Unsupervised Adaptation with Discriminative Mapping Transforms

Kai Yu, Mark Gales and Philip C. Woodland

***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 still used for unsupervised adaptation. This is because discriminatively trained transforms are highly sensitive to errors in the adaptation supervision hypothesis. This paper describes a new framework for estimating transforms that are discriminative in nature, but are less sensitive to this hypothesis issue. A speaker-independent discriminative mapping transformation (DMT) is estimated during training. This transform is obtained after a speaker-specific ML-estimated transform of each training speaker has been applied. During recognition an ML speaker-specific transform is found for each test-set speaker and the speaker-independent DMT then applied. This allows a transform which is discriminative in nature to be indirectly estimated, while only requiring an ML speaker-specific transform to be found during recognition. The DMT technique is evaluated on an English conversational telephone speech task. Experiments showed that using DMT in unsupervised adaptation led to significant gains over both standard ML and discriminatively trained transforms.**

***Index Terms*— unsupervised adaptation, discriminative training, criterion mapping function, discriminative mapping transform**

I. Introduction

Speaker adaptation is a widely used technique to build speaker-dependent models to recognise speech from unknown speakers. Given a well trained acoustic model, a small amount of data from the target speaker are used to modify the acoustic model parameters so that the resultant model is more suitable for recognising speech from the specific speaker. The most commonly used approaches for speaker adaptation are linear transformations of the acoustic model parameters as they can be robustly estimated given limited adaptation data [1], [2]. To estimate the linear transforms, both audio data and the associated transcriptions are required. If the correct transcriptions of the speaker-specific audio data are available, the adaptation operates in a *supervised* mode. However, in many applications, such as broadcast news transcription or conversational telephone speech, there is no transcription available for the test data. In this case, initial transcriptions must be generated using an unadapted model. Then linear transforms are estimated given the audio and these automatically generated transcription. The linear transforms are then used to adapt the acoustic model for a final recognition pass. This is *unsupervised adaptation* and the focus of this paper.

Originally, linear transforms were estimated using the maximum likelihood (ML) criterion and yielded significant gains over unadapted systems for both supervised and unsupervised adaptation [1], [2]. However, as most state-of-the-art systems use discriminative training criteria to reduce the word error rate (WER) [3], [4], [5], there has been interest in also using discriminative criteria for linear transform based adaptation [6], [7], [8], [9]. It has been shown that in supervised mode adaptation, the use of discriminative linear transforms (DLTs) can lead to significant performance improvements over ML transform estimation [6]. However, in unsupervised adaptation, the performance gain of DLT is greatly reduced [9], [10]. This is because discriminative criteria are more sensitive to errors in the hypotheses (or references) than the ML criterion. This sensitivity to hypothesis errors may be reduced using, for example, confidence scores [11], [12], [9] or lattice-based approaches [13], [14]. However, even for these approaches, gains over ML estimated transforms are still small. Thus despite gains in supervised adaptation, 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, *Maximum Likelihood Linear Regression* (MLLR) based discriminative speaker adaptive training (DSAT) [15], [8], [16], discriminative cluster adaptive training [17], and feature MPE (fMPE) [18] or region-dependent feature transforms [19] 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 estimating discriminative linear transforms. Here, ML is used to estimate all speaker-specific parameters in the recognition stage, whereas a speaker-independent discriminative transform is estimated during training.

The general 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 this paper, only MLLR adaptation of the means [1] will be examined. However, in theory this approach can be applied to any form of linear transforms, such as constrained MLLR [2]. During training, the speaker-independent DMT is estimated given ML-estimated transform of each training speaker. At recognition time, an ML speaker-specific transform

is found for each test-set speaker and the DMT applied to it. The combination of the DMT and the ML transform is then used for adaptation. As only the ML criterion is used during test data adaptation, the sensitivity to transcription errors in unsupervised adaptation will be greatly reduced. At the same time, due to the nature of the DMT, the combined transform will be discriminative in nature. Hence, the combined transform can be regarded as an approximation to a speaker-dependent DLT.

This paper is organised as follows. In section II, linear transforms for adaptation are reviewed, and how they may be used in combination with discriminative training discussed. The DMT framework is discussed in section III. Experiments on an English conversational telephone speech (CTS) task are presented in section IV followed by conclusions.

II. Linear Transforms For Adaptation

Linear transformations are the most commonly used approach to speaker adaptation with limited adaptation 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

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

where $\boldsymbol{\xi} = [\boldsymbol{\mu}^T \ 1]^T$ 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 for speaker s. The parameters of the transform, $\mathbf{W}_{\mathtt{ml}}^{(s)}$ are estimated using the ML criterion [1]

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

where $\mathbf{O}^{(s)}$ and $\mathcal{H}^{(s)}$ are the observations and reference/hypothesis of the adaptation data for speaker s respectively, and $\mathcal{M}$ are the HMM model parameters. An important issue is how the transcription/hypothesis, $\mathcal{H}^{(s)}$, is obtained. If it is known a-priori, this is *supervised adaptation* and the hypothesis is assumed to be error-free. When the transcription is not available, *unsupervised adaptation* must be used with recognised hypothesis from a speech recognition system. The basic procedure is:

1) Generate initial hypothesis $\mathcal{H}^{(s)}$ using an acoustic model, such as a speaker-independent (SI) model, possibly with an initial estimate of the transform.
2) Estimate the transform for speaker s given the audio $\mathbf{O}^{(s)}$ and initial hypothesis $\mathcal{H}^{(s)}$ as supervision. The process may be repeated.

An important aspect of unsupervised adaptation is that the hypothesis is, in general, errorful. Depending on the number of errors and the transform complexity, this may lead to an unreliable estimate of the transforms. Though affected by errorful hypotheses, it has been found that the ML estimated transforms are not very sensitive to the hypothesis errors and can yield good reductions in WER with unsupervised adaptation even at high error rates [20]. This is one of the main reasons for the wide-spread use of MLLR.

