Covariance Modelling for Noise-Robust Speech Recognition

R. C. van Dalen, M. J. F. Gales

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK
rcv25@cam.ac.uk, mjfg@eng.cam.ac.uk

Abstract

Model compensation is a standard way of improving speech recognisers' robustness to noise. Most model compensation techniques produce diagonal covariances. However, this fails to handle any changes in the feature correlations due to the noise. This paper presents a scheme that allows full-covariance matrices to be estimated. One problem is that full covariance matrix estimation will be more sensitive approximations, those for the dynamic parameters are known to crude. In this paper a linear transformation of a window of consecutive frames is used as the basis for dynamic parameter compensation. A second problem is that the resulting full covariance matrices slow down decoding. This is addressed by using predictive linear transforms that decorrelate the feature space, so that the decoder can then use diagonal covariance matrices. On a noise-corrupted Resource Management task, the proposed scheme outperformed the standard VTS compensation scheme.

Index Terms: Noise robust speech recognition, vector Taylor series, joint uncertainty decoding.

1. Introduction

Robustly handling changes in the background noise conditions is a major problem for speech recogniser systems. Resolving the mismatch between the training and test acoustic conditions has been an active area of research for many years. It is possible to use either feature enhancement or model compensation techniques. The latter have yielded good results and will be the focus of this paper. Standard model compensation methods produce diagonal covariance matrices for the corrupted speech distributions. However, feature correlations are known to change due to variations in the background noise. For example, in the limit as the noise masks the speech, the correlation pattern will be that of the noise. To date, full covariance matrix compensation has only been estimated using stereo data of clean and noise-corrupted speech [1]. However stereo data is seldom available. This paper examines full covariance compensation where the noise model is estimated from a small amount of noisy speech data.

The estimation of full covariance matrices is liable to be more sensitive to approximations in the compensation process than diagonal covariance matrices. It is standard practice in speech recognition to append to the static coefficients extracted from the audio signal dynamic coefficients that represent the changes in the statics. These are computed from a window of feature vectors around the current time instance. A popular approximation to compensate dynamic parameters is the continuous time approximation [2], which assumes dynamic coefficients are the time derivatives at that instance. Though this has been successfully applied to compensating diagonal covariance matrices [3, 4] it is not clear that is accurate enough for full covariance compensation. To improve the dynamic parameter compensation, this paper models the influence of the noise on the static coefficients of consecutive time frames. A linear transform of this window of features yields the dynamic parameters. By computing the distribution over the window of features, it is simple to derive the dynamic parameters distribution.

An additional problem is that when estimating full-covariance matrix compensation , compensation and decoding is computationally more expensive. To reduce the computational load during compensation, joint uncertainty decoding [3] may be used. Here components are grouped together into base classes and compensation is only required at this base class level. There are typically far fewer base classes than components. To handle the increase in computational load during decoding, predictive linear transforms [1] can be used. Base class-specific linear feature space transformations that reduce correlations, so that diagonal covariance matrices can be used for decoding. However, this does again increase the computational load during compensation [1].

The organisation of this paper is as follows. The next section describes the noise compensation methods used. Section 3 discusses how correlations can be compensated. Section 4 discusses experimental results on a noise-corrupted Resource Management task.

2. Model compensation

The additive noise $\mathbf{n}$ and the convolutional noise $\mathbf{h}$ transform the clean speech $\mathbf{x}$, resulting in noise-corrupted speech $\mathbf{y}$. In the mel-cepstral domain (i.e. for MFCCs) the mismatch between clean speech statics $\mathbf{x}_t^{\mathsf{s}}$ and the noise-corrupted speech statics $\mathbf{y}_t^{\mathsf{s}}$ at time t is expressed by

$$\begin{aligned}\mathbf{y}_t^{\mathsf{s}} &= \mathbf{x}_t^{\mathsf{s}} + \mathbf{h}_t^{\mathsf{s}} + \mathbf{C}\log\left(\mathbf{1} + e^{\mathbf{C}^{-1}\left(\mathbf{n}_t^{\mathsf{s}} - \mathbf{x}_t^{\mathsf{s}} - \mathbf{h}_t^{\mathsf{s}}\right)}\right) \\ &= \mathbf{x}_t^{\mathsf{s}} + \mathbf{h}_t^{\mathsf{s}} + f\left(\mathbf{x}_t^{\mathsf{s}}, \mathbf{n}_t^{\mathsf{s}}, \mathbf{h}_t^{\mathsf{s}}\right),\end{aligned} \quad (1)$$

