



# Prior Information for Rapid Speaker Adaptation

C. Breslin, K.K. Chin, M.J.F. Gales, K. Knill, H. Xu

Toshiba Research Europe Ltd, Cambridge, UK

{catherine.breslin, kkchin, mark.gales, kate.knill, haitian.xu}@cr1.toshiba.co.uk

## Abstract

Rapidly adapting a speech recognition system to new speakers using a small amount of adaptation data is important to improve initial user experience. In this paper, a count-smoothing framework for incorporating prior information is extended to allow for the use of different forms of dynamic prior and improve the robustness of transform estimation on small amounts of data. Prior information is obtained from existing rapid adaptation techniques like VTLN and PCMLLR. Results using VTLN as a dynamic prior for CMLLR estimation show that transforms estimated on just one utterance can yield relative gains of 15% and 46% over a baseline gender independent model on two tasks.

**Index Terms:** automatic speech recognition, speaker adaptation, VTLN, prior knowledge

## 1. Introduction

In many speech recognition applications, such as telephone dialogue systems and in-car satnavs, rapid adaptation to new speakers is vital as users' perception of a system can depend heavily on their initial experiences. However, adapting to a new speaker or environment given limited data remains a challenge.

The standard successful linear speaker adaptation techniques such as MLLR [1] and CMLLR [2] rely on sufficient adaptation data being available. Although they can be used with limited data by restricting the number of transform parameters, a robust estimate is not guaranteed. Existing rapid adaptation techniques such as CAT-CMLLR [3] and vocal tract length normalisation (VTLN) [7, 8, 9] for speaker adaptation, and PCMLLR [10, 11] for noise robustness only require a small number of parameters to be estimated. This enables them to perform well on limited data but they saturate quickly with, typically, only small gains seen over the baseline system.

Several authors have proposed using prior information to improve the robustness of adaptation transform estimates. MAP approaches such as MAP linear regression [4] have used prior information in the form of a distribution over transforms to constrain linear transform estimation. In [4], the prior distribution over transforms is estimated from training data. This prior distribution is static, not changing over utterances. Structured MAPLR (SMAPLR) is related and obtains prior information from the test utterances themselves [5, 6]. A regression class tree is used and the prior transform for any node is the transform estimated at its parent node. Thus transforms estimated higher up the tree using more frames are propagated down the tree and used as priors to obtain more robust estimates for transforms at nodes with few observations.

Rapid adaptation techniques such as VTLN and PCMLLR are potential sources of prior information for estimating adaptation transforms. Since the strength of these methods is their ability to adapt quickly to new data the use of such knowledge sources inevitably leads to a dynamic prior, which can change with each utterance. This is in contrast to previous work which

has used static prior information. The use of a dynamic prior estimated online should be beneficial as it will be better matched to the target environment.

In [11], a count-smoothing framework is used to combine standard adaptive statistics with dynamic prior information from PCMLLR. This paper builds on that framework and uses rapid adaptation methods as a dynamic prior to obtain more robust estimates of transforms with small amounts of adaptation data.

The paper is arranged as follows. VTLN and PCMLLR for rapid adaptation are discussed in section 2. A count-smoothing framework for incorporating dynamic prior information and two implementations of this framework are presented in section 3. Results obtained using one of the implementations with VTLN as a prior are presented in section 4, before conclusions are drawn in section 5.

## 2. Rapid Adaptation

This section discusses two forms of rapid adaptation which make use of knowledge about the mismatch due to both speaker variation and noise environment to robustly estimate adaptation transforms.

### 2.1. Speaker Adaptation - Linear VTLN

In quantised linear vocal tract length normalisation (VTLN) [8, 9] a set of possible linear transforms  $\mathbf{W}^{(\alpha)} = [\mathbf{b}^{(\alpha)} \mathbf{A}^{(\alpha)}]$  is pre-computed for a discrete number of different vocal tract lengths. These transforms are deterministic and are based on the known effect of vocal tract length on MFCC calculation, where the parameter  $\alpha$  represents the degree of frequency warping. Linear VTLN applies a transform to the feature vectors

$$p(\mathbf{o}_t|m) = |\mathbf{A}^{(\alpha)}| \mathcal{N}(\mathbf{A}^{(\alpha)} \mathbf{o}_t + \mathbf{b}^{(\alpha)}; \boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)}) \quad (1)$$

where component  $m$  has mean and variance  $\boldsymbol{\mu}^{(m)}$  and  $\boldsymbol{\Sigma}^{(m)}$ .

