EXTENDED VTS FOR NOISE-ROBUST SPEECH RECOGNITION

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

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK

ABSTRACT

Model compensation is a standard way of improving speech recognisers' robustness to noise. Currently popular schemes are based on vector Taylor series (VTS) compensation. They often use the *continuous time approximation* to compensate dynamic parameters. In this paper, the accuracy of dynamic parameter compensation is improved by representing the dynamic features as a linear transformation of a window of static features. A modified version of VTS compensation is applied to the distribution of the window of static features and, importantly, their correlations. These compensated distributions are then transformed to standard static and dynamic distributions. The proposed scheme outperformed the standard VTS scheme by about 10 % relative.

Index Terms— Speech recognition, acoustic noise, robustness

1. INTRODUCTION

Robustly handling changes in the background noise conditions is a major problem for speech recognition systems. Common approaches are to use either feature enhancement or model compensation techniques. The latter have been found to yield good results, particularly in conditions with low signal-to-noise ratios, and will be the focus of this paper. To achieve the best possible performance, model compensation schemes need to compensate the static as well as the dynamic parameters that are commonly used in HMM-based speech recognition systems. This paper describes a new approach for compensating the dynamic model parameters.

The first stage in developing a noise compensation scheme is to express how the noise conditions impact the "clean" speech. When cepstral-based parameters are used, the *mismatch function* between clean and noise-corrupted speech is non-linear. This non-linearity makes computing the exact distribution of the noise-corrupted speech intractable. A commonly used method that has yielded good results approximates the mismatch function with a first-order vector Taylor series (VTS) expansion. The mismatch function is simple to define for the static parameters. However, in HMM-based speech recognition systems dynamic features, for example delta and delta-delta coefficients, are appended to the static features to form the feature vector. The standard approach to compensate the associated dynamic parameters is to use the continuous time approximation [1]. It assumes that the dynamic coefficients are the time derivatives of the statics. The form of compensation for the dynamic parameters is then closely related to the static parameters.

In previous work the limitations of the continuous time approximation were highlighted and a modified version of data-driven parallel model combination (DPMC) proposed to improve the dynamic parameter compensation [2]. The dynamic coefficients can be expressed as a linear transformation over a window of static feature coefficients. The distribution over this "extended" feature vector is then computed. By linearly transforming the parameters of the distribution over the extended feature vector, the distribution of the static and dynamic parameters can be found. Though yielding reductions in word error rate, this form of compensation is highly inefficient as it uses many samples per distribution. In this paper, the distributions over extended feature vectors are computed by an extended version of VTS. This approach will be referred to as extended VTS. An important modification to the standard VTS approach is the compensation of the inter-frame correlations. Extended VTS implements the improved dynamic parameter parameter compensation more efficiently than extended DPMC.

Rogier van Dalen is funded by Toshiba Research Europe Ltd. Thanks go to Dr F. Flego and H. Liao for initial code and experiment configurations.

2. MODEL COMPENSATION

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

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

where $\mathbf{C}$ is the DCT matrix. It is standard practice in speech recognition to append dynamic features to the observation vector. They represent the change of the signal over time. Both first- and second-order coefficients ($\boldsymbol{y}_t^{\Delta}, \boldsymbol{y}_t^{\Delta^2}$ respectively) are normally used. Thus the observation feature vector is $\boldsymbol{y}_t = [\ \boldsymbol{y}_t^{\mathsf{s}\mathsf{T}}\ \ \boldsymbol{y}_t^{\Delta\mathsf{T}}\ \ \boldsymbol{y}_t^{\Delta^2\mathsf{T}}\ ]^{\mathsf{T}}$. For clarity of presentation only first-order, delta, coefficients $\boldsymbol{y}^{\Delta}$ 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 $\mathcal{N}\left(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\right)$, then

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

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 (1). There is no closed form for (2), so various approximations are used. The next sections briefly discuss two options, VTS and DPMC.

Prior to performing model compensation, the noise distributions are required. In this work, the noise model gives the distributions of $\boldsymbol{n}$ and $\boldsymbol{h}$. $\boldsymbol{n}$ (including the dynamic parameters) is assumed Gaussian with mean $\boldsymbol{\mu}_n$ and covariance $\boldsymbol{\Sigma}_n$; $\boldsymbol{h} = \boldsymbol{\mu}_h$ is assumed constant [3]. These distributions can be estimated using maximum-likelihood estimation and some data from the testing noise condition [4].

