ABSTRACT

Discriminative mapping transforms (DMTs) is an approach to robustly adding discriminative training to unsupervised linear adaptation transforms. In unsupervised adaptation DMTs are more robust to unreliable transcriptions than directly estimating adaptation transforms in a discriminative fashion. They were previously proposed for use with MLLR transforms with the associated need to explicitly transform the model parameters. In this work the DMT is extended to CMLLR transforms. As these operate in the feature space, it is only necessary to apply a different linear transform at the front-end rather than modifying the model parameters. This is useful for rapidly changing speakers/environments. The performance of DMTs with CMLLR was evaluated on the WSJ 20k task. Experimental results show that DMTs based on constrained linear transforms yield 3% to 6% relative gain over MLE transforms in unsupervised speaker adaptation.

Index Terms— Speaker Adaptation, Discriminative Training, CMLLR, Discriminative Linear Transforms, Minimum Phone Error, Discriminative Mapping Transforms

1. INTRODUCTION

Linear transforms have been used to improve ASR performance by adapting an acoustic model (AM) to a target speaker, often using limited adaptation data with an unreliable transcription, as in the unsupervised speaker adaptation scenario. Both model based (e.g. MLLR [1]) and feature based (e.g. CMLLR [2]) linear transforms based on MLE training have been used successfully in ASR. With the introduction of discriminative training criteria such as minimum phone error (MPE) [3], discriminative linear transforms (DLTs) have also been used [4]. DLTs have been shown to outperform MLE methods in supervised adaptation, but they have two key weaknesses when used in unsupervised online adaptation. First, the training process for DLTs is much more computationally expensive than MLE training. Second, the performance of DLTs is more sensitive to ASR errors in the transcriptions. Therefore, in unsupervised adaptation, MLE based linear transforms are still widely used.

Discriminative mapping transforms (DMT) [5] is a technique proposed to give a better balance between discriminative information and robustness to the ASR errors in unsupervised speaker adaptation. The idea is to factorise the linear transforms into two parts, the MLE transform which is dependent on the adaptation data and is used to capture the characteristics of the current speaker, and the offline discriminatively trained linear transform which captures speaker independent discriminative information. The discriminative mapping transforms can be estimated offline using training data, thus they are independent of the adaptation data and so are not affected by an incorrect hypothesis. Since ASR errors only affect the estimation of the MLE transform, the robustness of the system is maintained while capturing discriminative information offline. Furthermore, the CPU cost for the online part is very low compared to online DLT estimation because the MPE estimation is done offline. DMTs have been proposed based on MLLR [5]. In this work it is extended to support CMLLR transforms. Constrained DMTs share all the advantages of CMLLR, in particular operating in the feature space means that a new transform can be simply applied to rapidly changing environments or speakers. MLLR-based DMT requires every model parameter to be adapted.

DMTs are estimated based on the speaker dependent AM after speaker dependent transforms are applied. For the CMLLR case, the speaker dependent AM is effectively full covariance. This means that constrained DMT is more computationally expensive than MLLR based DMT. Estimation of CMLLR with full covariance matrices has been presented previously and the row-by-row updating method, which is similar to the estimation of standard CMLLR proposed by [6], is adapted for the estimation of constrained DMT.

Speaker adaptive training (SAT) [7] is a widely used method to train AMs based on non-homogeneous data. Discriminative speaker adaptive training (DSAT) [8, 9, 10] was also proposed for the training of discriminative canonical models based on speaker dependent transforms. In this work, DSAT is used to evaluate the performance of the linear transforms.

DMTs and their parameter estimation are presented in Sections 2 and 3, respectively. Experimental results are given in Section 4 followed by conclusions in Section 5.

