LEARNING DISCRIMINATIVE BASIS COEFFICIENTS FOR EIGENSPACE MLLR UNSUPERVISED ADAPTATION

Yajie Miao, Florian Metze, Alex Waibel

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
{ymiao, fmetze, ahw}@cs.cmu.edu

ABSTRACT

Eigenspace MLLR is effective for fast adaptation when the amount of adaptation data is limited, e.g., less than 5s. The general motivation is to represent the MLLR transform as a linear combination of basis matrices. In this paper, we present a framework to estimate a speaker-independent discriminative transform over the combination coefficients. This *discriminative basis coefficients transform* (DBCT) is learned by optimizing discriminative criteria over all the training speakers. During recognition, the ML basis coefficients for each testing speaker are firstly found, on which DBCT is applied to give the final MLLR transform discrimination ability. Experiments show that DBCT results in consistent WER reduction in unsupervised adaptation, compared with both standard ML and discriminatively trained transforms.

Index Terms— Speaker adaptation, discriminative training, speech recognition

1. INTRODUCTION

Speaker adaptation is widely used to build speaker-dependent models which can recognize speech from unknown speakers. The most commonly used approach for speaker adaptation is maximum likelihood linear regression (MLLR), which involves estimation of speaker-specific linear transforms over acoustic model parameters [1]. MLLR can perform robustly given limited adaptation data. However, when the amount of adaptation data becomes really small, say less than 5s, speaker adaptation based on MLLR does not always lead to improved recognition performance. This is because the estimation of MLLR transforms is too noisy and does not generalize well to the testing data. To solve this problem, eigenspace-based methods have been proposed [2, 3, 4]. Generally, there are two stages in this type of methods. During the training stage, an appropriate set of basis matrices are computed on the training data, using ML-fashion [4] or PCA-like algorithms [2, 3]. During testing, the adaptation transform of a specific speaker is represented as a combination of the basis matrices. Since the number of free parameters, i.e., combination coefficients, is reduced greatly, these methods can improve the robustness of MLLR.

Meanwhile, there has been considerable interest in exploiting discriminative criteria for improving MLLR adaptation. In the supervised mode, it has been shown that discriminative linear transforms (DLT) can bring significant improvement over ML transforms [5, 6]. However, the gains of DLT drop dramatically in unsupervised adaptation because DLT is very sensitive to supervision errors. A more recent work is to learn a global discriminative mapping transform (DMT) on the training data, which can map ML-estimated adaptation transforms to discriminative transforms [7, 8]. Since only ML estimation is performed during adaptation, this DMT method is found to be less sensitive to hypothesis errors and thus more suitable for unsupervised speaker adaptation.

In this paper, we combine these two lines of work and propose a framework which estimates a discriminative linear transform over the basis coefficients in eigenspace MLLR adaptation. This global speaker-independent transform, referred to as DBCT, acts as a linear mapping function from ML-estimated basis coefficients to discriminative ones. During training, this DBCT is learned by optimizing discriminative criterion on the training speakers. During recognition, DBCT is used to transform speaker-specific basis coefficients estimated in a normal ML manner. The final adaptation matrix derived from DBCT-transformed coefficients becomes discriminative in nature and at the same time robust to hypothesis errors. We use the maximum mutual information (MMI) criterion [9] for DBCT training and only examine MLLR adaptation of HMM-GMM means. However, this approach can be extended easily to the minimum phone error (MPE) criterion [9] and other forms of adaptation transforms such as fMLLR. Experiments with Switchboard data show the effectiveness of DBCT in improving unsupervised adaptation.

