Observation Process Adaptation for Linear Dynamic Models

Joe Frankel and Simon King

Centre for Speech Technology Research
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
Tel: +44 131 651 1769
Fax: +44 131 650 4587

Abstract

This work introduces two methods for adapting the observation process parameters of linear dynamic models (LDM) or other linear-Gaussian models. The first method uses the expectation-maximization (EM) algorithm to estimate transforms for location and covariance parameters, and the second uses a generalized EM (GEM) approach which reduces computation in making updates from $O(p^6)$ to $O(p^3)$, where p is the feature dimension. We present the results of speaker adaptation on TIMIT phone classification and recognition experiments with relative error reductions of up to 6%. Importantly, we find minimal differences in the results from EM and GEM. We therefore propose that the GEM approach be applied to adaptation of hidden Markov models which use non-diagonal covariances. We provide the necessary update equations.

Key words:
linear dynamic model, acoustic model adaptation, ASR, MLLR, GEM, HMM

1 Introduction

We first motivate the current study before outlining maximum likelihood linear regression (MLLR), the framework within which we develop adaptation for linear dynamic models (LDM).

Email address: joe@cstr.ed.ac.uk (Joe Frankel and Simon King).
URL: http://www.cstr.ed.ac.uk (Joe Frankel and Simon King).

The LDM, also known as the Kalman filter model, has been the subject of research and application by the engineering, control and machine learning communities. With $\mathbf{y}_t$ and $\mathbf{x}_t$ respectively denoting p and q dimensioned continuous-valued observation and state vectors at time t, the LDM is described by the following pair of equations:

$$\mathbf{y}_t = H\mathbf{x}_t + \boldsymbol{\epsilon}_t \qquad \boldsymbol{\epsilon}_t \sim N(\mathbf{v}, C) \tag{1}$$
$$\mathbf{x}_t = F\mathbf{x}_{t-1} + \boldsymbol{\eta}_t \qquad \boldsymbol{\eta}_t \sim N(\mathbf{w}, D) \tag{2}$$

and a distribution over the initial state, $\mathbf{x}_1 \sim N(\boldsymbol{\pi}, \Lambda)$. The LDM is a generative model, giving a time-varying multivariate Gaussian distribution over the observations. Underlying dynamics are modelled by the state evolution which is according to a first-order auto-regressive (AR) process. Equation 1 describes the *observation process* and Equation 2 describes the *state process*. See Frankel (2003) or Frankel and King (2007. In press) for more detail on the properties of these models, or for information on the wider class of linear Gaussian models see Roweis and Ghahramani (1999) or Rosti and Gales (2001).

Experiments reported in Digalakis (1992) and Frankel (2003) support the suitability of a linear predictor (such as given by the LDM) for modelling speech parameter dependencies *within* phone segments, and a number of authors have investigated the application of LDMs to acoustic modelling for automatic speech recognition (ASR) (Digalakis, 1992; Frankel, 2003; Ma and Deng, 2004a,b; Rosti, 2004; Frankel and King, 2007. In press). Motivation for this choice of model includes:

- first-order dynamics of state give a model of inter-frame correlations
- spatial correlations can be modelled fully or approximated via projection of lower dimensional state
- passing state information across phone boundaries relaxes the assumption of segmental independence
- continuous underlying representation reflects known properties of speech production

Frankel and King (2007. In press) demonstrated that the addition of a hidden dynamic state in LDMs gave rise to improved phone classification and recognition [1] accuracies over equivalent static models. The increases were statistically

[1] Phone recognition involves a joint search over phone sequence and segmentation; in classification the boundaries are given and the identity of each segment must be inferred.

significant although modest and, given the extra computational cost of LDMs compared to frame-based models, do not make a strong case for these models. However, we suggest that the averaging which occurs across speakers reduces the contribution of the LDM's state by reducing its ability to learn speaker-independent dynamics with the state process. A speaker-adaptive observation process should therefore lead to improved modelling.

With $\boldsymbol{\epsilon}_t \sim N(\mathbf{v}, C)$, the observation process dictates the location of the observed features with the parameters H and $\mathbf{v}$, and the distribution of measurement errors with the covariance [2] C. In this work we consider adaptation of these observation process parameters with a view to minimizing the differences between speakers *in the state space*.

1.2 MLLR adaptation for LDMs

Adaptation, whether to environment, channel, or speaker has become an integral part of modern ASR systems. One technique which has been successfully integrated into hidden Markov model (HMM) systems is maximum likelihood linear regression (MLLR) (Gales and Woodland, 1996). Linear transforms of the mean and/or covariance are estimated according to a maximum likelihood criterion via the expectation-maximization (EM) algorithm (Dempster et al., 1977; Bilmes, 1997). This involves defining an auxiliary function Q:

$$Q(\Theta^{(i+1)}, \Theta^{(i)}) = E_{\Theta^{(i)}}\left[l(\Theta^{(i+1)}|\mathcal{Y}, \mathcal{X})|\mathcal{Y}\right] \tag{3}$$

which is maximized at each iteration to step toward a maximum likelihood solution. $\mathcal{Y}$ is the observation sequence, $\mathcal{X}$ is the state sequence and $\Theta^{(i)}$ is the model parameters at iteration i.

