Extended VTS for Noise-Robust Speech Recognition

R. C. van Dalen,* *Student Member, IEEE,* and M. J. F. Gales, *Senior Member, IEEE*

Abstract

Model compensation is a standard way of improving the robustness of speech recognition systems to noise. A number of popular schemes are based on vector Taylor series (VTS) compensation, which uses a linear approximation to represent the influence of noise on the clean speech. To compensate the dynamic parameters, the *continuous time approximation* is often used. This approximation uses a point estimate of the gradient, which fails to take into account that dynamic coefficients are a function of a number of consecutive static coefficients. 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 distributions over standard static and dynamic features. With this improved approximation, it is also possible to obtain full-covariance corrupted speech distributions. This addresses the correlation changes that occur in noise. The proposed scheme outperformed the standard VTS scheme by $10\,\%$ to $20\,\%$ relative on a range of tasks.

Index Terms

Speech recognition, noise-robustness, model compensation.

I. INTRODUCTION

Changes in background noise conditions can severely impact the performance of speech recognition systems. Standard approaches to address this problem are to use either feature enhancement or model compensation techniques. The latter have been found to yield good performance, particularly in conditions with low signal-to-noise ratios, and will be the focus of this paper.

The first stage in developing a noise compensation scheme is to express how the noise affects the clean speech. When cepstral-based coefficients 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. There are a range of approximations that can be used to estimate the model parameters given the mismatch function [1], [2]. A commonly used method that has yielded good results approximates the mismatch function with a first-order vector Taylor series (VTS) expansion [3], [1]. Using this VTS approximation it is straightforward to compensate the parameters for the static parameters based on MFCC features. 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. A number of approaches to compensate parameters for these dynamic features have been proposed in the literature [4], [2], [5]. The standard is to use the continuous time approximation [4]. The continuous time approximation makes the assumption is 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. The continuous time approximation allows a mismatch function to be defined for any form of dynamic parameters, both those based on linear regression and simple differences. If only simple differences are considered then it is possible to find compensation by storing extra clean speech statistics [2]. This should be more precise than the continuous time approximation approach, but is only applicable to simple-difference-based dynamic parameters. Another scheme that attempts to improve compensation by using additional statistics, but in the log-spectral domain, is described in [5]. However, as section III-C will show, this approach involves approximations that negate any potential improvements and basically yields the same form as the continuous time approximation. Though there are known limitations to the use of the continuous time approximation it is still the form used in the vast majority of model-based compensation schemes [1], [6], [7].

This paper proposes a new approach for compensating the dynamic parameters that is applicable to both linear-regression and simple-difference based dynamic features. The dynamic coefficients can be expressed as a linear transformation over a window of static feature coefficients. The mismatch function for the static features can be used for each element of the window. Once the distribution over this "extended" feature vector is known, then the distribution of the static and dynamic parameters can be found by linearly transforming the parameters of the distribution over the extended feature vector. In the same fashion as standard model-based compensation there are a range of schemes that can be used to combine the extended clean speech and noise distributions to yield the extended corrupted speech distribution. In this work an extended version of VTS is introduced, called "extended VTS" (eVTS).

Improving the compensation of dynamic parameters also addresses a second problem. Standard model compensation methods diagonalise the covariance matrices for the corrupted speech distributions. This is consistent with the common form of the clean speech distribution, which allows robust parameter estimation and efficient decoding. However, the correlations between the features are expected to change due to variations in the background noise. For example, in the limit, when the noise masks the speech, the correlation pattern will be that of the noise. These effects could be modelled with full-covariance compensation. Though continuous time approximation compensation can generate block-diagonal covariance matrices (one block for static, one for delta and one for delta-delta features), these have not been used for recognition. In this work it is shown that the reason for this, apart from the computational cost, is that compensation with the continuous time approximation is not accurate enough. Full covariance matrix compensation is expected to be more sensitive to approximations in the dynamic parameter compensation than diagonal compensation. However, extended VTS should be more accurate than VTS with the continuous time approximation and thus enable effective full-covariance matrix compensation. Though the use of these full covariance matrices during decoding is computationally expensive, approaches such as predictive linear transforms [8] can be used. This paper only concentrates on the theoretical aspects of improved covariance compensation, rather than the computational load.

This paper introduces extended VTS and discusses how it can be used to generate full-covariance compensation. The organisation of this paper is as follows. The next section surveys model compensation techniques. Section III introduces extended VTS. Section IV discusses how to find the clean speech and noise statistics. Section V examines the accuracy of compensation with standard VTS and extended VTS. Section VI discusses results on a noise-corrupted Resource Management task, AURORA 2, and an in-car recorded corpus.

II. MODEL COMPENSATION

To compute the effect of the acoustic noise on the feature vectors of a speech recogniser, an expression for the mismatch between clean and corrupted speech is needed. The additive noise n and the convolutional noise h transform the clean speech x, resulting in noise-corrupted speech y. In the time domain this has the form [9]

$$y = h \star x + n \tag{1}$$

where $\star$ denotes convolution. Many speech recognition systems are based on MFCC features, which are in the cepstral domain. For time t, the static MFCC noise-corrupted speech vector will be denoted as $\mathbf{y}_t^s$.

Similarly, MFCC vectors of the other features will be written: $\mathbf{x}_t^s$ for the clean speech; $\mathbf{n}_t^s$ for the additive noise; and $\mathbf{h}_t^s$ for the convolutional noise. The mismatch function that relates the static corrupted speech with the sources is [1]

$$\mathbf{y}_t^s = \mathbf{x}_t^s + \mathbf{h}_t^s + \mathbf{C}\log\left(1 + \exp\left(\mathbf{C}^{-1}\left(\mathbf{n}_t^s - \mathbf{x}_t^s - \mathbf{h}_t^s\right)\right)\right)$$
$$= \mathbf{f}\left(\mathbf{x}_t^s, \mathbf{n}_t^s, \mathbf{h}_t^s\right), \tag{2}$$

where $\mathbf{1}$ is a vector of 1s, and $\log(\cdot)$ and $\exp(\cdot)$ indicate the element-wise logarithm and exponent, respectively. The superscript s will be used throughout this paper to denote the static coefficients and parameters.

It is standard practice in HMM-based speech recognition systems to augment the observation vector containing per-time slice (static) features, with dynamic features [10]. They represent the change of the static features over time. Both first- and second-order features ($\mathbf{y}_t^\Delta, \mathbf{y}_t^{\Delta^2}$ respectively) are normally used. The observation feature vector then becomes $\mathbf{y}_t = \left[\mathbf{y}_t^{s\mathsf{T}}\ \mathbf{y}_t^{\Delta\mathsf{T}}\ \mathbf{y}_t^{\Delta^2\mathsf{T}}\right]^\mathsf{T}$. For clarity of presentation only first-order, delta, coefficients $\mathbf{y}^\Delta$ will be considered, though the extension to delta-delta parameters or higher orders is simple. The first-order dynamics at time t are computed from a window $\pm w$ of static coefficients with linear regression [11]:

$$\mathbf{y}_t^\Delta = \frac{\sum_{\tau=1}^w \tau\left(\mathbf{y}_{t+\tau}^s - \mathbf{y}_{t-\tau}^s\right)}{2\sum_{\tau=1}^w \tau^2}. \tag{3}$$

Model compensation schemes combine clean speech and noise distributions with the mismatch function to find the parameters for the noise-corrupted speech model. The clean speech distributions are based on the HMM trained on clean speech data, with Gaussian components $\mathcal{N}\left(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\right)$. In this work, as in many others, the static convolutional (channel) noise is assumed constant $\mathbf{h}_t^s = \boldsymbol{\mu}_h^s$, and the additive noise Gaussian-distributed $\mathbf{n}_t \sim \mathcal{N}\left(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\right)$. These assumptions allow the noise model to be estimated in a maximum likelihood fashion on test data [6], [7] (section II-B will discuss this in detail). Also, since the noise is assumed independent and identically distributed, each clean speech Gaussian can be compensated separately. In this work each noise-corrupted speech component is also assumed Gaussian. The parameters of this Gaussian $\mathcal{N}\left(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\right)$ are[1]

$$\boldsymbol{\mu}_y = \mathcal{E}\left\{\mathbf{y}\right\}; \qquad\qquad \boldsymbol{\Sigma}_y = \mathcal{E}\{(\mathbf{y} - \boldsymbol{\mu}_y)(\mathbf{y} - \boldsymbol{\mu}_y)^\mathsf{T}\}, \tag{4}$$

---

[1]The dependence on the component has been dropped from the notation used in this paper for clarity.

