STRUCTURAL MAXIMUM A POSTERIORI LINEAR REGRESSION FOR FAST HMM ADAPTATION

Olivier Siohan *Tor André Myrvoll* * *Chin-Hui Lee*

Multimedia Communications Research Lab
Bell Laboratories – Lucent Technologies
600 Mountain Ave., Murray Hill, NJ 07974, USA
{siohan,chl}@research.bell-labs.com myrvoll@tele.ntnu.no

ABSTRACT

Transformation-based model adaptation techniques like maximum likelihood linear regression (MLLR) rely on an accurate selection of the number of transformations for a given amount of adaptation data. If too many transformations are used, the transformation parameters may be poorly estimated, can overfit the adaptation data, and offer poor generalization. On the other hand, if the number of transformations is too small, the adapted models can only provide a moderate improvement over the baseline models. An adaptation approach should therefore be flexible in order to estimate reliably a large number of transformations when the amount of adaptation data is large, and a small number of transformations when only a few adaptation utterances are available. In this work, we show that a significant improvement can be obtained over MLLR with dynamic regression classes, first by replacing the maximum likelihood estimation criterion by a maximum a posteriori criterion, then by introducing a tree-structure for the prior densities of the transformations. The effectiveness of the proposed approach is illustrated on the Spoke3 1993 test set of the WSJ task. Using the same regression classes as MLLR, it is shown that the proposed approach reduces the risk of overfitting and exploit the adaptation data much more efficiently than MLLR, leading to a significant reduction of the word error rate with as little as one adaptation utterance.

1. INTRODUCTION

Model adaptation techniques are an efficient way to reduce speaker-related fluctuations as well as acoustical environment discrepancies that typically occur between training and testing of speech recognition systems [1]. Under this very general family of techniques, two broad classes of approaches, namely *indirect* and *direct* adaptation, can be identified [2]. Indirect model adaptation approaches are traditionally transformation-based techniques where clusters of model parameters are transformed through a shared function $F_\eta(\cdot)$ whose parameters η are estimated from a set of adaptation data. Such approaches can be considered as 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 hidden Markov model (HMM) mean vectors estimated using maximum likelihood, this approach becomes the well-known maximum likelihood linear regression (MLLR) [3], which has been successfully applied to speaker and environment adaptation. It should be noted however that transformation-based adaptation techniques generally suffer from poor asymptotic properties leading to a quick saturation in performance as the amount of adaptation data increases.

Direct model adaptation techniques do not assume any underlying functional transformation but attempt to directly reestimate the model parameters. Acoustic units for which adaptation data is available are reestimated, leading to a *local* adaptation since unseen units are not modified. The most representative approach in direct adaptation is Bayesian learning, often implemented via maximum *a posteriori* (MAP) estimation [4, 5]. MAP adaptation combines under a well defined mathematical formulation the information provided by the adaptation data with some prior knowledge about the model parameters described by a prior distribution. A fundamental property of MAP adaptation is its nice asymptotic convergence to maximum likelihood estimation when the amount of adaptation data increases. This convergence is however fairly slow and a large amount of adaptation data is needed to observe a significant performance improvement.

Recent works have been focused on taking advantage of both direct and indirect adaptation approaches. A straightforward technique consists of applying an indirect adaptation method followed by a direct adaptation, as in [6] where a MLLR-like transformation is followed by MAP adaptation. Such a technique can be further refined by jointly reestimating the model and transformation parameters using a common MAP estimation criterion, as done in [7], and has been shown to outperform MAP and MLLR for small and large amount of adaptation data. Other approaches intend to provide additional information to constraint the adapta-

*This work was done while T. A. Myrvoll was on leave from the Department of Telecommunications, Norwegian University of Science and Technology, Norway.

tion procedure. In transformation-based model adaptation, this additional information can take the form of a prior distribution of the transformation parameters, which is then used to constraint the estimation via the use of a MAP criterion [8, 9, 10, 11]. For example, the traditional MLLR algorithm can be reformulated under a Bayesian framework by introducing a prior distribution for the affine transformation matrices leading to the MAPLR algorithm [10, 11]. Additional information can also be provided by a proper structuring of model parameters, transformation parameters and prior densities as in the structural MAP formulation (SMAP) [12, 13]. In SMAP, the model parameters are organized in a tree containing all the Gaussian distributions. Each node in that tree represents a cluster of Gaussian distributions sharing a common transformation representing the mismatch between training and testing conditions. The transformation parameters are represented by their probability density functions using hierarchical priors. Because of this highly constrained structure, efficient adaptation can be obtained when only a limited amount of adaptation data is available, while still preserving good asymptotic properties as the size of adaptation data increases.

This paper is a natural extension of the SMAP approach to a more complex class of transformation, namely an affine transformation of the mean vectors, similar to the transformation used in MLLR. Prior densities for the transformation matrices are hierarchically structured in a tree, and the transformation matrices are derived using a maximum *a posteriori* criterion. For that reason, the proposed approach is called Structural Maximum A Posteriori Linear Regression, or SMAPLR. Because of the use of structured priors, SMAPLR is expected to lead to a fast adaptation, reduce risk of overfitting, and improve convergence compared to MLLR. SMAPLR is also expected to outperform SMAP due to the more complex transformation family.

