FACTOR ANALYSIS BASED VTS DISCRIMINATIVE ADAPTIVE TRAINING

F. Flego and M.J.F. Gales

Cambridge University Engineering Department
Trumpington St., Cambridge CB2 1PZ, U.K.
{ff257,mjfg}@eng.cam.ac.uk

ABSTRACT

Vector Taylor Series (VTS) model based compensation is a powerful approach for noise robust speech recognition. An important extension to this approach is VTS adaptive training (VAT), which allows canonical models to be estimated on diverse noise-degraded training data. These canonical model can be estimated using EM-based approaches, allowing simple extensions to discriminative VAT (DVAT). However to ensure a diagonal corrupted speech covariance matrix the Jacobian (loading matrix) relating the noise and clean speech is diagonalised. In this work an approach for yielding optimal diagonal loading matrices based on minimising the expected KL-divergence between the diagonal loading matrix and "correct" distributions is proposed. The performance of DVAT using the standard and optimal diagonalisation was evaluated on both in-car collected data and the Aurora4 task.

Index Terms— Speech recognition, noise robustness, adaptive training, generative processes.

1. INTRODUCTION

There has been a large amount of interest in model compensation schemes for noise robust speech recognition. Among these, approaches such as VTS [1] and Joint Uncertainty Decoding (JUD) compensation [2] have been found to yield good recognition performance, particularly in low signal-to-noise ration (SNR) conditions. VTS and JUD adapt the "clean" acoustic models to a particular target noise condition. This is achieved using a noise model and a *mismatch function* which models its impact on the clean speech.

These model-based compensation forms have been successfully extended to adaptive training where the canonical model is estimated based on a set of training noise transforms. Experiments using training data with a wide range of noise conditions confirmed the advantages of such approaches compared to both multi-style and clean systems [2, 3, 4]. The noise and canonical model parameters are generally trained using Maximum Likelihood (ML) estimation. This can be done maximising an auxiliary function using either second-order, [2, 5, 3], or EM-based approaches [4, 6].

These EM-based approaches can be formulated within the Factor Analysis (FA) framework, where a generative process is used to model the relationship between the clean and the corrupted speech features. This allows EM-based update formulae for both the canonical model parameters and the noise transforms as in [7], to be obtained. Using this FA framework also enables a simple extension of ML-based adaptive training to discriminative adaptive training [6]. However, one problem with the above EM-style approaches is that to obtain a valid generative model and related updated expressions, constraints need to be placed on the generative process parameters. In particular the loading matrix needs to be diagonal so that diagonal compensated covariances are obtained [7].

In previous work the loading matrix (the Jacobian) was simply diagonalised to satisfy the constraints [6]. In this work an optimal diagonal loading matrix is found by minimising the KL divergence between the distributions that result from the diagonal loading matrix and the "correct" distribution. This optimal loading matrix can be applied for both ML and discriminative adaptive training.

In the next section VTS adaptive training is described. Both the second-order approach based on the mismatch function and the EM-based approach based on the FA generative process are described. The latter is then extended to provide DVAT and a KL divergence-based method is proposed to optimise the generative model parameters. The performance of such method is evaluated using both the *TREL* data with real noisy data in-car collected and the *Aurora4* database.

2. VTS ADAPTIVE TRAINING

VTS compensation bases on the following mismatch function (ignoring convolutional noise for simplicity)

$$\boldsymbol{y} = \boldsymbol{x} + \mathbf{C}\log\left(\mathbf{1} + \exp(\mathbf{C}^{-1}(\boldsymbol{z} - \boldsymbol{x}))\right) = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z}) \quad (1)$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the clean and corrupted speech static features, $\mathbf{C}$ is the DCT matrix, and $\boldsymbol{z}$ is the additive noise vector. Applying a first-order VTS approximation to Eq. 1 gives

$$\boldsymbol{y} \approx \boldsymbol{f}(\boldsymbol{\mu}_{\mathsf{x}}^{(m)}, \boldsymbol{\mu}_{\mathsf{z}}) + \mathbf{J}_{\mathsf{x}}^{(m)}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathsf{x}}^{(m)}) + \mathbf{J}_{\mathsf{z}}^{(m)}(\boldsymbol{z} - \boldsymbol{\mu}_{\mathsf{z}}) \quad (2)$$

This work was partly funded by Toshiba Research Europe Ltd (TREL) and the DARPA RATS program. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