In this work, we adapt only the observation process parameters H, $\mathbf{v}$ and C, though it would be straightforward to transfer the same techniques to the state process parameters.

Recalling that the observation process is defined by

$$\begin{aligned} \mathbf{y}_t &= H\mathbf{x}_t + \boldsymbol{\epsilon}_t \\ \boldsymbol{\epsilon}_t &\sim N(\mathbf{v}, C) \end{aligned} \tag{4}$$

we adapt model m as follows:

[2] With $\mathbf{v}$ estimated, we assume our errors to have zero mean.

$$\hat{H}_m = AH_m \tag{5}$$
$$\hat{\mathbf{v}}_m = \mathbf{v}_m + \mathbf{b} \tag{6}$$
$$\hat{C}_m = B_m^T \; G \; B_m \tag{7}$$

where B_m is the Cholesky decomposition of C_m, so that $C_m = B_m^T B_m$. Note that the adaptation parameters A, $\mathbf{b}$ and G are not subscripted by m: they are estimated as common parameters shared by all models within pre-defined clusters.

As stated in Gales and Woodland (1996), maximizing the mean and covariance together is problematic, and therefore the transforms are estimated in two stages such that

$$l(\check{\Theta}|\mathcal{Y},\mathcal{X}) \geq l(\hat{\Theta}|\mathcal{Y},\mathcal{X}) \geq l(\Theta|\mathcal{Y},\mathcal{X}) \tag{8}$$

where $\hat{\Theta}$ denotes a model set for which the transforms have been applied to the location parameters H and $\mathbf{v}$, and $\check{\Theta}$ to model sets where transforms have also been applied to the observation noise covariance C.

We propose two different approaches to estimating adaptation parameters for the location parameters. The first uses the EM algorithm, following the full covariance method of Gales (1997) [3], and the second uses a generalized EM (GEM) approach which reduces both processing and memory requirements.

2 Deriving updates for LDMs

With $\mathcal{Y} = \mathbf{y}_1^N$ and $\mathcal{X} = \mathbf{x}_1^N$ denoting sequences of p-dimensional observation and q-dimensional state vectors respectively, the LDM's Markovian structure means that the joint likelihood of state and observations can be written as:

$$p(\mathcal{Y},\mathcal{X}|\Theta) = p(\mathbf{x}_1|\Theta) \prod_{t=2}^{N} p(\mathbf{x}_t|\mathbf{x}_{t-1},\Theta) \prod_{t=1}^{N} p(\mathbf{y}_t|\mathbf{x}_t,\Theta) \tag{9}$$

The state is assumed to have a Gaussian initial density, and so the joint log-likelihood for the LDM is a sum of quadratic terms. We assume that the segmentation is given, and use the subscript m_r to denote the model used to generate segment r. Using the notation:

$$\boldsymbol{\delta}_{\mathbf{y}_t}^{m_r} = \mathbf{y}_t - H_{m_r}\mathbf{x}_t - \mathbf{v}_{m_r} \tag{10}$$

[3] The output distribution of the LDM has a non-diagonal covariance, meaning that the original MLLR (Gales and Woodland, 1996), which was tailored to diagonal-covariance HMMs, is not applicable

the portion of the log-likelihood function relating to the observations is given by:

$$l(\Theta|\mathcal{Y},\mathcal{X}) \propto \sum_{r=1}^{R}\sum_{t=a_r}^{b_r}\left\{\log|C_{m_r}| + \boldsymbol{\delta}_{\mathbf{y}_t}^{m_r\,T} C_{m_r}^{-1}\boldsymbol{\delta}_{\mathbf{y}_t}^{m_r}\right\} \tag{11}$$

where a_r and b_r denote the start and end time of segment r respectively, and R is the total number of segments. Since the state and observation parameters are linearly separable in the log-likelihood function, we need only consider Equation 11 in deriving updates for the adaptation parameters.

For convenience, we introduce the following notation:

$$\mathbf{x}_{t|N_r} = E_{\Theta^{(i)}}\left[\mathbf{x}_t|\mathcal{Y}_r\right] \tag{12}$$

$$\begin{aligned} P_t &= E_{\Theta^{(i)}}\left[\mathbf{x}_t\mathbf{x}_t^T|\mathcal{Y}_r\right] \\ &= \Sigma_{t|N_r} + \mathbf{x}_{t|N_r}\mathbf{x}_{t|N_r}^T \end{aligned} \tag{13}$$

These are the expectations of $\mathbf{x}_t$ and $\mathbf{x}_t\mathbf{x}_t^T$ evaluated using the parameter set $\Theta^{(i)}$ with respect to $\mathcal{Y}_r$, the N_r-length observation sequence corresponding to segment r. The estimates of the state mean and covariance, $\mathbf{x}_{t|N_r}$ and $\Sigma_{t|N_r}$ respectively, given observations $\mathcal{Y}_r$ can be computed using an RTS smoother (Rauch, 1963) as detailed in Frankel (2003).