VTLN adaptation is rapid as only  $\alpha$  must be estimated, which can be done based on the auxiliary function  $\mathcal{Q}(\alpha, \hat{\alpha})$ . To estimate  $\alpha$  given the current value  $\hat{\alpha}$

$$\alpha = \arg \max_{\hat{\alpha}} \{ \mathcal{Q}(\hat{\alpha}, \hat{\alpha}) \} \quad (2)$$

where

$$\mathcal{Q}(\alpha, \hat{\alpha}) = T \log |\mathbf{A}^{(\alpha)}| - \frac{1}{2} \left\{ \sum_r \sum_{i=1}^D \mathbf{w}_i^{(\alpha)} \mathbf{G}_{adi}^{(r)} \mathbf{w}_i^{(\alpha)\top} - 2 \mathbf{k}_{adi}^{(r)} \mathbf{w}_i^{(\alpha)\top} \right\} \quad (3)$$

and the associated linear transform  $\mathbf{W}^{(\alpha)}$  is selected using a brute-force search over the set of pre-computed transforms. The above expression assumes multiple regression classes, indexed using  $r$ , and a bias term in the transformation. For VTLN,  $\alpha$  is normally global and there is no bias term, i.e.  $\mathbf{b}^{(\alpha)} = \mathbf{0}$ .

The *adaptive* statistics needed to estimate  $\alpha$ ,  $\mathbf{G}_{adi}^{(r)}$  and  $\mathbf{k}_{adi}^{(r)}$ , are accumulated from adaptation data

$$\mathbf{G}_{\text{adi}}^{(r)} = \sum_{m \in r} \frac{1}{\sigma_i^{(m)2}} \sum_{t=1}^T \gamma_t^{(m)} \begin{bmatrix} 1 & \mathbf{o}_t^\top \\ \mathbf{o}_t & \mathbf{o}_t \mathbf{o}_t^\top \end{bmatrix} \quad (4)$$

$$\mathbf{k}_{\text{adi}}^{(r)} = \sum_{m \in r} \frac{\mu_i^{(m)}}{\sigma_i^{(m)2}} \sum_{t=1}^T \gamma_t^{(m)} \begin{bmatrix} 1 & \mathbf{o}_t^\top \end{bmatrix} \quad (5)$$

where  $\gamma_t^{(m)}$  is the posterior probability of component  $m$  generating the observation at time  $t$  using the current value of the transform  $\hat{\alpha}$ . Estimates of  $\hat{\alpha}$  can be iterated to achieve the maximum likelihood estimate of  $\alpha$ . Linear VTLN is constrained by the frequency warping function and the discrete values of  $\alpha$  chosen to precompute the set of transforms  $\{\mathbf{W}^{(\alpha)}\}$ . Hence, as the amount of adaptation data increases, VTLN does not improve and so its effect is limited.

## 2.2. Noise robustness - PCMLLR

Model-based noise robustness techniques can be used for rapid adaptation by making use of a noise mismatch function to model the effect of noise on speech. It is assumed that the observation  $\mathbf{o}_t$  is based on a clean speech observation  $\mathbf{x}_t$ , with mean and covariance matrix  $\boldsymbol{\mu}^{(m)}$  and  $\boldsymbol{\Sigma}^{(m)}$ , and additive noise  $\mathbf{n}_t$  with mean and covariance  $\boldsymbol{\mu}_n$  and  $\boldsymbol{\Sigma}_n$ . Convolutional noise is ignored for this discussion. For example, in VTS model-based compensation [12] the final distribution is given by

$$p(\mathbf{o}_t|m) = \mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_{\text{vts}}^{(m)}, \boldsymbol{\Sigma}_{\text{vts}}^{(m)}) \quad (6)$$

and the corrupted speech mean  $\boldsymbol{\mu}_{\text{vts}}^{(m)}$  can be found from