2.1. Vector Taylor series

Equation (1) can be approximated with a first-order vector Taylor series (VTS) [3]. Evaluating the partial derivatives of $\mathbf{f}$ at $\boldsymbol{\mu}^{\mathsf{s}}_n, \boldsymbol{\mu}^{\mathsf{s}}_x, \boldsymbol{\mu}^{\mathsf{s}}_h$, (1) becomes

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

with

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

The corrupted static mean and covariance become [5]

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

$$\boldsymbol{\Sigma}^{\mathsf{s}}_y = \operatorname{diag}\left(\mathbf{J}_x \boldsymbol{\Sigma}^{\mathsf{s}}_x \mathbf{J}_x^{\mathsf{T}} + \mathbf{J}_n \boldsymbol{\Sigma}^{\mathsf{s}}_n \mathbf{J}_n^{\mathsf{T}}\right). \quad (5b)$$

To compensate dynamic parameters, the continuous time approximation [1] is often used in conjunction with VTS. This approximation assumes that delta coefficients are derivatives of static coefficients with respect to time t, so that (for window width w)

$$\boldsymbol{y}^{\Delta}_t = \frac{\sum_{\tau=1}^{w} \tau(\boldsymbol{y}^{\mathsf{s}}_{t+\tau} - \boldsymbol{y}^{\mathsf{s}}_{t-\tau})}{2\sum_{\tau=1}^{w} \tau^2} \approx \left.\frac{\partial \boldsymbol{y}^{\mathsf{s}}}{\partial t}\right|_t; \quad (6)$$

$$\boldsymbol{\mu}^{\Delta}_y = \mathbf{J}_x \boldsymbol{\mu}^{\Delta}_x; \quad \boldsymbol{\Sigma}^{\Delta}_y = \operatorname{diag}\left(\mathbf{J}_x \boldsymbol{\Sigma}^{\Delta}_x \mathbf{J}_x^{\mathsf{T}} + \mathbf{J}_n \boldsymbol{\Sigma}^{\Delta}_n \mathbf{J}_n^{\mathsf{T}}\right). \quad (7)$$

2.2. Data-driven parallel model combination

Data-driven parallel model combination [6] (DPMC) is a Monte Carlo method for estimating the distribution of the corrupted speech. Samples are drawn from the distributions of $\boldsymbol{x}^{\mathsf{s}}$ and $\boldsymbol{n}^{\mathsf{s}}$. (1) then gives the value of $\boldsymbol{y}^{\mathsf{s}}$ for each sample. The expectations in (2) are estimated using the samples of $\boldsymbol{y}^{\mathsf{s}}$. A scheme that compensates dynamic coefficients computed with simple differences is possible.

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 and the speech and noise distributions, and could be viewed as an infinite-order VTS. 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.

3. EXTENDED VTS

The continuous time approximation does not yield accurate compensation. In some cases, performance decreases when it is used [7]. This work uses an alternative method, the key intuition to which is the following. Since dynamic coefficients are a linear combination of consecutive static feature vectors, a distribution over dynamic coefficients can be computed from a distribution over a window of static feature vectors.

For exposition, assume a window of ± 1 and only first-order dynamic coefficients. An *extended* feature vector $\boldsymbol{y}^{\mathsf{e}}_t$, containing the static feature vectors in the surrounding window, is given by $\boldsymbol{y}^{\mathsf{e}}_t = [\,\boldsymbol{y}^{\mathsf{s}}_{t-1}{}^{\mathsf{T}}\ \boldsymbol{y}^{\mathsf{s}}_t{}^{\mathsf{T}}\ \boldsymbol{y}^{\mathsf{s}}_{t+1}{}^{\mathsf{T}}\,]^{\mathsf{T}}$.[1] The transformation of the extended feature vector $\boldsymbol{y}^{\mathsf{e}}_t$ to a feature vector with static and dynamic parameters $\boldsymbol{y}_t$ can be expressed as a linear projection $\mathbf{D}$:

$$\boldsymbol{y}_t = \begin{bmatrix} \boldsymbol{y}^{\mathsf{s}}_t \\ \boldsymbol{y}^{\Delta}_t \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{I} & \mathbf{0} \\ -\frac{\mathbf{I}}{2} & \mathbf{0} & \frac{\mathbf{I}}{2} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}^{\mathsf{s}}_{t-1} \\ \boldsymbol{y}^{\mathsf{s}}_t \\ \boldsymbol{y}^{\mathsf{s}}_{t+1} \end{bmatrix} = \mathbf{D}\boldsymbol{y}^{\mathsf{e}}_t. \quad (8)$$

[1] It is straightforward to extend this to handle both second-order dynamics and linear-regression coefficients over a larger window of $\pm w$, so that $\boldsymbol{y}^{\mathsf{e}}_t = [\boldsymbol{y}^{\mathsf{s}}_{t-w}{}^{\mathsf{T}} \cdots \boldsymbol{y}^{\mathsf{s}}_{t+w}{}^{\mathsf{T}}]^{\mathsf{T}}$.