where $\mathbf{C}$ is the DCT matrix. It is standard practice in speech recognition to append dynamic features to the observation vector. Both first- and second-order coefficients ($\mathbf{y}_t^{\triangle}, \mathbf{y}_t^{\triangle^2}$ respectively) are normally used. Thus the observation feature vector is $\mathbf{y}_t = [\ \mathbf{y}_t^{\mathsf{s}\mathsf{T}} \quad \mathbf{y}_t^{\triangle\mathsf{T}} \quad \mathbf{y}_t^{\triangle^2\mathsf{T}}\ ]^{\mathsf{T}}$. For clarity of presentation only first-order, delta, coefficients $\mathbf{y}^{\triangle}$ will be shown.

Model compensation alters the speech recogniser parameters so they model the corrupted speech distribution. Each component in the clean speech model is usually handled separately. If the corrupted speech is distributed as $\mathcal{N}\left(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\right)$

$$\boldsymbol{\mu}_y = \mathcal{E}\left\{\mathbf{y}\right\}; \qquad \boldsymbol{\Sigma}_y = \operatorname{diag}\left(\mathcal{E}\left\{\mathbf{y}\mathbf{y}^{\mathsf{T}}\right\} - \boldsymbol{\mu}_y\boldsymbol{\mu}_y^{\mathsf{T}}\right). \quad (2)$$

Rogier van Dalen is funded by Toshiba Research Europe Ltd.

where the expectations are over the distribution of a component of the clean speech model and the noise distribution. The speech and noise are combined using equation (1). There is no closed form for (2), so various approximations are used.

If stereo data is available then single-pass retraining (SPR) may be used to approximate (2). This may be viewed as the "ideal" compensated system if the noise distributions are known [5].

2.1. Vector Taylor series

Equation (1) can be approximated with a first-order vector Taylor series (VTS) [6]. Evaluating the partial derivatives of f at $\boldsymbol{\mu}_n^{\mathsf{s}}, \boldsymbol{\mu}_x^{\mathsf{s}}, \boldsymbol{\mu}_h^{\mathsf{s}}$, (1) becomes

$$\begin{aligned}\mathbf{y}_t^{\mathsf{s}} \approx\ & \boldsymbol{\mu}_x^{\mathsf{s}} + \boldsymbol{\mu}_h^{\mathsf{s}} + f\left(\boldsymbol{\mu}_x^{\mathsf{s}}, \boldsymbol{\mu}_n^{\mathsf{s}}, \boldsymbol{\mu}_h^{\mathsf{s}}\right) \\ & + \mathbf{J}_x(\mathbf{x}_t^{\mathsf{s}} - \boldsymbol{\mu}_x^{\mathsf{s}}) + \mathbf{J}_n(\mathbf{n}_t^{\mathsf{s}} - \boldsymbol{\mu}_n^{\mathsf{s}}) + \mathbf{J}_h(\mathbf{h}_t^{\mathsf{s}} - \boldsymbol{\mu}_h^{\mathsf{s}}), \quad (3)\end{aligned}$$

with

$$\mathbf{J}_x = \frac{\partial \mathbf{y}^{\mathsf{s}}}{\partial \mathbf{x}^{\mathsf{s}}}; \qquad \mathbf{J}_n = \frac{\partial \mathbf{y}^{\mathsf{s}}}{\partial \mathbf{n}^{\mathsf{s}}}; \qquad \mathbf{J}_h = \frac{\partial \mathbf{y}^{\mathsf{s}}}{\partial \mathbf{h}^{\mathsf{s}}}. \quad (4)$$