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 the mismatch function in (2). However, no closed forms for the expectations in (4) exist, so approximations must be used.

A. Vector Taylor series compensation

The mismatch function in (2) can be approximated with a first-order vector Taylor series (VTS) [3]. The expansion point is normally set to the means of the clean speech and the noise. In that case, the approximated mismatch function for the static parameters becomes (assuming the convolutional noise is constant $\mathbf{h}_t^s = \boldsymbol{\mu}_h^s$)

$$\mathbf{y}_{t,\text{vts}}^s = \mathbf{f}\left(\boldsymbol{\mu}_x^s, \boldsymbol{\mu}_n^s, \boldsymbol{\mu}_h^s\right) + \mathbf{J}(\mathbf{x}_t^s - \boldsymbol{\mu}_x^s) + (\mathbf{I} - \mathbf{J})(\mathbf{n}_t^s - \boldsymbol{\mu}_n^s), \tag{5}$$

where $\mathbf{I}$ is the identity matrix, and $\mathbf{J}$, a full matrix, is the Jacobian of the clean speech

$$\mathbf{J} = \left.\frac{\partial \mathbf{y}^s}{\partial \mathbf{x}^s}\right|_{\boldsymbol{\mu}_n^s, \boldsymbol{\mu}_x^s, \boldsymbol{\mu}_h^s} = \mathbf{C}\mathbf{J}^{\log}\mathbf{C}^{-1}, \tag{6}$$

with $\mathbf{J}^{\log}$ the log-spectral domain Jacobian, which is diagonal with entries $j_{ii}^{\log} = 1/(1 + \exp([\mathbf{C}^{-1}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_x - \boldsymbol{\mu}_h)]_i))$. If the clean speech mean is much larger than the noise mean, $\mathbf{J}^{\log}$, and therefore $\mathbf{J}$, will tend to $\mathbf{I}$. Conversely, under high noise conditions, $\mathbf{J}$ will tend to $\mathbf{0}$ and $(\mathbf{I} - \mathbf{J})$ in (5) will tend to $\mathbf{I}$.

When the vector Taylor series approximation in (5) is applied to model compensation, the corrupted static mean and covariance of the compensated component become [1]

$$\boldsymbol{\mu}_y^s = \mathcal{E}\left\{\mathbf{y}_{t,\text{vts}}^s\right\} = \mathbf{f}\left(\boldsymbol{\mu}_x^s, \boldsymbol{\mu}_n^s, \boldsymbol{\mu}_h^s\right); \tag{7a}$$

$$\boldsymbol{\Sigma}_y^s = \mathcal{E}\{(\mathbf{y}_{t,\text{vts}}^s - \boldsymbol{\mu}_y^s)(\mathbf{y}_{t,\text{vts}}^s - \boldsymbol{\mu}_y^s)^\mathsf{T}\}$$

$$= \mathbf{J}\boldsymbol{\Sigma}_x^s\mathbf{J}^\mathsf{T} + (\mathbf{I} - \mathbf{J})\boldsymbol{\Sigma}_n^s(\mathbf{I} - \mathbf{J})^\mathsf{T}. \tag{7b}$$

To compensate the dynamic parameters, the continuous time approximation [4] 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

$$\mathbf{y}_t^\Delta \approx \left.\frac{\partial \mathbf{y}^s}{\partial t}\right|_t; \qquad \mathbf{x}_t^\Delta \approx \left.\frac{\partial \mathbf{x}^s}{\partial t}\right|_t; \qquad \mathbf{n}_t^\Delta \approx \left.\frac{\partial \mathbf{n}^s}{\partial t}\right|_t. \tag{8}$$

Combining this approximation and the VTS approximation in (5), the dynamic coefficients become

$$\mathbf{y}_t^\Delta \approx \left.\frac{\partial \mathbf{y}^s}{\partial \mathbf{x}^s}\frac{\partial \mathbf{x}^s}{\partial t}\right|_t + \left.\frac{\partial \mathbf{y}^s}{\partial \mathbf{n}^s}\frac{\partial \mathbf{n}^s}{\partial t}\right|_t \approx \mathbf{J}\mathbf{x}_t^\Delta + (\mathbf{I} - \mathbf{J})\mathbf{n}_t^\Delta = \mathbf{y}_{t,\text{vts}}^\Delta. \tag{9}$$

These mismatch functions can be used to yield the dynamic mean and covariance of the corrupted speech. Since the additive noise is assumed to be stateless, the expected value of its dynamic coefficients, $\boldsymbol{\mu}_\mathsf{n}^\Delta$, is zero. The compensated dynamic parameters are given by

$$\boldsymbol{\mu}_\mathsf{y}^\Delta = \mathcal{E}\left\{\mathbf{y}_{t,\mathrm{vts}}^\Delta\right\} = \mathbf{J}\boldsymbol{\mu}_\mathsf{x}^\Delta; \tag{10a}$$

$$\boldsymbol{\Sigma}_\mathsf{y}^\Delta = \mathcal{E}\{(\mathbf{y}_{t,\mathrm{vts}}^\Delta - \boldsymbol{\mu}_\mathsf{y}^\Delta)(\mathbf{y}_{t,\mathrm{vts}}^\Delta - \boldsymbol{\mu}_\mathsf{y}^\Delta)^\mathsf{T}\}$$
$$= \mathbf{J}\boldsymbol{\Sigma}_\mathsf{x}^\Delta\mathbf{J}^\mathsf{T} + (\mathbf{I} - \mathbf{J})\boldsymbol{\Sigma}_\mathsf{n}^\Delta(\mathbf{I} - \mathbf{J})^\mathsf{T}. \tag{10b}$$

The compensated mean and covariance matrix are formed by concatenating the static and dynamic parameters:

$$\boldsymbol{\mu}_\mathsf{y} = \begin{bmatrix} \boldsymbol{\mu}_\mathsf{y}^\mathsf{s} \\ \boldsymbol{\mu}_\mathsf{y}^\Delta \end{bmatrix}; \qquad \boldsymbol{\Sigma}_\mathsf{y} = \begin{bmatrix} \boldsymbol{\Sigma}_\mathsf{y}^\mathsf{s} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\mathsf{y}^\Delta \end{bmatrix}. \tag{11}$$

The resulting covariance matrix $\boldsymbol{\Sigma}_\mathsf{y}$ is block-diagonal in structure, as the Jacobian matrix $\mathbf{J}$ is full. Standard VTS does not yield full covariances. Decoding with this block-diagonal structure has two problems. First, it is computationally expensive. Second, the continuous time approximation for the dynamic parameters does not yield accurate block-diagonal compensation (section V-B will discuss this in more detail). Therefore, when decoding the following, standard, form for the output probability is used:

$$p\left(\mathbf{y}_t\right) = \mathcal{N}\left(\mathbf{y}_t\,;\,\boldsymbol{\mu}_\mathsf{y}\,,\,\mathrm{diag}\left(\boldsymbol{\Sigma}_\mathsf{y}\right)\right), \tag{12}$$

where $\mathrm{diag}\left(\cdot\right)$ denotes matrix diagonalisation.

B. Noise estimation

The discussion so far has assumed that the distribution of the noise is known. In practice, however, this is seldom the case. The noise model must therefore be estimated. The model $\mathcal{M}_n = \{\boldsymbol{\mu}_\mathsf{n}, \boldsymbol{\Sigma}_\mathsf{n}, \boldsymbol{\mu}_\mathsf{h}\}$ comprises the parameters of the additive noise, assumed Gaussian with $\mathcal{N}\left(\boldsymbol{\mu}_\mathsf{n}, \boldsymbol{\Sigma}_\mathsf{n}\right)$, and the convolutional noise $\boldsymbol{\mu}_\mathsf{h}$, which is assumed constant. The parameters are of the form

$$\boldsymbol{\mu}_\mathsf{n} = \begin{bmatrix} \boldsymbol{\mu}_\mathsf{n}^\mathsf{s} \\ \mathbf{0} \end{bmatrix}; \qquad \boldsymbol{\Sigma}_\mathsf{n} = \begin{bmatrix} \mathrm{diag}\left(\boldsymbol{\Sigma}_\mathsf{n}^\mathsf{s}\right) & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}\left(\boldsymbol{\Sigma}_\mathsf{n}^\Delta\right) \end{bmatrix}; \qquad \boldsymbol{\mu}_\mathsf{h} = \begin{bmatrix} \boldsymbol{\mu}_\mathsf{h}^\mathsf{s} \\ \mathbf{0} \end{bmatrix}. \tag{13}$$

The expected value of the dynamic coefficients of the additive noise are zero because the noise model has no state changes. Since the convolutional noise is assumed constant, its dynamic parameters are also zero. Using data from the target environment, it is possible to find a noise estimate with expectation–maximisation that maximises the likelihood assuming a particular form of model-based compensation,

for example VTS compensation [3], [12]. This iteratively updates the component-time posteriors (the expectation step) and the noise model (the maximisation step).

