iVector-Based Discriminative Adaptation for Automatic Speech Recognition

Martin Karafiát [#1], Lukáš Burget [*#2], Pavel Matějka [#3], Ondřej Glembek [#4], Jan "Honza" Černocký [#5],

[#] *Brno University of Technology, Speech@FIT, Božetěchova 2, Brno, 612 66, Czech Republic*
[1] karafiat@fit.vutbr.cz, [3] matejkap@fit.vutbr.cz, [4] glembek@fit.vutbr.cz [5] cernocky@fit.vutbr.cz

[*] *SRI International, 333 Ravenswood Avenue, Menlo Park, 94025, CA, USA*
[2] burget@speech.sri.com

***Abstract*—We presented a novel technique for discriminative feature-level adaptation of automatic speech recognition system. The concept of iVectors popular in Speaker Recognition is used to extract information about speaker or acoustic environment from speech segment. iVector is a low-dimensional fixed-length representing such information. To utilized iVectors for adaptation, Region Dependent Linear Transforms (RDLT) are discriminatively trained using MPE criterion on large amount of annotated data to extract the relevant information from iVectors and to compensate speech feature. The approach was tested on standard CTS data. We found it to be complementary to common adaptation techniques. On a well tuned RDLT system with standard CMLLR adaptation we reached 0.8% additive absolute WER improvement.**

I. Introduction

We propose new method for discriminative adaptation of automatic speech recognition (ASR) system, which is based on combination of two successful techniques: From speaker recognition field, we have borrowed the idea of representing speech segment using so called iVector. iVector is an information-rich low-dimensional fixed length vector extracted from the feature sequence. Recently, systems based on iVectors [1], [2], [3] extracted from cepstral features have provided excellent performance in speaker verification, which classifies iVectors as good candidates for representing information about speaker. Just like MLLR transformations for ASR adaptation became popular features in speaker recognition [4], we believe that iVectors — successful in speaker recognition — can be used as compact representations for ASR adaptation. For brevity, we will describe the proposed method only from the perspective of speaker adaptation. Keep in mind, however, that iVector represents information about both speaker and acoustic environment of the corresponding segment and therefore, the proposed technique is expected to effectively adapt ASR system to both speaker and acoustic environment.

In order to utilize information encoded in iVectors for adaptation of speech recognition system, we build on the idea of Region Dependent Linear Transform (RDLT) [5]. In the original version, RDLT is a nonlinear feature transformation, which is typically discriminatively trained using Minimum Phone Error (MPE) criterion [6]. More precisely, each feature vector is transformed by a linear transformation, which is selected from an ensemble of transformations depending on the acoustic region of the current frame. To apply this framework for discriminatively trained feature-level adaptation, we use the same form of frame-dependent transformation. However, the *fixed* iVector is transformed by such varying transformation and the resulting vector is added as a bias to the original feature vector.

The paper is organized as follows: the following section II presents the state-of-the-art in discriminative techniques for speaker adaptation and positions our proposal. Section III briefly introduces iVectors while IV defines the RDLT scheme and recipes. Section V suggests the the iVector adaptation. The following section VI describes the experimental setup including the baseline systems and VII presents the results with RDLT systems including the proposed iVector adaptation. Section VIII contains the conclusions and directions for future work.

II. Current techniques for discriminative adaptation and position of our proposal

The idea of using discriminative training criterion for adaptation is not new. In the early works on this topic [7], [8], [9], acoustic model parameters or features were adapted using transformations of the same form as in Maximum Likelihood Linear Regression (MLLR) or Constrained MLLR (CMLLR), where the adaptation transformations were estimated on adaptation data by optimizing discriminative rather than Maximum Likelihood (ML) criterion. While this approach provided excellent performance for supervised adaptation, it appeared to be too sensitive to the quality of the initial hypothesis in the case of unsupervised adaptation. Fortunately, our technique does not suffer form such problem as only the RDLT part is trained using MPE criterion on large amount of annotated training data. RDLT discriminatively adapts speech features based on the information encoded in the iVector. The iVectors estimated on adaptation data are, however, robustly obtained by optimizing Maximum a-posteriori (MAP) criterion. Moreover, there is no need for any initial hypothesis as iVectors are estimated using simple Gaussian Mixture Model (GMM).