$$\boldsymbol{\mu}_{\text{vts}}^{(m)} = \mathbf{C} \log(\exp(\mathbf{C}^{-1} \boldsymbol{\mu}^{(m)}) + \exp(\mathbf{C}^{-1} \boldsymbol{\mu}_n)) \quad (7)$$

where  $\mathbf{C}$  is the DCT. A similar expression can be derived for the noise covariance  $\boldsymbol{\Sigma}_{\text{vts}}^{(m)}$ . Compensation is rapid as only estimation of the noise distribution parameters,  $\boldsymbol{\mu}_n$  and  $\boldsymbol{\Sigma}_n$  is required.

It is possible to approximate a compensated acoustic model such as that in equation 6 by a linear transform using the predictive linear transform framework [10, 11]. The predictive transform parameters are obtained by minimising the KL divergence between a CMLLR adapted distribution and a target distribution. The statistics that are used to estimate the predictive transforms are

$$\mathbf{G}_{\text{pri}}^{(r)} = \sum_{m \in r} \frac{\gamma^{(m)}}{\sigma_i^{(m)2}} \begin{bmatrix} 1 & \mathcal{E}\{\mathbf{o}^\top|m\} \\ \mathcal{E}\{\mathbf{o}|m\} & \mathcal{E}\{\mathbf{o}\mathbf{o}^\top|m\} \end{bmatrix} \quad (8)$$

$$\mathbf{k}_{\text{pri}}^{(r)} = \sum_{m \in r} \frac{\gamma^{(m)} \mu_i^{(m)}}{\sigma_i^{(m)2}} \begin{bmatrix} 1 & \mathcal{E}\{\mathbf{o}^\top|m\} \end{bmatrix} \quad (9)$$

where  $\mathcal{E}\{\mathbf{o}|m\}$  and  $\mathcal{E}\{\mathbf{o}\mathbf{o}^\top|m\}$  are estimated from the target distribution in equation 6, and component occupancies,  $\gamma^{(m)}$ , are obtained from the training data. Thus

$$\mathcal{E}\{\mathbf{o}|m\} = \boldsymbol{\mu}_{\text{vts}}^{(m)} \quad ; \quad \mathcal{E}\{\mathbf{o}\mathbf{o}^\top|m\} = \boldsymbol{\Sigma}_{\text{vts}}^{(m)} + \boldsymbol{\mu}_{\text{vts}}^{(m)} \boldsymbol{\mu}_{\text{vts}}^{(m)\top} \quad (10)$$

PCMLLR uses these predictive statistics to estimate transforms directly from the standard CMLLR formulae. Decoding is then based on

$$p(\mathbf{o}_t|m) = |\mathbf{A}_{\text{pr}}^{(r_m)}| \mathcal{N}(\mathbf{A}_{\text{pr}}^{(r_m)} \mathbf{o}_t + \mathbf{b}_{\text{pr}}^{(r_m)}; \boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)}) \quad (11)$$

where component  $m$  belongs to regression class  $r_m$ . In the limit, as the number of regression classes increases to the number of components, the same compensation as VTS is possible.

## 3. Incorporating Prior Knowledge

The methods described in the previous section allow adaptation to new speakers or environments using small amounts of data, but are only approximate. Linear VTLN relies on a limited set of quantised transforms which quickly saturate, while PCMLLR relies on an accurate mismatch function being defined. An alternative is to use these rapid adaptation approaches as prior information for CMLLR transform estimation. This allows fast, robust estimation of adaptation transforms without the limitations of the previous approaches.

A conjugate prior distribution for CMLLR is not possible, so a count-smoothing framework for incorporating prior knowledge is used instead, as in [11]. Statistics for estimating the CMLLR transform,  $\mathbf{G}_i^{(r)}$  and  $\mathbf{k}_i^{(r)}$ , are based on interpolating adaptive and prior statistics, given by

$$\mathbf{G}_i^{(r)} = \mathbf{G}_{\text{adi}}^{(r)} + \tau \frac{\mathbf{G}_{\text{pri}}^{(r)}}{\sum_{m \in r} \gamma^{(m)}} \quad (12)$$

$$\mathbf{k}_i^{(r)} = \mathbf{k}_{\text{adi}}^{(r)} + \tau \frac{\mathbf{k}_{\text{pri}}^{(r)}}{\sum_{m \in r} \gamma^{(m)}} \quad (13)$$