The proposed approach has been evaluated on the Spoke3 part (non-native speakers) of the Wall Street Journal task, and compared to several standard adaptation techniques like MAP, MLLR and SMAP. The paper is organized as follows. Section 2 describes the SMAPLR algorithm, experimental results are given in section 3, and section 4 summarizes our findings.

2. STRUCTURAL *MAXIMUM A POSTERIORI* LINEAR REGRESSION

General principle

This section gives an informal description of the structural MAPLR algorithm. The tree-based MLLR algorithm [19] is first reviewed, outlining potential drawbacks and the need for additional information to constrain the estimation. We argue that this additional information can be derived from the tree structure, and incorporated into the estimation process using a Bayesian approach, leading to the SMAPLR algorithm.

Tree-based MLLR algorithm A crucial problem in transformation-based adaptation is to define the clusters of HMM parameters that will share a common transformation. We will assume that the transformation-based adaptation technique is MLLR, but the following discussion is general enough to be valid for any transformation-based adaptation technique.

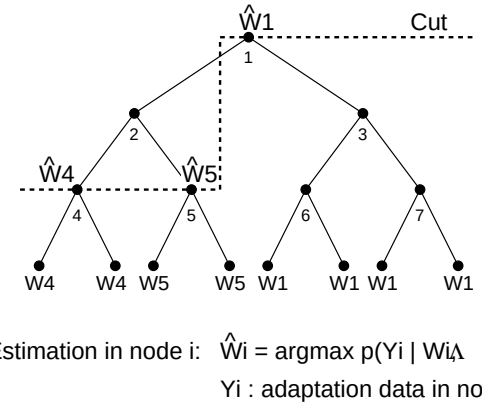

Figure 1: Tree-based MLLR algorithm. The cut is defined based on the amount of adaptation data available in each leaf. A transformation matrix is derived for each node along the cut. Each leaf node inherits the most specific transformation matrix derived at each cut-node, and the corresponding Gaussian density is adapted.

In early MLLR works, the clusters, also called regression classes, are defined based on broad phonetic classes [15], following the motivation is that similar classes of sounds should undergo the same transformation. These clusters are however defined statically, which significantly limit the effectiveness of the adaptation algorithm and lead to an early saturation of the adapted system accuracy as the amount of adaptation data increases.

This problem can be addressed by dynamically controlling the number of transformation clusters based on the available amount of adaptation data. Since it is reasonable to assume that acoustic units which are acoustically close to each other should share the same transformation, these acoustic units can be structured in a tree, and defining transformation classes consists of selecting a set of nodes on which to associate a transformation. This is the approach followed in [19], and illustrated in Fig 1. In this typical implementation of MLLR, a tree structure is first determined, whose leaves are the Gaussian probability density functions (pdf) of the continuous density HMMs. As the adaptation data is collected and aligned against its transcription[1], the sufficient statistics for each Gaussian pdf are accumulated. A "cut" (set of nodes) is then defined such that each subtree along the cut contains a number of frames larger than a predetermined threshold τ. A transformation is then associated to each node along the cut, and can be derived using for example maximum likelihood estimation as in MLLR. We should point out that the cut is obtained using a bottom-up approach so that the transformation associated to each cut node is the most specific transformation to apply on the Gaussian densities lying at the leaves of the corresponding

[1] assuming for simplicity a supervised adaptation scenario.

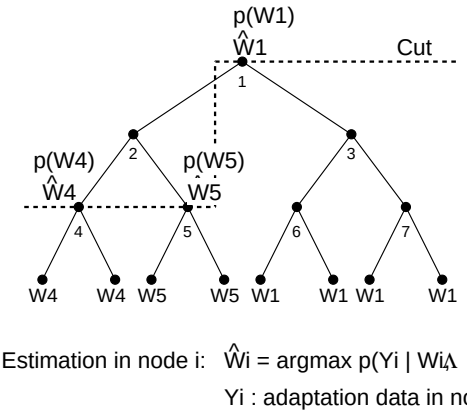

Figure 2: Tree-based MAPLR algorithm. As is MLLR, the cut is defined based on the amount of adaptation data available in each leaf. Assuming that a prior density $p(\mathbf{W}_i)$ is available in each node i along the cut, the transformation matrices $\mathbf{W}_i$ are estimated using a MAP criterion: $\hat{\mathbf{W}}_i = \text{argmax}\, p(\mathbf{Y}_i|\mathbf{W}_i, \mathbf{\Lambda})p(\mathbf{W}_i)$, where $\mathbf{Y}_i$ denotes the adaptation data available in node i.

subtree. If the amount of adaptation data is very small, the cut will lie near the root node since only nodes located at the higher level of the tree will gather enough adaptation data to define a transformation class. As the amount of adaptation data increases, the cut will slowly move toward the leaf nodes, leading to more and more transformation classes. The complexity of the transformations is therefore dynamically controlled based on the size of adaptation data and the threshold τ.

In order to get the best adaptation performance, a careful selection of the threshold τ is required. If τ is too small, too many transformation clusters might be defined, leading to model transformations that overfit the adaptation data. If τ is too large, the number of transformations might be to small, leading to limited improvement in recognition accuracy compared to the baseline system. Moreover, the recognition accuracy is quite sensitive to small changes in the location of the cut which can generate radically different transformation clusters.