978-1-4673-0046-9/12/$26.00

assuming clean speech and additive noise are Gaussian distributed with parameters $\mathcal{M}=\{\boldsymbol{\mu}_{\mathsf{x}}^{(m)},\boldsymbol{\Sigma}_{\mathsf{x}}^{(m)}\}$, and $\{\boldsymbol{\mu}_{\mathsf{z}},\boldsymbol{\Sigma}_{\mathsf{z}}\}$, respectively. $\mathbf{J}_{\mathsf{x}}^{(m)}$ and $\mathbf{J}_{\mathsf{z}}^{(m)}$ are the Jacobians of $\boldsymbol{y}$ with respect to vector $\boldsymbol{x}$, and $\boldsymbol{z}$, respectively. Taking the expectation of Eq. 2 provides the following VTS compensated model parameters for component m

$$\boldsymbol{\mu}_{\mathsf{y}}^{(m)} = \boldsymbol{f}(\boldsymbol{\mu}_{\mathsf{x}}^{(m)},\boldsymbol{\mu}_{\mathsf{z}}) \quad (3)$$

$$\boldsymbol{\Sigma}_{\mathsf{y}}^{(m)} = \mathrm{dg}(\mathbf{J}_{\mathsf{x}}^{(m)}\boldsymbol{\Sigma}_{\mathsf{x}}^{(m)}\mathbf{J}_{\mathsf{x}}^{(m)\mathsf{T}} + \mathbf{J}_{\mathsf{z}}^{(m)}\boldsymbol{\Sigma}_{\mathsf{z}}\mathbf{J}_{\mathsf{z}}^{(m)\mathsf{T}}) \quad (4)$$

where the $\mathrm{dg}()$ operator diagonalises the matrix. This is required so that diagonal covariance matrices can be used in decoding.

Though VTS has proved to be a powerful method for compensating clean models, in many practical situations only corrupted data is available for training the underlying clean models. In this case VTS adaptive training (VAT) can be used to handle the environment mismatch. Thus, the canonical model parameters $\hat{\mathcal{M}}$ are obtained given a set of training VTS transforms for each homogeneous block. Then, given $\hat{\mathcal{M}}$, the transforms can be refined and the process iterated until convergence. Two VAT forms are described in this section based on ML training: a second-order based scheme [2, 3], and an EM-based approach [6, 4], which also allows to easily extend VAT to VTS discriminative adaptive training (DVAT).

2.1. Maximum Likelihood training

To estimate the new VAT canonical model parameters $\hat{\mathcal{M}}$ the following auxiliary function can be used

$$\mathcal{Q}(\hat{\mathcal{M}};\mathcal{M}) = \sum_{s,t,m}\gamma_t^{(sm)}\log\left\{\mathcal{N}(\boldsymbol{y}_t^{(s)};\hat{\boldsymbol{\mu}}_{\mathsf{y}}^{(m)},\hat{\boldsymbol{\Sigma}}_{\mathsf{y}}^{(m)})\right\} \quad (5)$$

where it is assumed that a noise transform was estimated for each homogeneous block of data s and the compensated parameters are obtained from Eq. 3 and 4 based on $\hat{\mathcal{M}}$.

To estimate the new canonical model parameters $\hat{\mathcal{M}}$ standard second-order optimisation schemes can be used to maximise a quadratic approximation of the above auxiliary function. Additionally approximations are made to simplify the estimation of the first and second order derivatives [2, 3]. The effect of these approximations is that the estimated model parameters are not guarantee that the "real" auxiliary function, which is obtained using Eq. 3 and 4, is maximised, and a back-off procedure on the new estimates is generally applied [2].

An alternative EM-based VAT scheme can be obtained using the FA framework. It is first necessary to express the mismatch function in Eq. 2 as a FA-style generative model

$$\boldsymbol{y}|m = \boldsymbol{\Lambda}^{(m)}\boldsymbol{x} + \boldsymbol{\epsilon}^{(m)} \quad (6)$$

where $\boldsymbol{\Lambda}^{(m)}$ is the loading matrix and the following distributions are defined

$$\boldsymbol{x} \sim \mathcal{N}\big(\hat{\boldsymbol{\mu}}_{\mathsf{x}}^{(m)},\hat{\boldsymbol{\Sigma}}_{\mathsf{x}}^{(m)}\big), \qquad \boldsymbol{\epsilon}^{(m)} \sim \mathcal{N}\big(\boldsymbol{\mu}_{\epsilon}^{(m)}\boldsymbol{\Sigma}_{\epsilon}^{(m)}\big)$$

$$\boldsymbol{\mu}_{\epsilon}^{(m)} = \boldsymbol{\mu}_{\mathsf{y}}^{(m)} - \boldsymbol{\Lambda}^{(m)}\boldsymbol{\mu}_{\mathsf{x}}^{(m)}, \quad \boldsymbol{\Sigma}_{\epsilon}^{(m)} = \boldsymbol{\Sigma}_{\mathsf{y}}^{(m)} - \boldsymbol{\Lambda}^{(m)}\boldsymbol{\Sigma}_{\mathsf{x}}^{(m)}\boldsymbol{\Lambda}_{\mathsf{x}}^{(m)\mathsf{T}}$$