2. EIGENSPACE MLLR ADAPTATION

The idea of eigenspace MLLR adaptation is to estimate MLLR transforms in a subspace constrained by the basis matrices. These basis matrices are learned on the training set

either with PCA or iterative MLE. Specifically, the MLLR transform $\mathbf{W}^{(s)}$ for speaker *s* is represented as

$$\mathbf{W}^{(s)} = \sum_{n=1}^{N(s)} d_n^{(s)} \mathbf{W}_n \quad (1)$$

where $\mathbf{W}_n$ represents the *n-th* basis matrix. Speaker-specific basis coefficients $d_n^{(s)}$ are estimated to optimize the ML objective on the adaptation data. If the feature dimension is *D*, the number of basis coefficients $N^{(s)}$ is generally much smaller than *D*(*D*+1), i.e., the size of parameters in the conventional MLLR adaptation. Therefore, this type of eigenspace methods perform robustly under inadequate adaptation data.

In principle, our DBCT approach can be used with any eigenspace MLLR methods. In this paper, we deal with a specific implementation derived from the basis representation of fMLLR described in [10]. The basis matrices are obtained via singular value decomposition (SVD) on top of MLLR statistics collected from the training speakers, together with appropriate precon-ditioning. Totally we estimate *D*(*D*+1) basis matrices which are sorted by a decreasing order on their eigenvalues. For each testing speaker, an iterative line search algorithm is adopted to find the basis coefficients (equivalently the MLLR transform) which optimize the ML objective on the adaptation data. Compared with others, this implementation has the advantage that the number of basis matrices to be used, i.e., $N^{(s)}$, can be decided dynamically according to the amount of available adaptation data. For example, $N^{(s)}$ is set to the minimum of *D*(*D*+1) and ¦ $^{(s)}$, where $^{(s)}$ is the number of adaptation speech frames for this speaker and ¦ is a constant such as 0.2 [10]. In this paper, we call this method *basis-MLLR*. Interested readers can refer to [10] and the implementation of basis-fMLLR in the Kaldi toolkit[1].

3. DISCRIMINATIVE BASIS COEFFICIENTS TRANSFORM

This section formally describes how to learn the discriminative DBCT over the basis coefficients. After using basis-MLLR on the training set, we can get the ML basis coefficients for each training speaker *s*. Since basis matrices have been sorted by their importance [10], the first *P* coefficients represent the most important ones, i.e.,

$$\boldsymbol{\lambda}_1^{(s)} = [d_1^{(s)}, \cdots, d_P^{(s)}, 1] \quad (2)$$

which has been extended with an additional 1. The remaining coefficients form the second vector

$$\boldsymbol{\lambda}_2^{(s)} = [d_{P+1}^{(s)}, \cdots, d_{N(s)}^{(s)}] \quad (3)$$

[1] http://sourceforge.net/projects/kaldi/

Applying DBCT, denoted as $\mathbf{W}_{\text{dbct}}$, to $\boldsymbol{\lambda}_1^{(s)}$, the transformed coefficients vector is

$$\boldsymbol{\lambda}_{1,\text{dbct}}^{(s)} = \mathbf{W}_{\text{dbct}} \cdot \boldsymbol{\lambda}_1^{(s)} \quad (4)$$

where $\mathbf{W}_{\text{dbct}}$ has the size of *P*(*P*+1). Considering DBCT on a small subset of *P* coefficients has two notable benefits. First, in testing adaptation, the actual coefficient dimensions to be used may be much less than *D*(*D*+1). Thus, the DBCT transform modeled on the most important dimensions can still be applicable even under highly limited adaptation data. Second, if using the whole set of coefficients, the affine DBCT has the size of [*D*(*D*+1)+1][*D*(*D*+1)], which is expensive to manipulate.

To facilitate the estimation of $\mathbf{W}_{\text{dbct}}$, we isolate basis matrices from coefficients by integrating basis matrices into GMM mean vectors. For Gaussian component *m*, we have the *D*×*P* matrix $\mathbf{M}_1^{(m)} = [\mathbf{m}_1^{(m)}, \cdots, \mathbf{m}_P^{(m)}]$ and the *D*×(*N(s)* – *P*) matrix $\mathbf{M}_2^{(m)} = [\mathbf{m}_{P+1}^{(m)} \cdots, \mathbf{m}_{N(s)}^{(m)}]$, where the *i-th* column vector $\mathbf{m}_i^{(m)}$ represents the mean vector $\boldsymbol{\mu}^{(m)}$ transformed by the basis matrix $\mathbf{W}_i$, i.e.,