Adding transformation priors To handle some of the problems mentioned in the previous subsection, the estimation of the transformation in each cut-node can be constrained by using a MAP estimation criterion rather than the usual maximum likelihood. This is the basic idea behind the MAPLR algorithm which simply replaces the ML criterion by MAP in the MLLR formulation [10, 11]. For each cut-node, it is assumed that a prior distribution[2] of the transformation matrix is available, which helps getting a more reliable estimate of the transformation parameters. This is illustrated in Fig. 2, where the transformation matrices along each cut node i are now estimated using MAP, given the prior density $p(\mathbf{W}_i)$ constraining the estimation in each node.

This prior information provides an additional way to control the adaptation. Supposing that the prior has a very small variance, it means that whatever adaptation data is available, the estimate will be conditioned mainly by the prior information. A "good" prior information is therefore able to control, to some extend, the overfitting of the adaptation algorithm.

An important issue in the MAPLR formulation is to determine an estimate of the prior distribution $p(\mathbf{W}_i)$ in each cut-node. The solution adopted in [10, 11] is to derive the prior distribution directly from the speaker independent models. Despite providing improved results over MLLR, especially on an incremental unsupervised adaptation task, as well as reducing the sensitivity to the number of transformations, this is a very crude and ad-hoc approach, unable to provide very accurate prior information. This issue is addressed in the next section.

Adding structure to the transformation priors In the previous sections, the tree is only used to define the transformation clusters. All transformations $\mathbf{W}_i$ along the cut are estimated independently of each other, using the data available in the corresponding subtree. For example, in Fig. 2, the data lying in the subtree of node 4 used to estimate $\mathbf{W}_4$ does not affect in any way the estimation of $\mathbf{W}_5$.

Suppose now that a transformation is estimated in node 2, the parent of nodes 4 and 5. It is quite likely that the transformation derived in node 2 can provide some useful information to constraint the estimation its child nodes, thereby introducing a dependency in the estimation of $\mathbf{W}_4$ and $\mathbf{W}_5$ via their parent node.

An attractive way to introduce this constraint from the parent node is to use Bayesian statistics. Given a prior information $p(\mathbf{W}_2)$ in node 2 and some adaptation data $\mathbf{Y}_2$, the posterior distribution of $\mathbf{W}_2$ can be derived: $p(\mathbf{W}_2|\mathbf{Y}_2)$. Then, it is possible to define the prior distribution in the child node 4, $p(\mathbf{W}_4)$, as $p(\mathbf{W}_4) = p(\mathbf{W}_2|\mathbf{Y}_2)$ and carry out the MAP estimation of $\mathbf{W}_4$.

This process can be started at the root node and is illustrated in Fig. 3. Starting from an initial prior density $p(\mathbf{W}_1)$ centered at an identity transformation, $p(\mathbf{W}_1|\mathbf{Y}_1)$ is estimated and used as prior density in the immediate child nodes. This prior/posterior information is propagated using the same principle down to each node along the cut, where a MAPLR estimation of the transformations is then performed.

Since each node along the cut inherits its prior density $p(\mathbf{W})$ from the posterior density of its parent, $p(\mathbf{W}|\mathbf{Y})$, the estimation in a given cut-node is constrained via the posterior densities $p(\mathbf{W}_i|\mathbf{Y}_i)$ derived in each node on the path from the root node to the current node. Such a structural constraint provides an attractive way to control the complexity of the adaptation being much less sensitive to the location of the cut and therefore to a fine tuning of the number of transformations. If only a small amount of adaptation data is available, posterior distributions can still be reliably estimated near the top layers of the tree, and are less and less modified as they are propagated down the tree to the cut-nodes. That prevents overfitting even if the cut is located

[2]The reader should keep in mind that the prior distribution we mention henceforth is the prior distribution of the transformation matrix, not the prior distribution of the HMM parameters.

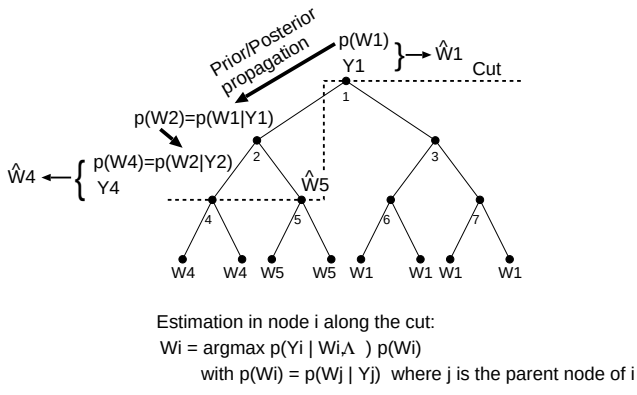

Figure 3: Tree-based SMAPLR algorithm. The adaptation data associated to a node i is denoted $\mathbf{Y}_i$. The corresponding prior density is denoted $p(\mathbf{W}_i)$. For each node i of parent j, the prior density $p(\mathbf{W}_i)$ is defined as $p(\mathbf{W}_j|\mathbf{Y}_j)$.

close to the leaf nodes since the prior information has been obtained from the top nodes. On the other hand, as more and more adaptation data is available, the prior densities are getting more and more refined as they are propagated down the tree, and can therefore provide more local transformations.

A direct consequence of this behavior is that SMAPLR can accommodate a cut that can be placed much lower in the tree compared to MLLR, since the complexity of the transformations can be controlled by the prior structure. Moreover, the SMAPLR adaptation makes a better use of the adaptation data than MLLR since the transformations are constrained from the root node to the cut nodes, and is therefore expected to outperform MLLR. The next section gives a more formal description of the SMAPLR algorithm, and focus especially on the propagation of the prior/posterior density in the tree.