In state-of-the-art speech recognition systems, discriminative training of acoustic models is commonly employed to obtain the best performance [3], [4], [5]. Inspired by the results, there has been interest in using discriminative criteria in linear transform based adaptation [6], [7], [8], [9]. The standard approach is to directly estimate linear transforms for each test-set speaker using a discriminative criterion. Speaker-dependent transforms estimated using these discriminative criteria are referred to as discriminative linear transforms (DLTs). The minimum Bayes-risk form of the DLT estimation formula can be expressed as

$$\mathbf{W}_{\mathtt{d}}^{(s)} = \arg\min_{\mathbf{W}} \left\{ \sum_{\mathcal{H}} P(\mathcal{H}|\mathbf{O}^{(s)}, \mathbf{W}; \mathcal{M}) \mathcal{L}(\mathcal{H}, \mathcal{H}_{\mathtt{sup}}^{(s)}) \right\} \tag{3}$$

where $P(\mathcal{H}|\mathbf{O}^{(s)}, \mathbf{W}; \mathcal{M})$ is the posterior probability of hypothesis $\mathcal{H}$ given the observation from speaker s and the model parameter $\mathcal{M}$ and the transform parameters $\mathbf{W}$, $\mathcal{L}(\mathcal{H}, \mathcal{H}_{\mathtt{sup}}^{(s)})$ is the loss function of $\mathcal{H}$ given the supervision $\mathcal{H}_{\mathtt{sup}}^{(s)}$. In this work, the minimum phone error (MPE) criterion is used, where $\mathcal{L}(\mathcal{H}, \mathcal{H}_{\mathtt{sup}}^{(s)})$ is defined as the number of incorrect phones[4][1]. From equation (3), as the posterior probability of each possible hypothesis is used in discriminative training, the correct transcription is required along with a compact representation of competing hypotheses. To estimate the DLT, in this work, lattices are used to represent competing hypotheses [21], [5]. Once the DLTs are estimated, the form of model adaptation remains the same,

$$\hat{\boldsymbol{\mu}}^{(s)} = \mathbf{A}_{\mathtt{d}}^{(s)}\boldsymbol{\mu} + \mathbf{b}_{\mathtt{d}}^{(s)} = \mathbf{W}_{\mathtt{d}}^{(s)}\boldsymbol{\xi} \tag{4}$$

where $\mathbf{W}_{\mathtt{d}}^{(s)} = [\mathbf{A}_{\mathtt{d}}^{(s)} \ \mathbf{b}_{\mathtt{d}}^{(s)}]$ is the DLT of speaker s.

As discriminative criteria aim to reduce the recognition error (or more generally the loss) of the training data with respect to the (assumed) "correct" transcription, it is not surprising that discriminative estimation is far more sensitive to the accuracy of the transcriptions than ML estimation. Furthermore, in unsupervised adaptation, the speech recognition system used to generate $\mathcal{H}_{\mathtt{sup}}^{(s)}$ and the competing hypotheses is often very similar to the system to be adapted. Then the competing hypotheses tend to be closer to the assumed "correct" hypothesis than if the actual transcription had been used. The discrimination ability of the trained transform may then be reduced as an underestimate of the "true" loss function is used. This is an inherent problem of directly using discriminative criteria in unsupervised mode[2]. Due to these sensitivities, although DLTs have been successfully used in supervised adaptation [6], only small gains over ML estimated transforms have been observed in unsupervised adaptation [9]. Various approaches, such as using confidence score to select high-quality supervision for transform estimation [9], have been investigated to improve the performance within the direct DLT estimation framework. However, the gains over ML-trained transforms are still disappointing [9]. Also, DLTs are much more computationally expensive to estimate due to the use of competing hypotheses. This is why ML estimated transforms

[1]Note that the definition of MPE criterion in [4] is an equivalent version based on phone accuracy rather than phone error. So the optimisation in [4] is to maximise the MPE criterion.

[2]Similar problems have also been found in unsupervised discriminative training for acoustic model parameters.

are still commonly used in unsupervised adaptation instead of DLTs.

ML-estimated transforms for unsupervised adaptation are often used in combination with discriminatively trained HMM parameters. In the most widely used form of discriminative speaker adaptive training (DSAT) [15], [16], 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 DSAT model. Discriminative cluster adaptive training (DCAT) [17] follows a similar procedure but uses multiple-cluster models as the canonical model. Thus, ML-estimated interpolation weights are found during training and recognition. In both DSAT and DCAT, 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) [18], [22] and region-dependent feature transforms (RDFT) [23] have also been used in combination with ML-estimated speaker-specific transforms. In these approaches, the acoustic space is partitioned into regions, region-dependent matrices are then discriminatively trained and used to transform the features. Though the matrices are acoustic region-dependent, they are independent of speakers. These discriminative transforms may be built on top of a speaker-specific ML-adapted feature-space to achieve further speaker adaptation gain [19]. All these schemes are relatively robust in unsupervised adaptation and adopt the same general strategy. Speaker independent parameters are discriminatively estimated while all speaker-specific parameters, the transforms, are ML trained.

III. Discriminative Mapping Transforms

The previous section has described that, due to the high sensitivity to initial hypotheses error, direct discriminative estimation of speaker-specific transforms does not work well for unsupervised adaptation. In this section, a new framework, the discriminative mapping transform, is proposed to address the sensitivity problem by *indirectly* estimating discriminative transforms.

A. Indirect discriminative adaptation using a criterion mapping function (CMF)

The criterion mapping function (CMF) uses the same general discriminative training strategy described in section II, i.e., discriminative criteria are only used for speaker-independent parameters during training while the ML criterion is used for speaker-dependent parameters during training and recognition. To achieve this purpose, a speaker-independent function, the CMF, is introduced to map, for example, ML-trained transforms into discriminative transforms [24]. The assumption here is that the effect of adaptation and discrimination can be factorised. Speaker-specific ML transforms are used to adapt the model to the speaker, while speaker-independent CMF is used to add discrimination power to the adapted parameters. As there is no discriminative estimation in the recognition stage, the sensitivity to hypothesis error should be reduced. Within the CMF framework, the final speaker-specific discriminative transform, similar to a DLT, is found using

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

where $\mathbf{W}_{\mathtt{ml}}^{(s)}$ is the speaker-dependent ML transform found using equation (2); and $\mathcal{F}_{\mathtt{dm}}(\cdot)$ is the mapping function with speaker-independent parameters, $\mathbf{\Lambda}$ to convert the ML space to discriminative space. As $\mathbf{W}_{\mathtt{d}}^{(s)}$ is not directly estimated from the data, this is an *indirect* discriminative adaptation scheme. The procedure for transform estimation is shown in figure 1.