Since $\mathbf{D}$ is a linear transformation, from the extended distribution $\mathcal{N}\left(\boldsymbol{\mu}^{\mathsf{e}}_y, \boldsymbol{\Sigma}^{\mathsf{e}}_y\right)$ the mean and covariance of the corrupted speech distribution for $\boldsymbol{y}$ are

$$\boldsymbol{\mu}_y = \mathbf{D}\boldsymbol{\mu}^{\mathsf{e}}_y; \qquad \boldsymbol{\Sigma}_y = \mathbf{D}\boldsymbol{\Sigma}^{\mathsf{e}}_y\mathbf{D}^{\mathsf{T}}. \quad (9)$$

The distribution of $\boldsymbol{y}^{\mathsf{e}} \sim \mathcal{N}\left(\boldsymbol{\mu}^{\mathsf{e}}_y, \boldsymbol{\Sigma}^{\mathsf{e}}_y\right)$ depends on the distributions over extended feature vectors of clean speech $\boldsymbol{x}^{\mathsf{e}} \sim \mathcal{N}\left(\boldsymbol{\mu}^{\mathsf{e}}_x, \boldsymbol{\Sigma}^{\mathsf{e}}_x\right)$ and additive noise $\boldsymbol{n}^{\mathsf{e}} \sim \mathcal{N}\left(\boldsymbol{\mu}^{\mathsf{e}}_n, \boldsymbol{\Sigma}^{\mathsf{e}}_n\right)$.

It is interesting to look at the structure of these extended distributions. The mean $\boldsymbol{\mu}^{\mathsf{e}}_y$ of the concatenation of consecutive static feature vectors is simply a concatenation of consecutive static means. The covariance $\boldsymbol{\Sigma}^{\mathsf{e}}_y$, however, has a structure with blocks giving the cross-covariances between static feature vectors for different time instances:

$$\boldsymbol{\mu}^{\mathsf{e}}_y = \begin{bmatrix} \boldsymbol{\mu}^{\mathsf{s}}_{y_{t-1}} \\ \boldsymbol{\mu}^{\mathsf{s}}_{y_t} \\ \boldsymbol{\mu}^{\mathsf{s}}_{y_{t+1}} \end{bmatrix}; \quad \boldsymbol{\Sigma}^{\mathsf{e}}_y = \begin{bmatrix} \boldsymbol{\Sigma}^{\mathsf{s}}_{y_{t-1}y_{t-1}} & \boldsymbol{\Sigma}^{\mathsf{s}}_{y_{t-1}y_t} & \boldsymbol{\Sigma}^{\mathsf{s}}_{y_{t-1}y_{t+1}} \\ \boldsymbol{\Sigma}^{\mathsf{s}}_{y_t y_{t-1}} & \boldsymbol{\Sigma}^{\mathsf{s}}_{y_t y_t} & \boldsymbol{\Sigma}^{\mathsf{s}}_{y_t y_{t+1}} \\ \boldsymbol{\Sigma}^{\mathsf{s}}_{y_{t+1}y_{t-1}} & \boldsymbol{\Sigma}^{\mathsf{s}}_{y_{t+1}y_t} & \boldsymbol{\Sigma}^{\mathsf{s}}_{y_{t+1}y_{t+1}} \end{bmatrix}. \quad (10)$$

As can be seen from (9), linear combinations of the blocks in the matrix $\boldsymbol{\Sigma}^{\mathsf{e}}_y$ form the covariance matrix over normal feature vectors $\boldsymbol{\Sigma}_y$. Therefore, for speech the cross-correlations between time instances cannot be assumed zero. Because the speech model switches between components, the distributions of the speech at different time instances for one component are not the same.

Since the extended feature vector is a concatenation of static feature vectors, it is possible to compensate each time instance separately to find a distribution over $\boldsymbol{y}^{\mathsf{e}}$. This requires distributions over extended feature vectors for the clean speech $\boldsymbol{x}^{\mathsf{e}}$ and additive noise $\boldsymbol{n}^{\mathsf{e}}$, with parameters similar to (10). How to find these will be discussed below.

