A special case of this is to simply use the standard ML auxiliary function [27], while ignoring the other two terms independent of λ in equation 6. 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. A related approximation scheme is variational approximation [8]. In the same fashion Jensen inequality is also applied to derive a evidence lower bound. The scheme approximates the joint posterior distribution over both model parameters and hidden variables $P(\mathcal{S}_{j,\tau}, \lambda|\mathcal{O}, \mathcal{W}, \mathcal{M})$, using a relatively simple and tractable *variational* distribution $\mathcal{P}(\psi, \lambda)$ in the modified E step.

$$\log p(\mathcal{O}|\mathcal{W}, \mathcal{M}) \geq \int \sum_{\psi \in \{\psi_r\}} \mathcal{P}(\psi, \lambda) \log \frac{p(\mathcal{O}, \psi, \lambda|\mathcal{M})}{\mathcal{P}(\psi, \lambda)} \mathrm{d}\lambda$$

Here $\{\psi_r\}$ denote valid hidden variable sequences allowable by the reference transcription. Maximizing the lower bound in the above equation, has been be shown equivalent to minimizing the Kullback-Leibler (KL) divergence between the variational distribution $\mathcal{P}(\psi, \lambda)$ and the true joint posterior distribution $P(\mathcal{S}_{j,\tau}, \lambda|\mathcal{O}, \mathcal{W}, \mathcal{M})$.

2.1.4 Markov chain Monte Carlo (MCMC) sampling

Another family of approximation methods for Bayesian evidence integral, are Markov chain Monte Carlo (MCMC) sampling schemes [9]. A simplest MCMC sampling based approximation to Bayesian evidence integral, is to average out a finite number of random samples taken in the parametric space as the following equation. This is often refereed to as simple Monte Carlo.

$$p(\mathcal{O}|\mathcal{W}, \mathcal{M}) \approx \frac{1}{N} \sum_i p(\mathcal{O}|\lambda_i, \mathcal{W}, \mathcal{M}) \tag{9}$$

Here λ_i is the ith sample of the model parameters and N is the total number of samples. The schemes therefore assumes that parameters prior is a uniform distribution, and samples are statistically independent against one other. Alternatively further assumptions may be further made upon the model parameter prior distribution. This means that samples will be either "rejected" via a binary decision as in *rejection sampling*, or appropriately weighted in *importance sampling*. However improperly weighting or rejection of samples can cause the Monte Carlo average to be dominated by a few samples with small variances. Furthermore all these few schemes assume statistical independence between samples. In contrast more complicated schemes like Gibbs sampling [9] may be used, which considers the statistical dependence between samples. The scheme iteratively picks up a model parameter sample, either in turn or randomly, which will then be replaced by a sample selected using a conditional distribution $P(\lambda_i|\lambda_1, ..., \lambda_{i-1}, \lambda_{i+1}, ..., \lambda_N)$. In general MCMC based sampling schemes do not provide a practical solution to evidence integral. This is because not only individual systems are required to be built for each candidate model structure being considered, but also for each parameter sample of the same structural configuration. Additionally the high dimensionality of the sampling space, and slow convergence of these schemes, make themselves computationally further infeasible to use for speech recognition systems.

2.2 Information Theory Methods

The second category of conventional complexity control techniques are based on *information theory*. These approaches regard the complexity control problem as finding an appropriate code length [4] for the underlying data generation process. Probabilistic distributions, that Bayesian methods use to model data, can be viewed as various means of generating codes. Assume both the sender and the receiver during a code transmission, know from which particular distribution $p(\mathcal{O}|\mathcal{W}, \mathcal{M})$, a code $\mathcal{O}$ is generated. Then according to Shannon Source Coding Theorem the entropy $-\log p(\mathcal{O}|\hat{\lambda}\mathcal{W}, \mathcal{M})$, penalized by a channel cost $\mathcal{C}(\mathcal{O}, \mathcal{M})$, forms a two-part code description length [3, 4, 5, 7].

$$\hat{\mathcal{M}} = \arg\min_{\{\mathcal{M}\}} \left\{ -\log p(\mathcal{O}|\hat{\lambda}, \mathcal{W}, \mathcal{M}) + \mathcal{C}(\mathcal{O}, \mathcal{M}) \right\} \tag{10}$$

The channel cost may be interpreted as the part of description length, which attributes to the complexity of the underlying code generator. The minimum description length (MDL), and a closely related minimum message length (MML) principle, select the model structure with the shortest two-part code description length. In literature this is sometimes also referred to as *stochastic complexity*. Various forms of code length have been studied. A most popular form is the mixture code length, which mathematical bears the same form as the Bayesian evidence integral in equation 2. Naturally this form of code length can be approximated via a first order asymptotic expansion, equivalent to BIC, or a Laplace approximation based second order expansion.

2.3 Limitations of Likelihood Based Schemes

One basic assumption made in the above two types of techniques is, that there is a strong correlation between WER and likelihood of unseen data. This means increasing likelihood on the unseen data can decrease WER in due course. However for complex patterns like human speech, such model correctness assumption is too strong for classical HMMs based speech recognition systems. Human speech does not have the statistics assumed by an HMM. A variety of strong assumptions, such as in modeling intra-frame and inter-frame correlation, are made on these models. These model correctness assumptions are known issues for conventional ML training for HMMs. Furthermore in recent research the correlation between WER and likelihood on unseen data, has been been fairly weak [27]. Therefore marginalizing the ML criterion, under the Bayesian or information theory framework, can be unsuitable for the complexity control of speech recognition systems. Once the correlation between WER and likelihood on unseen data is weak, various forms of approximation schemes visited earlier, are liable to fail in due course. It has also been found that BIC suffers particularly from this weak correlation, when used to simultaneously optimizing multiple complexity attributes. Naturally it would be preferable to marginalize a criterion, that is more closely related to the recognition error, without making an assumption of model correctness.

3 Discriminative Complexity Control

In contrast to conventional maximum likelihood learning, discriminative training criteria are closer to the recognition error. In this section, the marginalization of a discriminative measure is proposed as a novel complexity control technique. Two forms of popular discriminative training criteria, maximum mutual information (MMI) and minimum phone error (MPE) will be initially discussed. Various issues with directly marginalizing a discriminative training criterion, like MMI, will be discussed. Then a criterion related *discriminative growth function* is introduced, with largely reduced outliers sensitivity. A discriminative growth function computationally has a more tractable form to be marginalized over for complexity control.

3.1 Discriminative Training Criteria

In conventional ML training of HMM parameters, it is assumed that HMM is a genuine human speech generator. Given infinite amount of training data, the global ML estimate is optimal, in a sense that it is unbiased with minimum variance. However in practice such a model correctness assumptions is rarely true. Neither the training data quantity is unlimited, nor does human speech have the statistics assumed by an HMM. Furthermore, standard EM based ML learning of HMM model parameters, can only reach a local optimal estimate. In contrast model confusion over the training data is considered in discriminative training techniques. These techniques attempt to reduce the likelihood of alternative incorrect hypotheses, or more explicitly the recognition error. Hence discriminative training criteria are more closely related to recognition error, than the conventional ML criterion. Recently these have been successfully applied to LVCSR training [16, 18, 17]. In this section two popular forms of discriminative training criteria, maximum mutual information (MMI) and minimum phone error (MPE) will be introduced.

3.1.1 Maximum Mutual Information (MMI)

One of the most widely used discriminative training criteria, is the maximum mutual information (MMI) criterion. This is equivalent to maximizing the *a posteriori* probability of training data over the correct transcription $\mathcal{W}$. Following the expressions defined in the notation table, one can write

$$\mathcal{F}_{\mathrm{MMI}}(\lambda, \mathcal{M}) = \frac{p(\mathcal{O}, \mathcal{W}|\lambda, \mathcal{M})}{p(\mathcal{O}|\lambda, \mathcal{M})} \tag{11}$$

When the language model parameters $P(\mathcal{W})$ are fixed during MMI training, the MMI criterion is equivalent to conditional maximum likelihood (CML) [10]. In most research on MMI training, only acoustic model parameters are optimized, so both are equivalent to each other. Optimizing the MMI criterion implies, that not only the conventional ML criterion given the correct transcription $p(\mathcal{O}, \mathcal{W}|\lambda, \mathcal{M})$ is increased, but also decreased is the likelihood of a "composite" model [1]

$$p(\mathcal{O}|\lambda, \mathcal{M}) = \sum_{\tilde{\mathcal{W}}} p(\mathcal{O}|\lambda, \tilde{\mathcal{W}}, \mathcal{M}) P(\tilde{\mathcal{W}}) \quad (12)$$

which is summed over all possible hypotheses $\{\tilde{\mathcal{W}}\}$, including the correct word sequence. In literature these two are usually referred to as the *numerator* and *denominator* term respectively. For LVCSR MMI training it is infeasible to store all possible hypotheses for the denominator. In practice a finite number of confusable word sequences, are stored either in an N-best list or lattice, as a compact format to represent the model's confusion over the training data [15]. It has also been found that the MMI criterion is giving undue weight to outliers, which have very low posterior probability over the correct transcription [14]. MMI based discriminative training has been extensively used in most state-of-the-art LVCSR systems. Significant improvements over equivalent ML criterion trained models have been reported.

3.1.2 Minimum Phone Error (MPE)

In speech recognition the most commonly used recognition performance measurement is word error rate (WER). Using the MMI criterion the sentence level posterior probability over the correct word sequence is maximized. This well fits in the Viterbi search based decoding process. However it can be more useful to have training criteria more directly related to WER. The minimum word error (MWE) criterion is one such criterion more directly related to WER, in a sense that word level accuracy is computed via a recognition lattice for each training utterance [21, 22]. A closely related criterion is minimum phone error (MPE) criterion. Instead of evaluating recognition accuracy on word level, phone level accuracy over recognition lattices are computed in a word sense, under the constraint of the reference word transcription [17, 18]. The MPE criterion is expressed as the average correctness of all the paths $\{\tilde{\mathcal{W}}\}$ in a recognition lattice. The accuracy contribution from each path is simply weighted by its posterior probability. The computation of a word path accuracy measuring $\mathcal{A}(\tilde{\mathcal{W}}, \mathcal{W})$ normally involves a dynamic programming procedure.

$$\mathcal{F}_{\mathrm{MPE}}(\lambda, \mathcal{M}) = \sum_{\tilde{\mathcal{W}}} \frac{p(\mathcal{O}, \tilde{\mathcal{W}}|\lambda, \mathcal{M}) \mathcal{A}(\tilde{\mathcal{W}}, \mathcal{W})}{p(\mathcal{O}|\lambda, \mathcal{M})} \quad (13)$$

Empirical results on various LVCSR discriminative tasks have shown that MPE training typically outperforms MMI training. No significant has been found between MWE and MPE in terms of recognition WER, although MWE has been reported to be more powerful to fit the training data [18]. Most state-of-the-art LVCSR systems are trained either using MMI or MPE criterion.