The above generative model can be directly related to Eq. 2 when $\boldsymbol{\Lambda}^{(m)}=\mathbf{J}_{\mathsf{x}}^{(m)}$.

To obtain the FA style auxiliary function the following inequality can be used for the probability in Eq. 5

$$\log p(\boldsymbol{y}|\hat{\mathcal{M}}) = \log\int_{\boldsymbol{x}} p(\boldsymbol{y},\boldsymbol{x}|\hat{\mathcal{M}})d\boldsymbol{x} \geq \int_{\boldsymbol{x}} p(\boldsymbol{x}|\mathcal{M})\log p(\boldsymbol{y},\boldsymbol{x}|\hat{\mathcal{M}})d\boldsymbol{x} + \mathcal{H}(p(\boldsymbol{x}|\mathcal{M})) \quad (7)$$

where $\mathcal{H}(p(\boldsymbol{x}|\mathcal{M}))$ is the entropy of clean speech distribution and is independent of $\hat{\mathcal{M}}$. Substituting the above expression into Eq. 5 provides an alternative auxiliary function which yields the following EM-based solution (a similar expression can be found for $\hat{\boldsymbol{\Sigma}}_{\mathsf{x}}^{(m)}$)

$$\hat{\boldsymbol{\mu}}_{\mathsf{x}}^{(m)} = \frac{\sum_{t,m}\gamma_t^{(m)}\,\mathbb{E}\{\boldsymbol{x}|\boldsymbol{y}_t,m\}}{\sum_{t,m}\gamma_t^{(m)}} \quad (8)$$

The expectation above $\mathbb{E}\{\cdot\}$ is obtained from the $(\boldsymbol{x},\boldsymbol{y})$ joint distribution parameters provided by the FA generative model in Eq. 6.

For the EM-based approach, it is possible in principle to use the full Jacobian form for $\boldsymbol{\Lambda}^{(m)}$. Unfortunately this will yield full compensated covariances. Although it could be possible to diagonalise the final estimates, the likelihoods values obtained using full covariance statistics during training will differ considerably from those obtained after diagonalisation. Applying a back-off procedure as in the second-order approach is not feasible as it would cancel out the computational benefits of the EM-based approach.

The simplest approach to deal with this is to diagonalise the Jacobian

$$\boldsymbol{\Lambda}^{(m)} = \mathrm{dg}(\mathbf{J}_{\mathsf{x}}^{(m)}) \quad (9)$$

which, though discarding important information provided by the Jacobian off-diagonal terms, proved to be effective for both ML and discriminative training [6]. Unfortunately this introduces a mismatch between the resulting distributions and the mismatch function in Eq. 2.

2.2. Discriminative training

Minimum Phoneme Error (MPE) based discriminative training aims to minimise the following function

$$\mathcal{F}_{\mathtt{mpe}}(\mathcal{M}) = \sum_{s=1}^{S}\sum_{\mathcal{H}} P(\mathcal{H}|\boldsymbol{Y}^{(s)},\mathcal{M})\mathcal{L}(\mathcal{H},\mathcal{H}_{\mathtt{ref}}^{(s)}) \quad (10)$$

where $\mathcal{L}(\mathcal{H},\mathcal{H}_{\mathtt{ref}}^{(s)})$ is the "loss" measured at the phone-level between the hypothesis and reference $\mathcal{H}_{\mathtt{ref}}^{(s)}$.

To optimise this expression a weak-sense auxiliary function is generally used. An important stage in this is the setting of the component-specific constant $D^{(m)}$ that weights the

smoothing with the previous iteration model-parameters. For standard discriminative training this is usually set as

$$D^{(m)} = \max\left\{E\sum_{s=1}^{S}\sum_{t=1}^{T}\gamma_{\mathtt{den},t}^{(ms)}, 2D_{\mathtt{min}}^{(m)}\right\} \quad (11)$$

where $D_{\mathtt{min}}^{(m)}$ is the minimum value to ensure that the covariance matrix of component m is semi-positive definite and E is an empirically set constant [8].