2. DISCRIMINATIVE MAPPING TRANSFORMS

The aim of discriminative mapping transforms (DMT) [9, 5] is to modify a speaker-specific ML-estimated transform to be more discriminative in nature. Previously they have been used with MLLR mean transforms. The complete transformation of the mean in this case has the form, for component $m$ and speaker $s$:

$$
\mu_{s,m} = A_{ds} \mu_{s,m}^{(m)} + b_{ds}^{(s)} + b_{ds}
$$

where $W_{ds} = [A_{ds} b_{ds}]$ is the discriminatively estimated speaker-independent transform and $W_{ds}^{(s)} = [A_{ds}^{(s)} b_{ds}^{(s)}]$ the speaker-specific ML-estimated transform. As the DMT acts on the speaker-specific transformed mean, it is able to model attributes that standard discriminative training of the model parameters is not able to.

The aim in this paper is to apply the same concept to CMLLR transforms. CMLLR can be interpreted in two different ways. First it may be implemented as a regression class-specific transformation

---

1The number of regression classes for the DMT must be larger than the number used for the speaker-specific transform for this form of transform to yield gains. However for simplicity of notation the dependence on the regression classes has been dropped.
of the features. If this concept was applied to the DMT, following (1) this would yield the following transformation of the features
\[ o_t^{(sm)} = A_{ds} (A_{sl} o_t + b^{(s)}_{sl}) + b_{ds} \] (2)
This can be expressed in terms of transformations of the model parameters as
\[ \mu^{(sm)} = A_{ml}^{-1} A_{sl}^{-1} \left( \mu^{(m)} - b^{(s)}_{ml} \right) - b_{ds} \] (3)
\[ \Sigma^{(sm)} = A_{ml}^{-1} A_{sl}^{-1} \Sigma^{(m)} A_{sl} A_{sl}^T \] (4)
When discriminative training is performed, this form of DMT will yield the same form of model as if a discriminatively estimated semidiagonal covariance matrix was used, with standard CMLLR adaptation. Thus applying this form of transform does not increase the discriminative power of the adaptation transform.

An alternative view of CMLLR is as a constrained transformation of the model parameters. Applying a DMT transformation to this interpretation yields
\[ \mu^{(sm)} = A_{ml}^{-1} A_{sl}^{-1} \left( \mu^{(m)} - b^{(s)}_{ml} \right) - b_{ds} \] (5)
\[ \Sigma^{(sm)} = A_{ml}^{-1} A_{sl}^{-1} \Sigma^{(m)} A_{sl} A_{sl}^T \] (6)
In the same fashion as the MLLR DMT, this yields additional discriminative power as the DMT will attempt to make the transform more discriminative in nature. Implementing the transform directly as above is computationally expensive if full or block-diagonal transforms are used. Thus, for efficiency, the transform can be implemented as a transform of the features where
\[ o_t^{(sm)} = A_{sl} (A_{sl} o_t + b_{sl}) + b^{(s)}_{sl} \] (7)
The component likelihoods are computed as
\[ p(o_t | m, W_{ds}, W_{sl}^{(s)}) = |A_{sl}| A_{ds} X | \mathcal{V}(o_t^{(sm)}; \mu^{(m)}, \Sigma^{(m)}) | \] (8)
as usual requiring no transformation of the acoustic model parameters.

The estimation of the run-time speaker-specific CMLLR transform is the same as the standard CMLLR estimation, but is based on the DMT transformed features, \( A_{ds} o_t + b_{ds} \). Thus at run-time there is no additional cost after transforming the features when compared to standard CMLLR adaptation. The estimation of the DMT transform based on the MPE criterion is discussed in the next section.

### 3. DMT PARAMETER ESTIMATION

To estimate the DMT using MPE the following criterion is used
\[ W_{ds} = \arg \min_W \sum_{s,t} P(\mathcal{H}, W, W_{sl}^{(s)}) \mathcal{L}(\mathcal{H}, \mathcal{H}_{tsd}). \] (9)
where \( O^{(s)} \) is the adaptation data for speaker \( s \) and \( \mathcal{L}(\mathcal{H}, \mathcal{H}_{tsd}) \) is the loss, measured in phones between the reference \( \mathcal{H}_{tsd} \) for speaker \( s \) and the hypothesis \( \mathcal{H} \). The transform \( W_{sl}^{(s)} \) may be obtained using either MLLR or CMLLR.

