Adaptive Training using Discriminative Mapping Transforms

C. K. Raut, K. Yu and M. J. F. Gales

Cambridge University Engineering Department
Trumpington St., Cambridge CB2 1PZ, U.K.
{ckr21,ky219,mjfg}@eng.cam.ac.uk

Abstract

Speaker adaptive training (SAT) is a useful technique for building speech recognition systems on non-homogeneous data. When combining SAT with discriminative training criteria, maximum likelihood (ML) transforms are often used for unsupervised adaptation tasks. This is because discriminatively estimated transforms are highly sensitive to errors in the supervision hypothesis. In this paper, speaker adaptive training based on discriminative mapping transforms (DMTs) is proposed. DMTs are speaker-independent discriminative transforms that are applied to ML-estimated speaker-specific transforms. As DMTs are estimated during training, they are not affected by errors in the supervision hypothesis. The proposed method was evaluated on an English conversational telephone speech task. It was found to significantly outperform the standard discriminative SAT schemes.

Index Terms: speech recognition, speaker adaptive training, discriminative training and adaptation

1. Introduction

Speech recognition systems are increasingly being built with *found* data such as broadcast news and conversational telephone speech recordings. This data often comes from multiple speakers, channel or acoustic conditions and is inherently non-homogeneous in nature. One of the techniques to build a speech recognition system on this non-homogeneous data is *speaker adaptive training* (SAT) [1, 2]. In SAT, the speech and speaker/environment variabilities are modelled separately. The speech variabilities are represented by a set of *canonical models*, whereas the non-speech variabilities are usually modelled by a set of linear transforms.

Originally, the canonical models and transforms were both estimated using maximum likelihood (ML) criteria. However, state-of-the-art systems use discriminative training criteria such as minimum phone error (MPE) [3]. The use of these discriminative criteria has also been investigated for estimating the canonical models and transforms in SAT [4, 5, 6]. These discriminative SAT schemes have been found to be useful for supervised adaptation tasks, however, little if any gain has been observed for unsupervised adaptation. This is because discriminative transforms are highly sensitive to errors in the supervision hypothesis. An alternative approach uses ML transforms with discriminatively estimated canonical models [5, 7]. This scheme is applicable to both supervised and unsupervised tasks and has been found to yield consistent gains. However, in the same way as discriminative training of canonical models leads to performance gains, if discriminative transforms can be *robustly* estimated, additional gains should be possible by using them in a SAT framework.

In this paper, a discriminative mapping transform (DMT) [8] based adaptive training scheme is proposed. DMTs are speaker-independent discriminatively trained transforms that are applied to speaker-specific ML transforms. As they are speaker-independent, they are estimated using training data. During recognition, these speaker-independent DMTs are applied to ML-estimated speaker-specific test-set transforms. As the DMTs are estimated during training, they are not affected by any errors in the supervision hypothesis. Thus they are appropriate for use in a SAT framework during training when error free transcripts are available, and testing when the supervision hypothesis may contain errors.

This paper is organised as follows: In Section 2, commonly used speaker adaptive training techniques are reviewed. Section 3 describes the proposed DMT-based discriminative adaptive training scheme. Experimental results from the conversational telephone speech task are presented in Section 4. The paper is concluded with a discussion of the results.

2. Speaker Adaptive Training

Speaker adaptive training (SAT) is a useful technique for building speech recognition systems on non-homogeneous data. In SAT, a set of speaker transforms, as well as a set of canonical models, is trained. They are estimated in an iterative fashion: first, the speaker-specific transforms are found; and then the canonical models are updated given these transforms. Several forms of transforms are possible for SAT [1, 2]. However, only the transforms of the same form as maximum likelihood linear regression (MLLR) are considered in this paper. Thus the transform for speaker s is applied to the mean $\boldsymbol{\mu}$ of the model parameters to obtain the adapted mean $\hat{\boldsymbol{\mu}}^{(s)}$ as

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

where $\mathbf{W}^{(s)} = [\mathbf{A}^{(s)} \ \mathbf{b}^{(s)}]$ is the linear transform for speaker s and $\boldsymbol{\xi} = [\boldsymbol{\mu}^T \ 1]^T$ is the extended mean vector. The rest of this section describes various forms of SAT implemented in this paper. An ML-based SAT scheme is initially described. Two forms of discriminative SAT are then detailed.

2.1. Maximum Likelihood SAT

In ML-based SAT with MLLR [1], both the transforms and the canonical models are estimated using an ML criterion. The following iterative procedure is used:

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.