The prior statistics  $\mathbf{G}_{\text{pri}}^{(r)}$  and  $\mathbf{k}_{\text{pri}}^{(r)}$  are normalised so that they effectively contribute  $\tau$  frames to the final statistics. For transform estimation, the total occupancy count for a regression class  $\beta^{(r)} = \sum_{m \in r} \sum_{t=1}^T \gamma_t^{(m)} + \tau$ . As more data becomes available, the adaptive CMLLR statistics  $\mathbf{G}_{\text{adi}}^{(r)}$  and  $\mathbf{k}_{\text{adi}}^{(r)}$  will dominate, but for small amounts of data the prior statistics are more important.

Under this count-smoothing approach, the prior is not constrained to be static. It may be dynamic, and change across utterances. In this work, appropriate prior statistics are obtained by using the predictive statistics in equations 8 and 9 with a target distribution incorporating prior knowledge. Two frameworks for making use of this approach are discussed below.

### 3.1. Direct Transform Estimation

For direct transform estimation, decoding is based only on the transform  $\begin{bmatrix} \mathbf{b}^{(r_m)} & \mathbf{A}^{(r_m)} \end{bmatrix}$  estimated from smoothed statistics  $p(\mathbf{o}_t|m) = |\mathbf{A}^{(r_m)}| \mathcal{N}(\mathbf{A}^{(r_m)} \mathbf{o}_t + \mathbf{b}^{(r_m)}; \boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)}) \quad (14)$

For this form of model, all statistics for estimating the transform are based on the original feature-space and model parameters.

Appropriate prior statistics depend on the form of prior being used. If noise model compensation is used as a prior then the prior statistics are the same as those in equations 8, 9 and 10. If a linear transform prior is used, such as VTLN, then equation 1 can be rewritten as

$$p(\mathbf{o}_t|m) = \mathcal{N}(\mathbf{o}_t; \mathbf{A}^{(\alpha)-1}(\boldsymbol{\mu}^{(m)} - \mathbf{b}^{(\alpha)}), \mathbf{A}^{(\alpha)-1} \boldsymbol{\Sigma}^{(m)} \mathbf{A}^{(\alpha)-\top}) \quad (15)$$

This can be used as the target distribution to accumulate the prior statistics from equations 8 and 9 where

$$\mathcal{E}\{\mathbf{o}|m\} = \mathbf{A}^{(\alpha)-1}(\boldsymbol{\mu}^{(m)} - \mathbf{b}^{(\alpha)}) \quad (16)$$

It is not necessary to transform each model component using the prior. Equivalently, statistics that are pre-cached for each regression class can be transformed by the prior. For example the following statistics, based on the original model set, can be cached [13]

$$\mathbf{k}_{\text{cai}}^{(r)} = \sum_{m \in r} \frac{\gamma^{(m)} \mu_i^{(m)}}{\sigma_i^{(m)2}} \begin{bmatrix} 1 & \boldsymbol{\mu}^{(m)\top} \end{bmatrix} \quad (17)$$

Considering only the second term in equation 9, the predictive statistics,  $\mathbf{k}_{pri}^{(r)}$  can be written as

$$\mathbf{A}^{(\alpha)-1} \left( \left( \sum_{m \in r} \frac{\gamma^{(m)} \mu_i^{(m)}}{\sigma_i^{(m)2}} \boldsymbol{\mu}^{(m)} \right) - \left( \sum_{m \in r} \frac{\gamma^{(m)} \mu_i^{(m)}}{\sigma_i^{(m)2}} \right) \mathbf{b}^{(\alpha)} \right) \quad (18)$$

which is equivalent to the cached statistics  $\mathbf{k}_{cai}^{(r)}$  above transformed by the prior  $\begin{bmatrix} \mathbf{b}^{(\alpha)} & \mathbf{A}^{(\alpha)} \end{bmatrix}$ , and is true for all elements of  $\mathbf{k}_{pri}^{(r)}$  and  $\mathbf{G}_{pri}^{(r)}$ .