Our technique is similar in spirit to Discriminative Mapping transforms (DMT) [10], [11], where MLLR or CMLLR transformations are estimated on adaptation data using ML

criterion first. The adapted model parameters are further post-processed by an ensemble of discriminatively trained linear transformations (typically 64), where each transformation corresponds to a cluster of Gaussian components from the acoustic model. The transformations are discriminatively trained on large amount of annotated training data to refine the adapted models and to compensate for the discriminative power that could be taken away from discriminatively acoustic trained models when adapted using ML estimated transformations.

DMT can be seen as some form of region dependent transforms, where the regions in acoustic space are defined by the Gaussian clusters rather than by a dedicated GMM as it is in the case of RDLT. From this perspective, CMLLR-based DMT [11] is very similar to standard RDLT jointly trained with the following CMLLR adaptation as described in [5]. Therefore, it can be expected that, just like RDLT, DMT would bring improvements even without ML trained adaptation transformations. Unfortunately, the papers on DMT do not provide such analysis and it is not clear how much improvement is to be attributed to improved adaptation and how much to the improved discriminative acoustic model training.

In our approach, however, we do not estimate any feature or model transformations to adapt the acoustic model to the adaptation data. Instead, we estimate iVector summarizing information about the speaker and the acoustic environment of adaptation data independently of any ASR acoustic model. Also, the discriminatively trained transformation does no directly operate on speech features or model parameters. Instead, for each speech frame, it is trained to extract a correction bias vector from iVector. In our implementation, zero iVector, which is the expected value of iVector on training data, leads to zero correction bias and therefore to no adaptation. Therefore, it is easy to separately analyze the effect of RDLT used for adaptation and RDLT used, in the standard way, as a discriminative feature transformation.

III. iVECTORS

The iVector approach has become state of the art in the speaker verification field [1]. In this work, we show that it can be successfully applied to extract information useful for adapting ASR system. The approach provides an elegant way of reducing large-dimensional sequential input data to a low-dimensional fixed length feature vector while retaining most of the relevant information.

In the iVector framework, a GMM model is adapted to observation sequence representing a speech segment that we want to extract speaker information from. Only the mean parameters of a pre-trained GMM are adapted. The supervector of concatenated mean vectors for the adapted GMM is obtained as

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{i}, \tag{1}$$

where $\mathbf{m}$ is the segment-independent component of the mean supervector, $\mathbf{T}$ is a matrix of basis spanning the subspace covering the important variability (both useful and useless for adaptation) in the supervector space, and $\mathbf{i}$ is a low-dimensional latent variable representing coordinates in the subspace. We assume standard normal prior for the latent variable $\mathbf{i}$. GMM is adapted to the observation sequence by finding $\mathbf{i}$ that maximizes MAP criterion. This MAP point estimate of $\mathbf{i}$, which is obtained with single iteration of EM algorithm, is taken as the iVector representing the segment. The parameters of the GMM and the subspace are trained in unsupervised manner using EM algorithm on a collection of speech segment covering variety of speakers and acoustic environments. We use an efficient implementation of the training procedure suggested in [12].