1. **Initialise canonical model set and transforms.**
   The ML canonical model set, $\boldsymbol{\lambda}_{\mathtt{ml}}$, is initialised using the ML speaker-independent (SI) models, and speaker-specific transforms $\mathbf{W}_{\mathtt{ml}}^{(s)}$ for speaker s as $\mathbf{A}_{\mathtt{ml}}^{(s)} = \mathbf{I}, \mathbf{b}_{\mathtt{ml}}^{(s)} = \mathbf{0}$, where $\mathbf{I}$ is an identity matrix.
2. **Estimate transforms for each speaker.**
   MLLR [9] transforms $\mathbf{W}_{\mathtt{ml}}^{(s)}$ for each speaker s are found using

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

   where $\mathbf{O}^{(s)}$ and $\mathcal{H}^{(s)}$ are the observations and supervision for the adaptation data for speaker s, respectively and $\boldsymbol{\lambda}_{\mathtt{ml}}$ is the current canonical model set. Expectation maximisation (EM), which is an iterative scheme, is used to estimate the transform parameters. In this work, a single iteration of EM is used given the current speaker transform and canonical models.
3. **Update model parameters.**
   Given the set of estimated transforms, the model parameters are updated by maximising the log-likelihood over the training data from all speakers,

$$\boldsymbol{\lambda}_{\mathtt{ml}} = \arg\max_{\boldsymbol{\lambda}} \left\{ \sum_{s=1}^{S} \log p(\mathbf{O}^{(s)} | \mathcal{H}^{(s)}; \mathbf{W}_{\mathtt{ml}}^{(s)}, \boldsymbol{\lambda}) \right\} \quad (3)$$

   where S is the total number of speakers in the training data set. Again, a single iteration of EM estimation is used to update the model parameters.
4. **Go to step (2) unless converged.**

Canonical models estimated with SAT cannot be directly used for recognition. As unsupervised adaptation is being used in this work, an initial supervision hypothesis must be obtained. An SI model is often used for this purpose, trained with the same criterion as the SAT system being investigated [1]. Given this hypothesis, test-set speaker transforms are estimated in a similar fashion to the training procedure above, except that the model update stage in step (2) is omitted.

2.2. Discriminative SAT

The above ML-SAT approach has been found to yield gains over ML-SI systems. However state-of-the-art systems are commonly built using discriminative training criteria. A major concern with using discriminative criteria to estimate linear transforms is that the process is not robust to errors in the supervision hypothesis. This has led to two different forms of discriminative SAT (DSAT) being used. One based on ML speaker-specific transforms, the other on discriminatively estimated transforms. The implementations of both used in this paper are given below. MPE [3] is the form of the discriminative training criterion used in this work.

2.2.1. MLLR-based DSAT

This section describes the most commonly used form of DSAT for unsupervised adaptation tasks [5, 7]. In this approach, ML-based transforms are used in conjunction with the discriminatively trained models. During training, the ML-SAT scheme is initially run. A final set of speaker-specific MLLR transforms is estimated using the final ML canonical model set. These transforms are fixed and used for all subsequent discriminative canonical model updates. Thus in this scheme, the models are discriminatively updated, whereas the transforms use an ML criterion. As ML-based speaker-specific transforms are used, they should be relatively robust to errors in the supervision hypothesis

Once the transforms for each speaker are estimated, the set of canonical models are updated using the MPE criterion. The ML-SAT canonical models are used for initialisation. This may be expressed as

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

where $P(\mathcal{H} | \mathbf{O}^{(s)}; \mathbf{W}_{\mathtt{ml}}^{(s)}, \boldsymbol{\lambda})$ is the posterior probability of hypothesis $\mathcal{H}$ for the given observation and transform for speaker s, and $\mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)})$ is the phone-level loss function of $\mathcal{H}$ for the given supervision $\mathcal{H}^{(s)}$. The details of the canonical model estimation are described in [6].

The testing procedure has the same starting point as the ML-SAT scheme. The ML-SAT test procedure is first run to obtain initial ML speaker transforms. Additional ML-based transform estimations are then run using the final MPE-trained canonical models. In this work, two iterations were used; further iterations were found to yield no additional gains.