In the maximisation step, the static noise means can be updated at the same time using a fixed-point iteration [3]. The additive noise covariance, however, is more complex to estimate. It is possible to estimate it on the parts of the waveform known to contain noise without speech. Another options is to use gradient ascent to find an estimate for the additive noise variance for VTS with the continuous time approximation [12]. This needs to be alternated with the estimation of the noise mean.

The resulting noise model estimate maximises the likelihood of model compensation with VTS. Thus, the parameters do not necessarily correspond to the actual noise or to a consistent sequence of static observations.

III. EXTENDED VTS

The continuous time approximation yields a simple approach to compensating the dynamic feature model parameters. However, it relies on an approximation which has been found to degrade performance for some noise conditions [7]. This section describes an alternative method for compensating the dynamic model parameters called "extended VTS". The key intuition is that the dynamic coefficients are a linear combination of consecutive static feature vectors. Thus, if the corrupted speech distribution over the consecutive static feature vectors can be estimated then the distribution for the dynamic coefficients can be found.

A. Model compensation with extended feature vectors

For simplicity, a window of ± 1 and only first-order dynamic coefficients will be considered. An *extended* feature vector $\mathbf{y}_t^{\mathrm{e}}$, containing the static feature vectors in the surrounding window, is given by $\mathbf{y}_t^{\mathrm{e}} = \begin{bmatrix} \mathbf{y}_{t-1}^{\mathrm{s}\ \mathsf{T}} & \mathbf{y}_t^{\mathrm{s}\mathsf{T}} & \mathbf{y}_{t+1}^{\mathrm{s}\ \mathsf{T}} \end{bmatrix}^{\mathsf{T}}$.[2] The transformation of the extended feature vector $\mathbf{y}_t^{\mathrm{e}}$ to the standard feature vector with static and dynamic parameters $\mathbf{y}_t$ can be expressed as a linear projection $\mathbf{D}$:

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{y}_t^{\mathrm{s}} \\ \mathbf{y}_t^{\Delta} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{I} & \mathbf{0} \\ -\frac{\mathbf{I}}{2} & \mathbf{0} & \frac{\mathbf{I}}{2} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{t-1}^{\mathrm{s}} \\ \mathbf{y}_t^{\mathrm{s}} \\ \mathbf{y}_{t+1}^{\mathrm{s}} \end{bmatrix} = \mathbf{D}\mathbf{y}_t^{\mathrm{e}}. \tag{14}$$

---

[2]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 $\mathbf{y}_t^{\mathrm{e}} = \begin{bmatrix} \mathbf{y}_{t-w}^{\mathrm{s}\ \mathsf{T}} & \cdots & \mathbf{y}_{t+w}^{\mathrm{s}\ \mathsf{T}} \end{bmatrix}^{\mathsf{T}}$.

The second row of $\mathbf{D}$ applies the transformation from a window of statics to yield the standard delta features. $\mathbf{D}$ is a linear transformation, so that if the distribution of the extended corrupted speech $\mathbf{y}_t^e$ is Gaussian with mean $\boldsymbol{\mu}_y^e$ and covariance $\boldsymbol{\Sigma}_y^e$, it can be transformed to a distribution over statics and dynamics with

$$\boldsymbol{\mu}_y = \mathcal{E}\{\mathbf{y}\} = \mathbf{D}\mathcal{E}\{\mathbf{y}^e\} = \mathbf{D}\boldsymbol{\mu}_y^e; \tag{15a}$$

$$\boldsymbol{\Sigma}_y = \mathcal{E}\left\{\mathbf{y}\mathbf{y}^\mathsf{T} - \boldsymbol{\mu}_y\boldsymbol{\mu}_y^\mathsf{T}\right\} = \mathbf{D}\mathcal{E}\left\{\mathbf{y}^e\mathbf{y}^{e\mathsf{T}} - \boldsymbol{\mu}_y^e\boldsymbol{\mu}_y^{e\mathsf{T}}\right\}\mathbf{D}^\mathsf{T}$$

$$= \mathbf{D}\boldsymbol{\Sigma}_y^e\mathbf{D}^\mathsf{T}. \tag{15b}$$

It is interesting to look at the structure distribution of the extended feature vector, $\mathbf{y}^e$. The mean $\boldsymbol{\mu}_y^e$ of the concatenation of consecutive static feature vectors is simply a concatenation of static means at time offsets $-1, 0, +1$. For the corrupted speech, these will be written $\boldsymbol{\mu}_{y_{-1}}^s, \boldsymbol{\mu}_{y_0}^s, \boldsymbol{\mu}_{y_{+1}}^s$. The covariance $\boldsymbol{\Sigma}_y^e$ contains the covariance between statics at different time offsets. The covariance between offsets -1 and $+1$, for example, is written $\boldsymbol{\Sigma}_{y_{-1}y_{+1}}$. Thus, the full parameters of the extended distribution are

$$\boldsymbol{\mu}_y^e = \begin{bmatrix} \boldsymbol{\mu}_{y_{-1}}^s \\ \boldsymbol{\mu}_{y_0}^s \\ \boldsymbol{\mu}_{y_{+1}}^s \end{bmatrix}; \qquad \boldsymbol{\Sigma}_y^e = \begin{bmatrix} \boldsymbol{\Sigma}_{y_{-1}y_{-1}}^s & \boldsymbol{\Sigma}_{y_{-1}y_0}^s & \boldsymbol{\Sigma}_{y_{-1}y_{+1}}^s \\ \boldsymbol{\Sigma}_{y_0y_{-1}}^s & \boldsymbol{\Sigma}_{y_0y_0}^s & \boldsymbol{\Sigma}_{y_0y_{+1}}^s \\ \boldsymbol{\Sigma}_{y_{+1}y_{-1}}^s & \boldsymbol{\Sigma}_{y_{+1}y_0}^s & \boldsymbol{\Sigma}_{y_{+1}y_{+1}}^s \end{bmatrix}. \tag{16}$$

The standard parameters with statics and dynamics can be found from this extended distribution with (15). The problem is to find the correct form for the extended distribution. It is possible to use sampling methods to do this. This is discussed in more detail in [13]. However, this is slow, so that the next section will present a faster first-order approximation, extended VTS. A sampling methods that in the limit yields the best possible compensation, extended DPMC [14] , will be used to assess the performance of extended VTS.

B. Extended distribution compensation

The procedure to find the distribution over the statics and dynamics of the corrupted speech outlined in the previous section requires parameters for the extended corrupted speech distribution in (16). Since the extended feature vector is a concatenation of static feature vectors, it is possible to use the static mismatch function for each time offset to yield an overall mismatch function for $\mathbf{y}^e$.

An extension to static compensation using VTS can be used to find the extended corrupted speech distribution. The first-order vector Taylor series approximation in (5) is applied to each time instance

separately. Thus the expansion point for each time instance is given by the static means at the appropriate time offsets. These are obtained from the extended distributions over the clean speech, $\mathbf{x}^e$, and noise, $\mathbf{n}^e$. Thus using the form of the VTS approximation in (5) to time offset $+1$, for example:

$$\mathbf{y}^s_{t+1,\text{evts}} = \mathbf{f}(\boldsymbol{\mu}^s_{x_{+1}}, \boldsymbol{\mu}^s_{n_{+1}}, \boldsymbol{\mu}^s_{h_{+1}}) + \mathbf{J}_{+1}(\mathbf{x}^s_{t+1} - \boldsymbol{\mu}^s_{x_{+1}})$$

$$+ (\mathbf{I} - \mathbf{J}_{+1})(\mathbf{n}^s_{t+1} - \boldsymbol{\mu}^s_{n_{+1}}), \quad (17)$$

where the offset-dependent Jacobian is given by

$$\mathbf{J}_{+1} = \left.\frac{\partial \mathbf{y}^s}{\partial \mathbf{x}^s}\right|_{\boldsymbol{\mu}^s_{n_{+1}}, \boldsymbol{\mu}^s_{x_{+1}}, \boldsymbol{\mu}^s_{h_{+1}}}. \quad (18)$$

Note that $\mathbf{J}_0$ is equal to the Jacobian for standard VTS described in (6).

