UNSUPERVISED DISCRIMINATIVE ADAPTATION USING DISCRIMINATIVE MAPPING TRANSFORMS

K. Yu, M.J.F. Gales and P.C. Woodland

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK
Email: {ky219, mjfg, pcw}@eng.cam.ac.uk

ABSTRACT

The most commonly used approaches to speaker adaptation are based on linear transforms, as these can be robustly estimated using limited adaptation data. Although significant gains can be obtained using discriminative criteria for training acoustic models, maximum likelihood (ML) estimated transforms are used for unsupervised adaptation. This is because discriminatively trained transforms are highly sensitive to errors in the adaptation hypothesis. This paper describes a new framework for estimating transforms that are discriminative in nature, but are less sensitive to this hypothesis issue. A discriminative, speaker-independent, mapping transformation is estimated during training. This transform is obtained after a speaker-specific ML-estimated transform has been applied. During recognition an ML speaker-specific transform is found and the speaker-independent discriminative mapping transform then applied. This allows a transform which is discriminative in nature to be indirectly estimated, whilst only requiring an ML speaker-specific transform to be found during recognition. The scheme is evaluated on an English conversational telephone speech task, where it significantly outperforms both standard ML and discriminatively trained transforms.

Index Terms— Speaker adaptation, discriminative training.

1. INTRODUCTION

Linear transformations of the acoustic model parameters are the most commonly used approaches for speaker adaptation when there is limited training data [1, 2]. If unsupervised adaptation is required, for example in broadcast news transcription or conversational telephone speech, these transforms are usually found using Maximum Likelihood (ML) estimation. Though discriminative criteria are commonly used for training acoustic models [3], performance gains for speaker adaptation in an unsupervised mode have been limited [4, 5]. This is because discriminative criteria are more sensitive to errors in the hypotheses (or references) than the ML criterion. The sensitivity to the hypothesis may be reduced using, for example, confidence scores [6] or lattice-based approaches for improved hypotheses [7] but the gains are still small compared to ML estimated transforms. Thus despite gains in supervised adaptation [8], unsupervised discriminative adaptation is not commonly used.

A number of approaches have been proposed for combining ML-estimated transforms with discriminatively trained models. For example simplified discriminative speaker adaptive training (SAT) [9, 4], discriminative cluster adaptive training [10], and feature MPE (fMPE) [11] or region-dependent feature transforms [12] have all been successfully used in speech recognition. A general attribute of all these schemes is that all speaker-specific parameters of the system are estimated in an ML-fashion, whereas speaker-independent aspects of the system may be trained using discriminative criteria. This paper applies the same general approach to training discriminative linear transforms, whilst using ML to estimate all speaker-specific parameters.

The procedure adopted in this work is to use a speaker-independent mapping transform from one form of training criterion to another. This will be referred to as a *criterion mapping function* (CMF). The specific form examined in this work is to map a speaker-specific ML-estimated linear transform to be more similar to a Minimum Phone Error (MPE) discriminatively trained transform. A linear transform will be used, referred to as a *discriminative mapping transform* (DMT). In theory this approach can be applied to any form of linear transform, either mean, covariance or features. Here Maximum Likelihood Linear Regression (MLLR) adaptation of the means will be examined.

This paper is organised as follows. In section 2, linear transforms are reviewed, and how they are used in combination with discriminative training discussed. The new framework and estimation of DMT is detailed in section 3. Experiments on an English conversational telephone speech (CTS) task are described in section 4 followed by conclusions and future work.

2. LINEAR TRANSFORMS FOR ADAPTATION

Linear transformations are the most commonly used approaches to speaker adaptation with limited training data. Linear transform based speaker adaptation was initially investigated with ML estimation. For mean MLLR adaptation [1], the transformed mean for speaker s, $\hat{\boldsymbol{\mu}}^{(s)}$, can be expressed as

$$\hat{\boldsymbol{\mu}}^{(s)} = \mathbf{A}_{\mathtt{ml}}^{(s)}\boldsymbol{\mu} + \mathbf{b}_{\mathtt{ml}}^{(s)} = \mathbf{W}_{\mathtt{ml}}^{(s)}\boldsymbol{\xi} \tag{1}$$