2.1 *EM transformation of H and $\mathbf{v}$*

Replacing H and $\mathbf{v}$ in Equation 10 with their adapted equivalents, $\hat{H}$ and $\hat{\mathbf{v}}$, gives:

$$\hat{\boldsymbol{\delta}}_{\mathbf{y}_t}^{m_r} = \mathbf{y}_t - AH_{m_r}\mathbf{x}_t - \mathbf{v}_{m_r} - \mathbf{b} \tag{14}$$

and therefore an auxiliary function of:

$$Q(\Theta^{(i+1)},\Theta^{(i)}) \propto E_{\Theta^{(i)}}\left[\sum_{r=1}^{R}\sum_{t=a_r}^{b_r}\left\{\log|C_{m_r}| + \hat{\boldsymbol{\delta}}_{\mathbf{y}_t}^{m_r\,T} C_{m_r}^{-1}\hat{\boldsymbol{\delta}}_{\mathbf{y}_t}^{m_r}\right\}\right] \tag{15}$$

A maximum with respect to A is found by differentiating and setting to zero, giving:

$$\sum_{r=1}^{R}\sum_{t=a_r}^{b_r} C_{m_r}^{-1}\left(\hat{A}H_{m_r}P_tH_{m_r}^T + \hat{\mathbf{b}}\mathbf{x}_{t|N_r}^T H_{m_r}^T\right) = \sum_{r=1}^{R}\sum_{t=a_r}^{b_r} C_{m_r}^{-1}(\mathbf{y}_t - \mathbf{v}_{m_r})\mathbf{x}_{t|N_r}^T H_{m_r}^T \tag{16}$$

Similarly, differentiating with respect to to $\mathbf{b}$ yields:

$$\sum_{r=1}^{R}\sum_{t=a_r}^{b_r} C_{m_r}^{-1}\left(\hat{A}H_{m_r}\mathbf{x}_{t|N_r}+\mathbf{b}\right)=\sum_{r=1}^{R}\sum_{t=a_r}^{b_r} C_{m_r}^{-1}(\mathbf{y}_t-\mathbf{v}_{m_r}) \tag{17}$$

Letting $\boldsymbol{\gamma}_{\mathbf{y}_t}^{m_r}=\mathbf{y}_t-\mathbf{v}_{m_r}$, we introduce the following notation:

$$S^{(r)}=C_{m_r}^{-1} \tag{18}$$

$$Q^{(r)}=\begin{bmatrix}\sum_{t=a_r}^{b_r} H_{m_r}P_tH_{m_r}^T & \sum_{t=a_r}^{b_r} H_{m_r}\mathbf{x}_{t|N_r}\\ \sum_{t=a_r}^{b_r} \mathbf{x}_{t|N_r}^T H_{m_r}^T & b_r-a_r+1\end{bmatrix} \tag{19}$$

$$Z=\sum_{r=1}^{R} C_{m_r}^{-1}\sum_{t=1}^{N}\left[\boldsymbol{\gamma}_{\mathbf{y}_t}^{m_r}\mathbf{x}_{t|N_r}^T H_{m_r}^T\ \ \boldsymbol{\gamma}_{\mathbf{y}_t}^{m_r}\right] \tag{20}$$

$$\hat{\Phi}=\left[\hat{A}\ \hat{\mathbf{b}}\right] \tag{21}$$

which allows Equations 16 and 17 to be expressed as:

$$\sum_{r=1}^{R} S^{(r)}\hat{\Phi}Q^{(r)}=Z \tag{22}$$

Denoting element i,j of $S^{(r)}$, $\hat{\Phi}$, $Q^{(r)}$ and Z with $s_{i,j}^{(r)}$, $\hat{\phi}_{i,j}$, $q_{i,j}^{(r)}$ and $z_{i,j}$ respectively, we can write

$$\sum_{r=1}^{R}\sum_{k=1}^{p+1}\sum_{l=1}^{p} s_{i,l}^{(r)}\hat{\phi}_{l,k}q_{k,j}^{(r)}=z_{i,j} \tag{23}$$

$$\Rightarrow \sum_{k=1}^{p+1}\sum_{l=1}^{p}\hat{\phi}_{l,k}\sum_{r=1}^{R} s_{i,l}^{(r)}q_{k,j}^{(r)}=z_{i,j} \tag{24}$$

Equation 24 provides a set of $p(p+1)$ simultaneous equations in $p(p+1)$ unknowns which can be solved to find $\hat{\phi}$.

2.2 Generalized EM adaptation of H and $\mathbf{v}$

Choosing an alternative auxiliary function $Q(\Theta,\Theta^{(i)})$ such that

$$Q(\Theta^{(i+1)},\Theta^{(i)})\geq Q(\Theta,\Theta^{(i)}) \tag{25}$$

gives the generalized expectation maximization algorithm (GEM). By maximizing $Q(\Theta,\Theta^{(i)})$ at each iteration, we increase a lower bound on the original auxiliary function, and therefore step toward a maximum likelihood solution.