Strictly, the speaker-specific linear transforms, \( W_{sl}^{(s)} \), are constrained to be the ML-estimate given the DMT transform. However, in practice, to simplify optimisation the speaker-specific MLLR transforms are estimated first, followed by the DMT. This is the approach adopted for MLLR in [5] and will be used here.

Following the same approach as [4], a weak sense auxiliary function is used for MPE training of the DMT. Thus, again ignoring the dependence on the regression class and assuming that the speaker-specific ML transforms are known,
\[ Q_{\text{aux}}(W_{ds}; W_{ds}) = \] (10)
\[ Q_{\text{num}}(W_{ds}; W_{ds}) - Q_{\text{den}}(W_{ds}; W_{ds}) + Q_{\text{aux}}(W_{ds}; W_{ds}) \]
where \( W_{ds} \) and \( W_{ds} \) are the initial and updated transforms respectively and the terms on the RHS of the equation are the numerator, denominator and smoothing functions respectively. The smoothing function is used to ensure the concavity of the weak sense auxiliary function. The next sections describe each of these in more detail.

#### 3.1. Numerator and Denominator Functions

Both \( Q_{\text{num}}() \) and \( Q_{\text{den}}() \) have the same form but they contain the statistics from different phone arcs. For any arc, the choice is determined by the differential of the MPE objective function w.r.t. the arc log likelihood. If this value is positive, the statistics of the arc are accumulated to the numerator. Otherwise, they are accumulated to the denominator. The numerator term can be written as
\[ Q_{\text{num}}(W_{ds}; W_{ds}) = \sum_{s,t} \gamma^{(sm)}(t) |A_{ds}| \log N \left( A_{ds} o_t + b_{ds}; \mu^{(sm)}, \Sigma^{(sm)} \right) \]
where
\[ \mu^{(sm)} = A_{sl}^{-1} \left( \mu^{(m)} - b^{(s)}_{ml} \right) \] (12)
\[ \Sigma^{(sm)} = A_{ml}^{-1} A_{sl}^{-1} \Sigma^{(m)} A_{sl} A_{sl}^T \] (13)
and \( \gamma^{(sm)}(t) \) is the numerator occupancy at time \( t \) for component \( m \) using the current DMT parameters. The denominator term differs in the occupancy counts.

Compared to discriminative linear transform estimation, the statistics for the DMT estimation are more complex as the covariance matrix for adaptation, \( \Sigma^{(sm)}_{a_{li}} \), is full if a full CMLLR transform is used. To address this problem, the approach used for ML-estimation of transforms with full-covariance matrices is used [6]. The function can be mapped to a row-by-row parameter update method, where for row \( i \), ignoring constant terms
\[ Q^{(i)}(\hat{W}; \hat{w}) = \beta \log \left( c_i \hat{w}_i^T \right) - \frac{1}{2} \hat{w}_i^T G^{(i)} \hat{w}_i + \hat{w}_i k^{(i)} \] (14)
where \( \hat{w}_i \) is the \( i \)th row vector of matrix \( \hat{W} \) and \( c_i \) is the \( i \)th extended cofactor of \( \hat{W} \). The statistics \( \beta, G^{(i)} \) and \( k^{(i)} \) are defined as follows.
\[ \beta = \sum_{s,m,t} \gamma^{(sm)}(t), \quad G^{(i)} = G_{ii}, \] (15)
\[ k^{(i)} = k_i - \sum_{j \neq i} \hat{w}_j G_{ij}, \quad P^{(sm)}_{a_{li}} = \Sigma^{(sm)-1}_{a_{li}} \]
where
\[ G_{ij} = \sum_{s,t} \zeta_i \zeta_j^T \sum_{m} p_{a_{li}}^{(sm)} \gamma^{(sm)}(t) \]
\[ k_i = \sum_{s,m} p_{a_{li}}^{(sm)} \mu_{a_{li}}^{(sm)} \gamma^{(sm)}(t) \zeta_i^T \] (16)
In equation 16 \( \mu_{a_{li}}^{(sm)} \) and \( p_{a_{li}}^{(sm)} \) are the \( i \)th row and the \((i,j)^{th}\) element of \( P^{(sm)}_{a_{li}} \) respectively.
3.2. Smoothing Function