The mean and covariance of the static corrupted speech then become [7]

$$\boldsymbol{\mu}_y^{\mathsf{s}} = \boldsymbol{\mu}_x^{\mathsf{s}} + \boldsymbol{\mu}_h^{\mathsf{s}} + f\left(\boldsymbol{\mu}_x^{\mathsf{s}}, \boldsymbol{\mu}_n^{\mathsf{s}}, \boldsymbol{\mu}_h^{\mathsf{s}}\right); \quad (5)$$

$$\boldsymbol{\Sigma}_y^{\mathsf{s}} = \operatorname{diag}\left(\mathbf{J}_x \boldsymbol{\Sigma}_x^{\mathsf{s}} \mathbf{J}_x^{\mathsf{T}} + \mathbf{J}_n \boldsymbol{\Sigma}_n^{\mathsf{s}} \mathbf{J}_n^{\mathsf{T}}\right). \quad (6)$$

Here, the noise model gives the distributions of $\mathbf{n}$ and $\mathbf{h}$. $\mathbf{n}$ (including the dynamic parameters) is assumed Gaussian with mean $\boldsymbol{\mu}_n$ and covariance $\boldsymbol{\Sigma}_n$; $\mathbf{h}^{\mathsf{s}} = \boldsymbol{\mu}_h$ is assumed constant. [6, 3] These distributions may be estimated using maximum-likelihood estimation and some data in the testing noise condition.

To compensate dynamic parameters the continuous time approximation [2] is often used with VTS. This approximation assumes that delta coefficients are derivatives of static coefficients with respect to time t, so that

$$\mathbf{y}_t^{\Delta} \approx \left.\frac{\partial \mathbf{y}^{\mathsf{s}}}{\partial t}\right|_t; \quad (7)$$

$$\boldsymbol{\mu}_y^{\Delta} = \mathbf{J}_x \boldsymbol{\mu}_x^{\Delta}; \quad \boldsymbol{\Sigma}_y^{\Delta} = \operatorname{diag}\left(\mathbf{J}_x \boldsymbol{\Sigma}_x^{\Delta} \mathbf{J}_x^{\mathsf{T}} + \mathbf{J}_n \boldsymbol{\Sigma}_n^{\Delta} \mathbf{J}_n^{\mathsf{T}}\right). \quad (8)$$

2.2. Data-driven parallel model combination

Data-driven parallel model combination [5] (DPMC) is a Monte Carlo method for estimating the distribution of the corrupted speech. Samples are drawn from the distributions of $\mathbf{x}^{\mathsf{s}}$ and $\mathbf{n}^{\mathsf{s}}$. (1) then gives the value of $\mathbf{y}^{\mathsf{s}}$ for each sample. The expectations in (2) are estimated using the samples of $\mathbf{y}^{\mathsf{s}}$.

In the limit as the number of samples goes to infinity DPMC yields an accurate distribution for the noise-corrupted speech given the mismatch function. However, as a large number of samples are necessary to train the noise-corrupted speech distributions, the computational cost is much greater than for VTS.

For dynamic coefficients computed with simple differences a compensation scheme at alternative to (7) is possible [5]. Extra clean speech statistics are used. By adding the static coefficients from the previous time instance to the feature vector, so that it becomes $\mathbf{x}_t = \left[\ \mathbf{x}_t^{\mathsf{s}\,\mathsf{T}} \quad \mathbf{x}_t^{\Delta\,\mathsf{T}} \quad \mathbf{x}_{t-1}^{\mathsf{s}}{}^{\mathsf{T}}\ \right]^{\mathsf{T}}$, the dynamic coefficients for the noise-corrupted speech can be found by[1]

$$\begin{aligned}\mathbf{y}_t^{\Delta} = \mathbf{x}_t^{\Delta} + f\left(\mathbf{x}_t^{\Delta} + \mathbf{x}_{t-1}^{\mathsf{s}}, \mathbf{n}_t^{\Delta} + \mathbf{n}_{t-1}^{\mathsf{s}}, \mathbf{h}^{\mathsf{s}}\right) \\ - f\left(\mathbf{x}_{t-1}^{\mathsf{s}}, \mathbf{n}_{t-1}^{\mathsf{s}}, \mathbf{h}^{\mathsf{s}}\right) \quad (9)\end{aligned}$$

The shape of covariance matrices also needs an extension to provide enough data to compensate dynamics correctly. Matrices with non-zero entries for cross-covariances between the same coefficients in different time instances can be used [5]. This paper will refer to these as "striped".