As with the EM algorithm, GEM converges (Bilmes, 1997) to a (possibly local) maximum. Using this approach, a modification of the auxiliary function makes it possible to find estimates of $\hat{A}$ and $\hat{\mathbf{b}}$ which can be calculated at a reduced computational cost.

We first describe the method by which we define a lower bound, before showing how we use this to modify the auxiliary equation in Equation 15.

2.2.1 *Setting a lower bound*

Let $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be i.i.d Gaussian random variables with sample mean and covariance $\bar{\mathbf{y}}$ and $S_{\mathbf{y}}$ respectively. Similarly, let $\mathcal{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$ be i.i.d Gaussian random variables with sample mean and covariance $\bar{\mathbf{z}}$ and $S_{\mathbf{z}}$. With $\Phi = \{\bar{\mathbf{y}}, \bar{\mathbf{z}}, S_{\mathbf{y}}, S_{\mathbf{z}}\}$, the joint log-likelihood of $\mathcal{Y}$ and $\mathcal{Z}$ is then computed as:

$$\begin{aligned} l(\Phi|\mathcal{Y}, \mathcal{Z}) \propto & -1/2 \sum_{i=1}^{n} \left\{ \log |S_{\mathbf{y}}| + (\mathbf{y}_i - \bar{\mathbf{y}})^T S_{\mathbf{y}}^{-1} (\mathbf{y}_i - \bar{\mathbf{y}}) \right\} \\ & -1/2 \sum_{j=1}^{m} \left\{ \log |S_{\mathbf{z}}| + (\mathbf{z}_j - \bar{\mathbf{z}})^T S_{\mathbf{z}}^{-1} (\mathbf{z}_j - \bar{\mathbf{z}}) \right\} \end{aligned} \tag{26}$$

We modify the models for $\mathcal{Y}$ and $\mathcal{Z}$ by replacing $S_{\mathbf{y}}$ and $S_{\mathbf{z}}$ with a common covariance Σ. Writing $\Phi' = \{\bar{\mathbf{y}}, \bar{\mathbf{z}}, \Sigma\}$, the likelihood under this new model is given by:

$$\begin{aligned} l(\Phi'|\mathcal{Y}, \mathcal{Z}) \propto & -1/2 \sum_{i=1}^{n} \left\{ \log |\Sigma| + (\mathbf{y}_i - \bar{\mathbf{y}})^T \Sigma^{-1} (\mathbf{y}_i - \bar{\mathbf{y}}) \right\} \\ & -1/2 \sum_{j=1}^{m} \left\{ \log |\Sigma| + (\mathbf{z}_j - \bar{\mathbf{z}})^T \Sigma^{-1} (\mathbf{z}_j - \bar{\mathbf{z}}) \right\} \end{aligned}$$

By definition, the first and second terms of Equation 26 are maximized by their sample covariances, $S_{\mathbf{y}}$ and $S_{\mathbf{z}}$ respectively. Consequently

$$l(\Phi|\mathcal{Y}, \mathcal{Z}) \geq l(\Phi'|\mathcal{Y}, \mathcal{Z}) \tag{27}$$

which gives a lower bound on the joint likelihood of $\mathcal{Y}$ and $\mathcal{Z}$.

2.2.2 *Bounding the LDM log-likelihood function*

The bound as defined above holds over the data which the sample covariances are estimated on. In applying this to the observation likelihood function of Equation 11 for use in estimating adaptation parameters, we assume minimal mismatch between the original training and adaptation data. Given that the adaptation data represents a subset of the type of data used in the original training, this assumption may be flawed in certain cases. However, we did not observe problems in practice, as GEM training converged to increased likelihoods for all speakers on which it was used.

As above, we choose C to be a pooled covariance shared by all models, giving a modified auxiliary function of:

$$Q(\Theta^{(i+1)}, \Theta^{(i)}) \propto E_{\Theta^{(i)}} \left[\sum_{r=1}^{R} \sum_{t=a_r}^{b_r} \left\{ \log|C| + \hat{\boldsymbol{\delta}}_{\mathbf{y}_t}^{m_r T} C^{-1} \hat{\boldsymbol{\delta}}_{\mathbf{y}_t}^{m_r} \right\} \right] \tag{28}$$

Taking the partial derivative with respect to A and setting to zero yields:

$$\sum_{r=1}^{R} \sum_{t=a_r}^{b_r} \left(\hat{A} H_{m_r} P_t + \hat{\mathbf{b}} \mathbf{x}_{t|N_r}^T \right) H_{m_r}^T = \sum_{r=1}^{R} \sum_{t=a_r}^{b_r} \boldsymbol{\gamma}_{\mathbf{y}_t}^{m_r} \mathbf{x}_{t|N_r}^T H_{m_r}^T \tag{29}$$

where $\boldsymbol{\gamma}_{\mathbf{y}_t}^{m_r} = \mathbf{y}_t - \mathbf{v}_{m_r}$ as above. Similarly for $\mathbf{b}$:

$$\sum_{r=1}^{R} \sum_{t=a_r}^{b_r} \left(\hat{A} H_{m_r} \mathbf{x}_{t|N_r} + \hat{\mathbf{b}} \right) = \sum_{r=1}^{R} \sum_{t=a_r}^{b_r} \boldsymbol{\gamma}_{\mathbf{y}_t}^{m_r} \tag{30}$$