In theory both of the ML VAT estimation schemes in the previous section could be extended to DVAT. However, it is not possible to use Eq. 11 if the second-order approach described in the previous section is used for the optimisation. In this case the estimation of $D_{\mathtt{min}}^{(m)}$ is more problematic, further complicating the selection of an appropriate smoothing term. In contrast, using the FA based approach of Sec. 2.1 allows to use the same results obtained for MPE training and the following EM updates, similar to Eq. 8, are obtained [6] ($\hat{\boldsymbol{\Sigma}}_{\mathsf{x}}^{(m)}$ has a similar form)

$$\hat{\boldsymbol{\mu}}_{\mathsf{x}}^{(m)} = \frac{\sum_{s,t}\gamma_t^{(ms)}\mathcal{E}[\boldsymbol{x}|\boldsymbol{y}_t^{(s)},m] + D^{(m)}\boldsymbol{\mu}_{\mathsf{x}}^{(m)} + \tau_{\mathsf{p}}\boldsymbol{\mu}_{\mathsf{p}}^{(m)}}{\sum_{s,t}\gamma_t^{(ms)} + D^{(m)} + \tau_{\mathsf{p}}} \quad (12)$$

where $\{\boldsymbol{\mu}_{\mathsf{p}}^{(m)}, \boldsymbol{\Sigma}_{\mathsf{p}}^{(m)}\}$ and τ_{p} are the parameters of a prior which is used to reduce the risk of over-training [8]. It should be emphasised that this form still requires the diagonalisation in Eq. 9.

3. KL DIVERGENCE-BASED OPTIMISATION

It is unclear if Eq. 9 provides an optimal loading matrix form. The aim of this section is to derive a form which minimises the KL divergence between linearised estimates of the distributions using either the full Jacobian, or a diagonalised loading matrix. To simplify this problem fixed compensated covariances are used. Thus only the mean will shift based on

$$\hat{\boldsymbol{\mu}}_{\mathsf{y}}^{(m)} = \boldsymbol{\mu}_{\mathsf{y}}^{(m)} + \boldsymbol{\Lambda}_{\mathsf{x}}^{(m)}(\hat{\boldsymbol{\mu}}_{\mathsf{x}}^{(m)} - \boldsymbol{\mu}_{\mathsf{x}}^{(m)}) \quad (13)$$

Thus two distributions will be obtained, one based on $\boldsymbol{\Lambda}_{\mathsf{x}}^{(m)} = \mathbf{J}_{\mathsf{x}}^{(m)}$ and the second based on the unknown diagonal loading matrix, $\boldsymbol{\Lambda}$, to be estimated. As the new estimate of the mean is unknown, a distribution over the changes in the mean, $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}_{\mathsf{x}}^{(m)} - \boldsymbol{\mu}_{\mathsf{x}}^{(m)}$, must be used. The aim is thus to minimise the expected KL divergence over this mean distribution. This can be written as

$$\mathbb{E}\big\{\mathcal{KL}\big(\mathcal{N}(\boldsymbol{\mu}_{\mathsf{y}}^{(m)} + \mathbf{J}_{\mathsf{x}}^{(m)}\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathsf{y}}^{(m)}) || \mathcal{N}(\boldsymbol{\mu}_{\mathsf{y}}^{(m)} + \boldsymbol{\Lambda}\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathsf{y}}^{(m)})\big)\big\}$$

The optimal value for $\boldsymbol{\Lambda}_{\mathsf{x}}^{(m)}$ can be obtained from

$$\hat{\boldsymbol{\Lambda}}_{\mathsf{x}}^{(m)} = \arg\min_{\boldsymbol{\Lambda}}\left\{\mathtt{Tr}\left\{\boldsymbol{\Sigma}_{\mathsf{y}}^{(m)\text{-}1}(\mathbf{J}_{\mathsf{x}}^{(m)} - \boldsymbol{\Lambda})\,\mathbb{E}\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\}(\mathbf{J}_{\mathsf{x}}^{(m)} - \boldsymbol{\Lambda})^{\mathsf{T}}\right\}\right\}$$

which, differentiated and equated to zero, yields

$$\mathrm{dg}\left((\mathbf{J}_{\mathsf{x}}^{(m)} - \boldsymbol{\Lambda})\,\mathbb{E}\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\}\right) = \mathbf{0} \quad (14)$$