2.3. Joint uncertainty decoding

VTS and DPMC incur considerable computational cost since they compensate components individually. A technique that groups components into base classes and finds compensation per base class is called joint uncertainty decoding (JUD) [3]. Varying the number of base classes gives a trade-off between computational cost and accuracy. JUD compensation derives from the joint distribution of the clean speech and the noise-corrupted speech,

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x^{(r)} \\ \boldsymbol{\mu}_y^{(r)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_x^{(r)} & \boldsymbol{\Sigma}_{xy}^{(r)} \\ \boldsymbol{\Sigma}_{yx}^{(r)} & \boldsymbol{\Sigma}_y^{(r)} \end{bmatrix}\right). \quad (10)$$

This joint distribution can found using, for example, VTS or DPMC [8, 3]. The output distribution for component m in base class r follows from (10) and is of the form

$$p(\mathbf{y}|m) \propto \mathcal{N}\left(\mathbf{A}_{\mathsf{jnt}}^{(r)}\mathbf{y} + \mathbf{b}_{\mathsf{jnt}}^{(r)}; \boldsymbol{\mu}_x^{(m)}, \boldsymbol{\Sigma}_x^{(m)} + \boldsymbol{\Sigma}_{\mathsf{bias}}^{(r)}\right). \quad (11)$$

For the standard forms of VTS or DPMC diagonal $\boldsymbol{\Sigma}_x^{(r)}, \boldsymbol{\Sigma}_y^{(r)}, \boldsymbol{\Sigma}_{yx}^{(r)}$, are estimated, yielding diagonal $\boldsymbol{\Sigma}_{\mathsf{bias}}^{(r)}$ (and $\mathbf{A}_{\mathsf{jnt}}^{(r)}$). Note, full joint distributions have previously been estimated using stereo data, and give significant performance improvements [1].

3. Covariance matrix modelling

The approaches discussed in the previous section have used diagonal covariances for the output distributions, or estimated the full-covariance matrices from stereo data. In practice stereo data are rarely available, so schemes that allow full covariance matrix output distributions to be estimated from a noise model are needed. These noise models can be either known, or estimated from a small amount of noisy data [3]. This section describes the issues and approaches adopted to robustly performing model compensation to yield full covariance distributions for the corrupted speech. In addition, the issues of the increased computational cost of model compensation, statistics required, and decoding are discussed.

3.1. Covariance matrix estimation

In theory the compensation schemes discussed can be used to generate non-diagonal output distributions. VTS with the continuous time approximation, for example, gives

$$\boldsymbol{\Sigma}_y = \begin{bmatrix} \mathbf{J}_x \boldsymbol{\Sigma}_x^{\mathsf{s}} \mathbf{J}_x^{\mathsf{T}} + \mathbf{J}_n \boldsymbol{\Sigma}_n^{\mathsf{s}} \mathbf{J}_n^{\mathsf{T}} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_x \boldsymbol{\Sigma}_x^{\Delta} \mathbf{J}_x^{\mathsf{T}} + \mathbf{J}_n \boldsymbol{\Sigma}_n^{\Delta} \mathbf{J}_n^{\mathsf{T}} \end{bmatrix}. \quad (12)$$

[1] Normalisation of dynamic parameters is ignored for clarity of presentation.

This yields a block-diagonal structure (this is also true if the second-order dynamics are included).

It is interesting that this form of covariance structure has not been used, given the gains obtained using non-diagonal covariance matrices with stereo data. To illustrate why this may be the case it is useful to examine the compensated models in more detail. One approach to doing this is to use the average KL divergence over all the components between the compensated model-set and the "ideal" single-pass retrained system trained on stereo data [5].

Compensation	—	VTS	DPMC
$\mathbf{y}^{\mathsf{s}}$	42.28	0.93	0.88
$\mathbf{y}^{\Delta}$	2.52	4.32	0.49
$\mathbf{y}^{\Delta^2}$	2.45	11.29	0.46

Table 1: *Average* KL *divergence to a block-diagonal single-pass retrained system for* VTS *(continuous time) and* DPMC.

