In this paper, we extend the SMAP adaptation to a segment-based model: the Mixture Stochastic Trajectory Model (MSTM). SMAP approach is completed by the tree construction driven by adaptation data, a Minimum Description Length (MDL) structure definition of this tree and trajectory and state adaptations. On the Resource Management task, the segment-based noise adaptation and noise adaptation experiments show that the proposed SMAP approach gives a significant improvement compared to unadapted system.

1. INTRODUCTION

The adaptation techniques allow to reduce the differences between training and testing conditions: reduce the speaker variability, the noise mismatch, the channel or microphone mismatch and any other environment mismatch or combination of these factors. Two state of the art approaches should be mentioned for model adaptation: Maximum Likelihood Linear Regression adaptation (MLLR) [1] and Maximum a Posteriori adaptation [2]. The first one is based on linear transformations of acoustic model pdfs estimated using the adaptation data with Maximum Likelihood criterion. MAP approach can be viewed as extending the training but using only adaptation data and some knowledge about the estimation parameters. The parameter estimation is performed under the MAP criterion. Incorporating the prior information is particularly useful for dealing with the problems raised by sparse adaptation data for which ML approach gives inaccurate estimations. In order to cumulate the advantages of both approaches, several performant combinations of them have been proposed recently: Structural Maximum a Posteriori adaptation (SMAP) [3], Maximum a Posteriori Linear Regression adaptation (MAPLR) [4] and Structural Maximum a Posteriori Linear Regression adaptation (SMAPLIR) [5]. These methods show that the hierarchical organisation of models and hierarchical principle of parameter estimations under MAP criterion improve the estimation: the adaptation data is used more efficiently and some models without adaptation data can be nevertheless adapted.

The tree structure of the models can be chosen to improve the recognition accuracy under two different circumstances: either when there is a small amount of adaptation data or when the amount of adaptation data increases. Experimental estimation of the optimal tree structure (number of levels and number of nodes in each level) and the number of transformation parameters is very difficult because the optimal tree structure depends on the size of models and on the amount of adaptation data. An automatic estimation of the tree structure and of the number of transformation parameters has been proposed in [6, 7, 8]. These approaches are based on Minimum Description Length (MDL) principle: choose the model that gives the shortest description of data. The MDL allows to control the model complexity according to the available amount of adaptation data.

In this paper, the SMAP for a segment-based model, the Mixture Stochastic Trajectory Model (MSTM) [9] is proposed. Our SMAP uses the notion of trajectory (set of states) instead of the notion of state, as in classical HMM models. To illustrate the power of trajectory adaptation, the state adaptation version of SMAP is also proposed and evaluated. In our SMAP approach the hierarchical organisation of acoustic model trajectories is used (hierarchical organisation of state pdfs for state adaptation version of SMAP). The transformation parameter set is estimated level per level under the MAP criterion to integrate the prior information in the estimation. Using the tree organisation of the models and, so, transformation parameters, the posterior density of each node of transformation tree is used in prior distribution in its child nodes. We propose also a new approach to Gaussian tree building driven by adaptation data. To automatically determine the number of adaptation parameters in function of available adaptation data, we propose to use the MDL principle. Compared to [7, 8], in our approach the MDL is applied to each trajectory tree node in the trajectory adaptation version of SMAP.

The paper is organised as follows. We begin by a brief presentation of MSTM in section 2. The SMAP adaptation model for MSTM and the adaptation parameters estimation are presented in section 3. Section 4 gives the experimental results for continuous speech recognition task. Section 5 concludes the paper.

2. MSTM ACOUSTIC MODEL

The basic idea behind the Mixture Stochastic Trajectory Model (MSTM) recognizer [9] is that speech can be considered as a point that moves in the acoustic space as the articulatory system changes. A sequence of acoustic vectors corresponding to the subsequent positions of this moving point is called a trajectory of speech.
We assume that for each unit $u$ of the set of units $B$, a trajectory can be represented with sufficient accuracy by a fixed number $Q$ of points in the acoustic space. As observed segments representing the same phonetic unit can have different durations, the acoustic vectors in each speech segment $(z_1, \ldots, z_d)$ are resampled to $Q$ points, resulting in the fixed length segmental feature vector $X$:

$$(z_1, \ldots, z_d) \Rightarrow X = (x_1, \ldots, x_Q) \quad (1)$$

The most important part of the trajectory model is the statistical characterisation of the trajectory by means of the conditional pdf

$$P(X|d, u)$$

is to estimate the mismatch parameters

$$\gamma_{k,n} = Pr(t_k|X_n^u, d, u) = \frac{P(X_n^u|t_k, d, u)Pr(t_k|u)}{\sum_{i=1}^{K_u} P(X_n^u|t_i, d, u)Pr(t_i|u)} \quad (9)$$

This formula is different from the posterior state probability estimation of HMMs and allows to take several trajectory states into account. The re-estimation formulas for each adaptation parameter can be found in [10].