Structural MAPLR algorithm

There are basically two related issues in the SMAPLR algorithm. The first one is to derive an estimate of the transformation matrix at each cut-nodes according to a MAP criterion, while the second one is to approximate the posterior distribution of the transformation matrix given the adaptation data in each node. In order to address these issues, some notations should first be introduced. Then, we briefly describe the MAP estimation, which is nothing more than the MAPLR algorithm [10, 11], and finally we describe the hierarchical derivation of the prior densities.

Model and transformation description We adopt in this paper the same notations as in [10]. The set of HMM parameters is denoted $\mathbf{\Lambda} = \{\omega_{n,m}, \boldsymbol{\mu}_{n,m}, \mathbf{R}_{n,m}\}$ where $\omega_{n,m}$, $\boldsymbol{\mu}_{n,m}$, and $\mathbf{R}_{n,m}$ represent the mixture weight, mean vector and precision[3] matrix of the mth mixture component in the state n, with $\boldsymbol{\mu}_{n,m} \in \mathbb{R}^p$ and $\mathbf{R}_{n,m} \in \mathbb{R}^{p\times p}$. In a given state n, the distribution of an acoustic vector $\mathbf{y}$ is modeled with M Gaussian mixtures:

$$p(\mathbf{y}|s=n) = \sum_{m=1}^{M} \omega_{n,m}\mathcal{N}(\mathbf{y};\boldsymbol{\mu}_{n,m},\mathbf{R}_{n,m}), \qquad (1)$$

[3]The precision matrix is defined as the matrix inverse of the covariance.

where $\mathcal{N}(\mathbf{y};\boldsymbol{\mu}_{n,m},\mathbf{R}_{n,m})$ is a Normal distribution of mean $\boldsymbol{\mu}_{n,m}$ and precision matrix $\mathbf{R}_{n,m}$ defined as:

$$\mathcal{N}(\mathbf{y};\boldsymbol{\mu}_{n,m},R_{n,m}) \propto |R_{n,m}|^{1/2}\cdot\exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_{n,m})'\mathbf{R}_{n,m}(\mathbf{y}-\boldsymbol{\mu}_{n,m})\right\}. \qquad (2)$$

Regarding the model adaptation, we assume that the mean vector of a Normal distribution is transformed by applying an affine transformation defined by the parameter set $\boldsymbol{\eta} = \{\mathbf{A},\mathbf{b}\}$ where $\mathbf{A}$ is a $p\times p$ transformation matrix and $\mathbf{b}$ is a translation vector. As usually done in the MLLR formulation, we will prefer a more compact notation, $\boldsymbol{\eta} = \{\mathbf{W}\}$, where $\mathbf{W} = [\mathbf{A},\mathbf{b}]$ is a $p\times(p+1)$ matrix:

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_{n,m} &= \mathbf{A}\boldsymbol{\mu}_{n,m}+\mathbf{b} \qquad \text{or equivalently}\\ &= \mathbf{W}\hat{\boldsymbol{\mu}}_{n,m},\end{aligned} \qquad (3)$$

where $\hat{\boldsymbol{\mu}}_{n,m}$ is the extended mean vector defined as $\hat{\boldsymbol{\mu}}'_{n,m} = [\boldsymbol{\mu}'_{n,m},1]$.

MAPLR algorithm Let us assume that we want to estimate the transformation matrix $\mathbf{W}$ in a given node of the tree, given some adaptation data $\mathbf{Y}$. The MAP estimation problem can be simply written as follows:

$$\begin{aligned}\hat{\mathbf{W}} &= \underset{\mathbf{W}}{\operatorname{argmax}}\, p(\mathbf{W}|\mathbf{Y},\Lambda)\\ &\propto \underset{\mathbf{W}}{\operatorname{argmax}}\, p(\mathbf{Y}|\mathbf{W},\Lambda)p(\mathbf{W}),\end{aligned} \qquad (4)$$

where $p(\mathbf{W})$ is the *a priori* distribution of the transformation matrix $\mathbf{W}$. This corresponds to the MAPLR adaptation algorithm [10]. A convenient prior distribution family for $p(\mathbf{W})$ is the matrix variate Normal distribution, a matrix version of the multivariate Normal distribution, defined as follows [17]:

$$p(\mathbf{W}) \propto |\mathbf{\Sigma}|^{-(p+1)/2}|\mathbf{\Phi}|^{-p/2}\cdot \exp\left\{-\frac{1}{2}\operatorname{tr}(\mathbf{W}-\mathbf{M})'\mathbf{\Sigma}^{-1}(\mathbf{W}-\mathbf{M})\mathbf{\Phi}^{-1}\right\}, \qquad (5)$$

where $\mathbf{M}$, $\mathbf{\Sigma}$, and $\mathbf{\Phi}$ are the hyperparameters for that distribution family, with $\mathbf{M}\in\mathbb{R}^{p\times(p+1)}$, $\mathbf{\Sigma}\in\mathbb{R}^{p\times p}$, $\mathbf{\Sigma}\geq 0$, and $\mathbf{\Phi}\in\mathbb{R}^{(p+1)\times(p+1)}$, $\mathbf{\Phi}\geq 0$. To distinguish this prior density family from our generic pdfs' notation, $p(\cdot)$, let us represent the matrix variate Normal density by $g(\mathbf{W};\mathbf{\Psi})$ where $\mathbf{\Psi} = \{\mathbf{M},\mathbf{\Sigma},\mathbf{\Phi}\}$ is the set of hyperparameters.