$$\mathbf{m}_i^{(m)} = \mathbf{W}_i \cdot \boldsymbol{\xi}^{(m)} \quad (5)$$

where $\boldsymbol{\xi}^{(m)}$ is the extended vector of $\boldsymbol{\mu}^{(m)}$. Then, it's natural to derive the speaker-specific means $\boldsymbol{\mu}_s^{(m)}$, with the DBCT $\mathbf{W}_{\text{dct}}$ applied, as follows:

$$\boldsymbol{\mu}_s^{(m)} = \mathbf{M}_1^{(m)} \cdot \mathbf{W}_{\text{dbct}} \boldsymbol{\lambda}_1^{(s)} + \mathbf{M}_2^{(m)} \cdot \boldsymbol{\lambda}_2^{(s)} \quad (6)$$

Our goal is to obtain the speaker-independent $\mathbf{W}_{\text{dbct}}$ which can optimize the discriminative criterion on the training data. The standard MMI optimization scheme, based on the weak-sense auxiliary function [9], is used. The auxiliary function w.r.t. $\mathbf{W}_{\text{dbct}}$ is formulated as:

$$\begin{aligned} Q(\mathbf{W}_{\text{dbct}}, \hat{\mathbf{W}}_{\text{dbct}}) = & \sum_{s,t} \sum_m ¥_m^{num}(t_s) \cdot (\mathbf{x}_{t_s} - \boldsymbol{\mu}_s^{(m)})^T \boldsymbol{\Sigma}_m^{-1} (\mathbf{x}_{t_s} - \boldsymbol{\mu}_s^{(m)}) \\ & - \sum_{s,t} \sum_m ¥_m^{den}(t_s) \cdot (\mathbf{x}_{t_s} - \boldsymbol{\mu}_s^{(m)})^T \boldsymbol{\Sigma}_m^{-1} (\mathbf{x}_{t_s} - \boldsymbol{\mu}_s^{(m)}) \\ & + \sum_s \sum_m D_s^{(m)} \cdot (\hat{\boldsymbol{\mu}}_s^{(m)} - \boldsymbol{\mu}_s^{(m)})^T \boldsymbol{\Sigma}_m^{-1} (\hat{\boldsymbol{\mu}}_s^{(m)} - \boldsymbol{\mu}_s^{(m)}) \end{aligned} \quad (7)$$

where $¥_m^{num}(t_s)$ and $¥_m^{den}(t_s)$ are posterior occupancy of component *m* being at time t_s given the numerator and denominator lattices, t_s is the speech frame of speaker *s*, $D_s^{(m)}$ is a smoothing term with respect to speaker *s* and component *m* to ensure the convergence of the discriminative updates. Following [8], we set this term to be $D_s^{(m)} = E \sum_{t_s} ¥_m^{den}(t_s)$ where the constant *E* = 0.8. Also, $\hat{\boldsymbol{\mu}}_s^{(m)}$ is the adapted mean which is calculated using Eq. (6) and the current $\hat{\mathbf{W}}_{\text{dbct}}$, rather than the mean adapted by basis-MLLR.

The above DBCT can be estimated efficiently with an expectation maximization (EM) style algorithm. For limit of

space, we are omitting the detailed derivations. In the E step, the following two types of speaker-specific statistics are collected for all the training speakers:

$$\mathbf{G}^{(s)} = \sum_m \gamma_s^{(m)} \cdot \mathbf{M}_1^{(m)T} \cdot \boldsymbol{\Sigma}_m^{-1} \cdot \mathbf{M}_1^{(m)}$$
$$\mathbf{K}^{(s)} = \sum_m \sum_{t_s} \gamma_m(t_s) \cdot \mathbf{M}_1^{(m)T} \cdot \boldsymbol{\Sigma}_m^{-1} \cdot \mathbf{x}_{t_s}^{(m)} \cdot \boldsymbol{\lambda}_1^{(s)T} + \sum_m D_s^{(m)} \cdot \mathbf{M}_1^{(m)T} \cdot \boldsymbol{\Sigma}_m^{-1} \cdot \tilde{\boldsymbol{\mu}}_s^{(m)} \cdot \boldsymbol{\lambda}_1^{(s)T} \quad (8)$$