The expression for the corrupted speech in (17) requires distributions over extended feature vectors for the clean speech $\mathbf{x}^e$ (from training data) and noise $\mathbf{n}^e, \mathbf{h}^e$ (estimated). The forms of these distributions are analogous to those for the extended corrupted speech in (16). For example, the extended distribution of the clean speech contains static means $\boldsymbol{\mu}^s_{x_{-1}}, \boldsymbol{\mu}^s_{x_0}, \boldsymbol{\mu}^s_{x_{+1}}$, and cross-covariances between time offsets like $\boldsymbol{\Sigma}_{x_{-1}x_{+1}}$.

Approaches and approximations for these clean speech and noise statistics will be discussed in section IV-A.

The parameters of the extended corrupted speech distribution in (16) can be found by computing expectations over the distributions of $\mathbf{x}^e, \mathbf{n}^e, \mathbf{h}^e$. The mean for time offset $+1$, for example, is given by

$$\boldsymbol{\mu}^s_{y_{+1}} = \mathcal{E}\left\{\mathbf{y}^s_{t+1,\text{evts}}\right\} = \mathbf{f}(\boldsymbol{\mu}^s_{x_{+1}}, \boldsymbol{\mu}^s_{n_{+1}}, \boldsymbol{\mu}^s_{h_{+1}}). \quad (19)$$

The covariance matrix $\boldsymbol{\Sigma}^e_y$ (shown in (16)) requires the correlations between all time offsets in the window to be computed. The covariance between offsets 0 and $+1$, for example, is found by generalising (7b) to

$$\boldsymbol{\Sigma}^s_{y_0 y_{+1}} = \mathcal{E}\{(\mathbf{y}^s_{t,\text{evts}} - \boldsymbol{\mu}_{y_0})(\mathbf{y}^s_{t+1,\text{evts}} - \boldsymbol{\mu}_{y_{+1}})^\mathsf{T}\}$$

$$= \mathbf{J}_0 \boldsymbol{\Sigma}^s_{x_0 x_{+1}} \mathbf{J}^\mathsf{T}_{+1} + (\mathbf{I} - \mathbf{J}_0)\boldsymbol{\Sigma}^s_{n_0 n_{+1}}(\mathbf{I} - \mathbf{J}_{+1})^\mathsf{T}. \quad (20)$$

This is applied for each of the time offset blocks in $\boldsymbol{\Sigma}^e_y$.

C. Relationship between VTS *and e*VTS

Extended VTS uses a different vector Taylor series expansion point for every time offset to find more accurate compensation. The approximation that extended VTS applies to the mismatch function at the centre time offset is exactly the same as that of standard VTS. [13] gives a detailed derivation of the case

where that expansion point is used for all time instances. It turns out that in this case, extended VTS compensation becomes equivalent to standard VTS compensation.

A scheme related to eVTS was described in [5]. This proposed a similar form of linear transformation of a window of static parameters to obtain the compensated dynamic parameters. However, it used the same expansion point for all time offsets, negating the advantages in accuracy of explicitly modelling distributions over a window of features. This makes the scheme equivalent to standard VTS. A number of additional approximations were applied. Correlations between time instances were ignored. This discards information compared to standard speech recognisers' features: it becomes impossible to reconstruct the covariance over statics and dynamics.

D. Efficiency

The computational complexity of eVTS for full-covariance compensation is higher than of standard VTS. The main extra cost compared to standard VTS is in the number of blocks in the covariance matrix to compensate. For VTS, these blocks are the covariances of the statics, the deltas, and the delta-deltas (in (7b) and (10b)). For eVTS, they are the cross-covariances of all time instances (in (20)). Computing compensation for each of these blocks takes equal time. If the window of statics in the extended feature vectors is 9, then eVTS needs to perform this computation for $\frac{1}{2}(9 \times 10) = 45$ blocks against 3 for standard VTS. A more detailed analysis is given in [13].

In practice, however, per-Gaussian compensation is often too costly even when the standard version of VTS is used. Joint uncertainty decoding (JUD) [15] addresses this by computing compensation per base class rather than per Gaussian component. [13] discusses how eVTS can be used within the JUD compensation framework. Another issue is the computational cost of decoding with full covariance matrices. Predictive linear transformations [8] can solve this issue by applying transformation to the feature vectors, eliminating the need to decode with full covariance matrices. This is discussed in more detail in [13]. This paper concentrates on finding accurate compensation per component.

IV. EXTENDED STATISTICS

A practical issue when using eVTS is the form of the statistics for the clean speech and the noise. For standard VTS, the clean speech statistics are usually taken from the recogniser trained on clean speech and the noise model is usually estimated with maximum likelihood estimation, as discussed in section II-B. In contrast, eVTS requires distributions over the extended clean speech and noise vectors.

As these have more parameters than standard statistics, robustness and storage requirements need to be carefully considered.

A. Clean speech statistics

Model compensation schemes, such as VTS, use the Gaussian components from the uncompensated system as the clean speech distributions. For eVTS, however, distributions over the extended clean speech vector are required. For one extended clean speech Gaussian $\mathcal{N}\left(\boldsymbol{\mu}_{\mathsf{x}}^{\mathsf{e}}, \boldsymbol{\Sigma}_{\mathsf{x}}^{\mathsf{e}}\right)$, the parameters are

$$\boldsymbol{\mu}_{\mathsf{x}}^{\mathsf{e}} = \begin{bmatrix} \boldsymbol{\mu}_{\mathsf{x}_{-1}}^{\mathsf{s}} \\ \boldsymbol{\mu}_{\mathsf{x}_{0}}^{\mathsf{s}} \\ \boldsymbol{\mu}_{\mathsf{x}_{+1}}^{\mathsf{s}} \end{bmatrix} ; \qquad \boldsymbol{\Sigma}_{\mathsf{x}}^{\mathsf{e}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathsf{x}_{-1}\mathsf{x}_{-1}}^{\mathsf{s}} & \boldsymbol{\Sigma}_{\mathsf{x}_{-1}\mathsf{x}_{0}}^{\mathsf{s}} & \boldsymbol{\Sigma}_{\mathsf{x}_{-1}\mathsf{x}_{+1}}^{\mathsf{s}} \\ \boldsymbol{\Sigma}_{\mathsf{x}_{0}\mathsf{x}_{-1}}^{\mathsf{s}} & \boldsymbol{\Sigma}_{\mathsf{x}_{0}\mathsf{x}_{0}}^{\mathsf{s}} & \boldsymbol{\Sigma}_{\mathsf{x}_{0}\mathsf{x}_{+1}}^{\mathsf{s}} \\ \boldsymbol{\Sigma}_{\mathsf{x}_{+1}\mathsf{x}_{-1}}^{\mathsf{s}} & \boldsymbol{\Sigma}_{\mathsf{x}_{+1}\mathsf{x}_{0}}^{\mathsf{s}} & \boldsymbol{\Sigma}_{\mathsf{x}_{+1}\mathsf{x}_{+1}}^{\mathsf{s}} \end{bmatrix} . \tag{21}$$

As with standard model compensation schemes, when there is no noise the compensated system should be the original clean system. To ensure that this is the case, single-pass retraining [2] should be used to obtain the extended clean speech distributions. Here the posteriors associated with the complete data set for EM of the last standard clean speech training iteration (with static and dynamic parameters) are used to accumulate extended feature vectors around every time instance.

Another problem with using the extended statistics is ensuring robust estimation. The extended feature vectors contain more parameters than the standard static and dynamic ones. Hence, the estimates of their distributions will be less robust and take up more memory. If full covariance matrices for $\boldsymbol{\Sigma}_{\mathsf{x}}^{\mathsf{e}}$ are stored and used, both first- and second-order dynamic parameters use window widths of ± 2, and there are n static parameters, this requires estimating a $9n \times 9n$ covariance matrix for every component. This is memory-intensive and singular matrices and numerical accuracy problems can occur. One solution is to reduce the number of Gaussian components or states in the system. However, the precision of the speech model then decreases. Also, this makes it hard to compare the performance of compensation with extended VTS and standard VTS.