Extended VTS applies the first-order approximation in (3) to each time instance separately. The expansion point is given by the static means at the appropriate time instances, from the distributions over $\boldsymbol{x}^{\mathsf{e}}$ and $\boldsymbol{n}^{\mathsf{e}}$. The VTS approximation for time instance $t+1$ is

$$\begin{aligned}\boldsymbol{y}^{\mathsf{s}}_{t+1} \approx\ & \mathbf{f}\left(\boldsymbol{\mu}^{\mathsf{s}}_{x_{t+1}}, \boldsymbol{\mu}^{\mathsf{s}}_{n_{t+1}}, \boldsymbol{\mu}^{\mathsf{s}}_h\right) \\ & + \mathbf{J}_{x_{t+1}}(\boldsymbol{x}^{\mathsf{s}}_{t+1} - \boldsymbol{\mu}^{\mathsf{s}}_{x_{t+1}}) + \mathbf{J}_{n_{t+1}}(\boldsymbol{n}^{\mathsf{s}}_{t+1} - \boldsymbol{\mu}^{\mathsf{s}}_{n_{t+1}}), \quad (11)\end{aligned}$$

with the Jacobians

$$\mathbf{J}_{x_{t+1}} = \frac{\partial \boldsymbol{y}^{\mathsf{s}}_{t+1}}{\partial \boldsymbol{x}^{\mathsf{s}}_{t+1}}; \qquad \mathbf{J}_{n_{t+1}} = \frac{\partial \boldsymbol{y}^{\mathsf{s}}_{t+1}}{\partial \boldsymbol{n}^{\mathsf{s}}_{t+1}}. \quad (12)$$

The mean for that time instance is then given by

$$\boldsymbol{\mu}^{\mathsf{s}}_{y_{t+1}} = \mathbf{f}\left(\boldsymbol{\mu}^{\mathsf{s}}_{x_{t+1}}, \boldsymbol{\mu}^{\mathsf{s}}_{n_{t+1}}, \boldsymbol{\mu}^{\mathsf{s}}_h\right). \quad (13)$$

The covariance matrix $\boldsymbol{\Sigma}^{\mathsf{e}}_y$ contains the correlations between all time instances in the window. For example, the covariance between time instance t and $t+1$ is found by generalising (5b) to

$$\boldsymbol{\Sigma}^{\mathsf{s}}_{y_t y_{t+1}} = \mathbf{J}_{x_t} \boldsymbol{\Sigma}^{\mathsf{s}}_{x_t x_{t+1}} \mathbf{J}^{\mathsf{T}}_{x_{t+1}} + \mathbf{J}_{n_t} \boldsymbol{\Sigma}^{\mathsf{s}}_{n_t n_{t+1}} \mathbf{J}^{\mathsf{T}}_{n_{t+1}}. \quad (14)$$

This is applied for each block of (10). The computational cost is dominated by the calculation of the Jacobians, which are computed for each time instance, compared to just once for standard VTS.

A Monte Carlo approach to finding the extended noise-corrupted speech distribution, extended DPMC, has been introduced earlier [2]. Extended DPMC is based on DPMC as described in section 2.2, but it draws extended samples $\boldsymbol{x}^{\mathsf{e}}$ and $\boldsymbol{n}^{\mathsf{e}}$ from the distributions of the clean speech and additive noise. It then applies the mismatch function to each time instance to yield a sample for $\boldsymbol{y}^{\mathsf{e}}_t$. $\boldsymbol{\mu}^{\mathsf{e}}_y$ and $\boldsymbol{\Sigma}^{\mathsf{e}}_y$ are

estimated directly on these samples. This procedure is slow, but generates accurate compensation as the number of samples goes to infinity. This paper uses extended DPMC as a reference compensation method.

A practical issue needs to be considered when using compensation with extended feature vectors: the nature of the statistics for the clean speech and the noise.

For the clean speech, full covariance matrices for $\mathbf{\Sigma}_x^{\mathsf{e}}$ can be stored and used. However, if both 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: for each Gaussian component, the ith element of the static coefficients for a time instance is assumed to be correlated with only itself and the ith element of other time instances. This means that the blocks of $\mathbf{\Sigma}_x^{\mathsf{e}}$ in (10) are diagonalised. This causes $\mathbf{\Sigma}_x^{\mathsf{e}}$ 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, but if the noise is estimated, as in [4], this is complicated and computationally expensive. Unlike the speech model, the noise model has no structure, since it has no state changes. Thus, the blocks on the leading diagonal of $\mathbf{\Sigma}_n^{\mathsf{e}}$ (see (10)) are the same. With the assumption that the blocks are diagonal and the noise of different time instances is uncorrelated, then the estimation scheme in [4] can be used directly and the static elements duplicated for each time instance.