IV. REGION DEPENDENT LINEAR TRANSFORMS

In the RDLT framework, an ensemble of linear transformations is trained discriminatively. Each transformation corresponds to one region in partitioned feature space. Each feature vector is then transformed by a linear transformation corresponding to the region that the vector belongs to. The resulting (generally nonlinear) transformation has the following form:

$$F_{RDLT}(\mathbf{o}_t) = \sum_{r=1}^{N} \gamma_r(t)(\mathbf{A}_r\mathbf{o}_t + \mathbf{b}_r), \tag{2}$$

where o_t is input feature vector at time t, $\mathbf{A}_r$ and $\mathbf{b}_r$ are linear transformation and biases corresponding rth region and $\gamma_r(t)$ is probability that the vector $\mathbf{o}_t$ belongs to rth region. The probabilities $\gamma_r(t)$ are typically obtained using GMM (pre-trained on the input features) as mixture component posterior probabilities. Usually, RDLT parameters $\mathbf{A}_r$, $\mathbf{b}_r$ and ASR acoustic model parameters are alternately updated in several iterations. While RDLT parameters are updated using MPE criterion, ML update is typically used for acoustic model parameters. As proposed in [13] and described in RDLT context in [5], ML update of acoustic model parameters must be taken into account when optimizing RDLT parameters. Otherwise, the discriminative power obtained from MPE training of RDLT feature transformation is mostly lost after ML acoustic model re-training. In our experiments, we closely follow the training recipe described in [5].

In our experiments, we do not use the bias terms $\mathbf{b}_r$ (the number of their parameters would anyway be only a small proportion of parameters in matrices $\mathbf{A}_r$). In agreement with results reported in [5], we have found that omitting the bias terms has little effect on the performance.

RDLT can be seen as a generalization of previously proposed fMPE discriminative feature transformation. The special case of RDLT with square matrices $\mathbf{A}_r$ (i.e. without dimensionality reduction of input features) was shown [5] to be equivalent to fMPE with offset features as described in [14]. This is also the configuration used in our experiments. From fMPE recipe [13], we have also take the idea of incorporating context information by considering $\gamma_r(t)$ corresponding not only to the current frame but also to the neighboring frames (see section VII-A for more details). From our experience, this style of incorporation context information leads to significantly

better results compared to to the style previously considered in the context of RDLT [5], where feature vectors of multiple frames were stacked at the RDLT input and transformations with dimensionality reduction were used to recover the original feature dimensionality. Therefore, our RDLT baseline system configuration is very similar to the one described in the fMPE recipe. Still, we prefer to use the more general RDLT abstraction as it can be easily extended by the proposed iVector based adaptation.

V. iVector based adaptation

To utilize the RDLT framework for adaptation, we use transformation of the following form:

$$F_{ivec}(\mathbf{o}_t) = \mathbf{o}_t + \sum_{r=1}^{N} \gamma_r(t)\mathbf{A}_r\mathbf{i}_s, \quad (3)$$

where $\mathbf{i}_s$ is iVector estimated on adaptation data corresponding to speaker s. Typically, iVector dimensionality is larger than the dimensionality of feature vector, therefore $\mathbf{A}_r$ are matrices reducing the dimensionality of iVector to the one of feature vectors. The same MPE training framework as described in the previous section can be used to train RDLT to discriminatively extract the corrective term from iVector $\mathbf{i}_s$, which is added to the original feature vector $\mathbf{o}_t$ in order to adapt the features to the model. Note that, although the iVector stays constant, its transformation depends or region of current feature frame so that different pieces of information can be extracted from iVector to compensate feature frames from different regions of acoustic space.

We again use the iterative training scheme where, after updating RDLT parameters, acoustic model parameters are retrained on the compensated features. The resulting procedure can be seen as another form of speaker adaptive training (SAT) [15], [16].

Finally, we can combine both ideas of using RDLT for adaptation and discriminative feature transformation. Since the whole RDLT framework has to be implemented to deal with either of the two problems, it makes a little sense to use RDLT only for adaptation without using it also for feature transformation, which is expected to provide an additional significant gain. If the same data and the same region definitions are used to train RDLT for both problems, which is the case in our experiments, we can simply concatenate each feature vector with the appropriate iVector and process the resulting extended vectors

$$\tilde{\mathbf{o}}_t = \begin{bmatrix} \mathbf{o}_t \\ \mathbf{i}_s \end{bmatrix} \quad (4)$$

just as in the standard RDLT framework corresponding to equation (2). $\mathbf{A}_r$ will perform dimensionality reduction.