$\boldsymbol{\xi} = [\boldsymbol{\mu}^T \;\; 1]^T$ is the extended mean vector and $\mathbf{W}_{\mathtt{ml}}^{(s)} = [\mathbf{A}_{\mathtt{ml}}^{(s)} \;\; \mathbf{b}_{\mathtt{ml}}^{(s)}]$ is the extended linear transform of speaker s. The parameters of the transform, $\mathbf{W}_{\mathtt{ml}}^{(s)}$ are estimated using the ML criterion [1]

$$\mathbf{W}_{\mathtt{ml}}^{(s)} = \arg\max_{\mathbf{W}} \left\{ p(\mathbf{O}^{(s)}|\mathcal{H}^{(s)}, \mathbf{W}; \boldsymbol{\lambda}) \right\} \tag{2}$$

where $\mathbf{O}^{(s)}$ and $\mathcal{H}^{(s)}$ are the observations and reference/hypothesis of the adaptation data for speaker s respectively, and $\boldsymbol{\lambda}$ are the model parameters. An important issue is where the hypothesis, $\mathcal{H}^{(s)}$, is obtained from. If it is known a-priori, this is supervised adaptation. If it must be found using an acoustic model, possibly with an current estimate of the transform, this is unsupervised adaptation.

This work was supported in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.Thanks to Lan Wang for the code of standard DLT estimation.

Rather than using the ML criterion it is possible to use a discriminative training criterion, such as MPE. Transforms estimated using these discriminative criteria are referred to as discriminative linear transforms (DLTs). Here the form of adaptation remains the same,

$$\hat{\boldsymbol{\mu}}^{(s)} = \mathbf{A}_{\mathtt{dl}}^{(s)}\boldsymbol{\mu} + \mathbf{b}_{\mathtt{dl}}^{(s)} = \mathbf{W}_{\mathtt{dl}}^{(s)}\boldsymbol{\xi} \tag{3}$$

where $\mathbf{W}_{\mathtt{dl}}^{(s)} = [\mathbf{A}_{\mathtt{dl}}^{(s)} \ \mathbf{b}_{\mathtt{dl}}^{(s)}]$ is the DLT of speaker s. However DLTs are estimated using, for example, the MPE criterion which can be expressed as

$$\mathbf{W}_{\mathtt{dl}}^{(s)} = \arg\max_{\mathbf{W}} \left\{ \sum_{\mathcal{H}} P(\mathcal{H}|\mathbf{O}^{(s)}, \mathbf{W}; \boldsymbol{\lambda}) \mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)}) \right\} \tag{4}$$

where $P(\mathcal{H}|\mathbf{O}^{(s)}, \mathbf{W})$ is the posterior probability of hypothesis $\mathcal{H}$ given the observation from speaker s and the transform parameters, $\mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)})$ is the loss function of $\mathcal{H}$ given the supervision $\mathcal{H}^{(s)}$ measured at the phone level. DLTs have been successfully used in supervised adaptation [8] where $\mathcal{H}^{(s)}$ is known for the adaptation data. However gains have been limited using this form of transform in an unsupervised fashion. Discriminative criteria are far more sensitive to the reference used than ML-based schemes. Improved performance can be obtained using, for example, lattice adaptation but gains over ML-trained transforms have been disappointing.

ML-trained transforms for unsupervised adaptation are often used in combination with discriminatively trained model parameters, $\boldsymbol{\lambda}$. In simplified discriminative speaker adaptive training (SAT) [9, 4], the canonical HMMs are discriminatively updated given the ML estimated speaker transforms. During adaptation, ML transforms are estimated for each speaker and applied to the discriminative SAT model. Discriminative cluster adaptive training (CAT) [10] follows a similar procedure but uses multiple-cluster models as the canonical model. ML-estimated interpolation weights are estimated during training and recognition. In both discriminative SAT or CAT, the discriminative criterion is only used for the model parameters, not for the speaker-specific transform parameters. Discriminatively trained feature transforms such as Feature MPE (fMPE) [11] and region-dependent feature transforms (RDFT) [12] have also been used in combination with ML-estimated speaker transforms. In these approaches, the acoustic space is partitioned into regions, region-dependent matrices are discriminatively trained and used to transform the features. These discriminative transforms may be built on top of a speaker-specific ML-adapted feature-space. All these schemes adopt the same general strategy. Speaker independent parameters can be discriminatively estimated. All speaker-specific parameters, the transforms, are ML trained.