2.2.2. DLT-based DSAT

As previously mentioned, it is possible to estimate both the transforms and the canonical models using discriminative criteria. Again for training, the ML-SAT scheme is initially run and a set of ML speaker transforms estimated using the final ML canonical models. Starting with these ML canonical models and speaker transforms, steps (2) to (4) of the ML-based scheme are repeated. However, the transforms are now estimated using the MPE criterion,

$$\mathbf{W}_{\mathtt{d}}^{(s)} = \arg\min_{\mathbf{W}} \left\{ \sum_{\mathcal{H}} P(\mathcal{H} | \mathbf{O}^{(s)}; \mathbf{W}, \boldsymbol{\lambda}_{\mathtt{d}}) \mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)}) \right\}. \quad (5)$$

Transforms estimated in this fashion will be referred to as discriminative linear transforms (DLTs). For additional information about DLT estimation see [6]. The canonical model parameters are estimated using Equation 4 except that the ML speaker transforms are replaced by the DLTs.

The testing procedure again uses the ML-SAT testing to get initial MLLR speaker transforms. A modified version of the DLT-SAT training procedure is then run, omitting the model-update stage. This yields a set of speaker-specific test-set DLTs, which are then used for recognition.

As previously discussed, discriminative transforms are sensitive to errors in the supervision hypothesis. During training, the DLTs are estimated using the reference transcripts, so there are no supervision errors. If used in a supervised adaptation mode, DLTs can be robustly estimated and reductions in the error rate are obtained [6]. However, this is not the case for unsupervised adaptation. To reduce the impact of hypothesis errors, it is possible to use confidence scores and lattice-based adaptation as described in [6, 8]. Though these approaches yield slightly greater robustness to hypothesis errors, the improvements over MLLR-based DSAT are still normally small.

[1] Hence, the discriminative SAT experiments described in the paper use an MPE-SI model to generate the supervision hypothesis.

3. DMT-based DSAT

Recently, discriminative mapping transforms (DMTs) [8] have been proposed for estimating discriminative transforms. A DMT is a discriminatively estimated speaker-independent transform based on speaker-specific ML transforms. As DMTs are speaker-independent, the same transforms can be used for the training and test data. There is no need to estimate speaker-specific discriminative transforms on the test data. Thus the sensitivity to errors in the supervision hypothesis that has a severe impact on the performance of DLTs for unsupervised adaptation should not be a problem. In this section, the theory behind DMTs is discussed. This is followed by a description of how DMTs can be used for discriminative adaptive training.

A general form of the DMT [8] may be expressed as

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

where $\mathbf{W}_{\mathtt{d}}^{(s)}$ is the final discriminative-like speaker transform, $\mathbf{H}_{\mathtt{d}}$ and $\mathbf{c}_{\mathtt{d}}$ are the speaker-independent parameters of the DMT and $\mathbf{W}_{\mathtt{ml}}^{(s)}$ is the speaker-specific ML-transform. The operator 'vec()' maps a matrix to a vector. In this work, a simpler form of DMT is used where $\mathbf{H}_{\mathtt{d}}$ is restricted to be block-diagonal with each block being tied and $\mathbf{c}_{\mathtt{d}}$ is restricted to yield a bias on the mean. In this case, the final adapted mean obtained using MLLR-based DMT adaptation (MLLR+DMT) may be expressed as

$$\hat{\boldsymbol{\mu}}^{(s)} = \mathbf{A}_{\mathtt{d}}\hat{\boldsymbol{\mu}}_{\mathtt{ml}}^{(s)} + \mathbf{b}_{\mathtt{d}} = \mathbf{W}_{\mathtt{d}}\hat{\boldsymbol{\xi}}_{\mathtt{ml}}^{(s)} \quad (7)$$

where $\hat{\boldsymbol{\xi}}_{\mathtt{ml}}^{(s)} = [\hat{\boldsymbol{\mu}}_{\mathtt{ml}}^{(s)T} \ 1]^T$, and $\mathbf{W}_{\mathtt{d}} = [\mathbf{A}_{\mathtt{d}} \ \mathbf{b}_{\mathtt{d}}]$ is the DMT transform. The transformed mean $\hat{\boldsymbol{\mu}}_{\mathtt{ml}}^{(s)} = \mathbf{W}_{\mathtt{ml}}^{(s)}\boldsymbol{\xi}$ is the MLLR adapted mean with transform $\mathbf{W}_{\mathtt{ml}}^{(s)}$ estimated by maximising likelihood as given in Equation 2.