The estimates for $\hat{A}$ and $\hat{\mathbf{b}}$ can be combined in closed form as:

$$\begin{aligned} \left[\hat{A} \ \hat{\mathbf{b}} \right] = & \left[\textstyle\sum_{r=1}^{R} \sum_{t=a_r}^{b_r} \boldsymbol{\gamma}_{\mathbf{y}_t}^{m_r} \mathbf{x}_{t|N_r}^T H_{m_r}^T \quad \sum_{r=1}^{R} \sum_{t=a_r}^{b_r} \boldsymbol{\gamma}_{\mathbf{y}_t}^{m_r} \right] \\ & \times \begin{bmatrix} \sum_{r=1}^{R} \sum_{t=a_r}^{b_r} H_{m_r} P_t H_{m_r}^T & \sum_{r=1}^{R} \sum_{t=a_r}^{b_r} H_{m_r} \mathbf{x}_{t|N_r} \\ \sum_{r=1}^{R} \sum_{t=a_r}^{b_r} \mathbf{x}_{t|N_r}^T H_{m_r}^T & \sum_{r=1}^{R} b_r - a_r + 1 \end{bmatrix}^{-1} \end{aligned} \tag{31}$$

These estimates are composed of a few easily computed sufficient statistics.

2.3 EM adaptation of noise covariance C

A standard EM update provides a simple and efficient way to update C. With $\hat{H}_{m_r}$ and $\hat{\mathbf{v}}_{m_r}$ already estimated, replacing C with its adapted version $\hat{C} = B_m^T G B_m$ in the likelihood function of Equation 11 gives an auxiliary function of

$$Q(\Theta^{(i+1)}, \Theta^{(i)}) \propto E_{\Theta^{(i)}} \left[\sum_{r=1}^{R} \sum_{t=a_r}^{b_r} \left\{ \log|G| + \hat{\boldsymbol{\delta}}_{\mathbf{y}_t}^{m_r T} V_{m_r} G^{-1} V_{m_r}^T \hat{\boldsymbol{\delta}}_{\mathbf{y}_t}^{m_r} \right\} \right] \tag{32}$$

where we use V_{m_r} to denote $B_{m_r}^{-1}$. We now let

$$\hat{\boldsymbol{\delta}}_{\mathbf{y}_t,\Theta^{(i)}}^{m_r} = E_{\Theta^{(i)}}\left[\hat{\boldsymbol{\delta}}_{\mathbf{y}_t}^{m_r}\right] = \mathbf{y}_t - \hat{H}_{m_r}\mathbf{x}_{t|N_r} - \hat{\mathbf{v}}_{m_r} \tag{33}$$

so that taking a partial derivative with respect to G^{-1} and equating to zero gives:

$$\hat{G} = \frac{1}{N}\sum_{r=1}^{R}\sum_{t=a_r}^{b_r} V_{m_r}^T\left[\hat{\boldsymbol{\delta}}_{\mathbf{y}_t,\Theta^{(i)}}^{m_r}\hat{\boldsymbol{\delta}}_{\mathbf{y}_t,\Theta^{(i)}}^{m_r T} + \hat{H}_{m_r}\Sigma_{t|N_r}\hat{H}_{m_r}^T\right]V_{m_r} \tag{34}$$

where $N = \sum_{r=1}^{R}\{b_r - a_r + 1\}$ and $\Sigma_{t|N_r}$ denotes the smoothed estimate of the state covariance such that $P_t = \Sigma_{t|N_r} + \mathbf{x}_{t|N_r}\mathbf{x}_{t|N_r}^T$ as given in Equation 13. This can be calculated as above or split into a number of sufficient statistics as in Gales and Woodland (1996) to allow location and covariance parameter updates to be made in single pass.

2.4 Computational requirements

Computing EM updates for the location parameters requires solving a set of simultaneous equations. A standard Gaussian elimination approach takes $O(n^3)$ operations, where n is the number of unknowns. Assuming we wish to estimate a fully specified transform of the location parameters, we have $n = p(p+1)$ and so the computation is order $O(p^6)$. With the GEM approach, we require a matrix inversion and a matrix multiplication, both of which are order $O(n^3)$. In this case, we have $n = p+1$ and so the updates require $O(p^3)$ operations. The updates for adaptation of the observation noise covariance require matrix multiplications and so are simply $O(p^3)$.