3.2 Marginalizing Discriminative Training Criteria

One obvious form of discriminative complexity control, is to marginalize a discriminative criterion in the parametric space. This is in contrast to the marginalization of the ML criterion in conventional approaches. Analogically to Bayesian evidence integration, the idea is to examine how "discriminative" a particular model structure is, against competing incorrect recognition hypotheses, by considering all of its allowable parameterizations. Here if we take the MMI criterion for instance, this yields the following complexity control metric.

$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} \int \mathcal{F}_{\mathrm{MMI}}(\lambda, \mathcal{M}) p(\lambda|\mathcal{M}) \mathrm{d}\lambda \quad (14)$$

A similar form of MPE criterion integral may also be considered for complexity control. However there are important issues with a direct marginalization of discriminative criteria for complexity control. The primary issue is the criterion sensitivity to outliers. This is particularly a well known feature of the MMI criterion. The MMI criterion computation tends to give undue weights to outliers utterances with very low posteriors. State-of-the-art large vocabulary speech recognition systems are trained with hundreds of hours of data. Outliers,

[1] In practice language model probability is scaled by a constant $\alpha > 0$ to compensate for the dynamic range issue. Often the acoustic likelihood is also deweighted using the inverse of the language model probability scale. This will broaden the posterior distribution of different word paths in a lattice. Such increase in confusable data can improve generalization performance.

$$p(\mathcal{O}|\lambda, \mathcal{M}) = \sum_{\tilde{\mathcal{W}}} p(\mathcal{O}|\tilde{\mathcal{W}}, \lambda, \mathcal{M})^{\frac{1}{\alpha}} P(\tilde{\mathcal{W}})^{\alpha}$$

which are far from the decision boundary, can exist in the training corpora. They are often utterances with very low likelihood, or more explicitly, high recognition error rate. The sensitivity to outliers is a well known feature of the MMI criterion [12, 14]. In such a case the recognition performance ranking prediction, based on a direct MMI criterion marginalization, can be considerably distorted due to the presence of these outliers. The same issue exists with the MPE criterion. Consider the case if the MPE criterion over the whole training set, is expressed as a product of, instead of as a summation of, average correctness measuring over all training utterances. In a similar way to the MMI criterion, utterances with lower average correctness will be given undue weights. Therefore a direct marginalization of these discriminative training criteria, can be inappropriate for model complexity control.

Indeed in literature the MWE or MPE criterion is normally expressed in a form of summation of individual utterances' correctness. This may reduce the underlying criterion's sensitivity to outliers. This is also true for the MMI criterion, when it is expressed as a summation of utterance posteriors, instead of a product as given by its standard definition. Nevertheless for both criteria, if expressed in terms of a summation over all utterances, they would be mathematically non-trivial to optimize and integrate over. This is similar to the standard ML training, and also the marginalization of ML criterion in Bayesian learning techniques. The ML criterion, is expressed as the joint probability of all utterances given the reference transcription, model parameters and the structural configuration. By introducing the auxiliary function based lower bound, the optimization or marginalization of the ML criterion, can be largely simplified. Its convex auxiliary function based strict lower bound in equation 6 can be maximized or marginalized. If the ML criterion is expressed as the summation, of probability over all utterances, then its optimization and marginalization will become far more complicated.

3.3 Discriminative Growth Functions

One method to compensate the outliers sensitivity, is to explicitly deweight the posteriors of outliers utterances using a criterion smoothing function [14]. An alternative method investigated in this paper is to transform the underlying discriminative training criterion into an associated *growth function*, $\mathcal{G}(\lambda, \mathcal{M})$. Using this the original criterion's information is still retained, in a computational tractable form for marginalization in the parametric space. More importantly the underlying criterion's sensitivity to outliers can be largely removed. In this section firstly a general form of growth functions is introduced. Then two forms of discriminative growth functions are presented for the MMI and MPE criterion respectively. Both are in a computationally tractable form based on discriminative training auxiliary functions, for efficient marginalization. They can also be used to derive the standard extended Baum-Welch (EBW) based discriminative parameter re-estimation formula.

3.4 Generic Form of Growth Functions

A growth function is required be related to its associated standard discriminative training criterion, with largely reduced sensitivity to outliers. Furthermore it needs to be computationally more feasible to integrate over. The form of growth functions considered in this paper, is for a family of discriminative training criteria, which can be expressed as a ratio between two polynomials with positive coefficients. Both the MMI and MPE criteria are in this category. In the following equations $\mathcal{M}$ is omitted for a particular model structure being considered. Consider a discriminative training criterion expressed in the following form,

$$\mathcal{F}(\lambda) \quad = \quad \frac{\mathcal{F}_{\text{num}}(\lambda)}{\mathcal{F}_{\text{den}}(\lambda)} \tag{15}$$

a generic form of the associated growth function is proposed as,

$$\mathcal{G}(\lambda) \quad = \quad \mathcal{F}_{\text{den}}(\lambda)\left[\mathcal{F}(\lambda) - \mathcal{F}(\tilde{\lambda}) + C\mathcal{F}_{\text{sm}}(\lambda, \tilde{\lambda})\right] \tag{16}$$

where $\tilde{\lambda}$ is the *current* parameter estimate. The first two terms in the bracket contain information about the original criterion, and will thus give information about the curvature of the criterion surface in the parametric space. The third regularization term in the bracket, consists of some smoothing criterion or statistics, scaled by a constant $C > 0$. This removes the sensitivity to outliers, by penalizing highly unlikely word sequences. The exact form of the smoothing criterion $\mathcal{F}_{\text{sm}}(\lambda, \tilde{\lambda})$, depends on the underlying training criterion being considered. The criterion gradient information is also retained. The gradient of $\mathcal{G}(\lambda)$,

$$\frac{\partial \mathcal{G}(\lambda)}{\partial \lambda} \quad = \quad \left[\mathcal{F}(\lambda) - \mathcal{F}(\tilde{\lambda}) + C\mathcal{F}_{\mathtt{sm}}(\lambda, \tilde{\lambda})\right] \frac{\partial \mathcal{F}_{\mathtt{den}}(\lambda)}{\partial \lambda} + \mathcal{F}_{\mathtt{den}}(\lambda)\left[\frac{\partial \mathcal{F}_(\lambda)}{\partial \lambda} + C\frac{\partial \mathcal{F}_{\mathtt{sm}}(\lambda, \tilde{\lambda})}{\partial \lambda}\right] \tag{17}$$

when evaluated at the current parameter estimate $\tilde{\lambda}$, will be in the same turning point as the original criterion gradient when C approaches zero.

$$\lim_{C\to 0} \left.\frac{\partial \mathcal{G}(\lambda)}{\partial \lambda}\right|_{\lambda=\tilde{\lambda}} = \mathcal{F}_{\texttt{den}}(\tilde{\lambda}) \left.\frac{\partial \mathcal{F}(\lambda)}{\partial \lambda}\right|_{\lambda=\tilde{\lambda}} \tag{18}$$

In this paper such this form of discriminative growth functions are proposed as a model complexity control scheme, to reduce outliers sensitivity from the original criterion [28]. Another interesting property of a discriminative growth function, is it can be related the standard Extended Baum-Welch (EBW) discriminative updating formulae. Detailed derivations of using both MMI and MPE growth functions for complexity control will be given in the following sections.

3.4.1 MMI Growth Function

The form of the MMI growth function considered in this thesis is given by

$$\mathcal{G}(\lambda) = p(\mathcal{O}|\lambda) \left[\mathcal{F}_{\texttt{mmi}}(\lambda) - \mathcal{F}_{\texttt{mmi}}(\tilde{\lambda}) + C\mathcal{F}_{\texttt{sm}}(\lambda, \tilde{\lambda})\right] \tag{19}$$

where the smoothing criterion $\mathcal{F}_{\texttt{sm}}(\lambda, \tilde{\lambda})$ is given by

$$\mathcal{F}_{\texttt{sm}}(\lambda, \tilde{\lambda}) = P(\mathcal{W}|\mathcal{O}, \tilde{\lambda}) \tag{20}$$

The above smoothing criterion is equivalent to the MMI criterion, given in equation 11, evaluated at the current parameter estimate. It penalizes sentences with a very low posterior probability. As discussed in section 3.2, outliers for speech recognition systems are often sentences with very high recognition error or low likelihood. For the MMI criterion, the recognition error rate is approximated by the sentence posterior probability. Hence outliers may be included among these utterances with low posteriors. Giving undue weights to sentences with very low posteriors is an undesirable feature of the MMI criterion. Using the smoothing criterion in equation 20, the growth function's sensitivity to outliers may be reduced. Furthermore the term outside the bracket, $p(\mathcal{O}|\lambda)$, which is associated with the likelihood of a sentence, also helps reduce the sensitivity. Highly unlikely word sequences will be penalized.