A related scheme that also attempts to improve compensation for dynamic parameters, but in the log-spectral domain, is described in [8]. However, a large number of approximations were made to derive the VTS form, including ignoring correlations between time instances and parameter differences between time instances.

4. EXPERIMENTS

The compensation schemes described were evaluated on an artificially corrupted Resource Management task, and on a corpus collected by Toshiba Research Europe with in-car recorded data.

4.1. Resource Management task

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 extended DPMC. Since the additive background noise is known, it is possible to generate stereo data and use single-pass retraining [6] to obtain "ideally" compensated systems. It is also possible to extract the true noise model.

One approach to assess the quality of the compensation is to compare a compensated model set with the single-pass retrained system trained on stereo data. Figure 1 does this, using the average Kullback-Leibler divergence per diagonal entry of the covariance matrix over all the components. It uses the single-component system, which means that full extended statistics could be extracted. The known additive noise model had a full covariance matrix. The

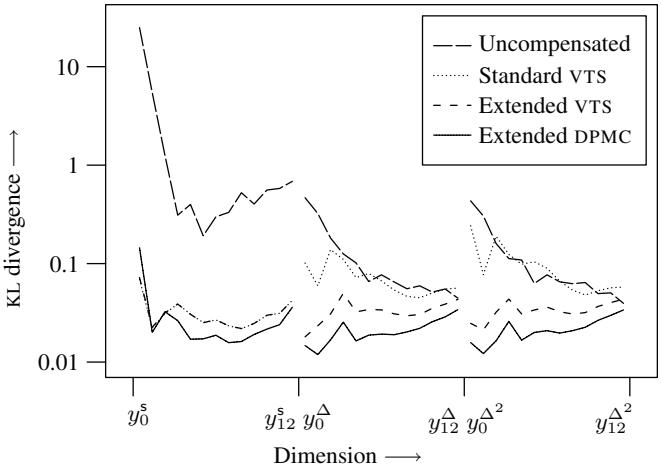

Fig. 1. Average Kullback-Leibler divergence between compensated systems and a single-pass retrained (ideal) system.

first group of coefficients is the statics $\boldsymbol{y}^{\mathsf{s}}$, the second the delta coefficients $\boldsymbol{y}^{\Delta}$, and the third the delta-delta coefficients $\boldsymbol{y}^{\Delta^2}$. As expected, the uncompensated system is furthest away from the single-pass retrained system, and extended DPMC provides the most accurate compensation given the speech and noise models. The difference between standard VTS and extended VTS is interesting. By definition, both yield the same compensation for the statics. For the dynamics, however, the continuous time approximation does not consistently decrease the distance to the single-pass retrained system. Extended VTS, though not as accurate as extended DPMC, provides a substantial improvement over standard VTS.

Scheme	20 dB	14 dB
—	38.1	83.8
VTS	7.3	13.8
eVTS	6.4	12.0
eDPMC	6.4	11.7

Table 1. Word error rates for eVTS compared with standard VTS and eDPMC. Unsupervised noise model estimation at the speaker level.

The previous experiment assumed the noise models were known. Table 1 shows results where noise parameters were estimated per speaker in an unsupervised fashion with a hypothesis from the uncompensated system [4]. The estimates were optimised for standard VTS. The extended noise model distribution was generated by simply duplicating the static estimated noise distribution for standard VTS (as described in section 3). For all cases the noise models had a diagonal covariance matrix structure. Because this was a system with 6 mixture components per state, robustness of the extended clean speech statistics was an issue, so striped covariance matrices were used.