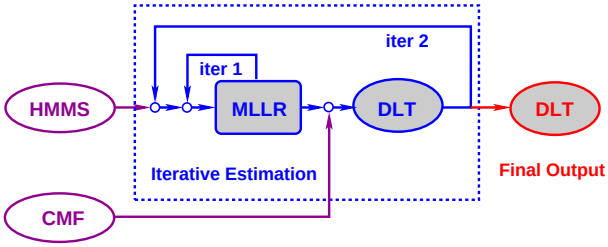

Fig. 1. Indirect discriminative transform estimation in testset adaptation

In figure 1, ellipses represent the parameters which are either known in advance or obtained without directly using supervision data, whereas squares represent the parameters that are directly estimated using the audio and supervision hypothesis of the adaptation data, a clear background represents speaker-independent parameters whereas shaded background represents speaker-dependent parameters. From figure 1, the only speaker-specific parameters to be directly updated are the MLLR parameters, the DLT is formed by applying a mapping without further parameter estimation. As in iterative MLLR [20], multiple iterations can be used to refine the estimation of MLLR. There are two ways to do this. The first way is as used in standard iterative MLLR, i.e. refine the MLLR transform estimate using the MLLR adapted model, as shown in "iter 1" in figure 1. This is consistent with normal unsupervised MLLR adaptation. The second approach is iterative DLT, i.e. estimate the MLLR transform using the DLT adapted model, as shown in "iter 2". In this paper, only "iter 1" is considered in the experiments as it gives a strict comparison to MLLR adaptation[3].

As shown in figure 1, the parameters of the CMF are required prior to testset adaptation. One way to obtain these parameters, $\mathbf{\Lambda}$, is to estimate them from the whole training data set. This has two advantages. First there is a large amount of training data to estimate the mapping function. This allows a large number of parameters to be robustly estimated. As described below, if a linear transform is used as the form of the CMF, a large number of regression base classes can be effectively used. Second, as the correct transcriptions are known for the training data, there are no hypothesis sensitivity issues. When using a CMF in recognition, an additional advantage is that during recognition only an ML-estimated transform is required to be estimated. This avoids need to

[3]Initial experiments using a multi-pass decoding framework have showed that using "iter 2" can obtain further improvement in WER.

generate competing hypotheses, which are required for direct estimation of DLTs. The rest of this section describes a specific implementation of the the CMF based on linear transforms.

B. Discriminative mapping transforms

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

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

where vec() maps the matrix to a vector form, for an n-dimensional feature vector, let the $n \times (n+1)$ matrix $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_{n+1}]$, where $\mathbf{w}_i$ is the i^th column vector of $\mathbf{W}$, then the column vector form of $\mathbf{W}$ is

$$\text{vec}(\mathbf{W}) = [\mathbf{w}_1^T, \cdots, \mathbf{w}_{n+1}^T]^T \tag{7}$$

$\mathbf{H}_{\mathtt{dm}}$ is an $n(n+1) \times n(n+1)$ matrix and $\mathbf{c}_{\mathtt{dm}}$ is a $n(n+1)$ column vector . In this work, a simpler form of transformation is used instead. $\mathbf{H}_{\mathtt{dm}}$ is restricted to be block-diagonal in structure with $n+1$ identical blocks $\mathbf{A}_{\mathtt{dm}}$. The transformation can then be expressed as

$$\mathbf{W}_{\mathtt{d}}^{(s)} = \mathbf{A}_{\mathtt{dm}}\mathbf{W}_{\mathtt{ml}}^{(s)} + \boldsymbol{\beta}_{\mathtt{dm}} \tag{8}$$

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

$$\hat{\boldsymbol{\mu}}^{(s)} = \left(\mathbf{A}_{\mathtt{dm}}\mathbf{W}_{\mathtt{ml}}^{(s)} + \boldsymbol{\beta}_{\mathtt{dm}}\right)\boldsymbol{\xi} = \mathbf{A}_{\mathtt{dm}}\hat{\boldsymbol{\mu}}_{\mathtt{ml}}^{(s)} + \mathbf{B}_{\mathtt{dm}}\boldsymbol{\mu} + \mathbf{b}_{\mathtt{dm}} \tag{9}$$

where $\boldsymbol{\beta}_{\mathtt{dm}} = [\mathbf{B}_{\mathtt{dm}}\ \mathbf{b}_{\mathtt{dm}}]$, $\mathbf{B}_{\mathtt{dm}}$ 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 $\mathbf{B}_{\mathtt{dm}} = \mathbf{0}$, this leads to

$$\hat{\boldsymbol{\mu}}^{(s)} = \mathbf{A}_{\mathtt{dm}}\hat{\boldsymbol{\mu}}_{\mathtt{ml}}^{(s)} + \mathbf{b}_{\mathtt{dm}} = \mathbf{W}_{\mathtt{dm}}\boldsymbol{\xi}_{\mathtt{ml}}^{(s)} \tag{10}$$

where $\boldsymbol{\xi}_{\mathtt{ml}}^{(s)} = [\hat{\boldsymbol{\mu}}_{\mathtt{ml}}^{(s)T}\ \ 1]^T$, $\mathbf{W}_{\mathtt{dm}} = [\mathbf{A}_{\mathtt{dm}}\ \mathbf{b}_{\mathtt{dm}}]$ is the DMT. This is the form used in this paper.