The above expression shows that the optimal value for $\boldsymbol{\Lambda}$ is dependent on the second-order moment of the $p(\boldsymbol{\mu})$ distribution. Two forms can be considered:

1. **Diagonal** $\mathbb{E}\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\}$: no specific information is provided on the correlation between dimensions, and all directions of the search space are considered independently. Using this distribution in Eq. 14 naturally provides the diagonal approximation in Eq. 9, irrespective of the diagonal matrix values.
2. **Full** $\mathbb{E}\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\}$: in this case the optimal $\boldsymbol{\Lambda}$ depends on the exact form of $\mathbb{E}\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\}$.

The structure of $\mathbb{E}\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\}$ is unknown and must be approximated. The approach used in this work is to consider $\boldsymbol{\mu}$ as a random variable β in the direction of the gradient of the auxiliary function in Eq. 5. This, assuming that $\boldsymbol{\Sigma}_{\mathsf{y}}^{(m)}$ is a scaled version of the residual outer product, yields $\mathbb{E}\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\} = \mathbb{E}\{\beta^2\}\,\mathbf{J}_{\mathsf{x}}^{(m)\mathsf{T}}\boldsymbol{\Sigma}_{\mathsf{y}}^{(m)\text{-}1}\mathbf{J}_{\mathsf{x}}^{(m)}$.

When this form is used, the following expression is obtained

$$\boldsymbol{\Lambda}^{(m)} = \mathrm{dg}(\mathbf{J}_{\mathsf{x}}^{(m)}\mathbf{J}_{\mathsf{x}}^{(m)\mathsf{T}}\boldsymbol{\Sigma}_{\mathsf{y}}^{(m)\text{-}1}\mathbf{J}_{\mathsf{x}}^{(m)})\mathrm{dg}(\mathbf{J}_{\mathsf{x}}^{(m)\mathsf{T}}\boldsymbol{\Sigma}_{\mathsf{y}}^{(m)\text{-}1}\mathbf{J}_{\mathsf{x}}^{(m)})^{\text{-}1} \quad (15)$$

The off-diagonal Jacobian terms are involved in the calculation before diagonalisation is applied, so will yield different loading matrices to the diagonal case.

One side-effect of using this approach is that the covariance of the generative model error term in Eq. 6 may become negative. As this would then yield an invalid generative process, none of the standard update formulae could be used. To avoid this issue, the result of Eq. 15 is smoothed as follows

$$\hat{\boldsymbol{\Lambda}}^{(m)} = \alpha^{(m)}\mathrm{dg}(\mathbf{J}_{\mathsf{x}}^{(m)}) + (1 - \alpha^{(m)})\,\boldsymbol{\Lambda}^{(m)} \quad (16)$$

with

$$\alpha^{(m)} = \max_{i=1:d}\left\{\frac{\sqrt{\boldsymbol{\sigma}_{\mathsf{y}i}^{(m)}/\boldsymbol{\sigma}_{\mathsf{x}i}^{(m)}} - \boldsymbol{\lambda}_i}{\mathbf{J}_{\mathsf{x}ii}^{(m)} - \boldsymbol{\lambda}_i}\right\} \quad (17)$$

where d is the size of the feature vector and lower case symbols, i.e. $\boldsymbol{\sigma}_{\mathsf{x}i}^{(m)}$, are use to indicate values of the diagonal matrices, while $\mathbf{J}_{\mathsf{x}ii}^{(m)}$ selects the i-th Jacobian diagonal element.

In the next section DVAT based on the proposed Eq. 16 is compared with standard DVAT based on Eq. 9.

4. EXPERIMENTS AND RESULTS

To evaluate the proposed approaches two different tasks were used in this work: the *Aurora4* and the *TREL* [6] configurations. For both tasks ML and discriminatively adaptively trained systems were trained. The tasks are briefly described below. For further details and contrasts with multi-style systems see the references.

The *TREL* training data (486 hours), [6], includes artificially corrupted clean speech data with car noise and in-car collected data. The recognition tasks consist of in-car

recorded data using a microphone mounted on the rear-view mirror, with either the engine-on (ENON, 35dB) or driving along a highway (HWAY, 18dB).

