CONSTRANGED MAXIMUM LIKELIHOOD LINEAR REGRESSION FOR SPEAKER ADAPTATION

Mohamed Afify* Olivier Siohan

Multimedia Communications Research Lab
Bell Laboratories, Lucent Technologies
600 Mountain Avenue, Murray Hill, NJ 07974, USA
afify, siohan@research.bell-labs.com

ABSTRACT
This paper proposes a new structure for use in MLLR adaptation aiming at constraining the transform for potentially better parameter estimation from sparse adaptation data. Motivations for the use of the new structure, and EM based parameter estimation are presented. Experimental results on Spoke3 of the Wall Street Journal task revealed that the proposed transformations outperform a full matrix for a small amount of adaptation data and performs equally well for large adaptation set. They also outperform diagonal transformations for all amounts of adaptation data.

1. INTRODUCTION

Maximum likelihood linear regression (MLLR) [1] has proved to be an effective way for model adaptation. In MLLR the system mean vectors are grouped into some regression classes, and a linear transformation is then estimated for each class so as to maximize the likelihood of the adaptation data. This transformation sharing allows a relatively fast adaptation from a small amount of data. However, for very short adaptation material MLLR may fail to achieve acceptable performance. Sometimes the transforms overtrain leading to even worse performance. One way to overcome this degradation is by reducing the number of transformations and hence the accuracy of the adaptation, another way is to consider constrained estimation of each transform. The recently proposed MAPLR [2], for example, uses a prior to derive appropriate constraints and shows superior performance over conventional MLLR especially in sparse data situations. Another way to look at the previous problem is to consider the elements of a transformation matrix to be constrained in some way, e.g. they are a certain function of a few parameters, or to assume that they have a certain structure. This helps in reducing the number of free parameters, and hence increases the robustness against the lack of sufficient adaptation data. We refer to the latter as structural constraints. Well known examples include diagonal [8], and block diagonal [3] transformations.

In this paper we address the extension of the above structural constraints by using band diagonal transformations, that is, only elements at a specified width from the diagonal are nonzero. This clearly offers another degree of freedom in handling the accuracy-robustness tradeoff, and may be used to improve the adaptation accuracy for a certain amount of data. We also provide an interesting theoretical motivation for introducing such band diagonal transformations starting from the recently proposed bilinear warping based adaptation [4]. While the EM algorithm used for MLLR estimation can be extended to the band diagonal case, we introduce an estimation procedure based on an iterative M-step which is very well suited for the sparse structure. It can be considered as an alternative way to the conventional estimation procedure.

The rest of the paper is organized as follows. In Section 2 we introduce and theoretically motivate the proposed transform structure. The estimation algorithm for the transform parameters is given in Section 3. The algorithm used in this work for distribution clustering and transformation selection is outlined in Section 4. Experimental evaluation for the Spoke3 task is presented in Section 5, followed by a conclusion in Section 6.

2. THE PROPOSED STRUCTURE

In this section we motivate and derive the proposed transformation structure. We first consider the cepstral coefficients, and then return to first and second order derivatives. Consider the bilinear transformation based adaptation [4]. In this method a bilinear (allpass) mapping [12] is applied to the cepstral sequence. The z-transform of this mapping is given in Eq. (1),

\[ z_{\text{new}} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} - 1 < \alpha < 1 \]  

(1)

The application of the above mapping to the cepstral sequence results in a transformed cepstral sequence given by

\[ c'_k = \sum_{n=1}^{N} \phi(n)c_n \]  

(2)
where $1 \leq k \leq N$, $c_k'$ is the new warped cepstral sequence, $c_n$ is the original cepstral sequence, $N$ is the cepstral order, and $\phi_b(n)$ is a special sequence expressable in terms of Laguerre sequence as discussed in [5]. It should be noted that this sequence, although not explicitly shown, is a function of the warping parameter $\alpha$. Applying the transformation of Eq. (3) to the cepstral means we can write [4]