The advantage of this form of simplification is that the speaker-independent DMT parameters, $\mathbf{W}_{\mathtt{dm}}$, can be estimated in a similar fashion to the standard DLTs in equation (3). Given equation (10), the estimation of the DMT, $\mathbf{W}_{\mathtt{dm}}$, using the MPE criterion can be expressed as

$$\mathbf{W}_{\mathtt{dm}} = \arg\min_{\mathbf{W}}\left\{\sum_{s,\mathcal{H}} P(\mathcal{H}|\mathbf{O}^{(s)}, \mathbf{W}; \mathcal{M}_{\mathtt{ml}}^{(s)})\mathcal{L}(\mathcal{H}, \mathcal{H}_{\mathtt{sup}}^{(s)})\right\} \tag{11}$$

where $\mathcal{M}_{\mathtt{ml}}^{(s)}$ is the MLLR adapted model parameters for speaker s, and all the other notation is the same as in equation (3). Note that the above summation is over all training speakers. Thus rather than accumulating statistics using the original HMM (as in equation (3)), the DMT estimation uses speaker-specific ML-adapted HMM parameters and sums over all training speakers. To optimise equation (11), the standard MPE optimisation scheme, based on the weak-sense auxiliary function [25], can be used. The derivation of the DMT update formulae is similar to the standard DLT [9]. Here, only the final update formulae are given. For more details, refer to [9]. The sufficient statistics required for DMT update are

$$\gamma_m(t_s) = \gamma_m^{\mathtt{n}}(t_s) - \gamma_m^{\mathtt{d}}(t_s) + \alpha\gamma_m^{\mathtt{ml}}(t_s) \tag{12}$$

$$\gamma_m^{(s)} = \sum_{t_s}\gamma_m(t_s) \tag{13}$$

$$\mathbf{G}_i = \sum_{m,s}\frac{\gamma_m^{(s)} + D_m^{(s)}}{\sigma_{ii}^{(m)}}\hat{\boldsymbol{\xi}}_{\mathtt{ml}}^{(sm)}\hat{\boldsymbol{\xi}}_{\mathtt{ml}}^{(sm)T} \tag{14}$$

$$\mathbf{k}_i = \sum_{m,s}\frac{\sum_{t_s}\gamma_m(t_s)o_{t_s,i} + D_m^{(s)}\hat{\mu}_i^{(sm)}}{\sigma_{ii}^{(m)}}\hat{\boldsymbol{\xi}}_{\mathtt{ml}}^{(sm)} \tag{15}$$

where t_s is the time index for speaker s, $\gamma_m^{\mathtt{n}}(t_s)$ and $\gamma_m^{\mathtt{d}}(t_s)$ are posterior occupancy of the Gaussian component m being at time s given the numerator and denominator lattices respectively [5]. The numerator and denominator lattices are specific representations of the correct transcription and competing hypotheses paths in MPE training respectively. The occupancy $\gamma_m^{\mathtt{n}}(t_s)$ and $\gamma_m^{\mathtt{d}}(t_s)$ are calculated using the lattice forward-backward algorithm [21] and $\gamma_m^{\mathtt{ml}}(t_s)$ is the normal ML posterior occupancy calculated given the correct transcription; α is a smoothing constant to balance the ML occupancy and the other occupancy to improve the generalisation ability of the discriminatively trained parameters and serves a similar function to the weight in the I-smoothing technique [4], [5][4]. $D_m^{(s)}$ is a smoothing term for each component m and speaker s to ensure the convergence of the discriminative update. In this work, it is set to be $D_m^{(s)} = E\sum_{t_s}\gamma_m^{\mathtt{d}}(t_s)$ where $E = 0.8$, which is a setup used in [9][5]. $\hat{\boldsymbol{\xi}}_{\mathtt{ml}}^{(sm)}$ is the ML adapted extended mean vector as defined in equation (1), $\sigma_{ii}^{(m)}$ is the i^{th} diagonal element of the covariance matrix of Gaussian component m, $o_{t_s,i}$ is the i^{th} element of the observation vector $\mathbf{o}_{t_s}$ for speaker s, $\hat{\mu}_i^{(sm)}$ is the i^{th} element of the current adapted mean $\hat{\boldsymbol{\mu}}^{(sm)}$, which is calculated using equation (10). Note that this is different from the ML adapted mean vector as the current DMT is also used to calculate $\hat{\boldsymbol{\mu}}^{(sm)}$.

Having obtained the above statistics, the i^{th} row vector of $\mathbf{W}_{\mathtt{dm}}$, $\mathbf{r}_i^T$, with the size of $1 \times (n+1)$, can be estimated by

$$\mathbf{r}_i = \mathbf{G}_i^{-1}\mathbf{k}_i \tag{16}$$

During training, the estimation of the DMTs is an iterative process. A DMT may be estimated in various ways as shown in figure 2. As in figure 1, ellipses represent parameters known in advance and squares denote parameters updated using the training data, a clear background represents speaker-independent parameters whereas shaded background represents speaker-specific parameters. There are also two ways of iteratively training DMTs. The first method, "iter 1", is

[4]In the experiments, α was set to 0.01 as in [9]. It was also found that there was no significant difference between setting $\alpha = 0$ and $\alpha = 0.01$. This is felt to be because the regression tree structure is used and there are sufficient data for each node in the tree. However, $\alpha = 0.01$ was used for all experiments as this should be a more robust configuration for situations where there is less training data.

[5]Setting E is to balance the update speed and convergence. Experiments showed that, with 1000 DMT regression base classes, using large E value, such as $E = 2.0$ will lead to very slow DMT parameter update. However, using very small E, such as $E = 0.5$, will lead to unstable update after several iterations. $E = 0.8$ is an appropriate value in practice.

the iterative DMT, i.e., given the HMM model and MLLR transforms for each training speaker, only DMT parameters are iteratively refined. The second approach, "iter 2", uses iterative DLT. Here MLLR and DMT are treated together, so not only the DMT, but also the MLLR transforms are iteratively updated using the HMM model and previously trained MLLR+DMT.

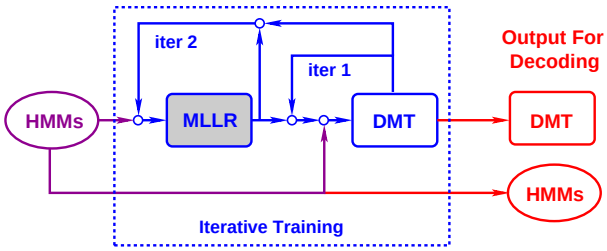

Fig. 2. Iterative DMT training procedure

In this paper, only iterative DMT is considered as this requires less training time[6]. Given a well trained set of HMMs $\mathcal{M}$, for example discriminatively trained SI HMMs, the iterative estimation procedure for DMT is summarised below:

1) Estimate the MLLR transforms $\mathbf{W}_{\texttt{ml}}^{(s)}$ for each training speaker s given $\mathcal{M}$.
2) Set $k = 0$ and $\mathbf{W}_{\texttt{dm}}^{(0)} = [\mathbf{I}\ \mathbf{0}]$, where $\mathbf{I}$ is the identity matrix.
3) Estimate $\mathbf{W}_{\texttt{dm}}^{(k+1)}$ using equation (16) given $\mathcal{M}$, $\mathbf{W}_{\texttt{ml}}^{(s)}$ and $\mathbf{W}_{\texttt{dm}}^{(k)}$
4) $k = k + 1$. Goto step 3 until converged.