Given this prior distribution, the MAP estimation problem can be solved under closed form via the EM algorithm. Under its most general formulation (full transformation matrix and full prior hyperparameters), the derivation leads to a system of $p\times(p+1)$ linear equations whose solution is the desired transformation matrix. Full details are given in [10, 11].

Hierarchical prior derivation The SMAPLR algorithm relies on the idea that the posterior distribution in a given node is used as prior distribution in its child nodes. In order

to carry out the MAPLR estimation in each node of the tree, it is essential to use a prior distribution belonging to the matrix variate Normal density used in the MAPLR formulation. However, the posterior distribution of $\mathbf{W}$ given the adaptation data $\mathbf{Y}$, $p(\mathbf{W}|\mathbf{Y})$, does not belong to the $g(\mathbf{W};\mathbf{\Psi})$ family, as illustrated below, where $S = \{s_t\}$ and $L = \{l_t\}$ is the state and mixture sequence associated to the observation sequence $\mathbf{Y} = \{\mathbf{y}_t\}$:

$$\begin{aligned} p(\mathbf{W}|\mathbf{Y},\mathbf{\Lambda}) &= \frac{p(\mathbf{Y}|\mathbf{W},\mathbf{\Lambda})g(\mathbf{W};\mathbf{\Psi})}{p(\mathbf{Y}|\mathbf{\Lambda})} \\ &= \frac{\sum_S \sum_L p(\mathbf{Y},S,L|\Lambda,\mathbf{W})g(\mathbf{W};\mathbf{\Psi})}{p(\mathbf{Y}|\mathbf{\Lambda})} \end{aligned} \quad (6)$$

The summation in (6) does not reduce to a matrix variate Normal distribution. Moreover, propagating the true posterior pdf down the tree would generate an ever expanding summation of terms. It is therefore necessary to approximate the true posterior distribution $p(\mathbf{W}|\mathbf{Y},\mathbf{\Lambda})$ by its "closest" (i.e. in the Kullback-Leibler sense) distribution $g(\mathbf{W};\mathbf{\Psi})$ from the matrix variate Normal prior family. This is however a very complex problem with no closed-form solution and an alternative solution is required.

As the amount of adaptation data increases, the posterior distribution $p(\mathbf{W}|\mathbf{Y},\mathbf{\Lambda})$ gets more and more concentrated around its mode, due to the dominant contribution of the likelihood. It becomes therefore reasonable to approximate the true posterior distribution $p(\mathbf{W}|\mathbf{Y},\mathbf{\Lambda})$ by a distribution from the matrix variate Normal family $g(\mathbf{W};\mathbf{\Psi})$ having the same mode. Hence, the problem reduces to deriving the mode of the posterior distribution $p(\mathbf{W}|\mathbf{Y},\mathbf{\Lambda})$, which is nothing more than the MAPLR estimate of the transformation matrix described at the previous section. Let i be a given node and j be one of its child node. Let $\mathbf{Y}_i$ be the adaptation data available in the node i and $\hat{\mathbf{W}}_i$ the corresponding MAPLR estimate of the transformation matrix. The prior density $p_j(\mathbf{W})$ simply approximate the true posterior distribution $p_i(\mathbf{W}|\mathbf{Y}_i)$ by a distribution from the matrix variate Normal density having $\hat{\mathbf{W}}_i$ as mode, in other words such that $\mathbf{M} = \hat{\mathbf{W}}_i$. Since there is an infinity of such distributions having $\hat{\mathbf{W}}_i$ as mode, we will assume as a simplification that the other hyperparameters, namely $\{\mathbf{\Sigma},\mathbf{\Phi}\}$ are not modified.

The whole SMAPLR adaptation algorithm can therefore be summarized as follows:

1. Start with an initial prior distribution from the matrix variate Normal density at the root node.

2. Explore the tree in a breadth-first manner. For each node j of parent i:

   (a) Set the hyperparameter $\mathbf{M}_j$ of the prior distribution to the MAPLR estimate $\mathbf{W}_i$ derived in the parent node i

   (b) Derive the MAPLR estimate $\mathbf{W}_j$ given the adaptation data $\mathbf{Y}_j$ associated to node j

   (c) If j is a leaf node, apply the transformation $\mathbf{W}_j$ to the corresponding Gaussian pdf

As a result, all Gaussian pdfs are adapted, regardless of the amount of adaptation data. If little adaptation data is available, the priors at the top of the tree have a dominant role and get propagated down the tree to the cut-nodes, without being significantly modified from layer to layer. As the amount of adaptation increases, the adaptation scheme is able to take full advantage of the information provided by the adaptation data while still benefitting from hierarchically derived prior densities.

3. EXPERIMENTS AND RESULTS

Database and System description

The proposed adaptation algorithm is evaluated on the non-native speaker adaptation part (Spoke3, November 93) of the Wall Street Journal (WSJ) task. The Spoke3 data consists of 5K-word read WSJ data collected from 10 non-native speakers of American English. Each speaker provided 40 utterances for fast model adaptation and 40 utterances for testing.