VI. Experimental setup

A. ASR training and testing data

The acoustic model was trained on ctstrain04 training set, a subset of the h5train03 set, defined at the University of Cambridge. It contains about 278 hours of well transcribed speech data from Switchboard I,II and Call Home English (see Table I).

Database	Amount of data [hours]
Switchboard I	248.52
Switchboard cellular	15.27
Call Home English	13.93
Total	277.72

TABLE I
CTS training data description.

Models	WER [%]
ML	34.7
ML - CMLLR	32.1
ML - CMLLR-SAT	31.9

TABLE II
Baseline: ML trained systems

All recognition results are reported on the Hub5 Eval01 test set (defined during 2001 NIST CTS evaluation) composed of 3 subsets of 20 conversations from Switchboard-1, Switchboard-2 and Switchboard-cellular corpora, for a total length of more than 6 hours of audio data.

A bigram language model was used for recognition. It was adopted from AMI speech recognition system for NIST Rich Transcriptions 2007 [17].

B. Baseline ASR systems

The speech recognition system is HMM-based cross-word tied-states triphones, with approximately 8500 tied states and 28 Gaussian mixtures per state. The features were 13 VTLN normalized Mel-Frequency PLP coefficients generated by HTK, augmented with their deltas, double-deltas and triple-deltas. Cepstral mean and variance normalization was applied with the mean and variance vectors estimated on each conversation side. HLDA was estimated with Gaussian components as classes and the dimensionality was reduced from 52 to 39. This model is denoted as ML in table II

Using this model, CMLLR adaptation transforms were generated for training and test data, one for each conversation side. This model also served for generating lattices, which were used for MPE training of RDLT. Only a single CMLLR transformation was used in our system, as we did not observe any significant gain from using multiple CMLLR or MLLR transformations with our system on this task. Table II shows 2.6% absolute improvement in Word Error Rate (WER)obtained from CMLLR adaptation and additional 0.2% WER improvement when the acoustic model was retrained in SAT fashion [16]. Unless stated otherwise, CMLLR SAT system forms the basis of all systems described in the following sections.

C. iVector extraction

In principle, both ASR acoustic models and iVector extraction could be based on the same features and trained on the same data. Also, iVector extraction and definition of regions

in RDLT could be based on the same GMM model. In our experiments, however, we use two different GMMs trained on different features, since we simply took iVectors extracted by our existing system optimized for speaker verification task [3].

The features used for the iVector extraction were 19 Mel frequency cepstral coefficients (with log-energy) calculated every 10 ms using 25 ms Hamming window. This 20-dimensional feature vector was subjected to short time mean and variance normalization using a 3s sliding window. Delta and double delta coefficients were then calculated using a 5-frame window giving 60-dimensional feature vectors. The iVector extraction was based on Semi-Tied Covariance (STC) GMM with 2048 mixture components, which was trained on NIST SRE 2004 and 2005 telephone data. The subspace matrix $\mathbf{T}$ was trained on more than 2500 hours of data from the following telephone databases: NIST SRE 2004, 2005, 2006, Switchboard II Phases 2 and 3, Switchboard Cellular Parts 1 and 2, Fisher English Parts 1 and 2. The results are reported with 400 dimensional iVectors. Similarly to CMLLR transformations, iVectors were generated per conversation side for training and test data.

One could object that the iVector extraction is trained on much more data than the baseline ASR system, which makes the comparison of systems unfair. However, the iVector extraction is trained in *unsupervised manner* on data that are mostly not transcribed and therefore unusable for ASR training. Also, while large amount of training data is necessary to obtain good performance is speaker verification, we believe that it is not the case in these experiments, as RDLT, which is trained to extract the adaptation information from the iVector, is still trained on the same data as baseline ASR system.