The potential of compensation with extended statistics shows in the difference between the performance of standard VTS and extended DPMC (eDPMC): 13.8 % versus 11.7 % for 14 dB. Two aspects are interesting to note. First, the diagonal extended noise model contains less information than the diagonal noise model for standard VTS. Even with that handicap, compensation with extended VTS (eVTS) and eDPMC is better. The second aspect is that, when compared with eDPMC, the first-order approximation in eVTS de-

grades the performance only slightly at a 14 dB SNR, and not at all at 20 dB. This may be in part because the noise profile is estimated to maximise the log-likelihood of a first-order approximation, albeit for standard VTS.

eVTS and eDPMC are able to produce full covariances as well, which is beneficial especially in low signal-to-noise ratios [2]. In the 14 dB condition, the word error rate for eVTS with full-covariance compensation is 11.4 % compared to 12.0 % for the diagonal case. However, this yields lower gains than going from VTS to eVTS, and decoding with full covariances is computationally expensive, though joint uncertainty decoding and predictive linear transformations [9] can be used. Therefore, this paper concentrates on diagonal covariance compensation.

4.2. Toshiba In-car corpus

Initial experiments were run on a task with real recorded noise: the Toshiba in-car database. This is a corpus collected by Toshiba Research Europe Limited's Cambridge Research Laboratory. It is a small/medium sized task with noisy speech collected in an office and in vehicles driving at various conditions. This work uses three test sets containing digit sequences (phone numbers) recorded in a car with a microphone mounted on the rear-view mirror. The ENON set, which consists of 835 utterances, is recorded with the engine idle, and has a 35 dB average signal-to-noise ratio. The CITY set, which consists of 862 utterances, is recorded driving in cities, and has a 25 dB average signal-to-noise ratio. The HWY set, which consists of 887 utterances, is recorded on the highway, and has a 18 dB average signal-to-noise ratio. Noise compensation was applied to a speech recogniser trained on clean data from the Wall Street Journal corpus. The system was based on the one described in [10], but the number of components was reduced to about 650, more appropriate for an embedded system. The number of components was about 7800. The language model was an open digit loop.

VTS iter.	Word error rate (%) ENON	CITY	HWY
—	3.85	31.81	66.18
0	3.35	8.87	13.11
1	1.24	3.09	3.78
2	1.37	2.65	3.15

Table 2. Iterations of estimating the noise model and finding a hypothesis for standard VTS on the Toshiba in-car task.

A noise model with diagonal additive noise covariance was estimated per utterance for standard VTS. An initial noise model was estimated from the first and last 20 frames and used to find hypothesis $\mathcal{H}^{(0)}$ (iteration 0). Two iterations of maximum likelihood estimation of the noise model and a decoding run were done. $\mathcal{H}^{(2)}$ was then scored. Table 2 shows this process. The extended noise model was found by repeating the static components of the noise model for standard VTS acquired in iteration 2.

Table 3 presents results for standard VTS and extended VTS. Because a per-utterance noise model was used, applying eDPMC was not feasible. For extended VTS, the speech statistics were striped as in table 1. Diagonal compensation is used. From table 3 it can be seen that extended VTS reduces the WER by about 10 % relative compared to standard VTS for all the noise conditions.

Scheme	Word error rate (%) ENON	CITY	HWY
VTS	1.37	2.65	3.15
eVTS	1.14	2.47	2.82

Table 3. Compensation with standard VTS and extended VTS on the Toshiba in-car task. Noise model diagonal and from iteration 2 of estimation for standard VTS in table 2.

5. CONCLUSION

Model-based noise robustness schemes based on VTS normally use the continuous time approximation for dynamic parameter compensation. This paper improves dynamic parameter compensation by introducing extended VTS. It applies a first-order approximation separately to consecutive static coefficients. The distribution over dynamic parameters is then computed with the linear transformation that dynamic coefficients are computed with. The new method was tested on a noise-corrupted Resource Management task, and a Toshiba in-car corpus. With a noise model estimated with maximum likelihood training for standard VTS, extended VTS obtained a 10 % relative reduction in error rate over standard VTS.

6. REFERENCES

[1] 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.

[2] R. C. van Dalen and M. J. F. Gales, "Covariance modelling for for noise robust speech recognition," in *Proceedings of Interspeech*, 2008, pp. 2000–2003.

[3] P. J. Moreno, *Speech Recognition in Noisy Environments*, Ph.D. thesis, Carnegie Mellon University, 1996.

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

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

[6] M. J. F. Gales, *Model-Based Techniques for Noise Robust Speech Recognition*, Ph.D. thesis, Cambridge University, 1995.

[7] 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.

[8] Á. de la Torre, D. Fohr, and J.-P. Haton, "Statistical adaptation of acoustic models to noise conditions for robust speech recognition," in *Proceedings of the* ICSLP, 2002, pp. 1437–1440.

[9] 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.

[10] H. Liao, *Uncertainty Decoding for Noise Robust Speech Recognition*, Ph.D. thesis, Cambridge University, 2007.