Using these estimations, the training Gaussian trajectories are adapted as follows: for all $(k, u) \in p$ and for $i = 1, \ldots, Q$ we have:

$$\mu_{k,i} = \frac{\mu_{k,i} + (\Sigma_{k,i})^{1/2} \mu_{p,i}}{(\Sigma_{k,i})^{1/2}} \quad (10)$$

The 10 is similar to those of [3], but in our case we use a mismatch trajectory instead of a mismatch pdf.

3.3. Tree-Structure of Estimation

The training trajectories and the corresponding mismatch trajectories are organised in a tree. An advantage of the tree-structure is the possibility to dynamically control the number of adaptation parameters in function of the amount of adaptation data. The mismatch parameters are estimated level per level from the root to the leaves using the MAP estimation: the posterior probability for each node of the tree is the product of ML estimate of parameters for this node and a prior density. We choose as prior density, a Normal-Wishart density, and assume that it depends on the MAP estimate of adaptation parameters of the parent node.
3.4. Tree Structure Estimation based on MDL

The Minimum Description Length principle (MDL) allows to select from different models \( \{ \tilde{a} \} \) with different complexity a probabilistic model \( \tilde{a} \) that maximise the likelihood of the observed data \( \{ x \} \) and in the same time has a small complexity:

\[
\hat{a} = \arg\min_a DL = \arg\min_a \left\{ -\log_\alpha P(x) + \frac{1}{2} \beta_a \log N \right\}
\]

(11)

where \( \beta_a \) is the number of the free parameters in the model \( a \), \( N \) is the amount of the observed data \( x \) and \( P(x) \) is the likelihood of the data \( x \). We propose to use the MDL to define automatically the tree depth, the optimal number of nodes for each level and, so, the optimal number of transformation parameters. Compared to \([7, 8]\), the likelihood of data is estimated using the MSTM models and not the HMM models. We use the recursive algorithm to calculate the MDL for each node of the tree and for the corresponding subtree. After the MDL definition of the tree structure, the acoustic models are adapted using the MAP estimated transformation of each node.

3.5. Tree Building Driven by Adaptation Data

If the amount of adaptation data is sufficiently important (several sentences) it is possible to use the information contained in the adaptation data during the acoustic model tree building. To do this, we propose to build two trees at the same time: the tree based on the adaptation data and the acoustic models and the tree based on the adaptation data. The advantage of the former tree is that it is based on a big amount of training data. The advantage of the latter tree is that it is based on the data specific to the test conditions. Our idea is to found a compromise between these two trees. The centroid \( \bar{m}_a \) of each node of the acoustic model tree is moved toward the centroid \( \bar{m}_a \) of the corresponding node of the adaptation data tree:

\[
\bar{m}_a = \sigma \bar{m}_a + (1 - \sigma) m_a
\]

(12)

where \( 0 \leq \sigma \leq 1 \) is a parameter determined experimentally. The computed \( m_a \) of the node is used as new centroid of this node. After this, each acoustic model (trajectory or state) is re-affected to the closest node according to some distance until the node centroid. In summary, the acoustic model tree is built taking into account the structure of the adaptation data. After building a model tree, the tree is used to estimate the transformation parameters in each node.

3.6. State Adaptation Version of SMAP

For the state adaptation version of SMAP, the notion of state is used instead of the notion of trajectory in all steps of the adaptation process:

- the model tree is built by clustering the states of MSTM;
- the adaptation data tree is built by using the sampled adaptation data (states);
- the transformation parameters are computed for each state independently of other states in the same trajectory;
- the adaptation is performed state per state using the transformation parameters for each state.

The state adaptation version of SMAP is close to classical SMAP for HMM [3].

4. EXPERIMENTS AND RESULTS

4.1. Experimental Conditions

The proposed approach has been validated on the DARPA Resource Management corpus (RM) for the task of speaker and noise adaptation. For the training of the speaker-independent acoustic models, the speaker-independent part of RM (RM1) is used (speech of 72 speakers, 2880 sentences). For the adaptation and the test, the speaker-dependent part of RM is used:

- 600 training sentences per speaker to train the speaker-dependent model and for adaptation;
- 100 development sentences per speaker for testing (1200 sentences).

The speech is sampled at 16 kHz and a 13th order mel-cepstral vector is computed every 10 ms using an analysis window of 32 ms. The long silence segments are removed automatically from speech using a speech/non speech detector (based on a speech energy threshold). The speech recognizer uses context-independent models for 47 phonetic units. In the current MSTM, a linear sampling is used. The segmental feature vector is composed of 5 acoustic vectors \( Q = 5 \). All covariance matrices \( \Sigma_{km} \) are assumed to be diagonal. Each mixture has a maximum of 16 components.

The statistical language model is based on [11] and determines the word sequences that can occur in a sentence. This language model has a word-pair equivalent perplexity of 69.

The binary tree for the SMAP adaptation is created using the LBG algorithm in a top-down manner using the Mahalanobis distance. For the refinement of classes, a k-means algorithm is used.

The proposed SMAP approach is compared to MLLR adaptation [12]. The MLLR uses broad phonetic classes to define the regression classes and diagonal transformation matrices with translations. For each amount of adaptation data the number of regression classes is estimated experimentally. The word error rate of the speaker-dependent system is 1.2%.