A standard Mel frequency cepstral coefficient (MFCC) front-end is used to create a feature vector of 39 components, consisting of 12 MFCC component plus the log-energy term and their first and second derivatives. Cepstral mean normalization is applied on each sentence.

The SI-84 training set (WSJ0) is used to build baseline triphone HMMs using the phonetic decision tree state tying algorithm described in [14]. The standard trigram language model provided by NIST for the WSJ corpus is used in all experiments, together with a 5K-word lexicon automatically generated using a general English text-to-speech system [18].

Experimental results

Adaptation experiments are carried out in supervised mode[4], for various amount of adaptation utterances (1, 5, 10, 20 and 40 utterances). For each each test speaker and adaptation configuration, the adapted models are saved and used to run recognition on the corresponding test data. Various adaptation techniques are used and compared to the proposed SMAPLR framework, including MAP adaptation [5], MLLR adaptation [3] and SMAP adaptation [13]. In each case, the HMM mean vectors are the only parameters to be adapted. Both SMAPLR and MLLR use full transformation matrices.

The tree of Gaussian densities needed by the MLLR, SMAPLR and SMAP algorithm is built from the baseline speaker independent HMMs using the algorithm described in [13], slightly modified to get a more balanced tree.

Rather than adjusting and specifying explicitly the number of transformation matrices (regression classes) used by MLLR for each size of adaptation utterances, the number of transformation matrices is determined using the tree based

[4]Supervised adaptation means that the text transcription of the adaptation data is known

on the amount of adaptation data available in each node, as in [19]. The number of transformations is therefore dynamically controlled using a single threshold specifying the minimum number of frames required in each node to define a regression class. This strategy is also used in the SMAPLR experiments, meaning that the same regression classes are used by MLLR and SMAPLR, the difference being that the SMAPLR transformations are hierarchically derived from the root node using a MAP criterion.

SMAPLR adaptation is controlled by the initial prior density $p(\mathbf{W})$ at the root node. We choose $p(\mathbf{W})$ such that its mode $\mathbf{M}$ represents an identity transformation, $\mathbf{M} = [\mathbf{I}\ \mathbf{0}]$. The scale of the prior distribution $p(\mathbf{W})$ is controlled by the two hyperparameters $\mathbf{\Sigma}$ and $\mathbf{\Phi}$. We fix $\mathbf{\Phi}$ to the identity matrix, $\mathbf{\Phi} = \mathbf{I}$, as done in [10, 11]. $\mathbf{\Sigma}$ is set to a scaled identity matrix, $\mathbf{\Sigma} = C \cdot \mathbf{I}$ so that the scaling is only controlled by a scalar coefficient C.

In a first series of experiments, the influence of the scaling factor C is studied. C is used to control the initial weight of the prior information compared to the weight of the adaptation data. For a given amount of adaptation data, as C decreases, the influence of the prior information increases. On the other hand, as C increases, the influence of the prior information decreases and the estimate of the transformation matrices relies mainly on the adaptation data. At the limit, if C is set to infinity, the prior information vanish and SMAPLR converges to MLLR; if C is set to 0, the estimate of $\mathbf{W}$ becomes equal to the mode $\mathbf{M}$ of the initial prior distribution and is an identity transformation. The influence of C is illustrated in Fig. 4 where the average word error rate (WER) is given for various amount of adaptation data, with $C = 1$, $C = \frac{1}{10}$ and $C = \frac{1}{100}$. If too little weight is given to the prior information ($C = 1$), the estimated transformations may overfit the adaptation data, as illustrated when using a single adaptation utterance. By increasing the weight of the prior ($C = 1/100$), the estimated transformations are constrained by the transformations lying near the root node of the tree, reducing the risk of overfitting. On the other hand, for a large amount of adaptation utterances, say 40, a strong prior information ($C = 1/100$) may be detrimental since it is constraining the transformation estimates too much with the estimates derived near the root node. It is reasonable to adjust the scaling factor C based on the depth of the tree. A simple heuristic consisting of setting a small C at the root note ($C = 1/100$) and increasing it for the nodes close to the leaves appears to be quite effective, leading to a performance close to $C = 1/100$ for small amount of adaptation data and close to $C = 1$ for larger amount of adaptation utterances.

This heuristic is used in the SMAPLR results presented in Fig 5, along with MAP, MLLR and SMAP adaptation results. Because of the limited amount of adaptation data, MAP performs quite poorly since only a fraction of the models are adapted. The MLLR results indicate a strong overfitting trend, especially when using a small amount of adaptation data. Since by design, MLLR and SMAPLR use to same tree, the regression classes and number of transformations are exactly the same for these two techniques. However, SMAPLR uses the adaptation data in a more efficient way and is able to reduce the overfitting because of the hierarchical prior structure. The SMAP algorithm which is also based on a hierarchical prior structure performs very well for a large amount of adaptation data, and is slightly better than the SMAPLR algorithm. However, for a small amount of adaptation data (1 and 5 utterances), SMAPLR outperforms SMAP, leading to 26.02% and 27.29% of WER respectively using a single adaptation utterance, compared to 29.1% of WER for the baseline (unadapted) system.

# Adapt. Utt.	1	5	10	20	40
Not tuned	39.25	27.80	23.14	19.72	17.89
Tuned	25.95	21.84	21.57	18.41	16.55

Table 1: Average word error rate for 2 MLLR settings: with (Tuned) and without (Not tuned) fine selection of the number of transformation matrices.