3 Experiments

Experimental work uses the TIMIT corpus (Lamel et al., 1986) following the standard train/test division. The base models for adaptation are as in Frankel and King (2007. In press) where an LDM is trained on the data corresponding to each of the 61 phone classes. A validation set comprising the utterances from 60 of the 462 training set speakers is set aside and used to determine the number of EM training iterations and language model scaling factor. Before final evaluation on the test set, new models are trained on the combined train and validation sets. The baselines in the experiments reported below use the set of LDMs prior to speaker adaptation.

The adaptation in these experiments is supervised, meaning that the TIMIT time-aligned phonetic labels are used when estimating relevant parameters. All 61 models are adapted according to transforms shared by all models, other than silence which is neither adapted nor used in estimating adaptation parameters.

The speech waveform is parameterized as 12 Mel-frequency cepstral coefficients (MFCC) and energy with 1^{st} and 2^{nd} derivatives appended. This gives a 39 dimensional feature vector, meaning that fully specified A, $\mathbf{b}$ and G require estimation of $39 \times 39 + 39 + 39 \times 40/2 = 2340$ parameters. Given the limited amount of data available for adaptation, it may be necessary to reduce the number of free parameters. The transforms A and G can be set to be diagonal or block-diagonal by enforcing zeros in the relevant portions of each matrix after the re-estimation step, and displacement by $\mathbf{b}$ need not be included in the adaptation scheme.

# adaptation	GEM			EM		
utterances	A	$\mathbf{b}$	G	A	$\mathbf{b}$	G
2	D	0	BD	D	0	D
4	D	0	BD	D	0	BD
6	$\surd$	$\surd$	I	$\surd$	$\surd$	I
8	$\surd$	$\surd$	I	$\surd$	$\surd$	I

Table 1
Results of choosing the form of adaptation parameters A, $\mathbf{b}$, G according to classification of held-out validation data, with transforms estimated on 2, 4, 6 or 8 utterances. Parameters are full ($\surd$), block-diagonal (BD), diagonal (D), identity (I) or zero (0).

3.1 Classification

There are 10 utterances from each of the 168 speakers in the TIMIT test set. Classification experiments were performed with adaptation transforms estimated on either 2, 4, 6 or 8 of these 10 utterances, and the remainder (8, 6, 4 or 2 utterances respectively) set aside for testing[4]. The `sa` sentences are the same for each speaker and are commonly omitted from TIMIT experiments to prevent bias in the distribution of segment types. However, we use them as the first two utterances when estimating the adaptation parameters. For each of the train set sizes, the form of the adaptation transforms and number of training iterations was determined using speakers from the validation set. Table 1 shows the forms which were chosen. Interestingly, for both GEM and EM adaptation, A is set to be diagonal, $\mathbf{b}$ excluded and G diagonal or block-diagonal where less adaption data is used. However, where more data is

[4] In all cases, baseline and adapted model results using matched data.

available the location parameters A and $\mathbf{b}$ are estimated fully and the covariance transform omitted.

# adaptation utterances	baseline accuracy	adapted accuracy		error reduction	
		GEM	EM	GEM	EM
2 (5.5s)	72.3%	72.6%	72.5%	1.1%	0.7%
4 (11.1s)	72.5%	73.1%	73.2%	2.2%	2.5%
6 (16.8s)	72.4%	73.4%	73.5%	3.6%	4.0%
8 (21.8s)	72.2%	73.4%	73.9%	4.3%	6.1%

Table 2
Test-set classification accuracies and relative error reductions for GEM and EM adaptation methods with transform parameters estimated on 2, 4, 6 or 8 utterances. The average per-speaker adaptation data duration in seconds is given in parentheses.

Test-set results are given in Table 2 and show that speaker adaptation yields increased classification accuracy. The performance improvement over the baseline increases as more utterances are used to estimate the adaptation transforms, with the highest being a 6.1% relative error reduction for EM adaptation based on 8 utterances. We find similar results for EM and GEM methods, with accuracies differing by 0.1% absolute for all but the 8-utterance adaptation where EM gives a 0.5% higher accuracy. Figure 1 shows the relative error reductions found with 2, 4, 6 or 8 adaptation utterances for both EM and GEM methods. The results do not appear to have reached a plateau, suggesting that were more data available for each speaker, adaptation might bring further accuracy increases.

3.2 Recognition

We also evaluate the speaker-adapted LDMs on the task of TIMIT phone recognition. Estimating speaker adaptation transforms on 8 utterances gives the highest classification validation accuracy for both EM and GEM methods, and these models are used for recognition. A language model scaling factor and phone insertion penalty are set using data from the validation speakers. The results of Table 3 show that speaker adaptation leads to relative accuracy increases of 4.5% and 4.6% using GEM and EM respectively.