The aim is to obtain a computationally efficient lower bound for the marginalized growth function. The growth function given in equation 19 may be re-written as

$$\mathcal{G}(\lambda) = p(\mathcal{O}, \mathcal{W}|\lambda) - P(\mathcal{W}|\mathcal{O}, \tilde{\lambda})p(\mathcal{O}|\lambda) + CP(\mathcal{W}|\mathcal{O}, \tilde{\lambda})p(\mathcal{O}|\lambda) \tag{21}$$

In the same fashion as the derivation of lower bound for Bayesian evidence, this can be rewritten in terms of the hidden state sequences $\{\psi\}$. A lower bound may be expressed using a generalized EM approach [19].

$$\begin{aligned}
\log \mathcal{G}(\lambda) &= \log \sum_{\psi} \mathcal{G}(\psi, \lambda) \\
&\geq \sum_{\psi} \mathcal{P}(\psi, \tilde{\lambda}) \log \frac{\mathcal{G}(\psi, \lambda)}{\mathcal{P}(\psi, \tilde{\lambda})} \\
&= \mathcal{L}_{\text{MMI}}(\lambda, \tilde{\lambda})
\end{aligned} \tag{22}$$

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

$$\mathcal{G}(\psi, \lambda) = p(\mathcal{O}, \psi, \mathcal{W}|\lambda) - P(\mathcal{W}|\mathcal{O}, \tilde{\lambda})p(\mathcal{O}, \psi|\lambda) + CP(\mathcal{W}|\mathcal{O}, \tilde{\lambda})p(\mathcal{O}, \psi|\lambda) \tag{23}$$

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

$$\mathcal{P}(\psi, \tilde{\lambda}) = \frac{\mathcal{G}(\psi, \tilde{\lambda})}{\sum_{\psi} \mathcal{G}(\psi, \tilde{\lambda})} \tag{24}$$

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

When C is big enough, such $\mathcal{P}(\psi, \tilde{\lambda})$ 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. It can be shown that the lower bound given in equation 22, can be re-expressed as

$$\mathcal{L}_{\text{MMI}}(\lambda, \tilde{\lambda}) \quad = \quad \log \mathcal{G}(\tilde{\lambda}) + \frac{\mathcal{Q}_{\text{MMI}}(\lambda, \tilde{\lambda}) - \mathcal{Q}_{\text{MMI}}(\tilde{\lambda}, \tilde{\lambda})}{\sum_{j,\tau} \gamma_j^{\text{MMI}}(\tau)} \tag{25}$$

where the MMI auxiliary function is defined as[3] More detailed proof can be found in appendix A.

$$\mathcal{Q}_{\text{MMI}}(\lambda, \tilde{\lambda}) \quad = \quad \sum_{j,\tau} \gamma_j^{\text{MMI}}(\tau) \log p(\mathbf{o}_\tau | \psi_\tau = \mathcal{S}_j, \lambda) \tag{26}$$

The MMI hidden variable occupancy $\gamma_j^{\text{MMI}}(\tau)$ is,

$$\gamma_j^{\text{MMI}}(\tau) \quad = \quad \gamma_j^{\text{num}}(\tau) - \gamma_j^{\text{den}}(\tau) + CP(\psi_\tau = \mathcal{S}_j | \mathcal{O}, \tilde{\lambda}) \tag{27}$$

where $\gamma_j^{\text{num}}(\tau) = P(\psi_\tau = \mathcal{S}_j | \mathcal{O}, \mathcal{W}, \tilde{\lambda})$ and $\gamma_j^{\text{den}}(\tau) = P(\psi_\tau = \mathcal{S}_j | \mathcal{O}, \tilde{\lambda})$ are the numerator and denominator posterior, as is referred in literature. The above equation is closely related to the standard form of MMI hidden variable occupancy,

$$\sum_\tau \gamma_j^{\text{MMI}}(\tau) \quad = \quad \sum_\tau \gamma_j^{\text{num}}(\tau) - \sum_\tau \gamma_j^{\text{den}}(\tau) + D_j \tag{28}$$

when D_j is a hidden variable specific smoothing constant. Given this form of MMI statistics, it can be trivially shown that the extended Baum-Welch (EBW) based discriminative parameter re-estimation formula can be derived [13]. Previous research on LVCSR MMI training shows that the selection of D_j, considerably affects the criterion convergence and test set generalization [16, 18]. A commonly used form of D_j is associated with the denominator occupancy $D_j = E \sum_\tau \gamma_j^{\text{den}}(\tau)$, where $E > 0$. This is equivalent to the MMI hidden state occupancy in equation 27 when $E = C$. The same form of smoothing function may be used for the growth function. 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 for standard MMI training.

The form of lower bound defined in equation 25 and the ML case given in equation 6 have similar forms. Given the current value of the criterion, the new value is estimated as the old value 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. The following marginalization of the MMI growth function lower bound is used in this paper for complexity control.

$$\hat{\mathcal{M}} \quad = \quad \arg\max_{\mathcal{M}} \int \exp\left(\mathcal{L}_{\text{MMI}}(\lambda, \tilde{\lambda})\right) p(\lambda | \mathcal{M}) \mathrm{d}\lambda \tag{29}$$

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

3.4.2 MPE Growth Function

The form of the MPE growth function considered in this paper is

$$\mathcal{G}(\lambda) \quad = \quad p(\mathcal{O}|\lambda) \left\{ \mathcal{F}_{\text{MPE}}(\lambda) - \mathcal{F}_{\text{MPE}}(\tilde{\lambda}) + C \sum_{\substack{\tilde{\mathcal{W}} \\ \mathcal{F}_{\text{MPE}}(\tilde{\lambda}) > \mathcal{A}(\tilde{\mathcal{W}}, \mathcal{W})}} P(\tilde{\mathcal{W}} | \mathcal{O}, \lambda) \left[\mathcal{F}_{\text{MPE}}(\tilde{\lambda}) - \mathcal{A}(\tilde{\mathcal{W}}, \mathcal{W}) \right] \right\} \tag{30}$$

The form of the smoothing criterion used, to compensate for the low accuracy contribution from hypotheses which are below the average level, $\mathcal{F}_{\text{sm}}(\lambda, \tilde{\lambda}) = \sum_{\substack{\tilde{\mathcal{W}} \\ \mathcal{F}_{\text{MPE}}(\tilde{\lambda}) > \mathcal{A}(\tilde{\mathcal{W}}, \mathcal{W})}} P(\tilde{\mathcal{W}} | \mathcal{O}, \lambda) \left(\mathcal{F}_{\text{MPE}}(\tilde{\lambda}) - \mathcal{A}(\tilde{\mathcal{W}}, \mathcal{W}) \right)$. Similar to the MMI growth function, this smoothing term in 30 removes the sensitivity to outliers, by penalizing highly

[3]Here only the part of MMI auxiliary function associated with hidden state densities are examined although a complete auxiliary function may also be derived for transition probabilities and hidden state priors.

unlikely word sequences. The other terms in the bracket contains information about the MPE criterion and will thus give information about the curvature of the criterion surface in the parametric space. Again $C > 0$ is a constant regularization term. The gradient of $\mathcal{G}(\lambda)$ has a similar property as the MMI growth function. When evaluated at the current parameter estimate $\tilde{\lambda}$, it will be in the same of direction as the MPE criterion gradient when C approaches zero. Using 13 equation 30 may be re-arranged as below.

$$\mathcal{G}(\lambda) \quad = \quad \sum_{\tilde{\mathcal{W}}} p(\mathcal{O},\tilde{\mathcal{W}}|\lambda)\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) - \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda})p(\mathcal{O}|\lambda) + C \sum_{\substack{\tilde{\mathcal{W}} \\ \mathcal{F}_{\mathrm{MPE}(\tilde{\lambda})>\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W})}}} P(\mathcal{O},\tilde{\mathcal{W}}|\lambda)\left[\mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) - \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W})\right] \quad (31)$$

To make the growth function marginalization more efficient, a lower bound of $\mathcal{G}(\lambda)$ may be derived using a generalized EM approach [19],

$$\begin{aligned}
\log \mathcal{G}(\lambda) &= \log \sum_{\psi} \mathcal{G}(\psi,\lambda) \\
&\geq \sum_{\psi} \mathcal{P}(\psi,\tilde{\lambda}) \log \frac{\mathcal{G}(\psi,\lambda)}{\mathcal{P}(\psi,\tilde{\lambda})} \\
&= \mathcal{L}_{\mathrm{MPE}}(\lambda,\tilde{\lambda}) \qquad (32)
\end{aligned}$$

where the hidden variable sequence ψ level expansion of the growth function is defined as below following equation 31.

$$\mathcal{G}(\psi,\lambda) = \sum_{\tilde{\mathcal{W}}} p(\mathcal{O},\psi,\tilde{\mathcal{W}}|\lambda)\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) - \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda})p(\mathcal{O},\psi|\lambda) + C\sum_{\substack{\tilde{\mathcal{W}} \\ \mathcal{F}_{\mathrm{MPE}(\tilde{\lambda})>\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W})}}} p(\mathcal{O},\psi,\tilde{\mathcal{W}}|\lambda)\left[\mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) - \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W})\right] \quad (33)$$

Applying Jensen's inequality for 32, similar to MMI growth functions, requires the hidden variable sequence "posterior" distribution $\mathcal{P}(\psi,\tilde{\lambda})$ must satisfy the non-negative and sum to one constraint. The form of posterior considered here is

$$\mathcal{P}(\psi,\tilde{\lambda}) \quad = \quad \frac{\mathcal{G}(\psi,\tilde{\lambda})}{\sum_{\psi}\mathcal{G}(\psi,\tilde{\lambda})} \qquad (34)$$

Again when C is big enough such non-negative and sum to one constraint can be satisfied. It can be shown that one can re-write 32 as the following, in a similar form to EM based standard ML training and the MMI growth function bound,

$$\mathcal{L}_{\mathrm{MPE}}(\lambda,\tilde{\lambda}) \quad = \quad \log \mathcal{G}(\tilde{\lambda}) + \frac{\mathcal{Q}_{\mathrm{MPE}}(\lambda,\tilde{\lambda}) - \mathcal{Q}_{\mathrm{MPE}}(\tilde{\lambda},\tilde{\lambda})}{\sum_{j,\tau}\gamma_j^{\mathrm{MPE}}(\tau)} \qquad (35)$$

where the MPE auxiliary function is given by [4]

$$\mathcal{Q}_{\mathrm{MPE}}(\lambda,\tilde{\lambda}) \quad = \quad \sum_{j,\tau} \gamma_j^{\mathrm{MPE}}(\tau) \log p(\mathbf{o}_\tau|\psi_\tau = \mathcal{S}_j,\lambda) \qquad (36)$$