When using MLLR, the discriminative criterion for the adapted model may be lower than for unadapted discriminative HMMs. The DMT may then require multiple iterations to converge for the discriminative criterion. From the training procedure, the DMT is dependent on specific HMMs. Hence, if the HMM model set changes, the DMT also needs to be re-estimated.

Once the DMT is trained, it is used in testset adaptation together with the HMM set. The procedure is similar to figure 1. As indicated in section III-A, in this paper, only standard iterative MLLR ("iter 1") is considered. To summarise, given a set of HMMs, $\mathcal{M}$, and DMT $\mathbf{W}_{\texttt{dm}}$, the discriminative adaptation for a test-set speaker s is performed as shown below:

1) Find initial transcriptions for s. In unsupervised mode, $\mathcal{M}$ may be used to decode the audio from speaker s.
2) Iteratively estimate MLLR transform $\mathbf{W}_{\texttt{ml}}^{(s)}$ given the initial hypothesis for speaker s
3) Adapt HMMs parameters using equation (10) using the DMT $\mathbf{W}_{\texttt{dm}}$ and the newly estimated $\mathbf{W}_{\texttt{ml}}^{(s)}$.
4) Use the adapted model to decode the audio.

The presentation of the DMT has so far only considered a single transformation for all Gaussian components. 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 [27]. The same approach to clustering Gaussians together to form multiple base-classes, either based on data-driven clustering in acoustic space or based on phonetic characteristics, can be used for DMT. As DMT estimation uses all the available training data, the number of transform classes may be made much larger than is usually used for standard speaker adaptation.

Though mean adaptation is considered in this paper, the DMT can also be applied to constrained MLLR (CMLLR) adaptation [28], [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 II, fMPE and RDFT both use a speaker-independent discriminatively trained transform given the speaker-dependent CMLLR adapted features. This is approach is similar to the 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-transformation and is dependent on the component being transformed.

IV. Experiments

In this section, the DMT technique is evaluated on a large vocabulary English conversational telephone speech task.

A. System description

The acoustic model training dataset consists of 5446 speakers, about 296 hours of data. The sources are the LDC Callhome English (`che`), Switchboard (`Swbd`) and Switchboard-Cellular (`SwCell`) datasets. The test set used to evaluate recognition performance is the `eval03` dataset, consisting of 144 speakers, about 6 hours. This test set has data from two different corpora. The `Swbd` corpus has a similar data type to the training data, while the `Fisher` corpus is not included in the training sources.

All systems used a 13-dimensional PLP front-end including `C0` and their first, second and third delta parameters. Side-level cepstral mean and variance normalisation and vocal tract length normalisation (VTLN) were used. An HLDA transform was applied to reduce the feature dimensionality to 39. State-clustered triphone HMMs with 6K distinct states and an average of 16 Gaussian components per state were used. The MPE [4] criterion was used to train all the acoustic models. In the MPE training process, the correct transcription was used to construct the numerator lattices, a heavily pruned bi-gram model was used with the ML model to generate the denominator lattices which contain competing hypotheses. Two different MPE systems were built. The first was a speaker-independent (SI) MPE system built from the ML-SI model. The second was a MPE trained mean-MLLR based discriminative speaker adaptive training (MPE-SAT) system [29]. This MPE-SAT system adopted the most commonly used discriminative adaptive training approach. An ML-SAT system was trained first and the HMM model parameters were discriminatively updated

[6]For the use of the iterative DLT method, refer to [26], where DMT is investigated in an adaptive training framework.

given the ML estimated transforms [15]. When using the MPE-SAT system in recognition, standard MLLR adaptation was normally used. Given the HMM models of the two systems, corresponding DMTs were estimated respectively.

Unless explicitly stated, in the recognition stage, mean-based linear transform adaptation was performed in unsupervised mode, and the default initial hypothesis was generated using the MPE-SI system. During the unsupervised adaptation, MLLR transforms were first iteratively estimated given the initial hypotheses. For the MPE-SI system, 4 iterations were used for MLLR update given the MPE-SI model. For the MPE-SAT system, 4 iterations of MLLR update were performed using the corresponding ML-SAT [30] models, then 2 more iterations were used to estimate MLLR given the MPE-SAT model. Once the testset speaker-specific MLLR transforms were estimated, DMT may then be used to implement discriminative adaptation. After adaptation, the final recognition was a single pass full Viterbi decoding using the adapted MPE systems and a tri-gram language model trained on 1044M words with a 58k dictionary.

As a contrast, standard direct DLTs were also estimated. In unsupervised adaptation, rather than using the correct transcription, an initial 1-best hypothesis was used to construct the numerator lattices for direct DLT estimation. This 1-best hypothesis was generated by the MLLR adapted MPE-SI model. Denominator lattices for testset speakers were generated using a similar way to the MPE training, where the ML-SI model and a heavily pruned bi-gram language model were used. Mean-based DLTs were then estimated using the MPE criterion as in [9].

It is worth noting that, for all experiments, two regression base classes, one for speech and one for silence, were used for MLLR and DLT to achieve robust estimate. However, for DMT, it is possible to use more transforms, which will be discussed later.

B. Effectiveness of DMT to improve discriminative criteria

As a specific form of criterion mapping function, DMT uses a linear transform to map the ML parametric space to the MPE parametric space. As shown in section III-B, the DMT is estimated on the whole training data with the MPE criterion. It is interesting to see how effective this mapping is in terms of the training data discriminative criterion. Rather than quoting the discriminative criterion in the minimum Bayes-risk form as in equation (3), the original MPE criterion [4], which is to be maximised, is used here. The original MPE criterion is the *expected phone accuracy* given the competing hypotheses (denominator lattices generated using heavily pruned bi-gram model) with respect to the numerator lattices. It is equivalent to equation (3) but defined in terms of phone accuracy rather than phone error. The experiments in this section used the MPE-SI models. Table I shows the expected phone accuracy of applying the standard MLLR and DMTs on training data. Note that, during training, the correct transcription was used to construct the numerator lattices.