where $\gamma_m(t_s) = \gamma_m^{num}(t_s) - \gamma_m^{den}(t_s)$, the accumulated component occupancy is

$$\gamma_s^{(m)} = D_s^{(m)} + \sum_{t_s} \gamma_m(t_s) \quad (9)$$

and the two component specific terms in $\mathbf{K}^{(s)}$ can be calculated as

$$\mathbf{x}_{t_s}^{(m)} = \mathbf{x}_{t_s} - \mathbf{M}_2^{(m)} \cdot \boldsymbol{\lambda}_2^{(s)}$$
$$\tilde{\boldsymbol{\mu}}_s^{(m)} = \hat{\boldsymbol{\mu}}_s^{(m)} - \mathbf{M}_2^{(m)} \cdot \boldsymbol{\lambda}_2^{(s)} \quad (10)$$

In the M step, it can be proved that DBCT estimation has the following updating formula to optimize the discriminative auxiliary function in Eq. (7):

$$\mathrm{vec}\left(\mathbf{W}_{\mathrm{dbct}}\right) = \left(\sum_s kron(\mathbf{G}^{(s)}, \mathbf{P}^{(s)})\right)^{-1} \mathrm{vec}\left(\sum_s \mathbf{K}^{(s)}\right) \quad (11)$$

where the vec(.) operator stacks the rows of a matrix into a single vector, kron(.) is the Kronecker product of two matrices, $\mathbf{P}^{(s)}$ is the scatter of the first-part coefficient vector defined in Eq. (2), that is, $\mathbf{P}^{(s)} = \boldsymbol{\lambda}_1^{(s)} \cdot \boldsymbol{\lambda}_1^{(s)T}$. Given a well trained acoustic model, for example discriminatively trained HMM-GMM, iterative estimation procedures for DBCT are summarized as follows:

(1) Estimate the ML basis coefficients $\mathbf{d}_{ml}^{(s)}$ for each training speakers *s* with basis-MLLR.
(2) Initialize $\mathbf{W}_{\mathrm{dbct}}^{(0)} = [\mathbf{I}; \mathbf{0}]$ and set $k = 0$.
(3) Collect statistics using Eq. (8) and estimate $\mathbf{W}_{\mathrm{dbct}}^{(k+1)}$ according to Eq. (11).
(4) $k = k + 1$. Go to step 3 if not converged.

After obtaining the DBCT, we can use it in recognition. For a testing speaker *s*, discriminative adaptation with DBCT is performed as follows:

(1) Perform first-pass decoding to generate the supervision hypothesis for the utterances of speaker *s*.
(2) Estimate ML basis coefficients $\mathbf{d}_{ml}^{(s)}$ with basis-MLLR.
(3) Adapt the acoustic model parameters using $\mathbf{W}_{\mathrm{dbct}}$ and the ML coefficients $\mathbf{d}_{ml}^{(s)}$ according to Eq. (6).
(4) Decode the testing set with the adapted model.

From the training process, we can see that DBCT depends on specific acoustic models. If the acoustic model changes, DBCT needs to be re-estimated.

4. EXPERIMENTS

The performance of DBCT is evaluated on an English conversational telephone speech task. The training data contains 898 speakers (conversation sides), around 72 hours, from the Switchboard-1 corpus. The testing data comes from a subset of the 2001 HUB5 evaluation set, consisting of 20 speakers and 1 hour of speech. Acoustic modeling is based on a 13-dimensional MFCC front-end including the C0 energy and its first, second derivatives with per-speaker mean normalization. An LDA transform reduces the feature dimension to 40, on which MLLT is applied. The ML model has 3000 clustered triphone states, with an average of 12 Gaussians per state. The MMI criterion [9] is used on top of the ML baseline and generates the speaker-independent MMI model. In MMI training, the numerator lattices are built from the reference, while the denominator lattices are produced by the ML model with a heavily pruned unigram language model. Our experiments are conducted with this MMI-SI model, which has a first-pass WER of 38.5% on the testing set. During unsupervised adaptation, basis-MLLR estimation is performed given the hypothesis output from MMI-SI. For all the decoding runs, we use a trigram language model built only with the training transcriptions.