Table 1 shows the average KL divergence between a VTS block-diagonal system using the continuous time approximation and the block-diagonal SPR system. The additive noise distribution (there is no convolutional noise) is known. Block-diagonal statistics are used for both the clean speech and noise models. It is clear from the table that VTS finds compensated parameters close to the SPR system for the static features, but that the dynamic parameters are not well compensated. Both the delta and delta-delta parameters are further from the SPR system than the uncompensated (clean) model set. The continuous time approximation is not sufficiently accurate to generate block-diagonal covariance matrices. The simple difference approximation in (9) could be used. However, this work uses an alternative, more general method.

The key intuition to this method is that the distribution of the dynamic coefficients can be computed exactly from the distribution over consecutive static coefficients. Dynamic coefficients are computed from a window of static coefficients with a linear transformation. The simplest form essentially re-expresses (9):

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{y}^{\mathsf{s}}_t \\ \mathbf{y}^{\Delta}_t \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{I} & \mathbf{0} \\ -\frac{\mathbf{I}}{2} & \mathbf{0} & \frac{\mathbf{I}}{2} \end{bmatrix} \begin{bmatrix} \mathbf{y}^{\mathsf{s}}_{t-1} \\ \mathbf{y}^{\mathsf{s}}_t \\ \mathbf{y}^{\mathsf{s}}_{t+1} \end{bmatrix} = \mathbf{D}\mathbf{y}^{\mathsf{e}}_t \quad (13)$$

where $\mathbf{D}$ is the dynamic coefficient matrix and $\mathbf{y}^{\mathsf{e}}_t$ is the vector of static coefficients in the appropriate window. It is straightforward to extend this to handle both linear-regression coefficients over a larger window, and second-order dynamics. If the distribution of the extended noise-corrupted vector $\mathbf{y}^{\mathsf{e}}$ is given by $\mathcal{N}\left(\boldsymbol{\mu}^{\mathsf{e}}_y, \boldsymbol{\Sigma}^{\mathsf{e}}_y\right)$ then the mean and covariance of the corrupted speech distribution for $\mathbf{y}$ are

$$\boldsymbol{\mu}_y = \mathbf{D}\boldsymbol{\mu}^{\mathsf{e}}_y; \qquad \boldsymbol{\Sigma}_y = \mathbf{D}\boldsymbol{\Sigma}^{\mathsf{e}}_y\mathbf{D}^{\mathsf{T}} \quad (14)$$

where $\mathbf{D}$ is the appropriate dynamic parameter matrix. This work uses DPMC to draw samples from the extended clean speech and noise distributions. These are combined together for each of the time instances using (1). The Gaussian distribution for $\mathbf{y}^{\mathsf{e}}$ can then be directly estimated.

Table 1 also shows the average KL divergence using DPMC with extended feature vectors. Static parameter compensation is similar to VTS. However, compensation for dynamic parameters is far closer to the single-pass retrained system. In addition, this approach can estimate full covariance matrices.

3.2. Practical implementation

A number of practical issues need to be considered when using DPMC with extended feature vectors: the nature of the statistics, noise model estimation, and the computational cost. The first issue is the form of statistics required for the clean speech and noise extended vectors. Clean full covariance matrices for $\boldsymbol{\Sigma}^{\mathsf{e}}_x$ can be stored and used. However, if first- and second-order dynamic parameters use window widths of ± 2 and there are d static parameters this requires estimating a $9d \times 9d$ covariance matrix for every component. This is memory intensive, and with large numbers of Gaussian components, singular matrices and numerical accuracy problems can occur. One approach to handling this problem is to use "striped" statistics (see section 2.2): for each Gaussian component, the ith element of the static coefficients for a time instance is assumed to be correlated with only the ith element of time instances. This causes $\boldsymbol{\Sigma}^{\mathsf{e}}_x$ to have a striped structure with only $45d$ parameters rather than $9d(9d+1)/2$ for the full case.

The noise model cannot be estimated a priori. If the noise is known, then it is possible to obtain a full covariance matrix. However, if the noise must be estimated, as in [3], this is complicated and computationally expensive. The simplest solution is to assume that the noise is independent and identically distributed for all time instances. If the noise distribution is also assumed to be diagonal, then the estimation scheme in [3] can be directly used and the static elements simply duplicated for each time instance.