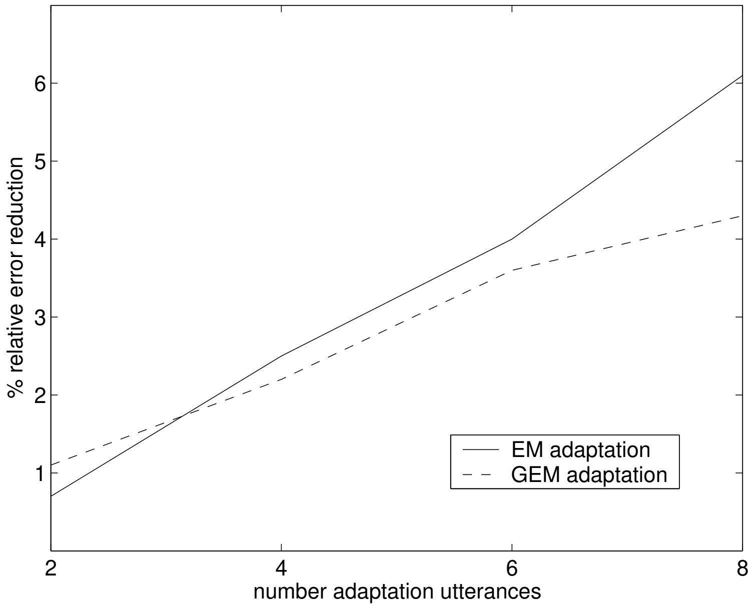

Fig. 1. Relative error reduction due to EM and GEM adaptation methods, with transforms estimated on 2, 4, 6 or 8 utterances.

# adaptation utterances	baseline accuracy	adapted accuracy		error reduction	
		GEM	EM	GEM	EM
8 (21.8s)	60.0%	61.8%	61.9%	4.5%	4.6%

Table 3
Test-set recognition accuracies and relative error reductions for GEM and EM adaptation methods with transform parameters estimated on 8 utterances. The average per-speaker adaptation data duration in seconds is given in parentheses.

4 Discussion

We have introduced two methods for adapting the observation process of linear dynamic models, which would be straightforward to apply to the state process of LDMs or to other linear Gaussian models. In section 4.2 we propose the use of the GEM method with HMM systems which use non-diagonal covariances, as a much lower complexity alternative to existing methods.

Error-rate reductions of around 5% relative were demonstrated under a variety of experimental conditions, even using a simple all-model speaker adaptation scheme. The EM approach gave the largest error reductions, but perhaps the most significant result is that the much lower complexity GEM method yielded almost the same improvements. Further benefits might be realized by incorporating the adaptation transforms into the model training phase.

4.1 Amount of adaptation data required

The experimental results show that the transforms can be estimated even when presented with very little adaptation data. Gunawardana and Byrne (2001) reported that $5s$ of data is insufficient for (unsupervised) estimation of a single global MLLR transform for an HMM-based switchboard system, and led to an increase in word error rate (WER). It was found that $10s$ of data yielded a slight improvement in WER, and that $30s$ was sufficient to give robust estimates of the MLLR transforms. In our experiments, we never find that error rates increase, even with only the smallest amounts of adaptation data.

Despite adaptation, the performance of LDMs for classification and recognition of speech remains lower than that found with hidden Markov model (HMM) systems. For example, Sun and Deng (2002) report an HMM-based TIMIT phone recognition accuracy of 72.95%. However, the techniques described in this paper are not restricted to LDMs, being generally applicable to models with non-diagonal Gaussian covariances.

4.2 Applicability to conventional HMM systems

Adaptation is an integral part of modern recognition systems, however the original MLLR (Gales and Woodland, 1996) was designed for models with diagonal covariances. A number of authors (Gales, 1999; Bilmes, 2000; Olsen and Gopinath, 2004) have investigated methods for extending covariances from diagonal to full in HMM systems. The approach is typically to employ covariance matrices with separate transform and magnitude components, thus allowing approximation of full covariances whilst introducing a minimum number of extra parameters. Two methods for adapting full covariance HMMs were described in Gales (1997), the first of which was full-covariance MLLR (the EM approach in this paper). Despite the proposal of an efficient means of computing the necessary statistics, full covariance MLLR remains computationally expensive since calculating the mean transform requires $O(p^6)$ operations. The second method, normalized-domain MLLR, was designed to alleviate this by rotating and scaling features in such a way that the Gaussian covariances were simply the identity matrix. Standard diagonal-covariance MLLR could then be applied. Evaluation using a semi-tied full covariance HMM system with Gaussian covariances trained from scratch (rather than being initialized with diagonal-covariance HMMs) found that full-covariance MLLR gave a 9% relative reduction in word error rate compared to a diagonal-covariance HMM system with standard MLLR. However, normalized-domain MLLR did not lead to error reduction over the diagonal-covariance system.

Given that we find similar performance for the EM and GEM methods presented in this paper, we propose that our GEM approach offers a significantly lower complexity method for adapting full-covariance HMMs. We present the necessary update equations for mean adaptation in Appendix A.

5 Acknowledgements

This work was supported by EPSRC grant GR/S21281/01.