Regarding the comparison between MLLR and SMAPLR, we should indicate that we did not try to adjust the number of transformations to get the best recognition performance. It is likely that by reducing the number of transformations for small amount of adaptation data, the MLLR performance can be improved and the overfitting can be reduced. This is illustrated in Table 1 where the MLLR experiments have been rerun for a much smaller number of transformations. While the overfitting can be eliminated, it still appears that MLLR adaptation with a careful selection of the number of transformations does not outperformed an out-of-the box SMAPLR where no fine tuning of the number of transformations has been performed. This illustrates the robustness of SMAPLR over MLLR, for which whenever the acoustic models are changed (or the tree structure is modified) it is necessary to readjust the threshold τ to select the best number of transformations. Moreover, we have shown in [11] that a carefully tuned MAPLR (with hierarchical priors) outperforms a carefully tuned MLLR. For these reasons, we believe that SMAPLR can offer improved performance and easier tuning than the standard MLLR.

4. CONCLUSION

A new adaptation technique called Structural *Maximum a Posteriori* Linear Regression, or SMAPLR, has been proposed. We have shown how the standard MLLR technique can be improved first by adding prior information regarding the transformation parameters to constraint the estimation (leading to the MAPLR algorithm), then by adding a hierarchical structure to the priors (leading to the SMAPLR algorithm). The advantages of this structured prior information is twofold. First, it provides a better use of the adaptation data since transformations are hierarchically derived, the global transformations being used to constraint the estimation of the more local transformations. Second, it significantly reduces the risk of overfitting the adaptation data. This was illustrated on the non-native speaker part (Spoke3)

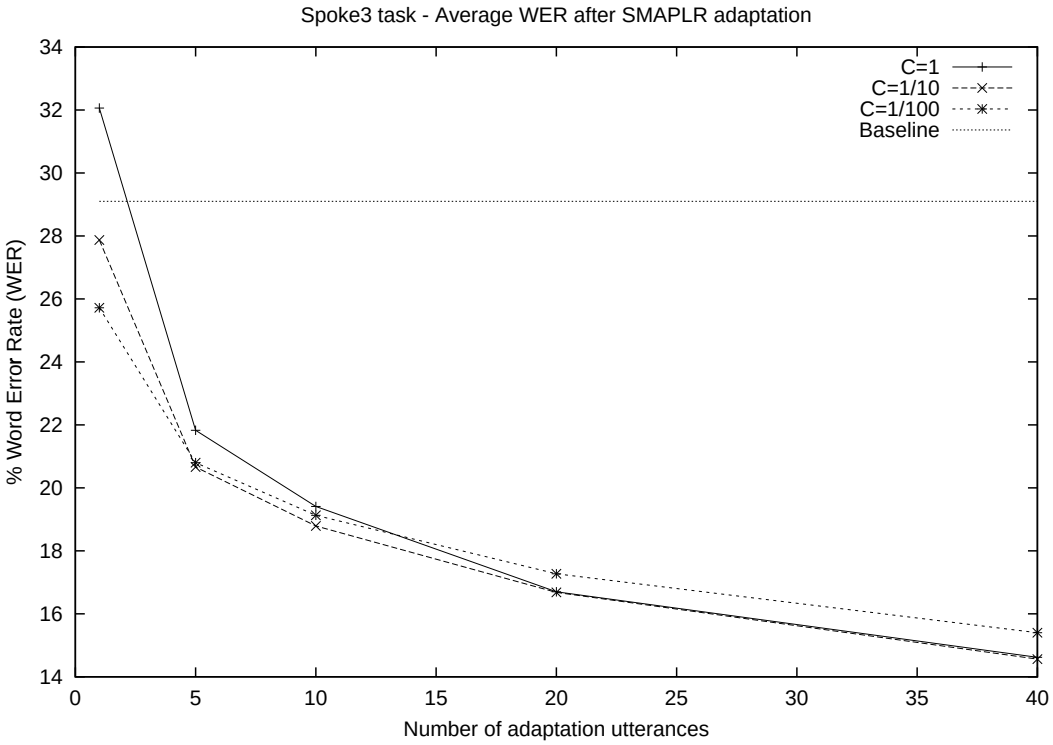

Figure 4: Average word error rate (WER) in % for SMAPLR adaptation for various amount of adaptation utterances and $C = 1$, $C = 1/10$ and $C = 1/100$. The baseline performance (unadapted models) is also given.