$$\mu_k' = \sum_{n=1}^{N} \phi_b(n) \mu_n$$  \hspace{1cm} (3)$$

where $\mu_k'$ is the new warped cepstral mean, and $\mu_n$ is the original cepstral mean. We note that Eq. (3) is a special form of MLLR with the transformation matrix parametrized using a single parameter $\alpha$. This puts a tight constraint on the transformation, and a method using line search can be used to find the optimal warping parameter which maximizes the likelihood of the adaptation data. In preliminary experiments with this approach we observed that the optimization leads to small values of $|k_1| \approx 0.1$. We show in [5] that for $|k_1| \ll 1$ the matrix formed from the sequence $\phi_b(n)$ is band diagonal, specifically, the elements of the matrix at a distance $\Delta$ from the main diagonal approach zero as $\Delta \rightarrow \infty$. In addition, we also show in [5] that for $|k_1| \ll 1$ the frequency warping induced by the mapping in Eq. (1) is such that

$$\omega' \leq (1 + 2\alpha')\omega$$  \hspace{1cm} (4)$$

where $\omega'$ is the warped frequency variable, and $\omega$ is the original frequency variable. The result in Eq. (4) comes in accordance with explicit linear frequency warping [13] where the warping factor is constrained to be very close to one. The above results imply the following:

- It is reasonable to expect a small warping factor $\alpha$ in agreement with explicit frequency warping procedures.
- For a small value of the warping parameter the transformation matrix defined by the sequence in Eq. (3) is band diagonal.

It is interesting to note that this band diagonal behaviour of the bilinear warping was also empirically observed in [6]. Since bilinear warping which mainly accounts for speaker normalization is shown to lead to a band diagonal transformation matrix, it can be expected that a band diagonal structure trained from data can efficiently capture speaker mapping effects. The above arguments motivate the use of a band diagonal structure for transforming the cepstral means. However, it should be noted that although this structure was based on a speaker adaptation argument it may be still used effectively for other types of environment adaptation. We will address this point in future studies. In [5] an argument is given in favor of using the same structure of the cepstra for the delta and the delta-delta. Hence, motivated by block diagonal transformations we form the transformation matrix $A$ from the blocks, $A^c$ for cepstrum, $A^d$ and $A^{dd}$ for the delta and delta-delta respectively, as follows

$$A = \begin{pmatrix} A^c & 0 & 0 \\ 0 & A^d & 0 \\ 0 & 0 & A^{dd} \end{pmatrix}$$  \hspace{1cm} (5)$$

where $A^c, A^d,$ and $A^{dd}$ satisfy band diagonal constraints of width $\Delta$, such that,

$$a_{ik} = \begin{cases} a_{ik} & i - \Delta \leq k \leq i + \Delta \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (6)$$

In our experiments we augment the cepstral features with energy and its delta and delta-delta. In this case we extend the transformation matrix in Eq. (5) by adding a block $A^e$ for the energy terms and also assume that it has a band diagonal structure.

### 3. PARAMETER ESTIMATION

While the traditional estimation algorithm of MLLR [1] can be extended to the band diagonal transformation we adopted a different approach here. The proposed estimation algorithm uses the usual E-step in MLLR estimation but uses an iterative M-step which optimizes element-wise each row of the transformation matrix. Our argument is that this algorithm is more suited to the sparse structure we use. During estimation zero elements are simply left out. In a more general sense, it does not require system solving and is thus free of matrix inversion problem. It can be considered as an alternative to the existing technique.

Assume we have $R$ regression classes, we want to estimate the set of regression matrices and bias terms given by $\{A_r, b_r : 1 \leq r \leq R\}$. For notational convenience the subscript $r$ will be dropped from the following equations. Maximum likelihood estimation in the framework of the EM algorithm results, after removing irrelevant terms, in the following likelihood function for $A$ and $b$ of each class

$$L \propto \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma(m)(x_t - A\mu_m - b)^T \Sigma^{-1}_m (x_t - A\mu_m - b)$$  \hspace{1cm} (7)$$

where $m$ ranges over Gaussian distributions belonging to the desired regression class, $\mu_m$ and $\Sigma_m$ are their means and covariances, and $\gamma(m) = p(q_t = m|X)$, where $q_t$ is the mixture Gaussian component at time $t$ and $X$ is the set of all observations, this quantity is usually calculated using the forward-backward algorithm. Restricting our attention to the case of diagonal covariance matrices, Eq. (7) reduces to