An alternative approach is to modify the structure of the covariance matrices, in the same fashion as diagonalising the standard clean speech covariance model. To maintain some level of inter-frame correlations, which may be useful for computing the dynamic parameters, each block is diagonalised. This yields the following structure:

$$\boldsymbol{\Sigma}_{\mathsf{x}}^{\mathsf{e}} = \begin{bmatrix} \operatorname{diag}\left(\boldsymbol{\Sigma}_{\mathsf{x}_{-1}\mathsf{x}_{-1}}^{\mathsf{s}}\right) & \operatorname{diag}\left(\boldsymbol{\Sigma}_{\mathsf{x}_{-1}\mathsf{x}_{0}}^{\mathsf{s}}\right) & \operatorname{diag}\left(\boldsymbol{\Sigma}_{\mathsf{x}_{-1}\mathsf{x}_{+1}}^{\mathsf{s}}\right) \\ \operatorname{diag}\left(\boldsymbol{\Sigma}_{\mathsf{x}_{0}\mathsf{x}_{-1}}^{\mathsf{s}}\right) & \operatorname{diag}\left(\boldsymbol{\Sigma}_{\mathsf{x}_{0}\mathsf{x}_{0}}^{\mathsf{s}}\right) & \operatorname{diag}\left(\boldsymbol{\Sigma}_{\mathsf{x}_{0}\mathsf{x}_{+1}}^{\mathsf{s}}\right) \\ \operatorname{diag}\left(\boldsymbol{\Sigma}_{\mathsf{x}_{+1}\mathsf{x}_{-1}}^{\mathsf{s}}\right) & \operatorname{diag}\left(\boldsymbol{\Sigma}_{\mathsf{x}_{+1}\mathsf{x}_{0}}^{\mathsf{s}}\right) & \operatorname{diag}\left(\boldsymbol{\Sigma}_{\mathsf{x}_{+1}\mathsf{x}_{+1}}^{\mathsf{s}}\right) \end{bmatrix} . \tag{22}$$

For each Gaussian component, the ith element of the static coefficients for a time instance is then assumed correlated with only itself and the ith element of other time instances. This type of covariance matrix will be called "striped". A striped $\boldsymbol{\Sigma}_\mathsf{x}^\mathsf{e}$ has only $45n$ parameters rather than $9n(9n+1)/2$ for the full case. A useful attribute of this structure is that it is possible to reconstruct a diagonal covariance over statics and dynamics from this, so no information is discarded compared to conventional clean speech models.

B. Noise model estimation

A noise model with extended feature vectors is necessary to perform compensation with extended VTS. This noise model is of the form $\mathcal{M}_n^\mathsf{e} = \{\boldsymbol{\mu}_\mathsf{n}^\mathsf{e}, \boldsymbol{\Sigma}_\mathsf{n}^\mathsf{e}, \boldsymbol{\mu}_\mathsf{h}^\mathsf{e}\}$. In this work, and the majority of other work, the noise model consists of a single Gaussian component for the additive noise, and the convolutional noise is assumed constant. The distribution for each time offset therefore is by definition the same. This means that the extended means for the additive and convolutional noise simply repeat the static means. The structure of the extended covariance $\boldsymbol{\Sigma}_\mathsf{n}^\mathsf{e}$ is also known. Since the noise is assumed identically distributed for all time instances at the same distance, the correlation between time instances is always the same. Thus, all diagonals of the covariance matrix repeat the same entries. Let $\boldsymbol{\Sigma}_{\mathsf{n}_0}^\mathsf{s}, \boldsymbol{\Sigma}_{\mathsf{n}_1}^\mathsf{s}, \boldsymbol{\Sigma}_{\mathsf{n}_2}^\mathsf{s}$ indicate the cross-correlation between noise that is 0, 1, or 2 time instances apart. The extended noise model then has the following form:

$$\boldsymbol{\mu}_\mathsf{n}^\mathsf{e} = \begin{bmatrix} \boldsymbol{\mu}_\mathsf{n}^\mathsf{s} \\ \boldsymbol{\mu}_\mathsf{n}^\mathsf{s} \\ \boldsymbol{\mu}_\mathsf{n}^\mathsf{s} \end{bmatrix} ; \qquad \boldsymbol{\Sigma}_\mathsf{n}^\mathsf{e} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathsf{n}_0}^\mathsf{s} & \boldsymbol{\Sigma}_{\mathsf{n}_1}^{\mathsf{s}\mathsf{T}} & \boldsymbol{\Sigma}_{\mathsf{n}_2}^{\mathsf{s}\mathsf{T}} \\ \boldsymbol{\Sigma}_{\mathsf{n}_1}^\mathsf{s} & \boldsymbol{\Sigma}_{\mathsf{n}_0}^\mathsf{s} & \boldsymbol{\Sigma}_{\mathsf{n}_1}^{\mathsf{s}\mathsf{T}} \\ \boldsymbol{\Sigma}_{\mathsf{n}_2}^\mathsf{s} & \boldsymbol{\Sigma}_{\mathsf{n}_1}^\mathsf{s} & \boldsymbol{\Sigma}_{\mathsf{n}_0}^\mathsf{s} \end{bmatrix} ; \qquad \boldsymbol{\mu}_\mathsf{h}^\mathsf{e} = \begin{bmatrix} \boldsymbol{\mu}_\mathsf{h}^\mathsf{s} \\ \boldsymbol{\mu}_\mathsf{h}^\mathsf{s} \\ \boldsymbol{\mu}_\mathsf{h}^\mathsf{s} \end{bmatrix} . \qquad (23)$$

In theory these noise parameters could be found using maximum likelihood estimation. However, this complicates the noise estimation process. It would be preferable to use the standard noise estimation schemes and map the parameters to the ones in the extended forms above. These "standard" noise parameters are

$$\boldsymbol{\mu}_\mathsf{n} = \begin{bmatrix} \boldsymbol{\mu}_\mathsf{n}^\mathsf{s} \\ \mathbf{0} \end{bmatrix} ; \qquad \boldsymbol{\Sigma}_\mathsf{n} = \begin{bmatrix} \mathrm{diag}\left(\boldsymbol{\Sigma}_\mathsf{n}^\mathsf{s}\right) & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}\left(\boldsymbol{\Sigma}_\mathsf{n}^\Delta\right) \end{bmatrix} ; \qquad \boldsymbol{\mu}_\mathsf{h} = \begin{bmatrix} \boldsymbol{\mu}_\mathsf{h}^\mathsf{s} \\ \mathbf{0} \end{bmatrix} . \qquad (24)$$

The extended noise means are straightforward functions of the static means of the standard noise model (24). Similarly, $\boldsymbol{\Sigma}_{\mathsf{n}_0}^\mathsf{s}$, the covariance between noise 0 time instances apart, is the static noise covariance $\boldsymbol{\Sigma}_\mathsf{n}^\mathsf{s}$. Computing the off-diagonals of the extended covariance, however, is not as straightforward.

A simple way of reconstructing the extended noise covariance from a standard noise model assumes that the noise is uncorrelated between time instances. This is done by setting the off-diagonal elements are set to zero, which yields

$$\boldsymbol{\Sigma}_n^e = \begin{bmatrix} \boldsymbol{\Sigma}_n^s & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_n^s & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma}_n^s \end{bmatrix}. \tag{25}$$

This only uses the static elements of the estimated noise covariance. For very low signal-to-noise ratio (SNR) conditions this form of extended noise distribution will not yield the standard noise distributions for the dynamic parameters.

It is also possible to reconstruct the off-diagonal elements in (25) from the dynamic parameters of the standard noise model. However, the ML-estimated dynamic parameters do not necessarily correspond to true sequences of noise samples. This is examined in detail in [13]. It turns out that reconstructing off-diagonals elements yields good performance on noise models directly trained on the artificially added noise, but poorer performance with ML-estimated noise models. The experiments in section VI will therefore use the diagonal reconstruction for the extended noise distribution.

1) Zeros in the noise variance estimate: An additional issue that can occur when estimating the noise model using maximum likelihood, is that noise variances estimates for some dimensions can become very small, or zero. Though this value may optimise the likelihood, it does not necessarily reflect the "true" noise variance. This can lead to the following problems in compensation.