The adaptation is used to adapt only means. Other experiments show that the variance adaptation has not given any improvement. The hyperparameters of Normal-Wishart density are determined experimentally on the small part of the testing speech of RM and increase then the tree level increases.

4.2. Speaker Adaptation Results

The recognition results using speaker-independent system, MLLR supervised speaker adaptation and SMAP supervised and unsupervised speaker adaptation is presented in table 1 in terms of word error rate (±0.5% confidence interval). For SMAP adaptation, the experiments show that using the MDL principle to define the tree structure gives the same recognition performance that best tree structure defined experimentally.

The proposed trajectory per trajectory adaptation (noted as SMAP (traj) in table 1) outperforms the state per state adaptation (cf. 3.6 and noted as SMAP (stat) in table 1). This can be explained by the segment-based nature of MSTM: all the model parameters are estimated in the trajectory per trajectory manner during the training, so it is better to estimate the model transformations in the same manner. In the following, we use only the trajectory adaptation version of MSTM.

The tree building driven by adaptation data, tested using 10 and 100 adaptation utterances, shows that, when \( \sigma \) decreases, the recognition accuracy decreases as well. Two reasons may be given:
either our tree combination defined by (12) is not adequate, or the acoustic model tree is sufficiently good.

<table>
<thead>
<tr>
<th>Nbr. of adapt. utt.</th>
<th>0</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>superv. MLLR</td>
<td>7.8</td>
<td>7.3</td>
<td>6.7</td>
<td>6.6</td>
<td>6.3</td>
</tr>
<tr>
<td>superv. SMAP (state)</td>
<td>7.8</td>
<td>6.7</td>
<td>6.7</td>
<td>6.5</td>
<td>6.3</td>
</tr>
<tr>
<td>superv. SMAP (traj)</td>
<td>7.8</td>
<td>6.8</td>
<td>6.5</td>
<td>5.9</td>
<td>5.2</td>
</tr>
<tr>
<td>unsup. SMAP (traj)</td>
<td>7.8</td>
<td>7.3</td>
<td>6.6</td>
<td>6.5</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 1. Comparison of SMAP speaker adaptation performance versus MLLR performance in function of adaptation data amount (0 adaptation utterance corresponds to unadapted system). Word error rate (%).

In all cases, the SMAP adapted system results are better than those of speaker-independent system and than those of MLLR adapted system. Furthermore, the unsupervised SMAP adaptation gives the same or better results than the supervised MLLR adaptation no matter how much adaptation data is used.

In summary, the SMAP adaptation reduces the word error rate significantly compared to the speaker-independent system if 5 or more adaptation sentences are used.

4.3. Speaker and Noise Adaptation Results

For noise adaptation, adaptation and test speech have been contaminated with a Gaussian noise at two levels: 15dB and 25dB. After this, Cepstral Mean Normalization is performed. Table 2 shows the performance of the SMAP trajectory adaptation in function of adaptation data amount and of noise level (for 15dB, the confidence interval is ±1%, for 25dB, the confidence interval is ±0.7%). Preliminary results shown that the performance of the MLLR adapted system in the case of noise adaptation is worse than the one of the SMAP adapted system and not presented here.

We can see that using an unsupervised adaptation with 1 adaptation sentence is not a good idea: the results are worse than those of unadapted system. There is an improvement with 5 or more adaptation sentences.

<table>
<thead>
<tr>
<th>Nbr. of adapt. utt.</th>
<th>0</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>15dB unsup.</td>
<td>50.2</td>
<td>52.1</td>
<td>50</td>
<td>48.5</td>
<td>44.5</td>
</tr>
<tr>
<td>15dB superv.</td>
<td>50.2</td>
<td>47.4</td>
<td>41.2</td>
<td>40.7</td>
<td>35</td>
</tr>
<tr>
<td>25dB unsup.</td>
<td>14.4</td>
<td>14.7</td>
<td>14.2</td>
<td>13.9</td>
<td>12.6</td>
</tr>
<tr>
<td>25dB superv.</td>
<td>14.4</td>
<td>15.2</td>
<td>13.6</td>
<td>12.7</td>
<td>10.2</td>
</tr>
</tbody>
</table>

Table 2. SMAP adapted system results in function of adaptation data amount and of noise level. Word error rate (%).

5. CONCLUSION

In this paper, the Tree-Structured Maximum a Posteriori adaptation is developed for a segment-based model: the Mixture Stochastic Trajectory Model. We propose two versions of adaptation: trajectory and state adaptation. To use the information contained in the adaptation data during the acoustic model tree building, the tree building driven by adaptation data is proposed. To automatically define the acoustic model tree structure and the number of transformation parameters, the MDL principle is used. A set of speaker adaptation and speaker and noise adaptation experiments in supervised and unsupervised modes show that SMAP adaptation significantly outperforms the speaker-independent system and outperforms the MLLR adapted system.

The study of the following topics is in progress: optimal estimation of the hyperparameters of Normal-Wishart density, the extension of SMAP adaptation to online adaptation and a study of the real-world noise adaptation.

6. REFERENCES