This efficient caching allows a dynamic prior such as VTLN to be used with very little additional computational cost. The adaptive statistics accumulated from adaptation data are used to select the VTLN prior for each utterance. The prior is then used to transform the cached statistics  $\mathbf{k}_{cai}^{(r)}$  and  $\mathbf{G}_{cai}^{(r)}$  and obtain the prior statistics. The same adaptive statistics that were used to select the prior are then interpolated with the prior statistics to estimate a new transform.

### 3.2. Cascades of Transforms

For situations where the prior is more complex than the smoothed transform being estimated, whether in terms of the structure of the transform or number of regression classes, direct estimation of the transform may degrade performance. To address this, the rapidly estimated prior transform can also be used as a *parent* transform in the count-smoothing framework.

For the case of using a feature space linear transform as parent, such as linearised VTLN, the following decoding expression is used

$$p(\mathbf{o}_t | m) = |\mathbf{A}^{(r_m)} \mathbf{A}^{(\alpha)}| \quad (19)$$

$$\mathcal{N}(\mathbf{A}^{(r_m)} (\mathbf{A}^{(\alpha)} \mathbf{o}_t + \mathbf{b}^{(\alpha)}) + \mathbf{b}^{(r_m)}; \boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)})$$

A similar expression can be derived using PCMLLR or any other form of feature transform prior as a parent. Now, both adaptive and prior statistics need to be obtained in the space defined by the parent,  $\mathbf{A}^{(\alpha)}$  and  $\mathbf{b}^{(\alpha)}$ . The correct adaptive statistics can be obtained by transforming previously accumulated adaptive statistics of equations 4 and 5 in the same fashion as the cached statistics above. Alternatively, adaptive statistics are simply accumulated after the parent transform has been applied.

In the parent transform domain the prior statistics,  $\mathbf{k}_{pri}^{(r)}$ , are simply the predictive statistics in 9 with an identity matrix prior. For the feature space linear transform parent case this is equivalent to the cached statistics in 17.

If the parent transform to be used is a general linear transform applied to the model parameters, the appropriate prior statistics are more complicated. For example, when using VTS as a prior, in place of equation 19 decoding would be based on

$$p(\mathbf{o}_t | m) = |\mathbf{A}^{(r_m)}| \mathcal{N}(\mathbf{A}^{(r_m)} \mathbf{o}_t + \mathbf{b}^{(r_m)}; \boldsymbol{\mu}_{vts}^{(m)}, \boldsymbol{\Sigma}_{vts}^{(m)}) \quad (20)$$

As both the mean and covariance matrix for each component have changed, the adaptive and prior statistics are also altered. For example the adaptive  $\mathbf{k}_{adi}^{(r)}$  and prior  $\mathbf{k}_{pri}^{(r)}$  would be given by

$$\mathbf{k}_{adi}^{(r)} = \sum_{m \in r} \frac{\mu_{vtsi}^{(m)}}{\sigma_{vtsi}^{(m)2}} \sum_{t=1}^T \gamma_t^{(m)} \begin{bmatrix} 1 & \mathbf{o}_t^T \end{bmatrix} \quad (21)$$

$$\mathbf{k}_{pri}^{(r)} = \sum_{m \in r} \frac{\gamma^{(m)} \mu_{vtsi}^{(m)}}{\sigma_{vtsi}^{(m)2}} \begin{bmatrix} 1 & \boldsymbol{\mu}_{vts}^{(m)T} \end{bmatrix} \quad (22)$$

As for the feature space transform discussed above, the prior statistics in the space of the VTS compensated model set use an identity matrix prior. However, this change to the prior statistics means they cannot be efficiently cached, as is the case when using a feature space transform as prior.

## 4. Experimental Results

This section presents results obtained using the methods described in section 3. The linear transform version of VTLN in [9] is used as a convenient prior which is expected to yield more information than a static identity transform. VTLN was not used in a cascade, described in section 3.2 as it is a weaker transform than the CMLLR transform being estimated, and hence will be subsumed. Future work will look at the interaction of noise robustness techniques such as JUD and VTS with the methods presented above.