The complete proof can be found in appendix B. The MPE Gaussian component occupancy $\gamma_j^{\mathrm{MPE}}(\tau)$ is given as below,

$$\gamma_j^{\mathrm{MPE}}(\tau) \quad = \quad \gamma_j^{\mathrm{num}}(\tau) - \gamma_j^{\mathrm{den}}(\tau) + C\sum_{\substack{\tilde{\mathcal{W}} \\ \mathcal{F}_{\mathrm{MPE}(\tilde{\lambda})>\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W})}}} P(\psi_\tau = \mathcal{S}_j|\mathcal{O},\tilde{\mathcal{W}},\tilde{\lambda})P(\tilde{\mathcal{W}}|\mathcal{O},\tilde{\lambda})\left[\mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) - \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W})\right] \qquad (37)$$

where

$$\begin{aligned}
\gamma_j^{\mathrm{num}}(\tau) &= \sum_{\tilde{\mathcal{W}}} P(\psi_\tau = \mathcal{S}_j|\mathcal{O},\tilde{\mathcal{W}},\tilde{\lambda})P(\tilde{\mathcal{W}}|\mathcal{O},\tilde{\lambda})\left[\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) - \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda})\right] \quad (\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) - \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) \geq 0) \\
\gamma_j^{\mathrm{den}}(\tau) &= -\sum_{\tilde{\mathcal{W}}} P(\psi_\tau = \mathcal{S}_j|\mathcal{O},\tilde{\mathcal{W}},\tilde{\lambda})P(\tilde{\mathcal{W}}|\mathcal{O},\tilde{\lambda})\left[\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) - \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda})\right] \quad (\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) - \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) < 0) \qquad (38)
\end{aligned}$$

[4] Here similar to MMI growth functions only the optimization of Gaussian means and variances are considered although a standard EBW based re-estimation formula may also be derived for transition probabilities and component priors.

Note that the above form of MPE occupancy is the same as in [17, 18]. The MPE statistics accumulated for a standard EBW update is given by

$$\sum_{\tau} \gamma_j^{\mathrm{MPE}}(\tau) \;=\; \sum_{\tau} \gamma_j^{\mathrm{num}}(\tau) - \sum_{\tau} \gamma_j^{\mathrm{den}}(\tau) + D_j \tag{39}$$

where $D_j = E \sum_\tau \gamma_j^{\mathrm{den}}(\tau)$. It can be easily shown that the smoothing constant in equation 37, derived using a growth function, is equivalent to the standard form of D_j when $C = E$. Given this forms of statisitcs an EBW re-estimation formula can be used for MPE training, as is proposed in [17, 18]. The following marginalization of the MPE growth function lower bound may be used for complexity control, via Laplace approximation.

$$\hat{\mathcal{M}} \;=\; \arg\max_{\mathcal{M}} \int \exp\left(\mathcal{L}_{\mathrm{MPE}}(\lambda, \tilde{\lambda})\right) p(\lambda|\mathcal{M}) \mathrm{d}\lambda \tag{40}$$

3.4.3 Discriminative Training Criteria and Growth Functions

So far it has been shown that for both MMI and MPE criteria, the EBW based discriminative updating formula may be viewed is a extension, to the standard Baum-Welch re-estimation formula used in conventional ML training. The state or Gaussian component occupancy is computed in a discriminative fashion instead. Discriminative training of HMM parameters may also be viewed as an EM based learning process. Sufficient statistics are accumulated in a discriminative fashion, in order to maximize an appropriate discriminative growth function, associated with the underlying training criterion. Note in both conventional ML and discriminative training, an increase in the auxiliary function guarantees an increase in the related ML, or discriminative growth function. However increasing a discriminative growth function, does not guarantee an increase in the underlying discriminative training criterion. The two are simply constrained to have the same gradient. Taking the MMI criterion as an example, the relationship between the criterion, the growth function and its lower bound is illustrated in figure 3. For complexity control, the marginalization of the ML criterion in Bayesian approaches, and discriminative growth functions, via their lower bounds, are related to each other. For both cases the underlying auxiliary functions being integrated over, differ from one another only in terms of the sufficient statistics required. Using a second order Laplace approximation a unified form of integral approximation may be derived. Detailed derivation of this will be given in the following section covering various implementation issues.

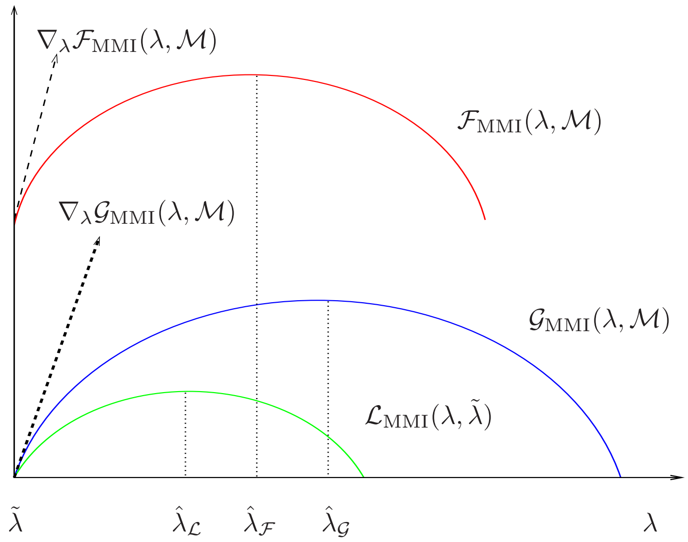

Figure 3: *Criterion, growth function and its lower bound*

4 Implementation Issues

In this section firstly a quantized method of evaluating complexity control schemes is presented. Other important implementation issues, when using marginalized discriminative growth functions for model complexity control, will be discussed. These issues can affect the performance of model structural optimization.

4.1 Evaluation of Complexity Control Criteria

When a complete set of systems for possible structural configurations are built and evaluated, for example if only one global complexity attribute is optimized, it is possible to compare the word error rate and performance ranking prediction for the complexity control criterion being used. A good complexity control criterion should give the correct rank ordering for all the systems to be compared. A measure of the distance between the two rankings is required to evaluate various complexity control schemes. In this paper, an empirical ranking prediction error metric is defined as

$$\text{RankErr}\% \quad = \quad \frac{\sum_{i,j} \delta(\mathbf{w}_i, \mathbf{w}_j) \times |\mathbf{w}_i - \mathbf{w}_j| \times |i-j|}{N \times \max_{i,j}\{|\mathbf{w}_i - \mathbf{w}_j|\} \times \max_{i,j}\{|i-j|\}} \tag{41}$$

Here $\{\mathbf{w}_1, ..., \mathbf{w}_N\}$ denotes a WER ranking prediction, of all N possible systems being considered, according to a particular complexity control criterion. 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. This means that such a difference is above a given *WER threshold*. This has a good intuitive feel, as penalizing systems that differ only slightly in error rate, is felt inappropriate. Differences between pairs of systems, which is less than a *WER threshold*, can be ignored. Hence the ranking error is related to the total amount of position shifts, weighted by WER difference, between all mis-ranked pairs of systems, if their WER difference exceeds a minimum threshold. The normalization term guarantees the ranking error will be positive and less than one.

4.2 Sharing Statistics among Model Structures

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 both conventional Bayesian complexity control and marginalized discriminative growth functions, this yields the form in equation 6, 25 and 35, requiring sufficient ML or discriminative statistics to be accumulated for a lower bound integration. 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, as a starting model structural configuration. 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, for various structural mutation from the current model structure, after this component merging operation. During such an operation the a pair of Gaussians being merged, are combined together to form a single component. Thus if to join components j and k to yield l, taking the MMI statistics as an example,

$$\gamma_l^{\text{MMI}}(\tau) \quad = \quad \gamma_j^{\text{MMI}}(\tau) + \gamma_k^{\text{MMI}}(\tau) \tag{42}$$

similarly for the first and second order statistics to be merged. All possible pairs of component merging are considered, and the one with the largest marginalized growth function value is 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 possible structural configurations, with varying number of retained dimensions.

4.3 Constrained Maximum Structural Mutation

When using discriminative growth functions for controlling complexity, the value of the growth function depends on the the confusion over the training data in recognition. However the sufficient statistics are accumulated given the a current model structure, and to be shared among multiple candidate structural configurations. If the configuration of the underlying model structure being considered, varies dramatically from the starting model structure, the shared sufficient statistics can be very different from the actual correct ones. In such a case the estimate of the growth function will be poor. Essentially an over-mutation of model structure from the current model, can loosen the growth function's lower bound in equation 25 and 35. This is because sufficient statistics are accumulated using a model, with a very different structural configuration, from the candidate model structure being considered. To overcome this problem, the overall structural optimization process can be set up in a multiple iteration mode. At each iteration a maximum change in the model complexity can be imposed. For the case of optimizing the number of components, the maximum number of Gaussian components that may be removed from any state is constrained to be 2 in these experiments. Between iterations of structural mutation model parameters are re-estimated using ML training. In practice this is found to yield a reasonable compromise between accuracy and speed.

4.4 Smoothing constant

For both the MMI and MPE growth functions there is a smoothing constant D_j. For extended Baum-Welch based standard discriminative training, this affects both the stability of the optimization and the rate of criterion convergence. For marginalized discriminative growth functions based complexity control, the setting of such smoothing constant will have a similar effect, on the speed of model structure change from the current model. A trade-off between structural mutation speed and structural optimization stability also has to be considered when setting D_j. This is for the same reason explained in the above section, such that the maximum structural change is constrained. In practice similar values to those used in standard LVCSR discriminative training, $D_j = E \sum_\tau \gamma_j^{\mathrm{den}}(\tau)$, where typically $E = 2.0$ was found to be a reasonable compromise.

4.5 Model Structure Optimization and Compression

Equation 25 and 35 require, 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, as is previously discussed, it is only possible to reduce the number of model parameters. Hence it is possible to modify this scheme for generating discriminatively determined compact systems, via a structural compression process. To allow a reduction in the complexity of the system, taking marginalized MMI growth functions for instance, it is required that

$$\int \mathcal{L}_{\mathrm{MMI}}(\lambda, \tilde{\lambda}) p(\lambda|\mathcal{M}) \mathrm{d}\lambda - \int \mathcal{L}_{\mathrm{MMI}}(\lambda, \tilde{\lambda}) p(\lambda|\hat{\mathcal{M}}) \mathrm{d}\lambda \quad \geq \quad -\alpha \int \mathcal{L}_{\mathrm{MMI}}(\lambda, \tilde{\lambda}) p(\lambda|\mathcal{M}_0) \mathrm{d}\lambda \tag{43}$$

where α is the compression factor, and $\mathcal{M}_0$ 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 structural mutation. By appropriately setting the values of the compression factor, a maximum affordable loss of marginalized discriminative growth functions can be used as a threshold during structural compression. This means that only structural compression operations, which cause a loss of the marginalized growth function value below a certain threshold, will be considered during the compression process.