One problem for the small noise variance estimates is that the clean speech "silence" models are never really estimated on silence. In practice even for clean speech there are always low levels of background noise. Thus the estimated noise is really only relative to this clean background level. At very high SNRs the noise may be at a similar level to the clean "silence". This will cause very small noise variance values. Another problem results from the form of the covariance matrix compensation. For the static parameters this is (repeated from (7b))

$$\boldsymbol{\Sigma}_y^s = \mathbf{J}\boldsymbol{\Sigma}_x^s\mathbf{J}^{\mathsf{T}} + (\mathbf{I} - \mathbf{J})\boldsymbol{\Sigma}_n^s(\mathbf{I} - \mathbf{J})^{\mathsf{T}}. \tag{26}$$

At low SNRs $\mathbf{J} \to \mathbf{0}$, so the corrupted distribution tends to the noise distribution. Conversely, at high SNRs $\mathbf{J} \to \mathbf{I}$ as the corrupted speech distribution tends to the clean speech distribution. The impact of this when estimating the noise covariance matrix $\boldsymbol{\Sigma}_n^s$ in high SNR conditions is that changes in the form of the noise covariance matrix have little impact on the final compensated distribution.

When VTS with the continuous time approximation is used with diagonal corrupted speech covariance matrices for both noise estimation and recognition, then the process is self-consistent. However if the noise estimates are used with eVTS to find full compensated covariance matrices this is not the case. This slight mismatch can cause problems. To address this issue a back-off strategy can be used. When the estimated noise variance has very low values rather than using full compensated covariance matrices, diagonal compensated variances can be used. This will occur at high SNRs, where the correlation changes compared to the clean speech conditions should be minimal. In this condition little gain is expected from full compensated covariance matrices.

An alternative approach to address this problem is to make the noise estimation and decoding consistent for eVTS. This is not investigated in this work. By using the same noise estimates for both VTS and eVTS, only differences in the compensation process are examined, rather than any differences in the noise estimation process. It should be emphasised that the results presented for eVTS may a slight underestimate of the possible performance if a fully integrated noise estimate was used. Integrated noise estimation with eVTS will be investigated in future work using, for example, the approach described in [16]. Here the joint distribution of the corrupted speech and extended noise is modelled using a Gaussian, and EM used to find the extended noise distribution.

V. PRELIMINARY EXPERIMENTS

Normally, word error rates are used to evaluate performance of compensation methods. However, this does not allow a detailed assessment of which aspects of the compensation process are working well and which poorly. An alternative approach is to compare compensated systems' distributions to their ideal counterparts. Experiments in this section used artificially corrupted training data to create an ideally compensated system.

A noise-corrupted version of the Resource Management task was used, the full details of which will be discussed in section VI-A. It is a 1000-word vocabulary task with 3.8 hours of training data. State-clustered cross-word triphone models with 6 components per state were built using the HTK RM recipe. To ensure robust extended clean speech statistics, per-state distributions were clustered to single components (as opposed to in section IV-A). The total number of Gaussians was 1600. NOISEX-92 Operations Room noise was artificially added at a 14 dB SNR. Hence it was possible to obtain the "correct" noise distribution, for both the standard and extended feature vector cases. For these experiments full covariance matrices

were used both for VTS and for the extended statistics in eVTS.[3]

The *known* noise situation allows the accuracy of the compensation scheme to be compared to the ideal, single pass retrained, system [2]. The Gaussian components in a single-pass-retrained system are estimated using noise-corrupted speech data and the component posteriors (associated with the complete data set) from the clean training data. This may be viewed as an ideal compensation scheme for the case where each Gaussian is adapted independently (as in this paper), as the corrupted speech distributions are directly based on corrupting the clean speech data using noise. It is then possible to examine how close the compensated Gaussians are to those in the single-pass retrained system. A useful comparison metric for this is the occupancy-weighted average of the component-for-component Kullback-Leibler (KL) divergence of the compensated system to the single-pass retrained system [2]. If $p^{(m)}$ is the mth Gaussian of the single-pass retrained system, and $q^{(m)}$ is the corresponding Gaussian of the compensated system, then this metric $\mathcal{D}$ is

$$\mathcal{D} = \sum_m \gamma^{(m)} \mathcal{KL}\big(p^{(m)} \| q^{(m)}\big) \Big/ \sum_m \gamma^{(m)}, \tag{27}$$

where $\gamma^{(m)}$ is the occupancy of component m in the last training iteration, for both the compensated and the single-pass retrained system.

A useful attribute of this form of metric is that, depending on the structure of the covariance matrices, it is possible to assess the compensation per coefficient or block of coefficients. When diagonal covariance matrices are used, each dimension may be considered separately. This allows the accuracy of the compensation scheme to be assessed for each dimension. Similarly, block-diagonal compensation can be examined per block of coefficients.

A. Diagonal compensation

Normally VTS-compensated covariance matrices are diagonalised. Thus it is interesting to initially examine this configuration. Using diagonal covariance matrices also allows each dimension to be assessed. Figure 1 contrasts the accuracy of an uncompensated system, and three forms of compensation: standard VTS, extended VTS, and, as an indicator of maximum possible performance, extended DPMC. This graph is for a 14 dB SNR, but graphs for other SNRs are very similar. The horizontal axis has the feature dimensions: 13 static MFCCs $\mathbf{y}^{\mathsf{s}}$, 13 first-order dynamics $\mathbf{y}^{\Delta}$, and 13 second-order dynamics $\mathbf{y}^{\Delta^2}$. As

---

[3]Similar trends were observed when striped noise statistics, consistent with diagonal standard noise models for VTS, were used for eVTS.
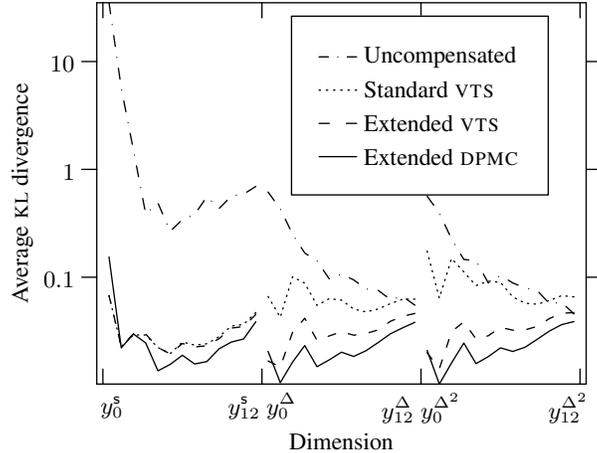
Fig. 1.   Average Kullback-Leibler divergence between compensated systems and a single-pass retrained (ideal) system.

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 deltas and especially the delta-deltas, 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.

B. Block-diagonal compensation

The previous section used diagonal covariance matrices. To compensate for changing correlations under noise more complex covariance matrix structures, such as full or block-diagonal, may be useful. Both VTS with the continuous time approximation, and eVTS, can also be used to generate block-diagonal covariance matrices for the output distributions. The form of this was shown in (11). The KL divergence to a single-pass retrained system at the block level can then be used. This allows the compensation of each of the following blocks of features to be individually assessed: the statics, and first- and second-order dynamics. VTS compensation uses block-diagonal statistics for both the clean speech and noise models. For eVTS the extended statistics have full covariance matrices.

Table I shows the average KL divergence between a system compensated with block-diagonal VTS with the continuous time approximation and the block-diagonal SPR system. Numbers for different SNRs were very similar. VTS finds compensated parameters close to the SPR system for the static features: the KL divergence goes from 58.8 to 1.0. However, the dynamic parameters are not compensated as

TABLE I

RESOURCE MANAGEMENT TASK: AVERAGE KL DIVERGENCE TO A BLOCK-DIAGONAL SINGLE-PASS RETRAINED SYSTEM

FOR VTS (CONTINUOUS TIME), EVTS AND DPMC AT 14 DB SNR.

Compensation	—	VTS	eVTS	eDPMC
$\mathbf{y}^{\text{s}}$	58.8	1.0	1.0	1.0
$\mathbf{y}^{\Delta}$	3.3	1.4	0.7	0.5
$\mathbf{y}^{\Delta^2}$	3.2	1.7	0.7	0.5

accurately, though both the delta and delta-delta parameters are still somewhat closer to the SPR system than the uncompensated model set (3.2 to 1.7 for the delta-deltas). Similarly to diagonal compensation (see figure 1), with block-diagonal covariances standard VTS finds good compensation for the static parameters, but less good for the deltas and delta-deltas.

eVTS has the same compensation as standard VTS for the statics. As in the diagonal-covariance case, however, for dynamic parameters compensation it is more accurate. It does yield a clear improvement over the uncompensated system (3.2 to 0.7 for the delta-deltas) and is close to eDPMC, which in the limit yields the best obtainable compensation.