### 4.1. Experimental Setup

Gender independent US English acoustic models were trained using a 39 dimensional MFCC feature vector, with static, delta and delta-delta parameters. A total of 312 hours of data from WSJ, TIDIGITS, TIMIT and internally collected noisy data was used for training triphone acoustic models. Decision tree clustering was used to yield 650 unique states. 12 Gaussian components were used per speech state and 24 Gaussian components per silence state, yielding approximately 8000 components. For adaptive training, transforms were estimated on a per-speaker basis, using a transform type consistent with decoding. Experiments were carried out on two tasks

- *Toshiba in-car task* - a database recorded in real driving conditions with phone numbers, 4 digits, command and control, and citynames subtasks. Each sub task includes two noisy conditions, engine on and highway, and there are a total of 8983 utterances spoken by native speakers with an average of 463 frames per utterance.
- *Multi-accent task* - a database recorded in studio conditions with additional noise. There are approximately 14k utterances split between telephone and TV control, spoken by users with a mixture of accents, with an average of 226 frames per utterance.

A separate transform is estimated for each test set utterance using two regression classes - speech and silence - to limit the number of parameters and allow for rapid adaptation. The baseline hypothesis was used for estimation of all transforms.

### 4.2. Results and Discussion

Experimental results are given in table 1. The first lines show results obtained for the baseline system without adaptation, and with standard VTLN and CMLLR. VTLN consistently yields small gains, e.g. on the multi-accent set the baseline error rate of 15.90% is improved to 15.44%. Diagonal CMLLR improves over VTLN on both test sets but full CMLLR does not impact on performance, suggesting that one utterance does not give enough data to robustly estimate the parameters.

Next, VTLN was used as a parent transform only when estimating a CMLLR child transform to be used in a cascade. The results show that using VTLN as a parent transform for estimating a diagonal CMLLR transform can give performance gains. For example, on the Toshiba in-car set, the error rates are 2.33% and 1.86% for VTLN and diagonal CMLLR respectively, but 1.79% when cascading VTLN and CMLLR. However, when used as a parent to estimate a full CMLLR transform, very little difference in error rate is seen. The error rates for VTLN and full CMLLR on the Toshiba set are 2.33% and 2.36% respectively, and 2.35% when cascading the two transforms. This suggests that incorporating prior knowledge as parent transform with no prior does not improve the robustness of poorly estimated CMLLR transforms on limited data.

Section 3 discussed combining prior and adaptive statistics for more robust transform estimates. Experiments were carried

	Parent	Prior	Standard		Adaptive		
			In-car	Mlt-acc	In-car	Mlt-acc	
Baseline	-	-	2.38	15.90	-	-	
VTLN	Block	-	2.33	15.44	2.17	15.11	
CMLLR	Diag	-	1.86	15.17	1.74	14.96	
CMLLR	Full	-	2.36	15.90	2.34	15.84	
CMLLR*	Diag	VTLN	1.79	14.98	1.77	14.94	
CMLLR*	Full	VTLN	2.35	15.90	2.35	15.90	
CMLLR	Full	-	Identity	1.92	15.11	2.11	14.58
CMLLR	Full	-	VTLN	1.87	14.82	1.63	13.54

Table 1: WER (%) on Toshiba in-car, and multi-accent tasks for standard and adaptive training (\*adaptive training uses only VTLN)

out using an identity matrix as prior and also using VTLN as a dynamic prior with the method described in section 3.1. A full CMLLR transform was trained for each utterance by combining the prior and adaptive statistics using equations 12 and 13.