The parameters of the DMT are estimated by using

$$\mathbf{W}_{\mathtt{d}} = \arg\min_{\mathbf{W}} \left\{ \sum_{s,\mathcal{H}} P(\mathcal{H}|\mathbf{O}^{(s)}; \mathbf{W}, \mathbf{W}_{\mathtt{ml}}^{(s)}, \boldsymbol{\lambda}_{\mathtt{d}}) \mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)}) \right\}. \quad (8)$$

This form of optimisation is related to the DLT optimisation in Equation 5. The parameter estimation for the DMT turns into a slightly modified version of DLT transform estimation [8]. In the same way as MLLR, DMTs can make use of multiple regression classes. An interesting aspect of DMTs is that the number of transform parameters can be made very large compared to the number used for speaker-specific linear transforms. This is because the parameters of the DMT are estimated using all the acoustic model training data, rather than just the data associated with a specific speaker. In this work, a thousand speaker-independent DMT transforms are used, compared to two (one speaker and one silence) for MLLR and DLT. DMTs have previously been applied to both discriminatively trained SI models and the MLLR-based DSAT models described in Section 2. The rest of this section describes how DMT-based DSAT can be implemented.

The starting point for training a DMT-based DSAT system is the same as the other DSAT approaches, the ML-SAT models and transforms are used. A final set of speaker-specific MLLR transforms is estimated and used as the initial set of speaker-specific transforms for the DMT build. In the same fashion as the DLT-based DSAT scheme, steps (2) to (4) of the ML-SAT scheme are repeated, where step (2) is replaced by an MLLR+DMT transform estimation and step (3) by a discriminative canonical model update.

Rather than using Equation 5, the transform estimation consists of two stages. First, given the current DMTs and speaker-specific MLLR transforms, the set of MLLR transforms is estimated using Equation 2. Given this new set of MLLR transforms and the current DMT, a new DMT is estimated using Equation 8.

The DMT-based DSAT canonical model is estimated using

$$\boldsymbol{\lambda}_{\mathtt{d}} = \arg\min_{\boldsymbol{\lambda}} \left\{ \sum_{s,\mathcal{H}} P(\mathcal{H}|\mathbf{O}^{(s)}; \mathbf{W}_{\mathtt{d}}, \mathbf{W}_{\mathtt{ml}}^{(s)}, \boldsymbol{\lambda}) \mathcal{L}(\mathcal{H}, \mathcal{H}^{(s)}) \right\}. \quad (9)$$

Using Equation 7, it is possible to combine the effects of the DMT and MLLR transform into a single linear transform of the means. Thus the standard MPE canonical model estimation schemes can be used. After the final canonical model update, additional iterations of MLLR and DLT estimation are performed. As all the available training data is used to estimate the DMT, it is possible to run more iterations of parameter estimation than, for example, the DLT-based scheme. In this work three additional iterations are used.

During the recognition stage, the procedure for estimating the test-set speaker transforms is similar to the DLT-based DSAT scheme. Rather than estimating a new DLT at each iteration, a new MLLR transform is estimated using the current canonical models and MLLR+DMT for that iteration. The DMT used is the one obtained during training, at the equivalent iteration.

4. Experimental Results

The evaluation experiments were conducted on a large vocabulary English conversational telephone speech (CTS) task. The acoustic training data consisted of about 296 hours of speech from 5446 speakers. It was taken from LDC Callhome English (`che`), Switchboard (`Swbd`) and Switchboard-Cellular (`swCell`) corpora. The `eval03` test-set was used for evaluation. It consists of about 6 hours of data from 144 speakers, taken from `Swbd` and `Fisher` corpora. The speech data was parameterised using 12 PLP Cepstral coefficients plus the 0th order (`C0`) coefficient. First, second and third derivatives were also appended. An heteroscedastic linear discriminant analysis transform was used to project this 52-dimensional feature-vector down to 39 dimensions. Speaker-level Cepstral mean and variance as well as a vocal track length normalisation was applied to the features. All HMM systems were based on state-clustered triphones with 6k distinct states. Each speech state had an average of 16 Gaussian components (32 Gaussian components for the silence models). The MPE criterion was used to estimate the discriminative acoustic models, the DLTs and DMTs. A trigram language model trained on 1044M words and a 58k words multiple pronunciation dictionary were used for decoding. Where significant differences in the performance are mentioned, this was assessed using the NIST pair-wised significance tests.

The ML-SAT and DSAT models were built using four iterations of ML training and then for the DSAT systems four MPE training iterations. MLLR-style adaptation (a linear transform of the means) was used in all experiments. All speaker-specific transforms used two base classes: one for speech and another for silence. For the DMT, 1000 regression base classes were used. As the CTS task is an unsupervised adaptation task, an initial hypothesis is required. This was obtained from an MPE-trained SI model and had a word error rate (WER) of 29.2%.