3. DISCRIMINATIVE MAPPING TRANSFORMS

Criterion mapping functions (CMFs) use the same general approach described above, but introduce a speaker independent transformation of the speaker-specific transforms. This CMF aims to map, for example, ML-trained transforms into discriminative transforms. As there is no discriminative estimation of speaker-specific transformations, used in the DLT, the sensitivity to the hypotheses should be reduced. Here the speaker-specific DLT is found using

$$\mathbf{W}_{\mathtt{dl}}^{(s)} = \mathcal{F}(\mathbf{W}_{\mathtt{ml}}^{(s)}; \boldsymbol{\Lambda}) \tag{5}$$

where $\mathbf{W}_{\mathtt{ml}}^{(s)}$ is the speaker-dependent ML transform found using equation (2), $\mathcal{F}(\cdot)$ is the mapping function with speaker-independent parameters, $\boldsymbol{\Lambda}$. As the parameters of the CMF are speaker independent they may be trained on all the training data. This has two advantages in training. First there is a large amount of training data to estimate the mapping function. Second the references are known for the training data, so there are no hypotheses sensitivity issues. An additional advantage is that during recognition only an ML-estimated transform is required to be estimated. This avoids the need to generate lattices, for example, which is required for directly estimating DLTs. The rest of this section describes a specific implementation of the the CMF based on linear transforms.

A simple form of the CMF is to use linear transformations of the ML transform parameters $\mathbf{W}_{\mathtt{ml}}^{(s)}$ to obtain the discriminative transformation. This is referred to as a discriminative mapping transform (DMT). One form of transformation is

$$\text{vec}(\mathbf{W}_{\mathtt{dl}}^{(s)}) = \mathbf{H}_{\mathtt{dl}}\text{vec}(\mathbf{W}_{\mathtt{ml}}^{(s)}) + \mathbf{c}_{\mathtt{dl}} \tag{6}$$

where vec() maps the matrix to a vector form, $\mathbf{H}_{\mathtt{dl}}$ is an $n(n+1) \times n(n+1)$ matrix and $\mathbf{c}_{\mathtt{dl}}$ is a $n(n+1)$ vector (for an n-dimensional feature vector). In this initial investigation a simpler form of transformation is used. $\mathbf{H}_{\mathtt{dl}}$ is restricted to be block-diagonal in structure. The transformation can then be expressed as

$$\mathbf{W}_{\mathtt{dl}}^{(s)} = \mathbf{A}_{\mathtt{dl}}\mathbf{W}_{\mathtt{ml}}^{(s)} + \boldsymbol{\beta}_{\mathtt{dl}} \tag{7}$$

where $\mathbf{A}_{\mathtt{dl}}$ and $\boldsymbol{\beta}_{\mathtt{dl}}$ are now the speaker-independent DMT parameters. For mean adaptation, this yields the following transformation

$$\hat{\boldsymbol{\mu}}^{(s)} = \left(\mathbf{A}_{\mathtt{dl}}\mathbf{W}_{\mathtt{ml}}^{(s)} + \boldsymbol{\beta}_{\mathtt{dl}}\right)\boldsymbol{\xi} = \mathbf{A}_{\mathtt{dl}}\hat{\boldsymbol{\mu}}_{\mathtt{ml}}^{(s)} + \mathbf{B}_{\mathtt{dl}}\boldsymbol{\mu} + \mathbf{b}_{\mathtt{dl}} \tag{8}$$

where $\boldsymbol{\beta}_{\mathtt{dl}} = [\mathbf{B}_{\mathtt{dl}} \ \mathbf{b}_{\mathtt{dl}}]$, $\mathbf{B}_{\mathtt{dl}}$ is a $n \times n$ matrix and $\hat{\boldsymbol{\mu}}_{\mathtt{ml}}^{(s)} = \mathbf{W}_{\mathtt{ml}}^{(s)}\boldsymbol{\xi}$. If the DMT is further restricted so that $\mathbf{B}_{\mathtt{dl}} = \mathbf{0}$, this leads to