In table I, using MLLR adaptation for the MPE-SI model degraded the expected phone accuracy. This is because applying MLLR transforms can be regarded as an approximation of performing one more iteration of speaker-dependent ML acoustic training. As the parameters were estimated to maximise the likelihood rather than the MPE criterion, the expected phone accuracy was reduced. Table I also shows the change in expected phone accuracy when using a DMT estimated using 1, 2, or 3 training iterations and with different number of regression base classes. Three sizes of regression classes were examined using the standard data-driven acoustic clustering approach [27], 2, 46 and 1000. A 46 base-class set was also estimated using phone information, i.e., each class used a distinct center-phone in the triphone models. All DMT training improved the expected phone accuracy. Increasing the number of regression base classes or the training iterations gave higher values. It can be found that with 1000 base classes, the DMT yielded a large increase in expected phone accuracy. This shows that given enough parameters and flexibility, the DMT is effective in improving the discrimination power of the adapted model on the training data.

Sys.	Gaussian Clustering	# Class	DMT Train Iteration		
			1	2	3
MPE-SI	—		0.826		
MLLR	Phone	2	0.817		
+ DMT	Acoustic	2	0.818	—	—
		46	0.819	—	—
		1000	0.829	0.836	0.841
	Phone	46	0.820	0.823	0.824

TABLE I

EXPECTED PHONE ACCURACY WITH RESPECT TO THE CORRECT TRANSCRIPTION ON THE TRAINING DATA

Having investigated the expected phone accuracy on the training data, it is also interesting to check it on unseen test data. As unsupervised adaptation is the focus of this work, the 1-best hypothesis generated from the MLLR adapted model was used as the reference transcription. The correct transcription of the test data was also used to give a contrast. A DMT with 1000 base classes and 3 training iterations was used. As an interesting comparison, DLTs were also directly estimated[7]. These expected phone accuracy values of the test set are shown in table II

Adaptation	Reference for Expected Phone Accuracy	
	1-best hyp.	correct trans.
MLLR	0.793	0.670
+ DMT	0.803	0.682
DLT	0.855	0.693

TABLE II

EXPECTED PHONE ACCURACY ON THE TEST DATA

From table II, the DMT improved the expected phone accuracy compared to the MLLR adapted model given the errorful hypothesis as the reference. This shows that the phone-level discrimination power of the DMT generalises to the

[7]Here only 1 iteration was used as more iterations degraded the adaptation performance.

test data when the ML-transform is estimated on error-full hypotheses[8]. The corresponding accuracy for DLT was much higher than for MLLR+DMT. This is expected as the DLT is able to tune to the reference hypotheses more precisely than the DMT. However, due to errors in the reference hypothesis, more than one DLT iteration led to parameter over-training and degraded the performance. Therefore, the WERs for only one iteration of DLT estimation are reported. When using the correct transcription as the reference, all adaptation approaches obtained degraded expected phone accuracy values. This is the effect of sensitivity to the errorful hypothesis. It is interesting to note that the accuracy reductions for MLLR and MLLR+DMT are similar, whereas the DLT yielded much larger accuracy reductions. This implies that the DLT is more sensitive to errors than the other two approaches.

C. Using DMT in unsupervised discriminative adaptation

The previous section shows the effectiveness of DMT in terms of the expected phone accuracy. This section investigates various aspects of DMT using the recognition performance of a full decoding framework. All experiments in this section were based on the MPE-SI system.

1) Number of base classes: As discussed in section III, to improve the power of the DMT, a large number of transforms may be used. Different numbers or types of regression base classes were investigated as described in section IV-B. Note that as there is sufficient training data, in all experiments, the actual number of transforms was always the same as the number of the base classes. The results, in terms of word error rate (WER), of MPE-SI system are shown in table III.

Sys.	Gaussian Clustering	# Class	DMT Train Iteration		
			1	2	3
MPE-SI	—		29.2		
MLLR	Phone	2	27.0		
+ DMT	Acoustic	2	27.0	—	—
		46	26.9	—	—
		1000	26.7	26.4	26.2
	Phone	46	26.8	26.7	26.7

TABLE III
%WER OF USING DMT WITH DIFFERENT BASE CLASSES

From table III, despite the drop in the expected phone accuracy values on the training data, performing MLLR adaptation on the MPE-SI model obtained significant[9] WER reductions. This shows that though the discrimination ability measured by the discriminative criterion may be limited by MLLR, the increased adaptation power on unseen test data can compensate for this and achieve overall improvements. Applying DMT in addition to MLLR is shown to yield further reduction in WER as it adds more discrimination ability to the adapted model. It can be observed that increasing the number of base-classes improved 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 showed slight gains after 1 DMT training iteration. The best performance was obtained using the 1000 base-classes. Performance with this system also improved with additional DMT iterations. Using three training iterations and 1000 base-classes[10], a significant absolute reduction in WER, 0.8%, was obtained over the MLLR adaptation[11]. It is worth noting that all the DMT performance changes are consistent with the criterion changes in table I. This implies that DMT can add discrimination ability without losing the already achieved adaptation power.

An interesting contrast is to see whether the gain of DMT comes from learning a criterion mapping or from simply increasing the number of transform parameters. To investigate this, 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, while the opposite is true for the ML-to-MPE DMT. In terms of recognition performance, the ML-to-ML mapping degraded the MLLR performance by 0.1%, which is statistically insignificant. This shows that the gain of DMT was not due to the increased number of transform parameters.

2) Sensitivity to hypothesis errors: One of the motivations for the use of DMT is that it should be less sensitive to errors in the adaptation hypothesis. To investigate this effect in detail, three forms of adaptation supervision were used to estimate the transforms. The baseline hypotheses used so far were generated by the unadapted MPE-SI model. As an alternative, this adapted model was used to generate lattices which were used in a lattice MLLR adaptation framework [14]. As alternative hypotheses in the lattice are used, this form of estimation should be less sensitive to hypothesis errors. Finally, the correct references were used for adaptation supervision. These three forms of hypotheses were used to generate MLLR transforms, to which DMT could then be applied. For the standard DLT estimation, the numerator was generated using the MLLR or lattice MLLR adapted MPE-SI model hypothesis respectively. For the reference supervision case, the correct transcription was used directly as the numerator for the DLT estimation[12].