#iter	DSAT Transform		
	MLLR	MLLR+DMT	DLT
0	0.783	0.803	0.821
1	0.817	0.840	0.863
2	0.836	0.861	0.887
3	0.848	0.874	0.902

Table 1: Expected phone correctness (one minus normalised MPE criterion) for different DSAT schemes during training

Table 1 shows the expected phone correctness (one minus normalised MPE criterion) for each of the DSAT schemes. At each iteration, the criterion value was obtained during the update of the model parameters. Thus the zeroth iteration shows the criterion after applying an MLLR, MLLR+DMT, or DLT to the final ML-SAT acoustic models. All schemes show an increase in the correctness as the number of iterations increases. The lowest correctness value was obtained with the MLLR-based DSAT scheme. Using a MLLR+DMT during adaptive training shows consistent gains. However, the largest correctness values were obtained with the DLTs. This indicates that the DLTs perform better on the training data than the other schemes.

Training Scheme	Transform		Supervision	
	Training	Testing	ref	hyp
SI (hyp)	—	—	—	29.2
SI	—	MLLR	24.3	27.0
DSAT	MLLR	MLLR	23.6	26.4
	DLT	DLT	18.4	28.1
	MLLR+DMT	MLLR+DMT	22.5	25.3

Table 2: Comparison of WER% of different DSAT schemes

The recognition performance on the `eval03` test-set for the various schemes is shown in Table 2. All systems used either the hypothesis from an MPE SI system, labelled *hyp* in the table, or the reference transcriptions, *ref*. Using MLLR adaptation on the SI system with the hypothesis shows large gains in performance, a reduction in WER of 2.2% absolute. MLLR-based DSAT gave an additional 0.6% absolute reduction in WER using the hypothesis. If the reference was used to estimate the transform instead, additional consistent gains are observed with both systems compared to using the hypothesis. The most striking result is the difference in performance of the DLT-based system, between using the reference or the hypothesis for the supervision. Using the reference, the DLT-based system yielded the best performance, whereas it had the worst performance among all DSAT schemes when using the hypothesis. This illustrates the sensitivity of DLTs to errors in the hypothesis. On the other hand, the DMT-based DSAT gave the best performance when using the hypothesis. A statistically significant 1.1% absolute gain over the MLLR-based, standard, DSAT approach was obtained.

Training Scheme	Transform		Supervision
	Training	Testing	hyp
SI	—	MLLR+DMT	26.2
DSAT	MLLR	MLLR+DMT	25.6
	DLT		25.6
	MLLR+DMT		25.3

Table 3: Comparison of WER% of different DSAT models with MLLR+DMT as testing transforms

From Table 2, using MLLR+DMT appears to be a good candidate for test-set adaptation. Therefore, the use of MLLR+DMT as a testing transform with other DSAT models was investigated. Table 3 shows the performance of the various DSAT schemes (and the MPE-SI model) using MLLR+DMT for test-set adaptation. As previously observed in [8], using DMTs in addition to MLLR yields a gain of about 0.8% absolute for both the MPE-SI model and the MLLR-based DSAT model compared to MLLR. However, the performance of both systems is still significantly worse than the DMT-based DSAT system. Using MLLR+DMT with the DLT-based DSAT system shows large gains over using the DLT as the test-set transform. Despite the DLT-based DSAT system having the best criterion on the training data, it is still significantly worse than the DMT-based DSAT system even with the robust MLLR+DMT test-set transform. This is felt to be because of the inconsistency between the training transforms, DLT, and the test-set transforms, MLLR+DMT.

5. Conclusion

This paper has presented a speaker adaptive training scheme based on discriminative mapping transforms (DMTs). DMTs are discriminatively trained speaker-independent transforms that are estimated during training based on maximum likelihood speaker-specific transforms. As the DMTs are speaker-independent, the same transforms can be used during recognition. Thus despite being discriminative in nature, they are not highly sensitive to the hypothesis errors, which is a known problem for discriminative linear transforms. This means they are useful for both supervised and unsupervised adaptation tasks. In this paper, DMTs with speaker-specific MLLR transforms are used in a discriminative speaker adaptive training framework. Update formulae for both the canonical models and DMT transforms are discussed. They may be implemented as simple modifications to the standard MLLR-based DSAT scheme and the DLT estimation. DMT-based adaptive training was evaluated on a large vocabulary English conversation telephone speech task, which is an unsupervised adaptation task. DMT-based adaptive training was found to significantly outperform standard approaches to discriminative speaker adaptive training.

6. References

[1] T. Anastasakos, J. Mcdonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

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

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

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

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

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

[8] K. Yu, M. J. F. Gales, and P. C. Woodland, "Unsupervised discriminative adaptation using discriminative mapping transforms," in *Proc. ICASSP*, 2008, pp. 4273–4276.

[9] 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.