4.6 Efficient Hessian Approximation

The marginalization of a lower bound of both standard ML criterion in Bayesian approaches in equation 6, and discriminative growth functions in equation 25 and 35, may be approximated via numerical methods such as Laplace approximation. However this requires the storage of a Hessian matrix with respect to model parameters. The number of model parameters in an LVCSR system can be in the millions. Computing and storing the Hessian matrix in a full structure for this is impractical. The solution used in this work, is to assume the Hessian matrix has a block diagonal structure [27]. The type of acoustic model structure being optimized in this paper, uses a mixture of Gaussians as a state density, and multiple HLDA based linear subspace projections. These linear projections are locally shared among clusters of Gaussian components, diagonalizing and compressing the original feature space [23]. It is therefore assumed that Gaussian components are independent of one other. Within each Gaussian component, the mean, variance and each row of the HLDA transform are further assumed independent against each other. Take the MMI growth function marginalization in equation 29 as an example, the block diagonal Hessian matrix may be partitioned into three sub-matrices with respect to

Gaussian means, variances $\{\check{\boldsymbol{\mu}}^{(j)}, \check{\boldsymbol{\Sigma}}^{(j)}\}$ in the projected subspaces and HLDA transforms $\{\mathbf{A}^{(r)}\}$ respectively,

$$\nabla^2_{\lambda} \mathcal{Q}_{\mathrm{MMI}}(\lambda, \tilde{\lambda}) \approx \begin{bmatrix} \mathbf{H}_{\mathbf{A}} & 0 & 0 \\ 0 & \mathbf{H}_{\boldsymbol{\mu}} & 0 \\ 0 & 0 & \mathbf{H}_{\boldsymbol{\Sigma}} \end{bmatrix} \tag{44}$$

where the sub-matrix associated with HLDA transform parameters is given as below. $\mathbf{a}_i^{(r)}$ denotes the ith row of HLDA transform $\mathbf{A}^{(r)}$.

$$\mathbf{H}_{\mathbf{A}} = \begin{bmatrix} \ddots & & 0 \\ & \frac{\partial^2 \mathcal{Q}_{\mathrm{MMI}}(\lambda,\tilde{\lambda})}{\partial^2 \mathbf{a}_i^{(r)}} & \\ 0 & & \ddots \end{bmatrix} \tag{45}$$

The other two sub-matrices with respect to transformed Gaussian means and variances are given by the following.

$$\mathbf{H}_{\boldsymbol{\mu}} = \begin{bmatrix} \ddots & & 0 \\ & \frac{\partial^2 \mathcal{Q}_{\mathrm{MMI}}(\lambda,\tilde{\lambda})}{\partial^2 \check{\boldsymbol{\mu}}^{(j)}} & \\ 0 & & \ddots \end{bmatrix} \qquad \mathbf{H}_{\boldsymbol{\Sigma}} = \begin{bmatrix} \ddots & & 0 \\ & \frac{\partial^2 \mathcal{Q}_{\mathrm{MMI}}(\lambda,\tilde{\lambda})}{\partial^2 \check{\boldsymbol{\Sigma}}^{(j)}} & \\ 0 & & \ddots \end{bmatrix} \tag{46}$$

For multiple HLDA systems the MMI auxiliary function given in equation 26 may be re-written as below. By definition the variances in the transformed space $\check{\boldsymbol{\Sigma}}^{(j)}$, is constrained to be diagonal. In the *nuisance* subspace Gaussian means and variances are tied across the whole HLDA transformation class.

$$\mathcal{Q}_{\mathrm{MMI}}(\lambda, \tilde{\lambda}) = \frac{1}{2}\sum_{j,\tau} \gamma_j^{\mathrm{MMI}}(\tau) \left\{ \log \left|\mathbf{A}^{(r_j)}\right|^2 - \log\left|\check{\boldsymbol{\Sigma}}^{(j)}\right| - (\mathbf{A}^{(r_j)}\mathbf{o}_\tau - \check{\boldsymbol{\mu}}^{(j)})^\top \check{\boldsymbol{\Sigma}}^{(j)-1}(\mathbf{A}^{(r_j)}\mathbf{o}_\tau - \check{\boldsymbol{\mu}}^{(j)}) \right\} \tag{47}$$

Here r_j denotes the transformation class that Gaussian component j is assigned to. From the above equation one can derive the second order gradient with respect to transform parameters,

$$\frac{\partial^2 \mathcal{Q}_{\mathrm{MMI}}(\lambda, \tilde{\lambda})}{\partial^2 \mathbf{a}_i^{(r)}} = -\frac{\sum_{j\in r,\tau} \gamma_j^{\mathrm{MMI}}(\tau)}{\left|\mathbf{A}^{(r)}\right|^2} \mathbf{c}_i^{(r)} \mathbf{c}_i^{(r)\top} - \mathbf{G}^{(r,i)} \tag{48}$$

where transformation class specific statistics is accumulated for each row of an HLDA transform,

$$\mathbf{G}^{(r,i)} = \begin{cases} \sum_{j\in r,\tau} \frac{\gamma_j^{\mathrm{MMI}}(\tau)}{\check{\boldsymbol{\Sigma}}^{(m)}_{[p]ii}} (\mathbf{o}_\tau - \boldsymbol{\mu}^{(j)})(\mathbf{o}_\tau - \boldsymbol{\mu}^{(j)})^\top & i \text{ is retained} \\ \sum_{j\in r,\tau} \frac{\gamma_j^{\mathrm{MMI}}(\tau)}{\check{\boldsymbol{\Sigma}}^{(r)}_{[n-p]ii}} (\mathbf{o}_\tau - \boldsymbol{\mu}^{(r)})(\mathbf{o}_\tau - \boldsymbol{\mu}^{(r)})^\top & i \text{ is nuisance} \end{cases} \tag{49}$$

and $\mathbf{c}_i^{(r)}$ is the cofactor vector of $\mathbf{a}_i^{(r)}$. $\{\boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma}^{(j)}\}$ are the Gaussian component specific mean and full covariances in the original feature space. $\{\boldsymbol{\mu}^{(r)}, \boldsymbol{\Sigma}^{(r)}\}$ the "global" mean and full covariance within some HLDA transformation class r. The second order gradient of the MMI auxiliary function with respect to the transformed Gaussian means and variances are computed as

$$\frac{\partial^2 \mathcal{Q}_{\mathrm{MMI}}(\lambda, \tilde{\lambda})}{\partial^2 \check{\boldsymbol{\mu}}^{(j)}} = -\frac{1}{2}\sum_{\tau} \gamma_j^{\mathrm{MMI}}(\tau) \check{\boldsymbol{\Sigma}}^{(j)-1} \tag{50}$$

$$\frac{\partial^2 \mathcal{Q}_{\mathrm{MMI}}(\lambda, \tilde{\lambda})}{\partial^2 \check{\boldsymbol{\Sigma}}^{(j)}_{[p]}} = -\frac{1}{2}\sum_{\tau} \gamma_j^{\mathrm{MMI}}(\tau) \left[2\mathbf{A}^{(r_j)}_{[p]} (\mathbf{o}_\tau - \boldsymbol{\mu}^{(j)})(\mathbf{o}_\tau - \boldsymbol{\mu}^{(j)})^\top \mathbf{A}^{(r_j)\top}_{[p]} \check{\boldsymbol{\Sigma}}^{(j)-3}_{[p]} - \check{\boldsymbol{\Sigma}}^{(j)-2}_{[p]} \right] \tag{51}$$

$$\frac{\partial^2 \mathcal{Q}_{\mathrm{MMI}}(\lambda, \tilde{\lambda})}{\partial^2 \check{\boldsymbol{\Sigma}}^{(r)}_{[n-p]}} = -\frac{1}{2}\sum_{j\in r,\tau} \gamma_j^{\mathrm{MMI}}(\tau) \left[2\mathbf{A}^{(r_j)}_{[n-p]} (\mathbf{o}_\tau - \boldsymbol{\mu}^{(r)})(\mathbf{o}_\tau - \boldsymbol{\mu}^{(r)})^\top \mathbf{A}^{(r_j)\top}_{[n-p]} \check{\boldsymbol{\Sigma}}^{(r)-3}_{[n-p]} - \check{\boldsymbol{\Sigma}}^{(r)-2}_{[n-p]} \right] \tag{52}$$

where the complete n dimensional feature space is split into a p dimensional retained and $n-p$ nuisance subspace. $\check{\boldsymbol{\Sigma}}^{(j)}_{[p]}$ and $\check{\boldsymbol{\Sigma}}^{(r)}_{[n-p]}$ are the projected variances in the retained and nuisance subspaces respectively.

Thus the Hessian matrix can be finally approximated as below.

$$\log\left|-\nabla^2_{\lambda}\mathcal{Q}_{\rm MMI}(\lambda,\tilde{\lambda})\right| \approx \sum_{r,i}\log\left|-\frac{\partial^2\mathcal{Q}_{\rm MMI}(\lambda,\tilde{\lambda})}{\partial^2\mathbf{a}_i^{(r)}}\right| + \sum_{j}\log\left|-\frac{\partial^2\mathcal{Q}_{\rm MMI}(\lambda,\tilde{\lambda})}{\partial^2\check{\boldsymbol{\mu}}^{(j)}}\right| + \sum_{j}\log\left|-\frac{\partial^2\mathcal{Q}_{\rm MMI}(\lambda,\tilde{\lambda})}{\partial^2\check{\boldsymbol{\Sigma}}^{(j)}_{[p]}}\right| + \sum_{r}\log\left|-\frac{\partial^2\mathcal{Q}_{\rm MMI}(\lambda,\tilde{\lambda})}{\partial^2\check{\boldsymbol{\Sigma}}^{(r)}_{[n-p]}}\right| \quad (53)$$

The approximated Hessian determinant in equation 53 can also be simplified for standard Gaussian density distributions, if ignoring the sub-matrix with respect to the HLDA transform and nuisance subspace parameters, or equivalently the first and fourth part of the right hand side of equation 53. The HLDA transforms in equation 51 also have to be treated as square and identity matrices.