VII. RDLT EXPERIMENTS

A. RDLT for discriminative feature extraction

In this section, we examine different configurations of RDLT used only in the usual way as a discriminative feature extraction. In the trivial case, where all feature frames are considers to belong to only one single region, RDLT comprises only one discriminatively trained linear transform. This configuration, which is also know as Discriminative HLDA [18], brings 0.5% absolute WER improvement compared to "ML CMLLR-SAT" baseline, as we can see in the first line of Table III.

The second line of the table reports additional 1.1% absolute WER improvement obtained from using 1000 regions. To define the regions in the acoustic space, all Gaussians from ML trained HMM model are pooled and clustered using agglomerative clustering to create GMM with desired number of components (see [19] for detailed description of the clustering algorithm).

In the following experiment, we incorporated also the information about context by using region posterior probabilities also from neighboring frames as suggested in [13]. Posterior probabilities of the GMM components for a current frame are stacked with the averages of posteriors for adjacent frames 1-2, 3-5 and 6-9 on the right and likewise for the left context (i.e. 7 groups spanning 19 frames in total). The resulting 7000 dimensional vector served as weights $\gamma_r(t)$ in equation (2) corresponding to 7000 transformations (39×39 matrices). Block diagram demonstrating such RDLT configuration is shown in Figure 1. The use of context brings large additional improvement (3% absolute) as can be seen in Table III in line denoted as "RDLT 7x1000 regions".

Models	WER [%]
RDLT 1 regions	31.4
RDLT 1000 regions	30.3
RDLT 7x1000 regions	**27.3**
RDLT 7x500 regions	27.6
RDLT 7x250 regions	27.7

TABLE III
RESULTS WITH RDLT USED AS FEATURE TRANSFORMATION FOR CMLLR-SAT ADAPTED SYSTEM.

Models	WER [%]
iVector RDLT 1 region	31.3
iVector RDLT 250 regions	30.2
iVector RDLT 500 regions	30.0
iVector RDLT 1000 regions	29.9

TABLE IV
RESULTS WITH RDLT USED ONLY FOR IVECTORS BASED ADAPTATION APPLIED ON TOP OF CMLLR-SAT ADAPTATION.

Next, we tested scaled-down systems to see a degradation of performance with smaller number of regions. A difference in WER between 1000 and 250 regions is 0.4%. This suggests that it is more important to invest parameters into context modeling than increasing the number of regions for the current frame.

B. iVector based adaptation

Table IV shows the behavior of the proposed adaptation approach with various number of transforms. To find the optimal configuration, we first considered the case corresponding to equation (3), where RDLT is used only for the adaptation. The optimal number of transformations saturates again on 1000 giving 2% absolute WER improvement over the CMLLR-SAT baseline. The differences between 500 and 1000 mixture components (and hence regions) is only 0.1% absolute.

We also experimented with incorporating the context information using the region posteriors form neighboring frames, but we found it ineffective when using RDLT for adaptation.

In table V, we compare the effect of CMLLR adaptation, iVector adaptation and combination of both for systems with and without RDLT used as discriminative feature transformation. For RDLT as feature transformation, we use the configuration with 7000 transformations as described in the previous section. For iVector adaptation, RDLT uses only 1000 transformations corresponding only to the regions for the current frame. This is the case even when both RDLT for feature transformation and RDLT for adaptation are combined. In this case, only 1000 transformations (39×439 matrices) corresponding to the current frame of GMM posteriors pro-