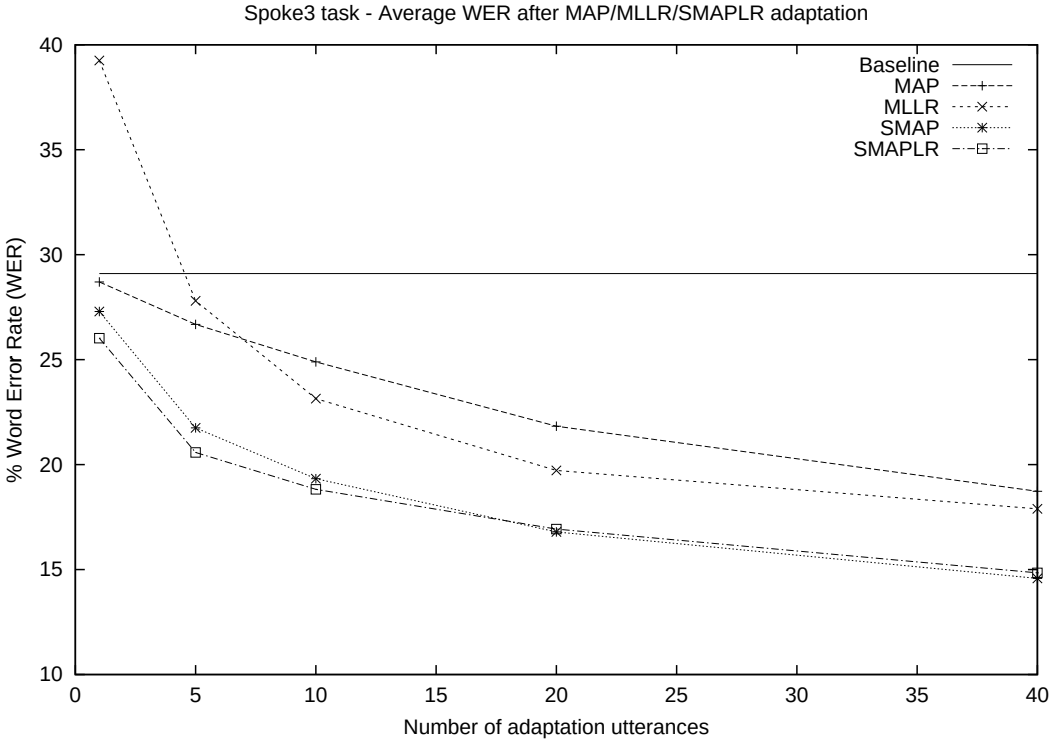

Figure 5: Average word error rate (WER) in % for MAP, MLLR, SMAP, and SMAPLR adaptation for various amount of adaptation utterances. The baseline performance (unadapted models) is also given.

of the Wall Street Journal task. Significant advantage is obtained over MLLR and MAP adaptation. Compared to SMAP adaptation, SMAPLR provides a small but significant improvement for very small amount of adaptation data (1 and 5 utterances) but their respective performances are statistically equivalent for larger amount of data. These are preliminary results and further work is needed to study and understand better the properties of this new adaptation technique.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their comments.

REFERENCES

[1] C.-H. Lee. On stochastic feature and model compensation approaches to robust speech recognition. *Speech Communication*, 25:29–47, 1998.

[2] C.-H. Lee. Adaptive classification and decision strategies for robust speech recognition. In *Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, May 1999.

[3] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden Markov models. *Computer Speech and Language*, 9:171–185, 1995.

[4] C.-H. Lee, C.-H. Lin, and B. H. Juang. A study on speaker adaptation of continuous density HMM parameters. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 145–148, Albuquerque, New Mexico, April 1990. ICASSP'90.

[5] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, April 1994.

[6] V. V. Digalakis and L. G. Neumeyer. Speaker adaptation using combined transformation and Bayesian methods. *IEEE Transactions on Speech and Audio Processing*, 4(4), July 1996.

[7] O. Siohan, C. Chesta, and C.-H. Lee. Joint maximum *a Posteriori* adaptation of transformation and HMM parameters. Submitted to IEEE Trans. on Speech and Audio Processing.

[8] J.-T. Chien, C.-H. Lee, and H.-C. Wang. A hybrid algorithm for speaker adaptation using MAP transformation and adaptation. *IEEE Signal Processing Letters*, 4(6):167–169, June 1997.

[9] J.-T. Chien, C.-H. Lee, and H.-C. Wang. Improved Bayesian learning of hidden Markov models for speaker adaptation. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Munich, Germany, 1997.

[10] O. Siohan, C. Chesta, and C.-H. Lee. Hidden Markov model adaptation using maximum a posteriori linear regression. In *Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, 1999.

[11] C. Chesta, O. Siohan, and C.-H. Lee. Maximum a posteriori linear regression for hidden Markov model adaptation. In *Proceedings of European Conference on Speech Communication and Technology*, volume 1, pages 211–214, Budapest, Hungary, 1999.

[12] K. Shinoda and C.-H. Lee. Structural MAP speaker adaptation using hierarchical priors. In *Proc. of IEEE Workshop on Speech Recognition and Understanding*, 1997.

[13] K. Shinoda and C.-H. Lee. Unsupervised adaptation using structural Bayes approach. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, Washington, USA, 1998.

[14] W. Reichl and W. Chou. A decision tree state tying based on segmental clustering for acoustic modeling. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 801–804, Seattle, Washington, USA, 1998.

[15] C. J. Leggetter and P. C. Woodland. Speaker adaptation of HMMs using linear regression. Technical Report F-INFENG/TR.181, CUED, Cambridge University Engineering Department, UK, June 1994.

[16] C. J. Leggetter and P. C. Woodland. Speaker adaptation of continuous density HMMs using multivariate linear regression. In *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 451–454, Yokohama, Japan, September 1994.

[17] A. K. Gupta and T. Varga. *Elliptically Contoured Models in Statistics*. Kluwer Academic Publishers, 1993.

[18] R. W. Sproat and J. P. Olive. Text-to-speech synthesis. *AT&T Technical Journal*, 74:35–44, 1995.

[19] C. J. Leggetter and P. C. Woodland. Flexible speaker adaptation for large vocabulary speech recognition. In *Proceedings of European Conference on Speech Communication and Technology*, volume 2, pages 1155–1158, Madrid, Spain, September 1995.