5 Results

In this section experimental results of marginalized discriminative growth functions based complexity control schemes are presented, on a large vocabulary spontaneous speech recognition task. A conversational telephone speech (CTS) transcription task was used. Initially various complexity control schemes were used, to optimize multiple model complexity attributes on a global level. This allowed all systems to be trained and evaluated. Then the measuring of the performance ranking prediction error, for each complexity control scheme, was presented. In the later part, multiple model complexity attributes were optimized on a local level. All model structures considered were trained using the ML criterion after the complexity has been determined.

5.1 Training Setup and Experimental Conditions

A series of experiments on a CTS task were conducted, to investigate the performance of a variety of model complexity control schemes. Three 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 Linguistic Data Consortium (LDC) Cellular conversation sides. Another two subsets, a 68 hour `h5etrain00sub` and 76 hour `h5etrain03sub`, were used for initial development. For evaluation a 3 hour subset of 2001 development data, `dev01sub`, was used. Cepstral features were extracted and further normalized on conversation side basis, using VTLN plus mean and variance normalization. A 52 dimensional acoustic feature was then generated by appending derivatives up to the third order. This was projected down using one or more HLDA projections to a 39 dimensional feature vector. For multiple HLDA systems all the silence Gaussians were assigned to one transform class, while all speech Gaussians were split into 64 distinct classes. The component assignment used a top-down splitting procedure, based on distance measuring of Gaussian components in the acoustic space. Continuous density, mixture of Gaussians, cross-word triphone, gender independent HMM systems were used. There are approximately 6k speech states after a phonetic decision tree based state tying. All recognition experiments used a 58k word trigram language model. The baseline system trained using the 296 hour full set has 16 Gaussian components per state. On the other two subsets the baseline system has 12 Gaussians per state. The lattices used for complexity control and further discriminative training of model parameters, on both the full set and subsets, were generated via full decoding. During decoding a baseline 39 dimensional global HLDA systems trained using ML criterion was used, with 16 and 12 Gaussians per state for the full set and subsets repetitively. A pruned bi-gram language model was also used in decoding. These word lattices were further marked with model alignment which is kept fixed for various systems training. This is the "exact match" approach described in [16].

5.2 Optimizing Global Complexity Attributes

The global complexity control experiments used a 68 hour subset `h5etrain00sub`, of all the available Switchboard I and Call Home English (CHE) conversation sides of the 296 hour full training set. Two global complexity attributes of a single transform HLDA system were optimized: the number of Gaussian components per state from the set $\{12, 16, 24\}$; the number of useful dimensions in the range $\{28, ..., 52\}$. The permutation of these two attributes led to 75 candidate model structures.

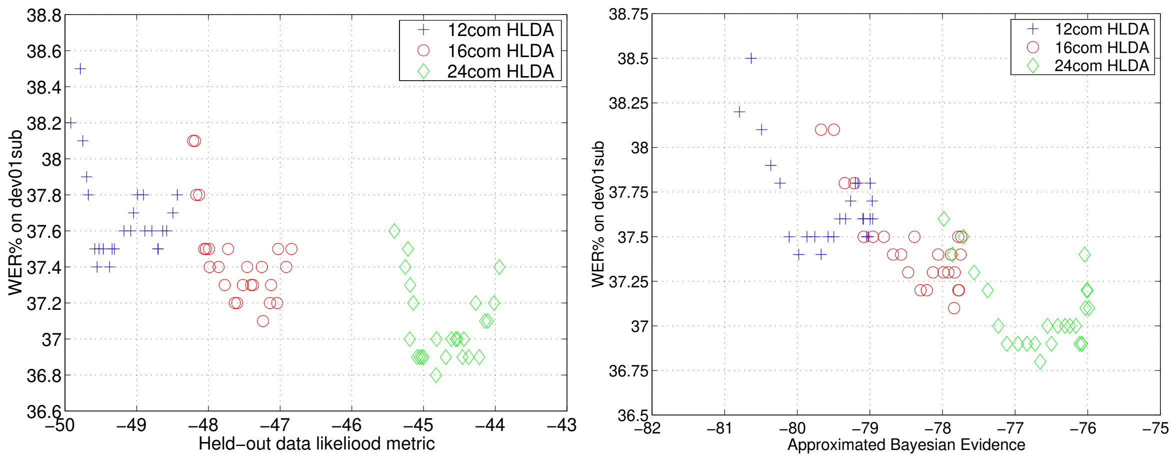

Figure 4: *Held out data likelihood and ML bound approximated Bayesian evidence vs. WER*

In previous work on this set-up the correlation between WER and held out data likelihood for the 75 systems was examined [27]. Though there was a very general trend, that error rate decreased with increased held-out data likelihood, the precise ordering of 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. This is significantly worse than the actual best system by 0.6% absolute as is shown in figure 4. Previous work also examined Laplace approximation and showed that the Bayesian evidence (which should be closely related to the held-out data log-likelihood), may be reasonably approximated as can be shown in figure 4. It has also be found that BIC is a poor criterion, when optimizing multiple complexity attributes simultaneously [27].

Figure 5 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. This will considerably restrict the power of the growth function, compared with using a state or component specific smoothing constant D_j. Figure 5 shows the marginalized MMI growth function

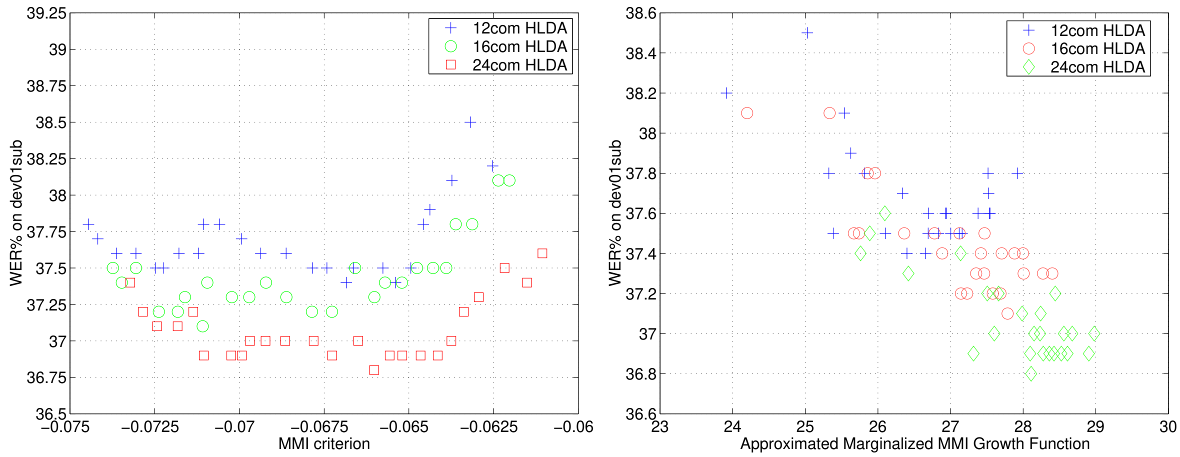

Figure 5: *Held out data MMI criterion and Marginalized MMI growth function vs. WER*

via its lower bound in equation 29 against WER, using Laplace approximation described in section 4. 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 4. The table consists of three sections.

complexity control	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 MMI GFunc	9.03	8.99	8.14
BIC ($\rho = 1$)	48.43	48.36	47.35
BIC ($\rho = 2$)	55.68	55.68	55.42
ML Bound	7.40	7.35	6.25
MMI GFunc	4.74	4.64	3.10

Table 1: *Performance ranking prediction error (%) on* `h5etrain00sub`

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 approximated Bayesian evidence. As described in [27] 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 8, 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 MMI growth function and is the score related to figure 5. It is interesting that the marginalized growth function slightly outperformed held-out growth function score. As expected if the WER threshold is increased then the ranking error decreases, though the general ranking of all complexity control schemes remains about the same.

5.3 Optimizing Local Complexity Attributes

In practice locally varying of structural configurations are preferred over a global level optimization, as more flexibility in the model structure can be introduced. The local complexity control experiments used a 76 hour subset `h5etrain03sub`, of all the available Switchboard I, Call Home English (CHE) and Cellular conversation sides of the 296 hour full training set, initially for locally optimizing a single complexity attribute. Finally on the 296 hour full set multiple complexity attributes will be optimized. Similar to the global complexity control experiments, the number of useful dimensions for HLDA systems to retained is in the range $\{28, ..., 52\}$.

5.3.1 Optimizing Number of Gaussians per State

	Complexity Control												
	Fixed				VarMix			MMI GFunc			MPE GFunc		
#Gauss	12	16	20	24	15.5	19.4	23.2	13.9	17.8	21.0	14.1	15.5	20.9
WER%	36.1	35.5	35.5	35.3	35.6	35.3	35.3	35.4	35.2	35.1	35.4	35.3	35.0

Table 2: *Optimizing #Gaussians on 76 hour* `h5etrain03sub`

Table 2 shows the word error rate for various complexity controlled global HLDA systems, used to determine the number of Gaussian components to train on each state. After the number of components was determined, 4 iterations of ML training were used to refine the parameter estimates. The *VarMix* technique is a simple occupancy based scheme. The number of components in a state is set to be 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, in contrast to the maximum allowable structural change imposed on all the marginalized MMI growth function systems. In the table VarMix systems typically yield a marginal gain. 0.1% to 0.2% absolute reduction in error rate was obtained from the equivalent baseline systems, which have a fixed number of Gaussian components per state. Comparing the baseline and VarMix systems the marginalized

MMI growth function gave slightly better performance though not being statistically significant. A 0.3% absolute WER reduction was obtained, using the 17.79 Gaussian per state MMI growth function system, over the equivalent 20 Gaussian baseline system. A similar trend of small performance gains is observed on all three MPE growth function systems. The 15.5 Gaussian per state MPE growth function system outperforms the equivalent VarMix system by 0.3% absolute. Comparing the 20 component baseline and the slightly bigger 20.9 component MPE growth function system, there is an absolute 0.5% WER reduction. It should be noted however that a maximum allowable change of 2 Gaussian components per state is used for all MMI growth function systems, while no such constraint is imposed on all VarMix systems. Another important feature of marginalized discriminative growth functions is that the scheme tend to generate a reasonably compact model structure. More than 10% smaller than all the equivalent baseline systems, marginal WER gains were still observed for all growth functions systems. It is also interesting to find, that the model structural configuration determined using marginalized discriminative growth functions, is very different from a standard complexity control scheme like VarMix. This is shown in figure 6, by taking marginalized MMI growth functions as an example.