VI. EXPERIMENTS

The performance of extended VTS was examined on three tasks. Two used artificial noise, and are therefore useful to test the noise estimation and behaviour at various signal-to-noise ratios. The first task is the Resource Management [17] corpus artificially corrupted with NOISEX data [18]. AURORA 2 [19] is a well-known digit recognition task with artificial noise. The final task used real, in-car recorded data collected by Toshiba Research Europe. Results on this task can give insight in performance when other effects, such as the Lombard effect, may impact performance.

For all tasks, clean training data was used to train the speech models. 39-dimensional feature vectors were used: 12 MFCCs and the zeroth coefficient, augmented with deltas and delta-deltas. Unless indicated otherwise, the MFCCs were found with HTK [11] and the deltas and delta-deltas were computed over a window of 2 observations left and 2 right, making the total window width 9.

Unlike results in the previous section, for all tasks the noise models were estimated, finding the maximum likelihood noise model [12], as described in section II-B, for compensating a clean system with VTS and the continuous time approximation. The initial noise model's Gaussian for the additive noise was the maximum-likelihood estimate from the first 20 and last 20 frames of the utterance, which

TABLE II

RESOURCE MANAGEMENT TASK: WORD ERROR RATES FOR APPROXIMATIONS OF eVTS. DIAGONAL-COVARIANCE

DECODING.

Jacobians	$\mathbf{\Sigma}_{\times}^{e}$	20 dB	14 dB
Standard VTS		6.8	13.7
Fixed	diag	7.8	13.5
Variable	diag	7.5	13.0
Variable	striped	6.2	12.0

were assumed to contain no speech. The initial convolutional noise estimate was **0**. Given this initial noise estimate for an utterance, a recognition hypothesis was found. This was used to find component-time posteriors. Then, the noise means and the additive noise covariance were re-estimated. For eVTS the extended noise model was reconstructed from the standard noise model with diagonal covariance, as described in section IV-B.

A. Resource Management task

The Resource Management task, which was also used in section V, is a medium-vocabulary task, with a 1000-word vocabulary. A noise-corrupted version of the this task was generated by adding Operations Room and Car noise from the NOISEX-92 database scaled to yield SNRs of 20 dB, 14 dB, and 8 dB. The training data contains 109 speakers reading 3990 sentences, 3.8 hours of data. State-clustered cross-word triphone models with 6 components per mixture were built using the HTK RM recipe. For this work the noise model was estimated per speaker. This made DPMC compensation feasible. The number of samples per distribution for extended DPMC was set to 10 000. (Performance did not improve with additional samples.) The noise covariance estimate did not contain any zero entries, so back-off as discussed in section IV-B1 was not necessary. 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.

Table II investigates the properties of extended VTS and striped statistics. The first row gives word error rates for standard VTS at 20 and 14 dB. As discussed in section III-C, extended VTS becomes standard VTS if the expansion point is chosen equal for all time instances. Varying the expansion point is expected to provide better compensation for dynamics. On the other hand, diagonalising extended clean speech statistics discards information compared to diagonalising standard features with statics and dynamics. For the second line in the table, the expansion points vary, but the Jacobian is fixed. At the lower SNR, the

TABLE III

RESOURCE MANAGEMENT TASK: WORD ERROR RATES FOR STANDARD VTS, EXTENDED VTS AND EXTENDED DPMC.

Scheme	Speech	Σ_y	Operations Room			Car		
			20	14	8	20	14	8
VTS	diag Σ_x	diag	6.8	13.7	30.0	5.2	9.1	18.7
	blck Σ_x	blck	7.0	14.2	31.6	5.3	9.7	20.1
eVTS	strpd Σ_x^e	diag	6.2	12.0	27.9	4.8	8.5	18.2
		full	6.3	11.2	26.7	5.0	8.3	17.9
eDPMC	strpd Σ_x^e	diag	6.3	11.9	27.9	4.8	8.2	16.4
		full	6.0	11.3	26.3	4.7	7.9	15.9

improved modelling helps, but at the higher SNR, where the corrupted speech distributions are closer to the clean speech, diagonalising the extended clean speech statistics discards vital information. For the third row, the Jacobians are allowed to vary, which gives complete eVTS, if with diagonal speech statistics. The bottom row uses striped statistics, as discussed in section IV-A, which discards no information compared to diagonal standard statistics. This leads to a consistent improvement over standard VTS. The following experiments will therefore use striped statistics for all extended feature vector systems.

Table III shows contrasts between compensation with standard VTS and with extended feature vectors using either eVTS or eDPMC. The results in the first row are from the standard scheme, diagonal-covariance compensation with VTS. Block-diagonal compensation with standard VTS was also implemented and block-diagonal clean speech statistics were used. The results for this approach are in the second row. The use of the block-diagonal compensation with VTS degraded performance, for example 13.7 % to 14.2 % for Operations Room noise at 14 dB.

Compensation with eVTS (shown in the middle two rows of the table) yielded better performance than standard VTS for both diagonal and full compensated covariance matrices. For diagonal-covariance compensation, the relative improvement is 5–10 % (6.8 to 6.2 %; 13.7 to 12.0 %, etc.) over standard VTS. Though at the higher SNR condition, 20 dB, full-covariance compensation did not improve performance over diagonal-covariance performance, gains were observed at the lower SNRs. For Operations Room noise at 14 dB, full-covariance compensation produces an 11.2 % word error rate, which is a 20 % relative improvement from standard VTS, and 7 % relative gain compared to the diagonal case.

In addition table III shows the performance of eDPMC, which in the limit can be viewed as the Gaussian compensation scheme. The results for this approach are shown in the bottom two rows of table III. When

TABLE IV

AURORA: DIAGONAL COMPENSATION WITH STANDARD VTS AND FULL COMPENSATION WITH EXTENDED VTS.

Scheme	VTS						eVTS		
Comp.	diagonal			block-diagonal			full		
SNR	A	B	C	A	B	C	A	B	C
00	28.2	26.2	25.9	24.3	22.9	23.6	23.5	24.3	22.6
05	10.5	9.3	9.9	8.2	7.9	8.2	7.1	7.2	6.9
10	4.3	3.9	4.4	3.3	3.2	3.4	2.5	2.4	2.8
15	2.2	2.2	2.3	1.9	1.8	1.8	1.2	1.2	1.4
20	1.6	1.4	1.6	1.3	1.2	1.2	0.8	0.7	1.0
Avg.	9.4	8.6	8.8	7.8	7.4	7.7	7.0	7.2	6.9

compared with eDPMC, the first-order approximation in eVTS degrades performance by up to 0.4 % absolute, except for 8 dB Car noise. However, eVTS is significantly faster than eDPMC.

B. AURORA task

AURORA 2 is a small vocabulary digit string recognition task [19]. Though it is less complex than the Resource Management corpus, it is a standard corpus for testing noise-robustness. Utterances are one to seven digits long and based on the TIDIGITS database with noise artificially added. The clean speech training data comprises 8440 utterances from 55 male and 55 female speakers. The test data is split into three sections. Test set A comprises 4 noise conditions: subway, babble, car and exhibition hall. Matched training data is available for these test conditions, but not used in this work. Test set B comprises 4 different noise conditions. For both test set A and B the noise was scaled and added to the waveforms. For the two noise conditions in test set C convolutional noise was also added. Each of the conditions has a test set of 1001 sentences with 52 male and 52 female speakers.

The feature vectors were extracted with the ETSI front-end [19]. The delta and delta-delta coefficients used 2 and 3 frames left and right, respectively, for a total window of 11 frames. The acoustic models were 16 emitting state whole word digit models, with 3 mixtures per state and silence. Since for this task, as for the Resource Management task, the noise estimates did not contain zero elements in the variance, the back-off strategy for the noise estimate discussed in section IV-B1 was not necessary.

Table IV shows results for compensation with VTS with the continuous time approximation and eVTS. Both diagonal and block-diagonal forms of VTS were used. VTS with diagonal compensation (trained on diagonal speech statistics) is the standard method. Results for this are shown in the first three columns

of table IV and are treated as the baseline performance figures. VTS can also be used to produce block-diagonal covariance matrices. With diagonal-covariance clean speech statistics (not in the table), this yielded no performance gain. However, performance gains were obtained when using block-diagonal clean-speech models, unlike for Resource Management. This difference in performance between the tasks is felt to be because of the additional complexity of the RM task compared to AURORA.