A smoothing function is used to ensure the concavity of the weak sense auxiliary function. Any function which satisfies the following requirement yields a valid smoothing function.

\[
\frac{\partial Q_{\text{sm}}(\mathbf{W}_{\text{dn}}; \mathbf{W}_{\text{dn}})}{\partial \mathbf{W}_{\text{dn}}} \bigg|_{\mathbf{W}_{\text{dn}} = \hat{\mathbf{W}}_{\text{dn}}} = 0 \tag{17}
\]

The standard approach, modified to reflect the DMT estimation, is to define the smoothing function as

\[
Q_{\text{sm}}(\mathbf{W}_{\text{dn}}; \mathbf{W}_{\text{dn}}) = \sum_{a,m} D_{m}^{(a)} \int p(o|m, \mathbf{W}_{\text{dn}}, \mathbf{W}_{\text{dn}}(s)) \log p(o|m, \mathbf{W}_{\text{dn}}, \mathbf{W}_{\text{dn}}(s)) \, d\alpha
\]

and

\[
D_{m}^{(a)} = E \sum_{t} \gamma_{\text{den}}^{(s,m)}(t) \tag{19}
\]

where \( E \) is a constant to control the convergence speed.

The computational cost of smoothing parameters based on equation 18 is high since the speaker dependent transforms have to be applied to every Gaussian in the HMMs for every speaker in the training data set. To reduce this cost, an approximation is made that the smoothing parameters are not dependent on the speaker dependent transforms. Based on this approximation, equation 18 can be simplified to

\[
Q_{\text{sm}}(\mathbf{W}_{\text{dn}}; \mathbf{W}_{\text{dn}}) = \sum_{a,m} D_{m}^{(a)} \int p(o|m, \mathbf{W}_{\text{dn}}, \mathbf{W}_{\text{dn}}(s)) \log p(o|m, \mathbf{W}_{\text{dn}}, \mathbf{W}_{\text{dn}}(s)) \, d\alpha
\]

Equation 20 satisfies the smoothing function requirement in equation 17, meanwhile all the operations of applying the speaker dependent transforms to Gaussians which are related to equation 18 are avoided. In equation 20, the “average” speaker transforms could also be used.

The smoothing function based on equation 20 can be re-written using the same form as equation 14, and the statistics are defined as follows:

\[
\beta_{\text{sm}}^{(i)} = \sum_{a,m} D_{m}^{(a)} m_{m} \mathbf{Y}_{m}^{(m)} \mathbf{Y}_{m}^{(m)T} \tag{21}
\]

\[

\mathbf{Z}_{m}^{(m)} = \left[ \mathbf{Y}_{m}^{(m)} \right]^{1/2} \left[ \mathbf{Z}_{m}^{(m)} \right]^{1/2} \tag{22}
\]

where \( \mathbf{Y}_{m}^{(m)} \) and \( \mathbf{Z}_{m}^{(m)} \) are as follows:

\[
\mathbf{Y}_{m}^{(m)} = \left[ \frac{\mathbf{Y}_{m}^{(m)} + \mu_{m}^{(m)} \mu_{m}^{(m)T}}{2} \right], \quad \mathbf{Z}_{m}^{(m)} = \left[ \frac{\mu_{m}^{(m)}}{1} \right]
\]

Equation 20 can be seen as equivalent to sampling the canonical AM with the initial transforms applied. Alternatively, sufficient statistics from which the initial transforms are generated, i.e. sufficient statistics from the previous iteration, can be used to smooth the current iteration.