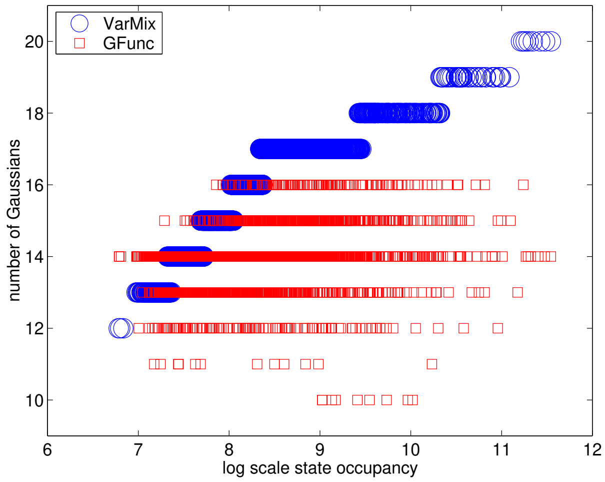

Figure 6: *#Gauss vs. log scale state occupancy on* `h5etrain03sub`

5.3.2 Optimizing Number of Retained Dimensions

Complexity Control	#Trans	#Dim	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	-	-
MMI GFunc	65	48.7	35.2	32.4	30.5
MPE GFunc	65	35.2	35.4	32.3	30.5

Table 3: *Optimizing number of retained dimensions on* `h5etrain03sub`

Experiments were also performed on the 76 hour `h5etrain03sub` for a 12 component HLDA system with 65 HLDA transforms. The aim is to optimize the transform class specific retained dimensionality. Table 3 shows the performance of various configurations. Appending the third derivatives and projecting back to 39 dimensions decreased the error rate by 1.4% absolute for ML training. Slight degradation of performance was observed when retaining all 52 dimensions for a single transform system. The number of transforms was then increased to 65. The use of all 52 dimensions gave the same performance as the HLDA projections to 39 dimensions using multiple transforms. For the projection to 39 dimension configuration MPE systems were built. The use of multiple transforms shows significant gains over a single transform. Furthermore the use of marginalized MMI growth function further reduced the error rate. At ML training stage, the complexity controlled multiple transform system, outperforms the two multiple transform systems with fixed dimensionality, by an absolute 0.3% WER reduction. Compared to global transform systems, there is an absolute WER reduction of 0.9%∼1.1%. More importantly the gain from this structural optimization process is found additive, to further four iterations of MPE criterion based discriminative training [18], 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, after MPE training and speaker adaptation. In the table similar performance gains are available on a system using marginalized MPE growth functions. An absolute WER reduction of 0.4% is obtained at the MPE training stage, compared with the standard multiple HLDA system. An overall gain of 0.8% absolute WER reduction is obtained over the 39 dimensional global transform system before speaker adaptation. The gain was also largely maintained after adaptation. It is also interesting to find that MPE growth function determined a much lower number of retained dimensions, 35.2 per Gaussian. This is lower than both the standard 39 dimensional configuration and the MMI growth function system.

5.3.3 Optimizing Multiple Complexity Attributes

Table 4 shows the performance of various structural configurations on the `h5etrain03` training data. Marginalized MMI growth functions based model 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. After four iterations of 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 changing significantly. Compared to the standard CUED evaluation set up (single transform, 39 dim projection), the complexity control yielded an absolute 0.5% gain. This performance gain was further maintained largely, after MLLR based speaker adaptation.

Complexity Control	#Gauss	#Trans	#Dim	WER%		
				MLE	MPE	MLLR
Std	28	-	39	34.7	-	-
Fixed	28	1	39	33.4	30.1	28.5
			52	33.2	-	-
Fixed	28	65	39	33.3	29.8	28.4
			52	32.9	-	-
MMI GFunc	25.6	65	41.5	32.7	29.6	28.0

Table 4: *Optimizing multiple complexity attributes on 296 hour* `h5etrain03`

5.4 Fast Model Structural Compression

A 16 component global HLDA system was generated on the `h5etrain03sub` training data, and then iteratively split until the number of components was 24. This was then used as the baseline model for the model structural compression experiments. In these experiments, using marginalized MMI growth functions, the number of Gaussian components in the system to be aggressively reduced. The aim is to at least retain the same WER, after the model structural compression. Table 5 shows the performance of various systems, with respect to setting of the compression factor α, as is defined in section 4. 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 rate. As a contrast a system with $\alpha = \infty$, where 2 components were removed from

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

Table 5: *Compressing #Gaussians on* `h5etrain03sub`

each state at each iteration is also shown. An equivalent VarMix 16 component system gave an error rate 35.5%. The compression gains on these experiments were marginal. This may be partly due to the small performance gain in the range considered. Only 0.2% absolute was gained, by increasing the number of components per state from 16 to 24.

6 Conclusion

A novel automatic model complexity control technique was proposed, using marginalized discriminative growth functions. The discriminative growth functions investigated, are closely related to maximum mutual information (MMI) and minimum phone error (MPE) criterion. They have a reduced sensitivity to outliers utterances, with either very low sentence posteriors or accuracy measuring. Model structure optimization and compression on a typical LVCSR task was performed. The number of Gaussians per state, and the number of useful dimensions of an HLDA system were determined. Both attributes were optimized at global and local level. Initial experiments indicate that this form of discriminative model complexity control may be useful for speech recognition.

In future work optimizing the complexity of discriminatively trained model structures will also be investigated, other rather than conventional ML training considered in this paper. In addition other forms of approximation schemes, for the marginalization of growth functions, will be investigated.

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] B. S. Clarke & A. R. Barron (1990). Information-Theoretic Asymptotic of Bayes Methods, *IEEE Transactions on Information Theory*, pp. 453–471, Vol. 30, No. 3, May 1990.

[4] 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.

[5] J. J. Rissanen (1996). Fisher Information and Stochastic Complexity, *IEEE Transactions on Information Theory*, pp. 40–47, Vol. 42, No. 1, January 1996.

[6] J. J. Rissanen (2001). Strong Optimality of the Normalized ML Models as Universal Codes and Information in Data, *IEEE Transactions on Information Theory*, Vol. 47, No 5, July 2001.

[7] M. H. Hansen & B. Yu (2001). Model Selection and the Principle of Minimum Description Length, *JASA*, pp. 746-774 Vol. 96, No. 454, 2001.

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

[9] Radford M. Neal (1993) Probabilistic Inference using Markov Chain Monte Carlo Methods, *Technical Report, CGT-TR-93-1*, Department of Computer Science, University of Toronto, 1993.

[10] A. Nádas (1983). A Decision Theorectic Formulation of a Training Problem in Speech Recognition and a Comparision of Training by Unconditional Versus Conditional Maximum Likelihood. *IEEE Transactions on Speech and Audio Processing*, pp. 814–817, Vol. 31, 1983.

[11] 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.

[12] M. M. Hochberg, L. T. Niles, J. T. Foote & H. F. Silverman (1991). Hidden Markov Model/Neural Network Training Techniques for Connected Alpha-Digit Speech Recognition. *Proc. ICASSP'91*, pp. 109 – 112, Toronto.

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

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

[15] V. Valtchev, J. J. Odell, P. C. Woodland & S. J. Young (1997). MMIE Training of Large Vocabulary Speech Recognition Systems. *Speech Communication*, pp. 303-314, Vol.22, 1997.

[16] 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.

[17] D. Povey & P. C. Woodland (2002). Minimum Phone Error and I-smoothing for Improved Discriminative Training, *Proc. ICASSP'02*, Florida, USA.

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

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

[20] 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.

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

[22] V. Goel, S. Kumar, & W. Byrne (2004). Segmental minimum Bayes-risk decoding for automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, To appear.

[23] 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.

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

[25] 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.

[26] D. Y. Kim, G. Evermann, T. Hain, D. Mrva, S. E. Tranter, L. Wang & P. C. Woodland (2003). Recent Advances in Broadcast News Transcription. *Proc. ASRU'03* , St. Thomas, U.S. Virgin Islands.

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

[28] 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.

A Derivations of MMI Growth Functions

Various forms of this growth function will be used in this work. First at the "current" model parameter value following 23 one can write

$$\mathcal{G}(\psi,\tilde{\lambda}) \quad = \quad p(\mathcal{O},\mathcal{W}|\tilde{\lambda})\left[P(\psi|\mathcal{O},\mathcal{W},\tilde{\lambda}) - P(\psi|\mathcal{O},\tilde{\lambda}) + CP(\psi|\mathcal{O},\tilde{\lambda})\right] \tag{54}$$

Let the MMI hidden variable sequence occupancy $\gamma_{\psi}^{\mathrm{MMI}}(\mathcal{O})$ be defined as

$$\gamma_{\psi}^{\mathrm{MMI}}(\mathcal{O}) \quad = \quad P(\psi|\mathcal{O},\mathcal{W},\tilde{\lambda}) - P(\psi|\mathcal{O},\tilde{\lambda}) + CP(\psi|\mathcal{O},\tilde{\lambda}) \tag{55}$$

This gives a simplified form of posterior in equation 34.

$$\mathcal{P}(\psi,\tilde{\lambda}) \quad = \quad \frac{\gamma_{\psi}^{\mathrm{MMI}}(\mathcal{O})}{\sum_{\psi}\gamma_{\psi}^{\mathrm{MMI}}(\mathcal{O})} \tag{56}$$