$$L \propto \sum_{m=1}^{M} \sum_{i=1}^{T} \frac{P(x_{ti} = \frac{\sum_j a_{ij}\mu_{mj} - b_j}{\sigma_{mi}})^2}{\sigma_{mi}}$$  \hspace{1cm} (8)$$

where $a_{ij}$ is the new warped cepstral sequence, $c_n$ is the original cepstral sequence, $N$ is the cepstral order, and $\phi_b(n)$ is a special sequence expressable in terms of Laguerre sequence as discussed in [5]. It should be noted that this sequence, although not explicitly shown, is a function of the warping parameter $\alpha$. Applying the transformation of Eq. (3) to the cepstral means we can write [4]
where $P$ is the feature vector dimension, $x_{tj}$ denotes the $j^{th}$ component of vector $x_t$, and $a_{ij}$, $b_i$, are the elements of $A$ and $b$.

Differentiating (8) with respect to $a_{ik}$ and equating to zero, and after some simplification we arrive at

$$a_{ik} = \frac{\sum_{m=1}^{M} n_m (\mu_{mk} \mu_{mi} - \sum_{j \neq k} a_{ij} \mu_{mj}) - b_i) / \sigma_{mi}^2}{\sum_{m=1}^{M} n_m \mu_{mk} / \sigma_{mi}^2}$$

(9)

Similarly estimation for bias terms can be derived as

$$b_i = \frac{\sum_{m=1}^{M} n_m (\mu_{mi} - \sum_j a_{ij} \mu_{mj}) / \sigma_{mi}^2}{\sum_{m=1}^{M} n_m / \sigma_{mi}^2}$$

(10)

It is interesting to note the similarity of the estimation in equations (9) and (10) to stochastic mathing [14]. The required statistics for performing the above estimation are calculated as

$$n_m = \sum_{t=1}^{T} \gamma_t(m)$$

(11)

and

$$\mu_{mi} = \frac{\sum_{t=1}^{T} \gamma_t(m) x_{ti}}{\sum_{t=1}^{T} \gamma_t(m)}$$

(12)

Thus, the steps of the EM algorithm reduce to calculating the sufficient statistics in Eqs. (11), and (12), and then applying the iterative M-step in Eqs. (9) and (10) for a specified number of iterations. The whole process consists of two embedded iterations, the usual one for collecting the statistics and an internal one for performing iterative maximization. In this work, as done in most adaptation scenarios, we use only one outer iteration, and four maximization iterations are applied based on some preliminary experiments. It is also worth noting that a similar approach was outlined in [7].

4. DISTRIBUTION CLUSTERING

This section briefly outlines the distribution clustering algorithm used in this work, and points out how the number of transformations is adjusted depending on the amount of adaptation data. The basic idea is to organize the distributions in the initial system, e.g. speaker independent system, in a tree structure, the root node contains all the distributions while the leaf nodes carry individual Gaussians. The clustering itself is based on acoustic similarity of the distributions and is a slightly modified version of the algorithm given in [9]. Once the tree is built it is used in the adaptation as follows. The adaptation data is aligned to the corresponding distributions, and thus each node in the tree receives a certain count, which is the sum of counts of its children. Using a threshold on the count creates a cut in the tree and a separate transformation is estimated for each node along the cut using the data of its children. The value of the count threshold controls the number of the estimated transformations, a small threshold shifts the cut down the tree and thus creates many transformations, while increasing the threshold value moves the cut up the tree and leads to decreasing the number of transformations.

The main idea of transformation generation is similar to [10], but the tree construction method is different.

5. EXPERIMENTAL EVALUATION

The proposed algorithm is tested on the non-native speaker part of the Wall Street Journal task, known as Spoke3. The data is collected from 10 non-native speakers, and each speaker provided 40 utterances for adaptation and 40 utterances for test. The SI-84 training set (WSJ0) is used to construct tree clustered triphone HMMs using the algorithm in [11], a 39 dimension feature vector which consists of 12 MFCC, log energy, and their first and second order derivatives is used by the system. A 5K lexicon and the standard trigram model provided by NIST are used in all experiments. The speaker independent word error rate error is 29.1%.