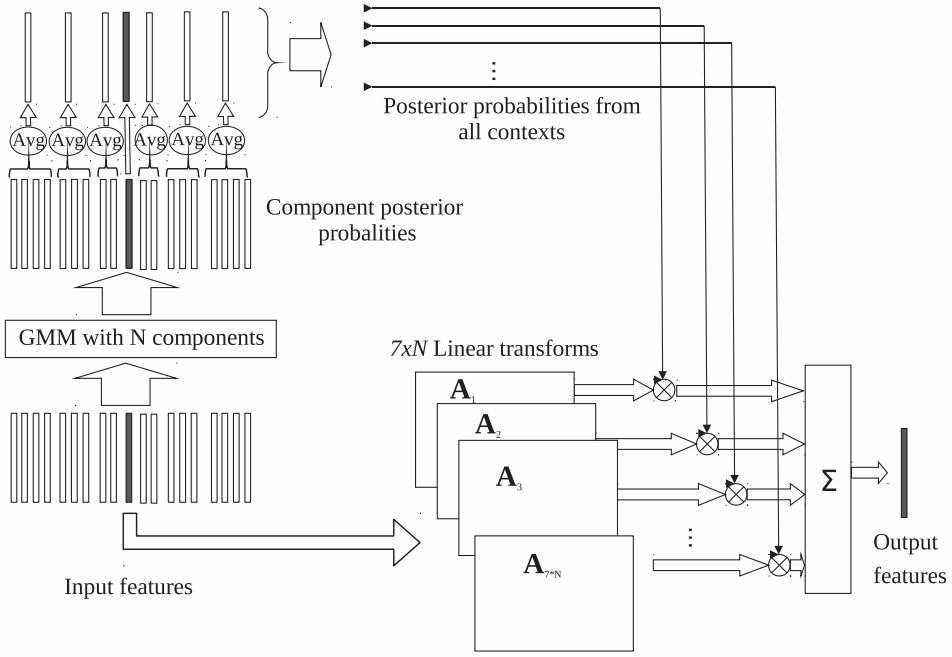

Fig. 1. RDLT with context transformations.

Adaptation	ML	RDLT
none	34.7	29.7
iVector	32.1	28.7
CMLLR-SAT	31.9	27.3
both	29.9	26.5

TABLE V
SUMMARY OF DIFFERENT TECHNIQUES.

cesses 39-dimensional feature vector concatenated with 400 dimensional iVector. The remaining transformations (39×39 matrices) corresponding to context posteriors process only the 39-dimensional feature vector.

The first line of table V shows the results without any adaptation. As can be seen, RDLT provides impressive improvement 5% absolute in this case. Comparing the following two lines, we see that iVector adaptation on its own appears to be slightly less effective than CMLLR transformation for this task. However, the two adaptation techniques seem to be complementary and the best result is obtained from their combination as can be seen from the last line in the table.

VIII. Conclusions and future work

We presented a novel technique for feature compensation based on iVectors — a popular technique in Speaker Recognition. We found it to be complementary approach to common adaptation techniques. On a well tuned RDLT system with standard CMLLR adaptation, we reached 0.8% additive absolute WER improvement. Without CMLLR adaptation, 1.0% absolute improvement was obtained.

Unsupervised estimation of CMLLR requires an additional decoding pass to obtain the adaptation hypothesis. On contrary, our approach only requires to extract the iVector from adaptation data which takes only a fraction of time necessary for decoding. Forwarding features through the set of transforms is also fast as only few transformations (usually only one or two) are applied per frame due to the sparsity of posterior probabilities. Therefore, our approach could be considered for decoding in the first pass of multi-pass systems or in one-pass systems.

This paper presents the first results of the proposed technique, in short-term, we will face the following issues:

1) Lattices used for discriminative training were generated using model with more than 8% higher WER compared to the performance of the final model. Further improvement could be obtain from lattices that would better reflect errors made by the final system.
2) iVector extraction was optimized for Speaker Recognition and the optimal configuration for speech recognition can be very different. Also, iVector extraction based on ASR features and GMM taken from RDLT would greatly simplify the system.
3) Finally, we would like to integrate the proposed approach into our full-featured system including other advanced techniques such as MPE model parameter training or neural network bottle-neck features.

This paper describe only one special instance of a more general scheme, where nonlinear transformation is trained discriminatively to compensate features based on external source of information useful for adaptation. Other forms of discriminatively trained nonlinear transformations can be considered (e.g. artificial neural networks), and different external sources of adaptation information can be found useful (e.g. noise spectrum estimate for noise robust speech recognition).