It is also possible to write

$$\mathcal{G}(\psi,\lambda) \quad = \quad \left[P(\mathcal{W}|\psi) - \mathcal{F}_{\mathrm{MMI}}(\tilde{\lambda}) + CP(\mathcal{W}|\mathcal{O},\tilde{\lambda})\right]p(\mathcal{O},\psi|\lambda) \tag{57}$$

as the observations are conditionally independent of the word sequence given the hidden state sequence ψ.

$$\begin{aligned} p(\mathcal{O},\psi,\mathcal{W}|\lambda) &= p(\mathcal{O},\psi|\lambda)P(\mathcal{W}|\mathcal{O},\psi,\lambda) \\ &= p(\mathcal{O},\psi|\lambda)\frac{p(\mathcal{O},\mathcal{W}|\psi,\lambda)}{p(\mathcal{O}|\psi,\lambda)} \\ &= p(\mathcal{O},\psi|\lambda)P(\mathcal{W}|\psi) \end{aligned} \tag{58}$$

Using equation 56 and 57, equation 22 can be re-written as

$$\mathcal{L}_{\mathrm{MMI}}(\lambda,\tilde{\lambda}) \quad = \quad \log\mathcal{G}(\tilde{\lambda}) + \sum_{\psi}\mathcal{P}(\psi,\tilde{\lambda})\log p(\mathcal{O},\psi|\lambda) - \sum_{\psi}\mathcal{P}(\psi,\tilde{\lambda})\log p(\mathcal{O},\psi|\tilde{\lambda}) \tag{59}$$

This lower bound can be re-expressed as

$$\mathcal{L}_{\mathrm{MMI}}(\lambda,\tilde{\lambda}) \quad = \quad \log\mathcal{G}(\tilde{\lambda}) + \frac{\mathcal{Q}_{\mathrm{MMI}}(\lambda,\tilde{\lambda}) - \mathcal{Q}_{\mathrm{MMI}}(\tilde{\lambda},\tilde{\lambda})}{\sum_{\psi}\gamma_{\psi}^{\mathrm{MMI}}(\mathcal{O})} \tag{60}$$

where the MMI auxiliary function is defined as

$$\begin{aligned} \mathcal{Q}_{\mathrm{MMI}}(\lambda,\tilde{\lambda}) &= \sum_{\psi}\gamma_{\psi}^{\mathrm{MMI}}(\mathcal{O})\log p(\mathcal{O},\psi|\lambda) \\ &= \sum_{\psi}\gamma_{\psi}^{\mathrm{MMI}}(\mathcal{O})\log p(\mathcal{O}|\psi,\lambda) + \sum_{\psi}\gamma_{\psi}^{\mathrm{MMI}}(\mathcal{O})\log P(\psi|\lambda) \\ &= \sum_{j,\tau}\gamma_{j}^{\mathrm{MMI}}(\tau)\log p(\mathbf{o}_{\tau}|\psi_{\tau}=\mathcal{S}_j,\lambda) + \sum_{\psi}\gamma_{\psi}^{\mathrm{MMI}}(\mathcal{O})\log P(\psi|\lambda) \end{aligned} \tag{61}$$

and the MMI hidden variable occupancy $\gamma_j^{\mathrm{MMI}}(\tau)$ is,

$$\gamma_j^{\mathrm{MMI}}(\tau) \quad = \quad \gamma_j^{\mathrm{num}}(\tau) - \gamma_j^{\mathrm{den}}(\tau) + CP(\psi_{\tau}=\mathcal{S}_j|\mathcal{O},\tilde{\lambda}) \tag{62}$$

B Derivations of MPE Growth Functions

Following 33 one can write

$$\mathcal{G}(\psi,\tilde{\lambda}) \quad = \quad \sum_{\tilde{\mathcal{W}}}p(\mathcal{O},\psi,\tilde{\mathcal{W}}|\tilde{\lambda})\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) - \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda})p(\mathcal{O},\psi|\tilde{\lambda}) + C\sum_{\substack{\tilde{\mathcal{W}} \\ \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda})>\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W})}}p(\mathcal{O},\psi,\tilde{\mathcal{W}}|\tilde{\lambda})\left[\mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) - \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W})\right]$$

$$
\begin{aligned}
&= p(\mathcal{O}|\tilde{\lambda}) \left\{ \sum_{\tilde{\mathcal{W}}} P(\psi|\mathcal{O},\tilde{\mathcal{W}},\tilde{\lambda}) P(\mathcal{W}|\mathcal{O},\tilde{\lambda}) \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) - \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) P(\psi|\mathcal{O},\tilde{\lambda}) \right. \\
&\quad \left. + C \sum_{\substack{\tilde{\mathcal{W}} \\ \mathcal{F}_{\mathrm{MPE}(\tilde{\lambda}) > \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W})}}} P(\psi|\mathcal{O},\tilde{\mathcal{W}},\tilde{\lambda}) P(\mathcal{W}|\mathcal{O},\tilde{\lambda}) \left[\mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) - \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) \right] \right\} \qquad (63)
\end{aligned}
$$

Let us define $\gamma_{\psi}^{\mathrm{MPE}}(\mathcal{O})$ as the MPE hidden variable sequence occupancy,

$$
\begin{aligned}
\gamma_{\psi}^{\mathrm{MPE}}(\mathcal{O}) &= \sum_{\tilde{\mathcal{W}}} P(\psi|\mathcal{O},\tilde{\mathcal{W}},\tilde{\lambda}) P(\tilde{\mathcal{W}}|\mathcal{O},\tilde{\lambda}) \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) - \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) P(\psi|\mathcal{O},\tilde{\lambda}) \\
&\quad + C \sum_{\substack{\tilde{\mathcal{W}} \\ \mathcal{F}_{\mathrm{MPE}(\tilde{\lambda}) > \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W})}}} P(\psi|\mathcal{O},\tilde{\mathcal{W}},\tilde{\lambda}) P(\mathcal{W}|\mathcal{O},\tilde{\lambda}) \left[\mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) - \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) \right] \qquad (64)
\end{aligned}
$$

the hidden variable sequence posterior in equation 34 may be simplified as below.

$$
\mathcal{P}(\psi,\tilde{\lambda}) = \frac{\gamma_{\psi}^{\mathrm{MPE}}(\mathcal{O})}{\sum_{\psi} \gamma_{\psi}^{\mathrm{MPE}}(\mathcal{O})} \qquad (65)
$$

Note that equation 33 may also be re-written as

$$
\mathcal{G}(\psi,\lambda) = p(\mathcal{O},\psi|\lambda) \left\{ \sum_{\tilde{\mathcal{W}}} P(\tilde{\mathcal{W}}|\psi) \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) - \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) + C \sum_{\substack{\tilde{\mathcal{W}} \\ \mathcal{F}_{\mathrm{MPE}(\tilde{\lambda}) > \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W})}}} P(\tilde{\mathcal{W}}|\psi) \left[\mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) - \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) \right] \right\} \qquad (66)
$$

given that the likelihood of an observation sequence is conditionally dependent on the HMM state sequence, following equation 58. Combining 34, 65 and 66, equation 32 may now be re-written as

$$
\mathcal{L}_{\mathrm{MPE}}(\lambda,\tilde{\lambda}) = \log \mathcal{G}(\tilde{\lambda}) + \sum_{\psi} \mathcal{P}(\psi,\tilde{\lambda}) \log p(\mathcal{O},\psi|\lambda) - \sum_{\psi} \mathcal{P}(\psi,\tilde{\lambda}) \log p(\mathcal{O},\psi|\tilde{\lambda}) \qquad (67)
$$

If we define the MPE auxiliary function as

$$
\begin{aligned}
\mathcal{Q}_{\mathrm{MPE}}(\lambda,\tilde{\lambda}) &= \sum_{\psi} \gamma_{\psi}^{\mathrm{MPE}}(\mathcal{O}) \log p(\mathcal{O},\psi|\lambda) \\
&= \sum_{\psi} \gamma_{\psi}^{\mathrm{MPE}}(\mathcal{O}) \log p(\mathcal{O}|\psi,\lambda) + \sum_{\psi} \gamma_{\psi}^{\mathrm{MPE}}(\mathcal{O}) \log P(\psi|\lambda) \\
&= \sum_{j,\tau} \gamma_{j}^{\mathrm{MPE}}(\tau) \log p(\mathbf{o}_{\tau}|\psi_{\tau} = \mathcal{S}_j,\lambda) + \sum_{\psi} \gamma_{\psi}^{\mathrm{MPE}}(\mathcal{O}) \log P(\psi|\lambda) \qquad (68)
\end{aligned}
$$

and the MPE Gaussian component occupancy $\gamma_{j}^{\mathrm{MPE}}(\tau)$ may be written as below following equation 64,

$$
\gamma_{j}^{\mathrm{MPE}}(\tau) = \gamma_{j}^{\mathrm{num}}(\tau) - \gamma_{j}^{\mathrm{den}}(\tau) + C \sum_{\substack{\tilde{\mathcal{W}} \\ \mathcal{F}_{\mathrm{MPE}(\tilde{\lambda}) > \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W})}}} P(\psi_{\tau} = \mathcal{S}_j|\mathcal{O},\tilde{\mathcal{W}},\tilde{\lambda}) P(\tilde{\mathcal{W}}|\mathcal{O},\tilde{\lambda}) \left[\mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) - \mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) \right] \qquad (69)
$$

where

$$
\begin{aligned}
\gamma_{j}^{\mathrm{num}}(\tau) &= \sum_{\tilde{\mathcal{W}}} P(\psi_{\tau} = \mathcal{S}_j|\mathcal{O},\tilde{\mathcal{W}},\tilde{\lambda}) P(\tilde{\mathcal{W}}|\mathcal{O},\tilde{\lambda}) \left[\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) - \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) \right] \quad (\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) - \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) \geq 0) \\
\gamma_{j}^{\mathrm{den}}(\tau) &= -\sum_{\tilde{\mathcal{W}}} P(\psi_{\tau} = \mathcal{S}_j|\mathcal{O},\tilde{\mathcal{W}},\tilde{\lambda}) P(\tilde{\mathcal{W}}|\mathcal{O},\tilde{\lambda}) \left[\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) - \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) \right] \quad (\mathcal{A}(\tilde{\mathcal{W}},\mathcal{W}) - \mathcal{F}_{\mathrm{MPE}}(\tilde{\lambda}) < 0) \qquad (70)
\end{aligned}
$$