The multi-condition and multichannel (Mic1 and Mic2) data of the *Aurora4* database (16kHz, 12 hours, 7138 utterances) was used. The recognition task is a 5K-word dictation task with 14 test sets, 330 utterances each. Sets 01-07 were recorded with Mic1 adding to sets 02-07 different noises with random SNR from 5 to 15 dB. The same approach was used to obtain sets 08-14 but Mic2 was used instead. In the following, letters A, B, C, and D will be used to indicate test sets 01, 02-07, 08, 09-14, respectively.

For both systems 12 MFCCs plus zeroth cepstrum, delta and delta-delta and the systems were built using the same configuration in [6]. The number of components in the TREL task was about 7800 (to mimic a compact in-car system) and for the Aurora4 task 50,000.

For each training configuration, a VAT system was trained using the second-order approach described in Sec. 2.1 and it was then used as initial system to train two DVAT systems: one based on Eq. 9, DVAT_{dg}, the other on Eq. 15, DVAT_{opt}.

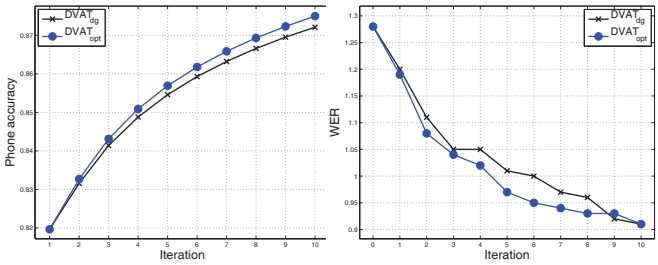

Fig. 1. DVAT on *TREL*: phone level accuracy and WER%.

Initial experiments were run using *TREL* configuration to evaluate the training performance of the two DVAT systems. In the left graph of Fig. 1 the Phone accuracy is plotted for the 10 iterations of canonical model estimation. The results show that the proposed optimal loading matrix approach achieves a higher phone-level accuracy with respect to the standard approach DVAT_{dg}. The effect of this is that a faster converge is achieved as it is shown in the right plot. Here the average WERs are plot for each iteration showing that the proposed scheme achieves lower WERs with respect to DVAT_{dg} in almost all iterations. After 10 iterations of canonical model re-estimation DVAT provided a WER of 0.5% and 1.4% for the ENON and HWAY condition, respectively, compared to the VAT system (iteration 0) which obtained 0.7% and 1.8%, for the two conditions, respectively. The same phone-level accuracy and WERs trends were observed for *Aurora4*. The final comparison between VAT and DVAT in this case is shown in Table 1 for each noise/channel conditions. There is little difference in WER performance between the two approaches, though again for all iterations the MPE criterion was higher for the optimal loading matrix.

System	A	B	C	D	avg
VAT	8.1	13.5	11.6	20.5	15.98
DVAT_{dg}	7.4	12.9	11.3	19.8	15.34
DVAT_{opt}	7.4	12.8	11.4	19.8	15.31

Table 1. VAT and DVAT on *TREL*: WER(%).

5. CONCLUSIONS

An extension of the previous work on canonical model parameter estimation for DVAT was presented. Rather than using a diagonal loading matrix in the FA generative process used to obtain EM-based model updates, an optimal value was derived basing on the KL divergence criterion. This provided a higher training phone-level accuracy which also gave improved WERs at each iteration of model re-estimation. However, after several iterations, both approaches provided similar WERs. Though the current implementation has not yielded performance gains, it provides a framework for obtaining optimal loading matrices. Applying this approach to more complex tasks with larger amounts of data, or alternative distribution for the mean shift, may yield performance gains.

6. REFERENCES

[1] A. Acero, L. Deng, T. T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. ICSLP*, 2000.

[2] H. Liao and M. J. F. Gales, "Adaptive Training with Joint Uncertainty Decoding for Robust Recognition of Noisy Data," in *Proc. ICASSP*, 2007, vol. 4.

[3] O. Kalinli, M.L. Seltzer, and A. Acero, "Noise adaptive training using a vector Taylor series approach for noise robust automatic speech recognition," in *Proc. ICASSP*, 2009.

[4] Q. Huo and Y. Hu, "Irrelevant variability normalization based HMM training using VTS approximation of an explicit model of environmental distortions," in *Proc. Interspeech*, 2007.

[5] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series," in *Proc. ASRU*, 2007.

[6] F. Flego and M. J. F. Gales, "Discriminative adaptive training with VTS and JUD," in *Proc. ASRU*, 2009.

[7] F. Flego and M. J. F. Gales, "Factor analysis based VTS and JUD noise estimation and compensation," in *Proc. ICASSP*, 2011.

[8] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge University, 2003.