$$\hat{\boldsymbol{\mu}}^{(s)} = \mathbf{A}_{\mathtt{dl}}\hat{\boldsymbol{\mu}}_{\mathtt{ml}}^{(s)} + \mathbf{b}_{\mathtt{dl}} = \mathbf{W}_{\mathtt{dl}}\boldsymbol{\xi}_{\mathtt{ml}}^{(s)} \tag{9}$$

where $\boldsymbol{\xi}_{\mathtt{ml}}^{(s)} = [\hat{\boldsymbol{\mu}}_{\mathtt{ml}}^{(s)T} \ 1]^T$. The advantage of this form of simplification is that the speaker-independent linear transform parameters, $\mathbf{W}_{\mathtt{dl}}$, can be estimated in a similar fashion to the standard DLTs in equation (3). The training criterion can be expressed as

$$\mathbf{W}_{\mathtt{dl}} = \arg\max_{\mathbf{W}} \left\{ \sum_{s} \sum_{\mathcal{H}} P(\mathcal{H}|\mathbf{O}^{(s)}, \mathbf{W}; \boldsymbol{\lambda}_{\mathtt{ml}}^{(s)}) \mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)}) \right\} \tag{10}$$

where $\boldsymbol{\lambda}_{\mathtt{ml}}^{(s)}$ is the ML-transform adapted model parameters for speaker s. Thus rather than accumulating statistics using the original HMM (as in the DLT), the DMT estimation uses speaker-specific ML-adapted HMM parameters and sums over all training speakers. The update formulae of DMT are similar to the standard DLT ones, which can be found in [6]. This is the form of DMT investigated in this paper.

The presentation of the DMT has so far only considered a single transformation. Given the simplifications from the more powerful transform in equation (6), it would be useful to have multiple DMT linear transforms, in the same fashion as having multiple MLLR transforms [13]. The same approach to clustering Gaussians together to form multiple base-classes, clustering in acoustic space or based on phonetic characteristics, can be used. Note, as the DMT transform estimation uses all the available training data, the number of transform classes may be made larger than is usually used for standard speaker adaptation.

Though mean adaptation is considered in this paper, the DMT can also be applied to variance or constrained MLLR (CMLLR)

adaptation [14, 2]. When using DMTs with CMLLR, it becomes a speaker-independent discriminative feature mapping. It is interesting to contrast this DMT transformation with fMPE or RDFT. As discussed in section 2, fMPE and RDFT both use a speaker-independent discriminatively trained transform on top of the speaker-dependent CMLLR adapted features. This is similar to the idea of DMT. However, fMPE and RDFT both use posteriors of the adapted features and directly estimate the discriminative transforms. In contrast, DMT trains a mapping from a ML feature-transformation to a discriminative feature space and is dependent on the component being transformed.

4. EXPERIMENTS

The performance of discriminative mapping transforms was evaluated on a large vocabulary English conversational telephone speech (CTS) task. The training dataset consisted of 5446 speakers, about 296 hours of data. The sources were the LDC Call-home English (`che`), Switchboard (`Swbd`) and Switchboard-Cellular (`SwCell`) datasets. The test set was the `eval03` dataset, consisting of 144 speakers, about 6 hours. All systems used a 12-dimensional PLP front-end with the `C0` energy and its first, second and third derivatives with side-level Cepstral mean and variance normalisation. An HLDA transform was applied to reduce the feature dimension to 39. VTLN was also used. State-clustered triphone HMMs with 6K distinct states and an average of 16 Gaussian components per state were used. All adaptation was carried out in an unsupervised mode (there was no supervised adaptation data available).

Minimum phone error (MPE) [3] was used to train all the acoustic models. Two MPE systems were built. One was a speaker-independent (SI) MPE system. The second was a mean-MLLR based MPE speaker adaptive training (SAT) system [15]. Here an ML-based MLLR SAT system was built and then only the canonical model was updated using the MPE criterion given the ML transforms. During unsupervised adaptation, 4 iterations of MLLR estimation was first performed given the hypotheses from the MPE SI system. As a contrast standard DLTs were also estimated. Here MLLR was initially applied and used to generate the 1-Best hypothesis as the numerator for DLT estimation. The ML-SI model and a heavily pruned bi-gram language model were used to generate denominator lattices. For all experiments separate speech and silence transforms were used for MLLR and DLT.