A GEM mean adaptation for HMMs

The goal of a mean transform for HMMs is to estimate

$$\hat{\mu}_m = \hat{\mathbf{W}}_m \xi_m \tag{A.1}$$

where ξ_m represents an extended mean vector

$$\xi_m = \begin{bmatrix} 1 \ \mu_1 \ \ldots \ \mu_n \end{bmatrix} \tag{A.2}$$

We quote Gales and Woodland (1996) and define the following auxiliary function:

$$\begin{aligned}&\mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}) = \\ &K_1 - \frac{1}{2}\mathcal{L}(\mathbf{0}_T|\mathcal{M}) \sum_{m=1}^{M}\sum_{\tau=1}^{T} L_m(\tau)\left[K_m + \log(|\Sigma_m|) + (\mathbf{o}(\tau) - \hat{\mu}_m)^T \Sigma_m^{-1} (\mathbf{o}(\tau) - \hat{\mu}_m)\right]\end{aligned} \tag{A.3}$$

where $\mathbf{0}_T = \{\mathbf{o}(1), \ldots, \mathbf{o}(T)\}$ is the adaptation data, K_1 is a constant dependent only on the transition probabilities, K_m is the normalization constant associated with Gaussian m, and

$$L_m(\tau) = p(q_m \tau | \mathcal{M}, \mathbf{0}_T) \tag{A.4}$$

where $q_m(\tau)$ indicates Gaussian m at time τ.

Following the GEM method of Section 2.2, we construct a lower bound on A.3 by replacing model-specific Σ_m with a common Σ, giving

$$\begin{aligned}&\mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}) = \\ &K_1 - \frac{1}{2}\mathcal{L}(\mathbf{0}_T|\mathcal{M}) \sum_{m=1}^{M}\sum_{\tau=1}^{T} L_m(\tau)\left[K_m + \log(|\Sigma|) + (\mathbf{o}(\tau) - \hat{\mu}_m)^T \Sigma^{-1} (\mathbf{o}(\tau) - \hat{\mu}_m)\right]\end{aligned} \tag{A.5}$$

With $\mathbf{W}_m$ tied across a set of R Gaussians $\{m_1, \ldots, m_R\}$, Equation A.5 is maximized with respect to $\mathbf{W}_m$ by

$$\hat{\mathbf{W}}_m = \left(\sum_{r=1}^{R} \sum_{\tau=1}^{T} L_{m_r}(\tau)\mathbf{o}(\tau)\xi_{m_r}^T \right) \left(\sum_{r=1}^{R} \sum_{\tau=1}^{T} L_{m_r}\xi_{m_r}\xi_{m_r}^T \right)^{-1} \tag{A.6}$$

References

Bilmes, J., 1997. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Tech. Rep. ICSI-TR-97-021, University of Berkeley.

Bilmes, J., 2000. Factored sparse inverse covariance matrices. In: Proc. ICASSP.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B (39), 1–38.

Digalakis, V., 1992. Segment-based stochastic models of spectral dynamics for continuous speech recognition. Ph.D. thesis, Boston University Graduate School.

Frankel, J., 2003. Linear dynamic models for automatic speech recognition. Ph.D. thesis, The Centre for Speech Technology Research, Edinburgh University.

Frankel, J., King, S., January 2007. In press. Speech recognition using linear dynamic models. IEEE Transactions on Speech and Audio Processing.

Gales, M., 1997. Adapting semi-tied full-covariance matrix HMMs. Technical Report CUED/F-INFENG/TR298, Cambridge University Engineering Department.

Gales, M., May 1999. Semi-tied covariance matrices for hidden Markov models. IEEE Transactions on Speech and Audio Processing 7 (3), 272–281.

Gales, M., Woodland, P., 1996. Mean and variance adaptation within the MLLR framework. Computer, Speech and Language 10, 249–264.

Gunawardana, A., Byrne, W., 2001. Discounted likelihood linear regression for rapid speaker adaptation. Computer, Speech and Language. 15 (1), 15–38.

Lamel, L., Kassel, R., Seneff, S., February 1986. Speech database development: design and analysis of the acoustic-phonetic corpus. In: Proc. Speech Recognition Workshop. Palo Alto, CA., pp. 100–109.

Ma, J., Deng, L., 2004a. A mixed-level switching dynamic system for continuous speech recognition. Computer Speech and Language 18, 49–65.

Ma, J., Deng, L., 2004b. Target-directed mixture linear dynamic models for spontaneous speech recognition. IEEE Transactions on Speech and Audio Processing 12 (1), 47–58.

Olsen, P., Gopinath, R., January 2004. Modeling inverse covariance matrices by basis expansion. IEEE Transactions on Speech and Audio Processing 12 (1), 37–46.

Rauch, H. E., 1963. Solutions to the linear smoothing problem. IEEE Transactions on Automatic Control 8, 371–372.

Rosti, A., Gales, M., 2001. Generalised linear Gaussian models. Tech. Rep. CUED/F-INFENG/TR.420, Cambridge University Engineering.

Rosti, A.-V., 2004. Linear gaussian models for speech recognition. Ph.D. thesis, Machine Intelligence Laboratory, University of Cambridge.

Roweis, S., Ghahramani, Z., 1999. A unifying review of linear Gaussian models. Neural Computation 11 (2).

Sun, J., Deng, L., 2002. An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition. Journal of the Acoustical Society of America 111 (2), 1086–1101.