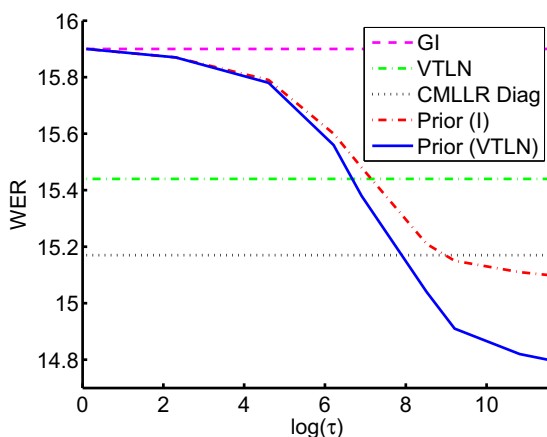


Figure 1: Effect of prior on Multi-accent set WER

Figure 1 shows the performance on the multi-accent set as the value of  $\tau$  is increased - note that full CMLLR transforms are equivalent to  $\tau = 0$ . As can be seen, the resulting transform can give better performance than either the full CMLLR or VTLN transforms alone. For some values of  $\tau$ , the transforms from interpolated statistics give gains over the robust diagonal CMLLR transform. VTLN appears to be a relatively weak prior as it does not give large gains by itself, and only gives small improvements over an identity prior. Results are given in table 1 for  $\tau = 50000$  for the two test sets. On the in-car set, the identity prior yields an error rate of 1.92% and the VTLN prior gives 1.86%, which are relative gains of 20% and 22% respectively over the baseline model.

Finally, adaptive training was carried out using VTLN, CMLLR, and CMLLR with a prior. As expected, gains are seen from adaptive training using both VTLN and CMLLR. The use of the VTLN prior in adaptive training yields further improvement in error rate. On the Toshiba in-car set, adaptive training using VTLN and diagonal CMLLR yielded results of 2.17% and 1.77% WER respectively, but a further improvement to 1.63% was seen using the approach proposed in section 3.1. This is an improvement of 46% relative over the baseline performance. Relative improvement on the multi-accent task is 15%.

## 5. Conclusions

This paper has addressed the problem of producing robust estimates of complex adaptation techniques such as CMLLR from limited data. A count-smoothing framework was extended to

use dynamic prior statistics, derived from rapid adaptation techniques, are interpolated with the main adaptive statistics. Two implementations were proposed: direct transform estimation and cascades of transforms. Using VTLN in the direct transform estimation framework as a prior for estimating CMLLR transforms was shown to be more robust than standard CMLLR to small amounts of data, and more robust than using a cascade of transforms where no prior was used. Where the prior is more complex than the smoothed transform being estimated a cascade of transforms is more powerful than the direct estimation approach. In future work the ability to efficiently use a complex parent transform as a prior to estimate a simpler child transform will be investigated.

## 6. References

- [1] C. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs," *Computer Speech and Language*, vol. 9, 1995.
- [2] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1997.
- [3] M.J.F. Gales, "Cluster Adaptive Training of Hidden Markov Models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 417–428, 2000.
- [4] C. Chesta, O. Siohan, and C.H. Lee, "Maximum A Posteriori Linear Regression for Hidden Markov Model Adaptation," in *Eurospeech*, 1999.
- [5] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained Structural Maximum A Posteriori Linear Regression for Average-Voice-Based Speech Synthesis," in *Interspeech*, 2006.
- [6] O. Siohan T.A. Myrvoll and C.H. Lee, "Structural Maximum A Posteriori Linear Regression for Fast HMM Adaptation," in *ICSA ITRW ASR2000*, 2000.
- [7] L. Lee and R.C. Rose, "A Frequency Warping Approach to Speaker Normalisation," *IEEE Transactions on Speech and Audio Processing*, vol. 6, 1998.
- [8] D.Y. Kim, S. Umesh, M.J.F. Gales, T. Hain, and P.C. Woodland, "Using VTLN for Broadcast News Transcription," in *Interspeech*, 2004.
- [9] P.T. Akhil, S.P. Rath, S. Umesh, and D.R. Sanand, "A Computationally Efficient Approach to Warp Factor Estimation in VTLN using EM Algorithm and Sufficient Statistics," in *Interspeech*, 2008.
- [10] M.J.F. Gales and R.C. van Dalen, "Predictive Linear Transforms for Noise Robust Speech Recognition," in *ASRU*, 2007.
- [11] F. Flego and M.J.F. Gales, "Incremental Predictive and Adaptive Noise Compensation," in *ICASSP*, 2009.
- [12] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. of ICSLP*, Beijing, China, Oct. 2000.
- [13] H. Xu, M.J.F. Gales, and K.K. Chin, "Improving Joint Uncertainty Decoding Performance by Predictive Methods for Noise Robust Speech Recognition," in *ASRU*, 2009.