4.1. Effectiveness of DMT

As the DMT performs a mapping of the ML-transform into the MPE space it is interesting to see how effective this mapping is in terms of increasing the criterion value on the test data using the estimated ML transform. These experiments used the MPE-SI models. For the DMT 1000 regression classes were used. The MPE criterion[1] values of the test data (the 1-Best hypothesis generated from MLLR adapted MPE-SI model was used as numerator) are shown in table 1 for the standard MLLR transform as well as when combined with the DMT,

The table shows the change in criterion when using the DMT estimated using 1, 2, or 3 training iterations. From table 1, DMT improved the MPE criterion values compared to the MLLR adapted model. Increasing the training iteration gave higher values. This shows that the discrimination power of the DMT generalises to the test data and when the ML-transform is estimated on error-full hypotheses. As a contrast it is also possible to compare these MPE values with those obtained when estimating a DLT directly. After 1-iteration of test-set DLT estimation, the criterion is 0.855, and after 2 iterations, 0.889. These higher criterion values are expected as the DLT is able to tune to the test hypotheses more than the DMT. However, this tuning means that after more than one DLT iteration the performance degrades. In these results only the WERs of one iteration of DLT estimation are reported.

[1]This is one minus the normalised form of equation (4).

Adaptation	Training Iteration		
	1	2	3
MLLR	0.793		
+DMT	0.800	0.802	0.803

Table 1. MPE Criterion values of MLLR and MLLR+DMT

4.2. Number of Base-Classes

As discussed in section 3, to improve the power of DMTs, large number of transforms, base-classes, may be used. Three sizes of base-class were examined, 2, 46 and 1000. The 46 base-classes were estimated using either acoustic clustering or phone information. The results are shown in table 2. The first line in the table gives the baseline MLLR performance.

Gaussian Clustering	# Class	Train Iteration		
		1	2	3
—-		27.0		
Acoustic	2	27.0	—	—
	46	26.9	—	—
	1000	26.7	26.4	26.2
Phone	46	26.8	26.7	26.7

Table 2. WER% using DMT with different base-class sizes

From table 2 increasing the number of base-classes improves performance. For the 2 base-class system there is no gain over the baseline MLLR system. Both the 46-class phone and acoustic clustered systems show slight gains after 1 DMT training iteration. The best performance was obtained using the 1000 base-classes. Performance with this system also improved with the number of DMT iterations, in-line with the MPE-criterion gains in table 1. Using three training iterations and 1000 base-classes a 0.8% absolute reduction in WER was obtained over the MLLR baseline.

An interesting contrast is to see whether the DMT is really learning a criterion mapping rather than a transformation of a MLLR-transform to more phone-specific transforms. An ML-to-ML mapping transform was estimated using the 46 phone base classes. This increased the test-set ML criterion but decreased the MPE criterion compared to MLLR, the opposite is true for the ML-to-MPE DMT. The ML-to-ML mapping degraded the MLLR performance by 0.1%.

4.3. Effect of Hypothesis Quality

One of the motivations for the use of the DMT was that it should be less sensitive to errors in the adaptation data hypotheses. To investigate this effect in detail, three adaptation supervisions were used to estimate the transforms. The baseline hypotheses used to date were generated by the unadapted MPE-SI model. As an alternative, this adapted model was then used to generate lattices which were used in a lattice MLLR adaptation framework [7]. As alternative hypotheses in the lattice are used, this form of estimation is less sensitive to hypothesis errors. Finally, the correct reference was used. These

three forms of transform were used to generate MLLR transforms, to which DMTs could then be applied. For the DLT the numerator was generated using the MLLR or lattice MLLR adapted MPE-SI model. However, for the reference hypotheses case, this was used directly as the numerator for the DLT.

Adaptation	Supervision		
	1-Best Hyp.	Lattice Hyp.	Reference
MLLR	27.0	26.7	24.3
+ DMT	26.2	25.9	23.4
DLT	26.8	26.6	21.7

Table 3. WER% using different supervision hypotheses

Table 3 gives the WER comparison using these different supervision hypotheses. For MLLR, using the reference obtained 2.7% absolute gain over the 1-Best hypothesis and 2.4% over the lattice supervision. This is similar to DMT performance differences. In contrast, for DLT, the reference gained 5.1% over 1-Best and 4.9% over lattice supervision, which are far larger than MLLR with and without DMT. This confirms that the DMT is less sensitive to the quality of supervision and suitable for unsupervised adaptation. It is also interesting to note that with error-full hypotheses, either 1-Best or lattice, DMT always significantly outperformed DLT and MLLR. But with reference supervision, DLT was significantly better than DMT. This is expected because DMT is estimated on the training data set and is not tuned to the reference as heavily as DLT.