The results for this are shown in the middle three columns of table IV. Compared to the standard diagonal VTS scheme, this gave, for example, relative reductions in word error rate of 15 % to 22 % at 5 dB SNR.

The results for eVTS are shown in the last three columns of table IV. Here full covariance matrix extended clean speech models were used to produce compensated full covariance matrices for decoding. The improved compensation for dynamics causes extended VTS to perform better than to block-diagonal standard VTS in all but one noise conditions. At 5 dB again, relative improvements are an extra 3 % to 10 %.

The results presented here used the simple AURORA back-end. Large gains over the standard VTS approach (similar results for VTS are given in [7]) were obtained. Using the simple back-end recogniser, rather than one with more Gaussian components per state, has ensured that block-diagonal and full covariance matrix clean speech models can be robustly used. Though not always practical this indicates the possible gains from schemes such as eVTS on a standard task.

C. Toshiba in-car task

Experiments were also 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 comprises a set of small/medium sized tasks with noisy speech collected in an office and in vehicles driving at various conditions. This work used three of the 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, was recorded with the engine idle, and has a 35 dB average signal-to-noise ratio. The CITY set, which consists of 862 utterances, was recorded driving in the city, and has a 25 dB average signal-to-noise ratio. The HWY set, which consists of 887 utterances, was recorded on the highway, and has a 18 dB average signal-to-noise ratio. The clean speech models were trained on the Wall Street Journal corpus, based on the system described in [15], but the number of states was reduced to about 650, more appropriate for an embedded system. The acoustic models used were cross-word triphones decision-tree clustered per state, with three emitting states per HMM, twelve components per GMM and diagonal covariance matrices. The

TABLE V

EXTENDED VTS ON THE TOSHIBA IN-CAR TASK.

Scheme	Decoding	ENON 35 dB	CITY 25 dB	HWY 18 dB
VTS	diag	1.2	2.5	3.2
eVTS	diag	1.1	2.4	2.8
	full	1.7	2.5	2.4
eVTS	back-off	1.1	2.2	2.4
% utterances		87 %	38 %	11 %

number of components was about 7800. For the eVTS scheme extended clean speech statistics were again striped for robustness. The language model was an open digit loop. In the noise estimation stage, the noise model was re-estimated twice on a new hypothesis.

Table V shows results on the Toshiba task. The top row contains word error rates for the standard compensation method: VTS trained on diagonal speech statistics. The performance of eVTS using diagonal covariance matrices is shown in the second row. Again eVTS shows gains over VTS, especially at the lowest SNR condition, HWY. In the HWY condition about a 12% relative reduction in error rate was obtained.

Initially full-covariance matrix compensation with eVTS was evaluated without the use of the back-off scheme described in section IV-B1. Using eVTS with full-covariance decoding yielded additional gains compared to diagonal compensation at low SNRs (2.8 % to 2.4 %). However the performance was degraded at higher SNR conditions, for example ENON where performance was degraded from 1.1 % to 1.7 %.

In contrast to the previous tasks, at high SNRs there were found to be zeros in the noise variance estimate. The back-off scheme, labelled "eVTS back-off" in table V, was therefore used. Here diagonal covariance matrix compensation was used if any noise variance estimate fell below 0.05 times the variance floor used for clean speech model training (results were consistent over a range of values from 0.0 to 0.1). The bottom line in table V shows the percentage of utterances for each of the task where the system was backed off to diagonal covariance matrix compensation. As expected the percentage at high SNRs, 87%, was far higher than at lower SNRs, 11%. Using this back-off approach gave consistent gains over using either diagonal or full compensation eVTS alone. Note as the back-off is based on the ML-estimated noise variances it is fully automated. Compared to standard VTS, eVTS with back-off gave relative reductions

of 25% in the HWY condition, 12% in CITY, and 8% in ENON.

VII. CONCLUSION

A popular form of model compensation scheme to handle changes in background noise is VTS. In order to achieve the best performance it is necessary to compensate all the model parameters to reflect the impact of the background noise on the speech. Though it is easy to define a mismatch function for the static parameters, it is non-trivial to specify one for the dynamic parameters. To address this the continuous time approximation is often used. This paper has introduced an alternative, more accurate, approximation, extended VTS (eVTS). Here the distribution over dynamic parameters is computed based on a linear transformation of a window of static parameters. By using this more accurate dynamic parameter compensation scheme it is possible to use more complex covariance matrices in the compensated system. Computing full-covariance matrix compensation is more sensitive to approximations of the mismatch function. eVTS allows accurate full covariance matrices to be generated from the compensation process. This enables the acoustic models to better reflect changes in the correlations that result from the varying noise conditions.

The new method was tested on a noise-corrupted Resource Management task, AURORA 2, and a Toshiba in-car corpus. With a noise model estimated with maximum likelihood training for standard VTS, extended VTS obtained about 10 % relative reduction in error rate over standard VTS for diagonal compensation at higher signal-to-noise ratios, and about 20 % for full-covariance compensation at lower signal-to-noise ratios.

Though eVTS yields reductions in word error rate with full covariance matrices, there are a number of refinements that can be implemented. First to address the computational cost, JUD or predictive approaches can be used. The current implementation of eVTS uses noise model estimates based on VTS with the standard continuous time approximation. Improved performance may be obtained by using noise estimates based on eVTS. Also, canonical model parameters could be estimated with adaptive training, similar to [20]. Finally, improved efficient approximations, for example using unscented transformations, may also be used. In initial experiments this was found to yield gains with extended feature vector approaches. All these approaches will be examined in future work.

REFERENCES

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

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

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

[4] 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.

[5] 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* ICSLP, 2002, pp. 1437–1440.

[6] H. Liao and M. J. F. Gales, "Adaptive training with joint uncertainty decoding for robust recognition of noisy data," in *Proceedings of* ICASSP, vol. IV, 2007, pp. 389–392.

[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* ASRU, 2007, pp. 65–70.

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

[9] A. Acero, "Acoustical and environmental robustness in automatic speech recognition," Ph.D. dissertation, Carnegie Mellon University, 1990.

[10] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE *Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.4)," 2006. [Online]. Available: http://htk.eng.cam.ac.uk/docs/docs.shtml

[12] 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.

[13] R. C. van Dalen and M. J. F. Gales, "Extended VTS for noise-robust speech recognition," Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR.636, September 2009. [Online]. Available: http://mi.eng.cam.ac.uk/~rcv25/pdf/van_dalen_tr636_extended_vts.pdf

[14] ——, "Covariance modelling for noise robust speech recognition," in *Proceedings of Interspeech*, 2008, pp. 2000–2003.

[15] H. Liao, "Uncertainty decoding for noise robust speech recognition," Ph.D. dissertation, Cambridge University, 2007.

[16] D. Kim, C. Un, and N. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," *Speech Communication*, vol. 24, pp. 39–49, 1998.

[17] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *Proceedings of* ICASSP, vol. 1, 1988, pp. 651–654.

[18] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[19] H.-G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions," in *Proceedings of* ASR, 2000, pp. 181–188.

[20] O. Kalinli, M. Seltzer, and A. Acero, "Noise adaptive training using a vector Taylor series approach for noise robust automatic speech recognition," in *Proceedings of* ICASSP, April 2009, pp. 3825–3828.

Rogier van Dalen (S'07) graduated with an M.Sc. degree in Computer Science from Delft University of Technology and an M.A. in English Literature and Linguistics from Leiden University, both in the Netherlands. In 2007 he received an M.Phil. in Computer Speech, Text and Internet Technology from Cambridge University, UK. Since October 2007 has been pursuing the Ph.D. degree at Cambridge University, on noise-robust speech recognition.

From 2005 to 2006, he was a Lecturer in artificial intelligence at Delft University for a year. In summer 2007 he was a Research Assistant at the Engineering Department, Cambridge University.

Mark Gales (M'01–SM'09) received the B.A. degree in electrical and information sciences and the Ph.D. degree from the University of Cambridge, Cambridge, UK, in 1988 and 1995, respectively.

Following graduation, he worked as a Consultant at Roke Manor Research, Ltd. In 1991, he took up a position as a Research Associate in the Speech, Vision and Robotics Group, Engineering Department, Cambridge University. From 1995 to 1997, he was a Research Fellow at Emmanuel College, Cambridge. He was then a Research Staff Member in the Speech Group, IBM T.J. Watson Research Center, Yorktown Heights, NY until 1999 when he returned to the Engineering Department, Cambridge University, as a University Lecturer. He is currently a Reader in Information Engineering and a Fellow of Emmanuel College.