ACKNOWLEDGMENT

This work was partly supported by Technology Agency of the Czech Republic grant No. TA01011328, Czech Ministry of Education project No. MSM0021630528, Grant Agency of Czech Republic project No. 102/08/0707, and by Czech Ministry of Trade and Commerce project No. FR-TI1/034.

REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front–end factor analysis for speaker verification," in *IEEE Trans. on Audio, Speech and Lang. Process.*, 2010.

[2] P. Kenny, "Bayesian speaker verification with heavy–tailed priors," keynote presentation, Proc. of Odyssey 2010, Brno, Czech Republic, June 2010.

[3] P. Matějka, O. Glembek, F. Castaldo, J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Černocký, "Full-covariance UBM and heavy-tailed PLDA in I-vector speaker verification," in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011*. IEEE Signal Processing Society, 2011, pp. 4828–4831.

[4] A. Stolcke, L. Ferrer, and S. Kajarekar, "Improvements in mllrtransform-based speaker recognition," in *in Proc. IEEE Odyssey 2006 Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, June 2006.

[5] B. Zhang, S. Matsoukas, and R. Schwartz, "Recent progress on the discriminative region-dependent transform for speech feature extraction," in *Proc. of Interspeech2006*, Pittsburgh, PA, USA, Sep 2006, pp. 2977–2980.

[6] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, 2003.

[7] A. Gunawardana and W. Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," in *Proc. Eurospeech 2001*, Aalborg, Denmark, Sep. 2001.

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

[9] S. Tsakalidis, V. Doumpiotis, and W. J. Byrne, "Discriminative linear transforms for feature normalization and speaker adaptation in hmm estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 13, pp. 367–376, 2005.

[10] K. Yu, M. J. F. Gales, and P. C. Woodland, "Unsupervised discriminative adaptation using discriminative mapping transforms," in *Proc. ICASSP, LAS VEGAS, NV*, Las Vegas, NV, USA, 2008, pp. 4273–4276.

[11] L. Chen, M. J. F. Gales, and K. K. Chin, "Constrained discriminative mapping transforms for unsupervised speaker adaptation," in *Proc. ICASSP*, Prague, Czech Republic, 2008.

[12] O. Glembek, L. Burget, P. Kenny, M. Karafiát, and P. Matějka, "Simplification and optimization of i-vector extraction," in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011*. IEEE Signal Processing Society, 2011, pp. 4516–4519.

[13] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fmpe: Discriminatively trained features for speech recognition," in *in Proc. IEEE ICASSP*, 2005.

[14] D. Povey, "Improvements to fMPE for discriminative training of features," in *Proc. of Interspeech2005*, Lisbon, Portugal, Sep 2005, pp. 2977–2980.

[15] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP '96*, vol. 2, Philadelphia, PA, 1996, pp. 1137–1140. [Online]. Available: citeseer.ist.psu.edu/anastasakos96compact.html

[16] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," 1997. [Online]. Available: citeseer.ist.psu.edu/gales97maximum.html

[17] T. Hain *et al.*, "The 2007 AMI(DA) system for meeting transcription," in *Proc. Rich Transcription 2007 Spring Meeting Recognition Evaluation Workshop*, Baltimore, Maryland USA, May 2007.

[18] B. Zhang and S. Matsoukas, "Minimum phoneme error based heteroscedastic linear discriminant analysis for speech recognition," in *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal processing (ICASSP)*, Philadelphia, PA, USA, march 2005, pp. 925–929.

[19] D. Povey, L. Burget, M. Agarwal, P. Akyazi, A. Ghoshal, O. Glembek, K. N. Goel, M. Karafiát, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "The subspace gaussian mixture model-a structured model for speech recognition," *Computer Speech and Language*, vol. 25, no. 2, pp. 404–439, 2011.