4.4. DMT on adaptively trained system

The previous experiments were based on the MPE-SI model. Using DMTs with MPE-SAT models was also investigated. The comparison between different adaptation approaches on MPE-SI and MPE-SAT models are shown in table 4 using a 1000 base-class DMT obtained with 3 training iterations.

Adaptation	MPE-SI	MPE-SAT
MLLR	27.0	26.4
+ DMT	26.2	25.6
DLT	26.8	26.3

Table 4. WER% using DMT with MPE-SI and MPE-SAT models

From table 4, MLLR with and without DMT, and the DLT on the MPE-SAT system both significantly outperformed the corresponding MPE-SI systems. The gains of using the DMT with MLLR over the baseline MLLR system and DLT were retained for the MPE-SAT system. Using MLLR with DMT gave a 0.8% absolute reduction in WER over the standard MLLR system and 0.7% absolute over the DLT system. For these experiments the DMT was only used during test, not during the SAT training stage. This will be investigated in future work.

5. CONCLUSION AND FUTURE WORK

This paper has described a new framework for robust discriminative unsupervised adaptation. In this framework, a speaker-independent criterion mapping function (CMF) is estimated during training and used to map the ML estimated speaker-dependent transforms to a more discriminative form. As only ML-adapted speaker-specific transforms are estimated on the adaptation data, the transform is not highly sensitive to the adaptation hypotheses, which is a major issue with standard discriminative estimation of linear transforms. A simple initial implementation of the CMF based on linear transforms is described. This is referred to as a discriminative mapping transform (DMT). The approach is applied to MLLR adaptation in this paper. Experiments on a CTS English task illustrated that DMT can significantly outperform standard DLT for both discriminatively trained SI and SAT models in unsupervised adaptation.

This paper has only described initial experiments on CMFs. A number of alternative transforms and applications, such as during adaptive training and more complex transforms, will be investigated in future work.

6. REFERENCES

[1] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

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

[3] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002, Orlando.

[4] L. Wang and P. C. Woodland, "Discriminative adaptive training using the MPE criterion," in *Proc. ASRU*, 2003.

[5] S. Tsakalidis, V. Doumpiotis, and W. Byrne, "Discriminative linear transforms for feature normalisation and speaker adaptation in HMM estimation," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 13, no. 3, pp. 367–376, 2005.

[6] L. Wang, *Discriminative linear transforms for adaptation and adaptive training*, Ph.D. thesis, Cambridge University, 2006.

[7] L. F. Uebel and P. C. Woodland, "Speaker adaptation using lattice-based MLLR," *Proc. ISCA ITR-Workshop on Adaptation Methods for Speech Recognition*, 2001.

[8] L. F. Uebel and P. C. Woodland, "Discriminative linear transforms for speaker adaptation," *Proc. ISCA ITR-Workshop on Adaptation Methods for Speech Recognition*, 2001.

[9] A. Ljolje, "The AT&T LVCSR-2001 system," in *Proc. the NIST LVCSR Workshop*, NIST, 2001.

[10] K. Yu and M. J. F. Gales, "Discriminative cluster adaptive training," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 5, pp. 1694–1703, 2006.

[11] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig, "The IBM 2004 conversational telephony system for rich transcription," in *Proc. ICASSP*, 2005.

[12] B. Zhang, S. Matsoukas, and R. Schwartz, "Recent progress on the discriminative region-dependent transform for speech feature extraction," in *Proc. INTERSPEECH*, 2006.

[13] C. J. Leggetter, *Improved acoustic modelling for HMMs using linear transformations*, Ph.D. thesis, Cambridge University, 1995.

[14] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 2, pp. 357–366, 1995.

[15] K. Yu, *Adaptive training for large vocabulary continuous speech recognition*, Ph.D. thesis, Cambridge University, 2006.