Table 1 shows recognition results for adaptation data varying from 1 to 40 sentences, and for different transform structures, bd-n stands for the transform width about the main diagonal, for example, bd-0 is a diagonal transform, bd-11 is equivalent to the block diagonal transform [3], and the row full stands for conventional MLLR. The threshold value used to create a tree cut in these experiments is 100. From the table we observe that for small amounts of adaptation data (e.g. 1 and 5 utterances) all used structures significantly outperform the full transform, which severely overtrains for 1 utterance leading to worse performance (39.3%) than the baseline. As the amount of training data increases the full MLLR improves but even for 40 utterances some band diagonal structures still perform quite similarly. For example the bd-6 transform has about one twelfth the number of parameters of the full transform, and their recognition performance is almost similar. The diagonal transform is worse than the band diagonal for all sizes of adaptation data. Its performance rather saturates with an increasing amount of data which can be attributed to its simple structure.

We repeated the same set of experiments with a threshold value of 50, effectively increasing the number of transforms. The results given in Table 2 show the same trend as the previous results, but all recognition rates improve. While it is clear that the bd-11 is superior at 40 utterance, it is outperformed by other bands (e.g. bd-4) for one utterance, for other amounts of adaptation data all bands perform quite similarly. Note that the full transform results are not shown in this case.
Table 1: Recognition results for varying amounts of adaptation data and transform structure. Tree cut threshold = 100

<table>
<thead>
<tr>
<th># Utter</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>bd-0</td>
<td>26.7</td>
<td>24.4</td>
<td>23.3</td>
<td>22.0</td>
<td>20.6</td>
</tr>
<tr>
<td>bd-2</td>
<td>25.2</td>
<td>22.3</td>
<td>21.3</td>
<td>19.5</td>
<td>18.3</td>
</tr>
<tr>
<td>bd-4</td>
<td>24.7</td>
<td>21.6</td>
<td>20.8</td>
<td>18.9</td>
<td>17.6</td>
</tr>
<tr>
<td>bd-6</td>
<td>24.7</td>
<td>21.5</td>
<td>20.3</td>
<td>18.8</td>
<td>17.1</td>
</tr>
<tr>
<td>bd-11</td>
<td>24.9</td>
<td>21.4</td>
<td>21.0</td>
<td>18.3</td>
<td>17.0</td>
</tr>
<tr>
<td>Full</td>
<td>39.3</td>
<td>27.8</td>
<td>23.1</td>
<td>19.7</td>
<td>17.9</td>
</tr>
</tbody>
</table>

Table 2: Recognition results for varying amounts of adaptation data and transform structure. Tree cut threshold = 50

<table>
<thead>
<tr>
<th># Utter</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>bd-0</td>
<td>26.1</td>
<td>23.6</td>
<td>22.4</td>
<td>20.8</td>
<td>19.8</td>
</tr>
<tr>
<td>bd-2</td>
<td>24.8</td>
<td>21.4</td>
<td>20.4</td>
<td>18.5</td>
<td>17.4</td>
</tr>
<tr>
<td>bd-4</td>
<td>24.6</td>
<td>21.3</td>
<td>20.3</td>
<td>18.0</td>
<td>16.6</td>
</tr>
<tr>
<td>bd-6</td>
<td>24.8</td>
<td>21.2</td>
<td>19.9</td>
<td>17.9</td>
<td>16.3</td>
</tr>
<tr>
<td>bd-11</td>
<td>25.4</td>
<td>21.1</td>
<td>19.7</td>
<td>17.7</td>
<td>15.5</td>
</tr>
</tbody>
</table>

6. CONCLUSION

We have presented a band diagonal structure to be used in the framework of MLLR adaptation, aiming at constraining the transformation parameters for better estimates, especially for sparse adaptation data. We motivated the use of this structure by an argument based on speaker warping. We developed an iterative maximization technique embedded in the EM estimation algorithm which is very convenient for the proposed sparse transform. Experimental results showed that the proposed structure outperformed the full transform for small amounts of adaptation data, and performed equally well for larger amounts, it also outperformed the diagonal transform for all amounts of adaptation data. No clear conclusion about the best value of the transform band, but the bd-11 transform (equivalent to the so called block diagonal transform) performed slightly better for 40 adaptation utterances and slightly worse for 1 adaptation utterance.

7. REFERENCES