Table IV gives the WER comparison using these different supervision hypotheses for adapting the MPE-SI system. As a general trend, lattice MLLR outperformed 1-best MLLR and using the reference as supervision always got significantly better performance. This shows that supervision quality does

[8]Note that DLTs also had a higher expected phone accuracy when using the correct transcriptions as the reference. However, expected phone accuracy is not the 1-best phone accuracy and does not necessarily consistently correlate with WER. The improved performance of the DLTs is felt to be because they are changing the posterior distributions, in particular sharpening posteriors for those phone correctly classified. Hence, decoding experiments are always required to illustrate the improvement in WER even with test set expected phone accuracy values.

[9]Wherever the term "*significant*" is used for experiment results, a pair-wise significance test was done using the Matched-Pair Sentence-Segment Word Error (MAPSSWE) test at a significance level of 5%, or 95% confidence [31].

[10]Using very large number of transform base-classes with appropriate smoothing values, such as 5600 base-classes, gave a WER of 26.1%. Though there is slight improvement, the computational cost was increased a lot. Therefore, in this paper, the setup of 1000 transforms was used.

[11]More DMT update iterations were also performed. The 4^{th} iteration gave a WER of 26.1%, indicating the convergence of the update. Therefore, in the following experiments, 3 iterations were used for DMT update.

[12]Note that in this case, there are no OOV words.

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 IV

%WER OF USING DIFFERENT SUPERVISION HYPOTHESES

have an impact on adaptation. For MLLR, using the reference obtained a 2.7% absolute gain over the 1-best hypothesis and 2.4% over the lattice supervision. This is similar to performance differences obtained with DMT. In contrast, for DLT, using the reference yielded a WER reduction of 5.1% absolute over 1-best and 4.9% over lattice based supervision. This is far larger than the difference for MLLR with and without DMT. This confirms that the DMT is less sensitive to the quality of supervision and is thus suitable for unsupervised adaptation. It is also interesting to note that with errorful 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 test set reference as heavily as DLT.

Table IV shows the overall robustness of MLLR+DMT with respect to the supervision hypothesis quality. It is also worth investigating the detailed pattern of the adaptation gains with respect to different WER regions of the supervision hypothesis. The WERs for each speaker in the test set were calculated. Depending on the WERs of the unadapted MPE-SI system, i.e., the 1-best supervision, speakers were grouped into several WER regions and each group has similar amount of data. For each group, the corresponding WER of MLLR and MLLR+DMT adaptation (first column of table IV) were then calculated. Figure 3 shows the adaptation gains over the unadapted system with respect to the supervision WERs.

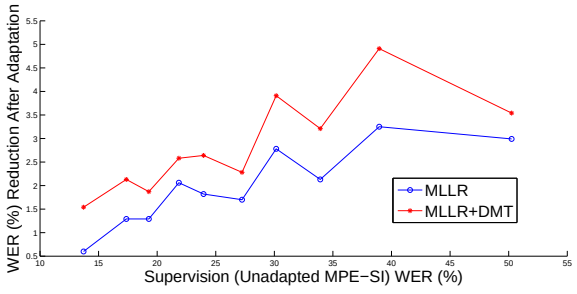

Fig. 3. %WER reduction of adaptation v.s. supervision %WER

From figure 3, as the WER of the supervision increases, the adaptation gains generally become larger other than at the very high WER. This is expected as for very good supervision, the room for improvement is small. In contrast, for high WER supervision where there is far larger possible improvement, the estimated parameters are less reliable. It is worth noting that MLLR+DMT had very similar trend of supervision sensitivity to MLLR adaptation and always outperformed MLLR adaptation. This is consistent with the observation from table IV.

As shown in equation (5), in the CMF framework, the final combined linear transform actually comes from transforming the ML estimated parameters rather than being directly estimated from data. Thus, in addition to illustrating the sensitivity of MLLR+DMT to the supervision hypothesis in figure 3, it is also interesting to show the sensitivity of the approximate DLT (MLLR+DMT) with respect to the WER improvement of applying MLLR transforms. The quality of the MLLR parameters may be measured by the gain of MLLR adaptation over the unadapted MPE-SI model. To investigate the relationship, the additional WER reduction of MLLR+DMT over the MLLR adaptation was also calculated. In this case, the speakers were grouped according to the MLLR adaptation gains. The results are plotted in figure 4.

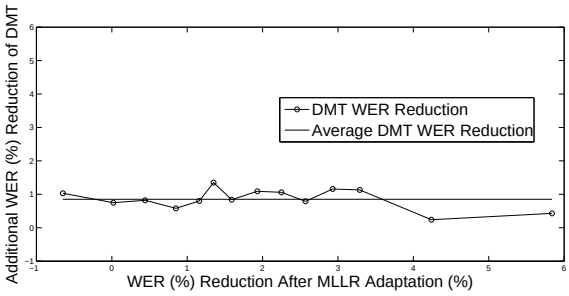

Fig. 4. Additional %WER reduction of DMT v.s. MLLR adaptation %WER reduction

From figure 4, the additional gains of MLLR+DMT over MLLR adaptation are relatively stable for different MLLR improvements. Statistical significance tests showed that almost all additional gains with different MLLR improvements were significant[13]. The average additional DMT gain is 0.8%, and the standard deviation of the DMT gains in figure 4 is 0.3% across all MLLR improvements whose range was from -0.7% to 5.8%. This shows that the gain from DMT is relatively independent of the gain from MLLR adaptation, which is consistent with the assumption that it is possible to factorise the effect of discrimination and adaptation. This independence also explains why MLLR+DMT has a similar robustness to supervision quality as MLLR. This advantage in robustness is felt to be a nature of the indirect estimation of DLTs in equation (5) because no *test* data is involved in the DMT estimation.

To further investigate the generalisation ability of DMT, the breakdown of the unsupervised adaptation (1-best hypothesis as supervision) performance is shown in table V.