As previously mentioned DPMC is computationally more expensive than VTS. If the samples are drawn from the extended vectors then this is even more expensive than standard DPMC. To reduce the impact of this, joint uncertainty decoding with DPMC can be used rather than DPMC. This means that distributions for the base classes, rather than individual components, are required. This speeds up the compensation process. However, the decoding stage will still be expensive: a full $\boldsymbol{\Sigma}^{(r)}_{\mathsf{bias}}$ results in a full-covariance matrix decode. One option to address this is to use predictive linear transforms [1]. Here a linear transform $\mathbf{A}^{(r)}_{\mathsf{pst}}$ for each regression class r is estimated in an maximum likelihood fashion using the JUD statistics. This paper uses predictive semi-tied transforms (PST). (11) becomes

$$p(\mathbf{y}|m) \propto \mathcal{N}\left(\mathbf{A}^{(r)}_{\mathsf{pst}}(\mathbf{A}^{(r)}_{\mathsf{jnt}}\mathbf{y} + \mathbf{b}^{(r)}_{\mathsf{jnt}}); \mathbf{A}^{(r)}_{\mathsf{pst}}\boldsymbol{\mu}^{(m)}_x, \tilde{\boldsymbol{\Sigma}}^{(m)}_{\text{diag}}\right), \quad (15)$$

where

$$\tilde{\boldsymbol{\Sigma}}^{(m)}_{\text{diag}} = \text{diag}\left(\mathbf{A}^{(r)}_{\mathsf{pst}}\left(\boldsymbol{\Sigma}^{(m)}_x + \boldsymbol{\Sigma}^{(r)}_{\mathsf{bias}}\right)\mathbf{A}^{(r)\mathsf{T}}_{\mathsf{pst}}\right). \quad (16)$$

With PST, the model compensation stage becomes more costly, because all models must be updated, but decoding uses diagonal covariance matrices and is thus fast.

4. Experiments

The compensation schemes described were evaluated on the 1000 word Resource Management database to which operations Room noise from the NOISEX-92 database was added at 20 dB. This task contains 109 training speakers reading 3990 sentences, 3.8 hours of data. All results are averaged over three of the four available test sets, Feb89, Oct89, and Feb91, a total of 30 test speakers and 900 utterances. State-clustered triphone models with either 1 or 6 components per mixture were built using the HTK RM recipe. 10 000 samples per distribution were used for DPMC. Since the additive background noise is know, it is possible to generated stereo data and use single-pass retraining to obtain "ideal" model compensated systems. It is also possible to extract the true noise model.

	Statistics		Compensation	
	$\mathbf{\Sigma}_x$	$\mathbf{\Sigma}_x^{\mathsf{e}}$	$\mathbf{\Sigma}_y$	
Scheme			Diag.	Full
—	—	—	38.2	64.6
SPR	—	—	12.4	7.5
VTS	diag.	—	15.5	18.1*
VTS	block	—	14.4	15.5*
DPMC	—	striped	13.6	12.3
DPMC	—	full	13.0	10.8

Table 2: *Word error rates for* VTS *and* DPMC*. Noise model from known noise. * block-diagonal compensation.*

Initial experiments were run using the single-component system. This meant that full extended statistics could be extracted for DPMC. The known additive noise model used had a diagonal covariance matrix (this gave slightly poorer performance than a full covariance matric, but is simpler to estimate in practice). Table 2 shows the word error rates of VTS and DPMC compensation with different forms of clean speech statistics, outputting diagonal or full models. VTS with diagonal clean speech statistics and diagonal compensation (15.5 %) may be viewed as the standard approach. When VTS generates block-diagonal covariance compensation, performance decreases because of the continuous time approximation. The performance of VTS improves slightly by using block-diagonal statistics. As expected from table 1, DPMC produces better compensation than VTS in all cases. Moreover, when DPMC estimates full covariance matrices, performance improves over the diagonal case. When striped statistics (see section 2.2) are used the gains are not as large as the full system, especially when estimating full covariance matrices. However these statistics are more compact and can be robustly for larger systems. The diagonal system estimated from full clean speech statistics with DPMC comes close to the SPR system, though when generating full covariances, not all of the performance gain that SPR displays is seen. DPMC's compensation (10.8 %) clearly performs better than the standard diagonal VTS (15.5 %).