3.3. Parameter Updates

Since all 3 terms of the MPE weak sense auxiliary function, i.e. the numerator function, denominator function and smoothing function, have the same form as equation 14, based on equation 10 the sufficient statistics of constrained DMT training can be expressed as

\[
\beta_{\text{sm}} = \beta_{\text{an}} - \beta_{\text{den}} + \beta_{\text{an}}
\]

\[
\mathbf{G}_{\text{sm}}^{(i)} = \mathbf{G}_{\text{sm}}^{(i)} - \mathbf{G}_{\text{den}}^{(i)} + \mathbf{G}_{\text{an}}^{(i)}
\]

\[
\mathbf{k}_{\text{sm}}^{(i)} = \mathbf{k}_{\text{sm}}^{(i)} - \mathbf{k}_{\text{den}}^{(i)} + \mathbf{k}_{\text{an}}^{(i)}
\]

I-smoothing based on MLE statistics can also be added to equation 23 [4]. Differentiating equation 14 w.r.t. \( \mathbf{w} \), using the redefined sufficient statistics from equation 23, and setting equal to zero, \( \mathbf{w} \) can be computed as

\[
\mathbf{w}_{i} = (\alpha c_{i} + k_{\text{sm}}^{(i)}) G_{\text{sm}}^{(i)-1} \tag{24}
\]

\( \alpha \) is a root of a quadratic equation as described in [2].

Thus the estimation of the DMT transform has the same form as full covariance CMLLR transform estimation. Though this process is iterative, as rows are inter-dependent due to both the cofactors and the full-covariance form of \( \mathbf{K}_{\text{sm}}^{(i)} \) and \( \mathbf{K}_{\text{den}}^{(i)} \), the discriminative optimisation is only required at training time. As previously discussed, at run-time only the ML-estimate of the CMLLR transform is required. This ML estimate uses diagonal covariance matrices and is estimated in the DMT-transformed space.

4. EXPERIMENTAL RESULTS

In this work, the DMT with constrained linear transforms was evaluated based on the WSJ 20k vocabulary system. The training data consists of 194 hours of speech data which includes WSJ0 and WSJ1 training data. There are 1822 speakers in the complete training set. A 39 dimensional speech feature vector was used, comprising 12 MFCCs, log energy and their first-order and second-order derivatives. Cepstral mean normalisation (CMN) and semi-tied covariance matrices (STC) were also applied to the feature vectors. The HMMs were state-clustered triphone with 4.3K states and 14 Gaussians per state. A 20K closed vocabulary was used, with a standard trigram language model trained on the WSJ87-89 text corpus with about 40M words. The test set contains 215 utterances from 10 speakers. For all MPE trained systems, E was chosen between 0.5 and 1. I-smoothing was also applied.

Both standard acoustic models and CMLLR-SAT ML acoustic models were trained. In addition, discriminative training was applied to the standard acoustic models. Three forms of DSAT model were trained based on the ML-SAT system. They differ in the form of CMLLR-style transform used. First, the baseline system used standard CMLLR transforms estimated on the final ML acoustic models (CMLLR). Second, using the CMLLR transforms a constrained DMT transform was trained (CMLLR+DMT). Finally, a discriminative linear transform (DLT) was estimated. In each case, MPE training was performed in the space defined by the corresponding transforms.

The training process for the constrained DMT is similar to the process for MLLR mean DMT training described in [9]. The constrained DMTs were estimated using equation 9 given the current

\[\text{in addition to the standard WSJ training data, clean data from Speecon, TIMIT data, and internally collected data was included for consistency with internal system development.}\]
training speaker specific CMLLR transform. Ideally, DMT and the training speaker CMLLR transform should be updated iteratively, but due to CPU time limitations the CMLLR transforms were not updated. To maintain consistency, the CMLLR transforms for the test speakers were estimated with feature vectors that had not been transformed by DMT. For consistency, DLT and CMLLR transforms were also not updated. All CMLLR and DLT transforms used 2 regression classes, one for speech and the other for silence. 64 regression classes were used for the DMT. The number of classes here is significantly less than that used in [9]. Though more transforms could be used, this increases the recognition cost as it increases the number of transforms that must be applied to the feature vector.

