Computer Speech and Language (2002) **16**, 5–24
doi: 10.1006/csla.2001.0181
Available online at http://www.idealibrary.com on IDEAL®

Structural maximum *a posteriori* linear regression for fast HMM adaptation

Olivier Siohan, Tor André Myrvoll and Chin-Hui Lee

Multimedia Communications Research Lab, Bell Laboratories—Lucent Technologies, 600 Mountain Ave., Murray Hill, NJ 07974, U.S.A.

Abstract

Transformation-based model adaptation techniques have been used for many years to improve robustness of speech recognition systems. While the estimation criterion used to estimate transformation parameters has been mainly based on maximum likelihood estimation (MLE), Bayesian versions of some of the most popular transformation-based adaptation methods have been recently introduced, like MAPLR, a maximum *a posteriori* (MAP) based version of the well-known maximum likelihood linear regression (MLLR) algorithm. This is in fact an attempt to constraint parameter estimation in order to obtain reliable estimation with a limited amount of data, not only to prevent overfitting the adaptation data but also to allow integration of prior knowledge into transformation-based adaptation techniques. Since such techniques require the estimation of a large number of transformation parameters when the amount of adaptation data is large, it is also required to define a large number of prior densities for these parameters. Robust estimation of these prior densities is therefore a crucial issue that directly affects the efficiency and effectiveness of the Bayesian techniques. This paper proposes to estimate these priors using the notion of hierarchical priors, embedded into the tree structure used to control transformation complexity. The proposed algorithm, called structural MAPLR (SMAPLR), has been evaluated on the Spoke3 1993 test set of the WSJ task. It is shown that SMAPLR reduces the risk of overtraining and exploits the adaptation data much more efficiently than MLLR, leading to a significant reduction of the word error rate for any amount of adaptation data.

1. Introduction

It is well known that automatic speech recognition (ASR) systems are highly sensitive to variations between training and testing conditions such as changes in speaker characteristics, channel conditions or acoustic environment (Lee, 1998). As a result, many techniques have been developed to reduce or limit such discrepancies (Gong, 1995; Lee & Huo, 2000), either by finding robust and invariant speech features, by modifying features or recognition parameters using compensation or adaptation techniques (Lee, 1998), or by using robust decision strategies (Lee, 1999).

The present work focuses on model adaptation techniques where, given some small amount of adaptation data reflecting the testing environment, the objective is to modify the model

0885–2308/02/010005 + 20 $35.00/0

parameters to better match the test data. The reader is referred to a recent article (Lee & Huo, 2000) for an in-depth overview of adaptation strategies for speech recognition. As expected, such an objective faces conflicting requirements since a large amount of adaptation data is required to accurately re-estimate the large number of parameters that constitute acoustic models used in state-of-the-art large vocabulary speech recognition systems. This problem is typically addressed by making use of various forms of prior knowledge to constrain the parameters' estimation. In the past, such constraints have been implemented in different ways, typically by using some form of constrained estimation criterion or by sharing related parameters.

The use of constrained estimation has been typically implemented using a Bayesian approach like maximum *a posteriori* (MAP) estimation, which combines under a well defined mathematical formulation the information provided by the adaptation data with some prior knowledge about the parameters to be estimated. In the seminal work of Lee, Lin, and Juang (1991) and Gauvain and Lee (1994), hidden Markov model (HMM) parameters are estimated using a MAP criterion, leading to an accurate estimation of these parameters when a small amount of data is available. An attractive property of MAP adaptation is its asymptotic convergence to MLE when the amount of adaptation data increases. However, the convergence can be quite slow since only a small fraction of the parameters are adapted when the amount of adaptation data is limited.

On the other hand, indirect transformation-based model adaptation relies on parameter sharing: clusters of model parameters are transformed through a shared function $F_\eta(\cdot)$ whose parameters have to be estimated, typically using a MLE criterion. This corresponds to a global adaptation scenario since all model parameters Λ_c belonging to a common cluster c are simultaneously transformed to $\tilde{\Lambda}_c = F_{\eta_c}(\Lambda_c)$. Due to this sharing, all acoustic units can be adapted which makes transformation-based adaptation techniques especially attractive in situations where only a limited amount of adaptation data is available. When the transformation $F_\eta(\cdot)$ is an affine transformation of HMM mean vectors estimated using MLE, this approach becomes the well-known MLLR (Leggetter & Woodland, 1995*b*). Several related variants have also been proposed, such as the multiple stochastic transformations approach (Diakoloukas & Digalakis, 1999), where several linear transformations are applied to the same Gaussian density. It should be noted however that transformation-based adaptation techniques may suffer from poor asymptotic properties leading to a quick saturation in performance as the amount of adaptation data increases.

It is worth mentioning that most transformation-based approaches have been initially formulated using a ML estimation criterion. However, it does not have to be so, and it is of course possible to replace ML by any other criterion, as for example MAP, which incorporates prior knowledge about the transformation parameters in the estimation process. Such an approach is used in Chien, Wang, and Lee (1996, 1997*c*) where several channel biases are estimated based on MAP adaptation, effectively making use of prior channel statistics. A similar principle can be used for MLLR adaptation, which can be reformulated under a Bayesian framework by introducing a prior distribution for the affine transformation matrices, leading to the MAPLR algorithm (Siohan, Chesta, & Lee, 1999; Chesta, Siohan, & Lee, 1999). In the context of MAPLR, the use of prior distribution for the transformation matrices can significantly improve performance over MLLR in two ways. First, the regularization effect of the priors may prevent overtraining the transformation matrices which typically occurs when too many transformations are estimated from a small amount of data (Siohan *et al.*, 1999), leading to a poor generalization on unseen test data. Second, a careful estimation of the transformation priors used in MAPLR can lead to improved accuracy over MLLR, as reported