4.1. Effectiveness of DBCT

As discussed in Section 3, DBCT is applied on a small subset of the basis coefficients. Therefore, it can be robustly estimated with limited training data. To verify this, we perform DBCT learning on various training sets with different sizes. Table 1 presents the results for the adapted MMI model, using standard basis-MLLR and basis-MLLR +DBCT respectively. The coefficient dimension *P*, on which DBCT is applied, is set to 10. On each training set, we run DBCT estimation for 4 iterations and give the final MMI objective. Note that we are not reporting the actual MMI objective as computed in [9]. Instead, the objective here equals the parts in Eq. (7) dependent on $\mathbf{W}_{\mathrm{dbct}}$.

We can see that using DBCT in addition to basis-MLLR adaptation yields further reduction on WER. Reducing the amount of training data results in superior recognition performance and MMI objectives. The best WER is achieved on the 9-hour set, where DBCT brings 0.7% absolute improvement to basis-MLLR. With more training data available, we observe decreased MMI objectives, which indicate that DBCT may not reach the optimal point. Thus, for the 72-hour set, we run DBCT estimation for more iterations and achieve the best WER in the 6th iteration. Despite a larger MMI objective, the recognition performance is only 0.1% better than using 9 hours data. This confirms that we are able to learn DBCT only with a small training set. If we shrink the training set further to 3 hours, the performance of DBCT degrades significantly.

Table 1. Performance (WER%) of basis-MLLR with DBCT in unsupervised adaptation.

	WER	MMI Obj
Basis-MLLR	37.1	---
+ DBCT 72Hrs	36.6	1.415
+ DBCT 36Hrs	36.6	1.463
+ DBCT 18Hrs	36.8	1.706
+ DBCT 9Hrs	**36.4**	2.033
+ DBCT 3Hrs	37.2	1.959
+ DBCT 72Hrs (6th iteration)	36.3	2.129

4.2. Sensitivity to Supervision Errors

The following two subsections examine properties of DBCT. Unless stated otherwise, we show the results of DBCT trained with 9 hours data. As a global transform, DBCT should be less sensitive to supervision errors compared with DLT [5, 6]. To investigate this point, three types of adaptation supervisions are used. The baseline hypothesis are from MMI-SI in the first-pass decoding. Hypothesis of worse quality are generated by the ML model. Finally, the correct reference is also taken as supervision. These supervisions are used to estimate basis-MLLR coefficients, on which DBCT is applied. During DLT training, the numerator lattices are built from these supervisions and the denominator lattices are generated by MMI-SI.

Table 2 presents the WER comparison with various supervisions. For basis-MLLR, using the reference obtains 1.4% absolute gains over the ML-SI and MMI-SI supervisions. This is similar to DBCT performance differences. In contrast, for DLT, the reference has 3.9% absolute improvement over ML-SI and 3.5% over MMI-SI supervisions. This shows that DBCT is less sensitive to the quality of supervisions and thus suitable for unsupervised adaptation. It can also be observed that DBCT always outperforms DLT under erroneous supervisions. But with reference supervision, DLT is significantly better than DBCT. This is because DBCT is learned on the training set and is not tuned to the reference during adaptation.