Adaptation	Swbd	Fisher	Overall
MLLR	31.0	22.6	27.0
+ DMT	30.1	21.7	26.2

TABLE V

%WER BREAKDOWN BY CORPUS IN `eval03`

[13]Only one point was not significant. This had a gain of 0.24%.

As indicated in section IV-A, the Fisher corpus is not included in training, while Swbd is included. From table V, the gains of DMT are evenly distributed between the two corpora. This again indicates that DMT has a good generalisation to data of different types.

3) DMT in MLLR-based discriminative adaptive training: The previous experiments were based on the MPE-SI model. Using DMTs with MLLR-based MPE-SAT models was also investigated. The comparison between different adaptation approaches on MPE-SI and MPE-SAT models are shown in table VI 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 VI
%WER USING DMT WITH MPE-SI AND MPE-SAT MODELS

From table VI, MLLR with and without DMT, and the DLT on the MPE-SAT system both significantly outperformed the corresponding MPE-SI systems. The significant 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. DMTs can also be used during adaptive training. In [26], this was found to yield gains over using MLLR-based discriminative adaptive training.

V. CONCLUSION

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 maximum likelihood estimated speaker-dependent transforms to a more discriminative form. The final transform can be regarded as an approximation to a discriminative transform directly estimated on the adaptation data. As only ML-adapted speaker-specific transforms are estimated on the adaptation data, the transform is not very sensitive to errors in 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 and MLLR for both discriminatively trained SI and SAT models in unsupervised adaptation.

ACKNOWLEDGEMENT

This work was supported in part under the GALE program of the Defence Advanced Research Projects Agency (DARPA), Contract No. HR0011-06-C-0022. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred. Thanks to Lan Wang for the code for DLT estimation.

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] L. R. Bahl, P. F. Brown, P. V. D. Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. ICASSP*, Tokyo, 1986.
[4] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, Orlando, 2002.
[5] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, vol. 16, pp. 25–48, 2002.
[6] L. F. Uebel and P. C. Woodland, "Discriminative linear transforms for speaker adaptation," *Proc. ISCA ITR-Workshop on Adaptation Methods for Speech Recognition*, 2001.
[7] A. Gunawardana and W. J. Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," in *Proc. EuroSpeech*, Aalborg, 2001.
[8] J. McDonough, T. Schaaf, and A. Waibel, "On maximum mutual information speaker adapted training," in *Proc. ICASSP*, Orlean, 2002.
[9] L. Wang and P. C. Woodland, "MPE-based discriminative linear transforms for speaker adaptation," *Computer Speech & Language*, vol. 22, no. 3, pp. 256–272, 2008.
[10] 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.
[11] T. Anastasakos and S. V. Balakrishnan, "The use of confidence measures in unsupervised adaptation of speech recognisers," in *Proc. ICSLP*, Sydney, 1998.
[12] F. Wallhoff, D. Willett, and G. Rigoll, "Frame-discriminative and confidence-driven adaptation for LVCSR," in *Proc. ICASSP*, Istanbul, 2000.
[13] M. Padmanabhan, G. Saon, and G. Zweig, "Lattice-based unsupervised MLLR for speaker adaptation," *Proc. ISCA ITRW ASR2000*, pp. 128–131, 2000.
[14] L. F. Uebel and P. C. Woodland, "Speaker adaptation using lattice-based MLLR," *Proc. ISCA ITR-Workshop on Adaptation Methods for Speech Recognition*, 2001.
[15] A. Ljolje, "The AT&T LVCSR-2001 system," in *Proc. NIST LVCSR Workshop*, NIST, 2001.
[16] L. Wang and P. C. Woodland, "Discriminative adaptive training using the MPE criterion," in *Proc. ASRU*, St. Thomas, 2003.
[17] 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.
[18] 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*, Philadelphia, 2005.
[19] B. Zhang, S. Matsoukas, and R. Schwartz, "Recent progress on the discriminative region-dependent transform for speech feature extraction," in *Proc. INTERSPEECH*, Pittsburgh, 2006.
[20] P. C. Woodland, D. Pye, and M. J. F. Gales, "Iterative unsupervised adaptation using maximum likelihood linear regression," in *Proc. ICSLP*, Philadelphia, 1996.
[21] V. Valtchev, J. J. Odell, P. C. Woodland, and S. J. Young, "Lattice-based discriminative training for large vocabulary speech recognition systems," *Speech Communication*, vol. 22, no. 4, pp. 303–314, 1996.
[22] D. Povey, B. Kingsbury, L. Mangu, G. Saon, and G. Zweig, "fMPE: discriminatively trained features for speech recognition," in *Proc. ICASSP*, Philadelphia, 2005.
[23] B. Zhang, S. Matsoukas, and R. Schwartz, "Discriminatively trained region dependent feature transforms for speech recognition," in *Proc. ICASSP*, Toulouse, 2006.
[24] K. Yu, M. J. F. Gales, and P. C. Woodland, "Unsupervised discriminative adaptation using discriminative mapping transforms," in *Proc. ICASSP*, Las Vegas, 2008.

[25] D. Povey, “Discriminative training for large vocabulary speech recognition,” Ph.D. dissertation, Cambridge University, 2003.
[26] C. K. Raut, K. Yu, and M. J. F. Gales, “Adaptive training using discriminative mapping transforms,” in *Proc. INTERSPEECH*, Brisbane, 2008.
[27] C. J. Leggetter and P. C. Woodland, “Flexible speaker adaptation using maximum likelihood linear regression,” in *Proc. ARPA Spoken Language Technology Workshop*, 1995, pp. 104–109.
[28] 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.
[29] K. Yu, “Adaptive training for large vocabulary continuous speech recognition,” Ph.D. dissertation, Cambridge University, 2006.
[30] T. Anastasakos, J. Mcdonough, R. Schwartz, and J. Makhoul, “A compact model for speaker adaptive training,” in *Proc. ICSLP*, Philadelphia, 1996.
[31] L. Gillick and S. J. Cox, “Some statistical issues in the comparison of speech recognition,” in *Proc. ICASSP*, Glasgow, 1989.

Kai Yu Biography text here.

Mark Gales Biography text here.

Phil Woodland Biography text here.