	Statistics		Comp.	
Scheme	$\mathbf{\Sigma}_x$	$\mathbf{\Sigma}_x^{\mathsf{e}}$	$\mathbf{\Sigma}_y$	WER
VTS	diag.	—	diag.	8.5
DPMC	—	striped	diag.	7.5
DPMC	—	striped	full	6.9
VTS-JUD	diag.	—	diag.	9.5
DPMC-JUD	—	full	diag.	8.6
DPMC-JUD	—	full	full	7.9
PST	—	full	—	7.8

Table 3: *Word error rates for* VTS *and* DPMC*, and* JUD *and predictive semi-tied. Unsupervised noise model estimation.*

The previous experiments assumed the noise models were known. Table 3 shows results from a system built with the 6 mixture components per state system, where the all the noise parameters were estimated in an unsupervised fashion on the test data for each speaker using VTS [3]. These estimates were directly used for VTS. However for DPMC the extended noise model distribution was generated by simply duplicating the static VTS estimated noise distribution. For all cases the noise models had a diagonal covariance matrix structure. For robustness striped clean speech statistics were used in DPMC. The top half of table 3 compares VTS and DPMC. Compared to the uncompensated clean system performance (38.0 %), VTS gave large gains. However, DPMC produces better diagonal compensation than VTS. Further gains are obtained using DPMC to produce full covariance matrices (6.9 %). This is an absolute reduction of 1.6,% (19,% relative) compared to standard VTS.

The use of JUD to decrease the computational load was then investigated. Here, the 9.5K components were clustered into 16 base classes. Robust estimates of full clean speech statistics for these base-classes can be used DPMC, rather than striped statistics. The bottom half of table 3 shows that full DPMC-JUD (also diagonal DPMC-JUD) outperforms diagonal VTS-JUD for estimating the joint distribution. To reduce the cost of full DPMC-JUD decoding PST was used. Though more computationally expensive at the compensation stage, the decoding cost is about the same as standard diagonal JUD. In line with results in [1], PST performed about the same as full JUD.

5. Conclusion

Standard model-based compensation schemes, such as VTS, normally only produce diagonal covariance noise corrupted speech distributions. This ignores any correlation changes in the feature vector due to background noise. In this paper a variant of DPMC that uses extended feature vectors is used to generate full corrupted speech distributions that allows correlation changes to nbe modelled. An important aspect of achieving performance gains is a more accurate dynamic parameter compensation scheme and the associated additional clean speech statistics. To address the additional computational lod of the scheme both JUD, to handle compensation costs, and PST, to handle decoding costs, are described. Large performance gains on a noise-corrupted Resource Management task were obtained.

6. References

[1] M. J. F. Gales and R. C. van Dalen, "Predictive linear transforms for noise robust speech recognition," in *Proceedings of the ASRU Workshop*, 2007, pp. 59–64.

[2] R. A. Gopinath, M. J. F. Gales, P. S. Gopalakrishnan, S. Balakrishnan-Aiyer, and M. A. Picheny, "Robust speech recognition in noise - performance of the IBM continuous speech recognizer on the ARPA noise spoke task," in *ARPA Workshop on Spoken Language System Technology*, 1995, pp. 127–130.

[3] H. Liao and M. J. F. Gales, "Joint uncertainty decoding for robust large vocabulary speech recognition," Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR.552, November 2006.

[4] 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 *Proceedings of the ASRU Workshop*, 2007, pp. 65–70.

[5] M. J. F. Gales, "Model-based techniques for noise robust speech recognition," Ph.D. dissertation, Cambridge University, 1995.

[6] P. J. Moreno, "Speech recognition in noisy environments," Ph.D. dissertation, Carnegie Mellon University, 1996.

[7] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proceedings of the ICSLP*, vol. 3, 2000, pp. 229–232.

[8] H. Xu, L. Rigazio, and D. Kryze, "Vector Taylor series based joint uncertainty decoding," in *Proceedings of Interspeech*, 2006, pp. 1125–1128.