4.3. Training Stability

Another notable observation is that DBCT may encounter instability in discriminative updates. In Fig. 1, we plot the MMI objectives for 10 iterations of DBCT estimation and the corresponding WER. With E=0.8, there is a dramatic rise in the MMI objective after 5 iterations, while the performance of DBCT on the testing set begins to drop. This instability can be relieved by setting the const E to a larger value, which corresponds to a smaller step size in each iteration. Fig. 1 also shows the MMI objectives and WER with E=1.2. In this case, DBCT estimation remains stable within the 10 iterations. But we need more iterations to reach the optimal recognition results. Moreover, we observe that increasing the coefficient dimension P to 20 or 30 introduces more instability into DBCT training. That is, the testing WER goes up quickly to 50% after several iterations. That's why we set P to 10 in our experiments.

Table 2. WER% of DBCT and DLT with different supervisions.

	Supervision		
	ML-SI	MMI-SI	Ref
WER	41.3	38.5	----
Basis-MLLR	37.1	37.1	35.7
+DBCT	36.5	36.4	35.2
DLT	37.1	36.7	33.2

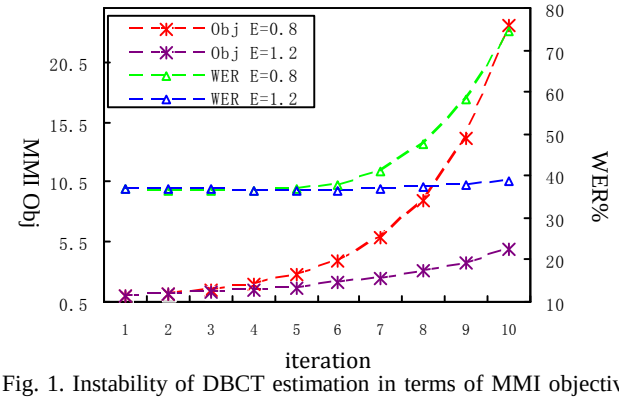

Fig. 1. Instability of DBCT estimation in terms of MMI objective and WER% in 10 iterations.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose an approach to estimating DBCT for eigenspace MLLR adaptation. The DBCT transform is learned on top of basis-MLLR and applied in recognition to improve the ML adaptation with additional discrimination. Experiments show the effectiveness of DBCT in improving unsupervised adaptation. In our future work, we will focus on the extension of this method to cluster adaptive training (CAT) [11], where we can learn the discriminative transforms on the cluster combination coefficients.

6. ACKNOWLEDGMENTS

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD / ARL) contract number W911NF-12-C-0015. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

7. REFERENCES

[1] M. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[2] K. T. Chen, W. W. Liau, H. M. Wang, and L. S. Lee, “Fast speaker adaptation using eigenspace-based maximum likelihood linear regression,” in *Proc. ICSLP*, pp. 742–745, 2000.

[3] N. Wang, S. Lee, F. Seide, and L. S. Lee, “Rapid speaker adaptation using a priori knowledge by eigenspace analysis of MLLR parameters,” in *Proc. ICASSP*, pp. 345–348, 2001.

[4] K. Visweswariah, V. Goel, and R. Gopinath, “Maximum Likelihood training of bases for rapid adaptation,” unpublished manuscript, 2002.

[5] L. Wang and P. C. Woodland, “Discriminative adaptive training using the MPE criterion,” in *Proc. ASRU*, pp. 279–284, 2003.

[6] L. Wang and P. C. Woodland, “MPE-based discriminative linear transforms for speaker adaptation,” *Computer Speech and Language*, vol. 22, pp. 256-272, 2008.

[7] K. Yu, M. Gales, and P. C. Woodland, “Unsupervised discriminative adaptation using discriminative mapping transforms,” in *Proc. ICASSP*, pp. 4273–4276, 2008.

[8] K. Yu, M. Gales, and P. C. Woodland, "Unsupervised adaptation with discriminative mapping transforms,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 714-723, 2009.

[9] D. Povey, “Discriminative training for large vocabulary speech recognition,” Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., 2003.

[10] D. Povey and K. Yao, “A basis representation of constrained MLLR transforms for robust adaptation,” *Computer Speech and Language*, vol. 26, pp. 35-51, 2012.

[11] M. Gales, "Cluster adaptive training of Hidden Markov Models,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 8, no. 4, pp. 417-428, 2000.