The MPE criterion of the training data is shown in Table 1 for DSAT iterations 0 and 6 with different linear transforms. Iteration 0 represents the MPE criterion based on the MLE canonical models. Constrained DMT yields a better MPE criterion than CMLLR. This is a good indication that DMT does indeed add discriminative information to the MLE based CMLLR. DLT gives the best MPE criterion. Since the DLT optimises the MPE criterion directly instead of acting on top of CMLLR, this is not surprising. After 6 iterations, the MPE criterion for all cases are very similar.

<table>
<thead>
<tr>
<th># iter</th>
<th>CMLLR</th>
<th>CMLLR+DMT</th>
<th>DLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.9518</td>
<td>0.9548</td>
<td>0.9568</td>
</tr>
<tr>
<td>6</td>
<td>0.9838</td>
<td>0.9838</td>
<td>0.9847</td>
</tr>
</tbody>
</table>

The performance on the WSJ20K task for the system at iteration 0, using the MLE SAT canonical AM, is shown in Table 2. Unsupervised adaptation for CMLLR, CMLLR+DMT and DLT was carried out using supervision hypotheses from the MLE SI AM. Constrained DMT achieved about 6% relative gain over CMLLR in the framework of MLE SAT. Smaller gains are seen with the DLT.

<table>
<thead>
<tr>
<th>model</th>
<th>transforms</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE SI</td>
<td>-</td>
<td>6.47</td>
</tr>
<tr>
<td>MLE SI</td>
<td>CMLLR</td>
<td>6.00</td>
</tr>
<tr>
<td>MLE SAT</td>
<td>CMLLR</td>
<td>5.61</td>
</tr>
<tr>
<td>MLE SAT</td>
<td>CMLLR+DMT</td>
<td>5.27</td>
</tr>
<tr>
<td>MLE SAT</td>
<td>DLT</td>
<td>5.45</td>
</tr>
</tbody>
</table>

Similarly, unsupervised adaptation was carried out using DSAT for the three different types of linear transforms. The supervision hypotheses were generated using the MPE SI AM and 6 iterations of DSAT updates were performed. Results are shown in Table 3. The DSAT systems show significant gains over the ML-SAT systems. Despite the fact that DLT yields the best MPE criterion, it achieves slightly worse performance than CMLLR. This could be due to DLT being sensitive to supervision errors. In contrast, the constrained DMT (CMLLR+DMT) system achieves 3% relative gain over the CMLLR based DSAT.

The relative gain from the DSAT constrained DMT system is less than that achieved with the MLE SAT DMT system. This is because the DSAT canonical AM has been trained with MPE so there is less room for DMT to improve the discriminative power of the AM.

<table>
<thead>
<tr>
<th>model</th>
<th>transforms</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPE SI</td>
<td>-</td>
<td>4.88</td>
</tr>
<tr>
<td>MPE SI</td>
<td>CMLLR</td>
<td>4.60</td>
</tr>
<tr>
<td>DSAT</td>
<td>CMLLR</td>
<td>4.28</td>
</tr>
<tr>
<td>DSAT</td>
<td>CMLLR+DMT</td>
<td>4.15</td>
</tr>
<tr>
<td>DSAT</td>
<td>DLT</td>
<td>4.49</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

In this work, the DMT technique was extended to support constrained linear transforms. This allows robust unsupervised “discriminative like” speaker adaptation transforms to be estimated. Though the training of the DMT is computationally expensive, by using appropriate approximations, it is practical even for large training sets. The constrained DMT can be simply integrated into the standard discriminative SAT framework for updating the canonical models. Experimental results based on WSJ 20k task show that constrained DMTs yield 3% to 6% relative gain over MLE transforms in unsupervised speaker adaptation.

6. REFERENCES