FACTOR ANALYSED HIDDEN MARKOV MODELS

A-V.I. Rosti and M.J.F. Gales

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, United Kingdom
e-mail: {avir2, mjfg}@eng.cam.ac.uk

ABSTRACT

This paper presents a general form of acoustic model for speech recognition. The model is based on an extension to factor analysis where the low dimensional subspace is modelled with a mixture of Gaussians hidden Markov model (HMM) and the observation noise by a Gaussian mixture model. Here the HMM output vectors are the latent variables of a general factor analyser. The model combines shared factor analysis with a dynamic version of independent factor analysis. This factor analysed HMM (FAHMM) provides an alternative, compact, model to handle intra-frame correlation. Furthermore, it allows variable dimension subspaces to be explored. A variety of model configurations and sharing schemes are examined, some of which correspond to standard systems. The training and recognition algorithms for FAHMMs are described and some initial result with Switchboard are presented.

1. INTRODUCTION

It is hard to find a single transform that decorrelates speech feature vectors for all states in an HMM system. One solution to this problem is to use full covariance matrices. However this dramatically increases the number of model parameters. Alternatively Gaussian mixture models may be used to model each state. This is the most common approach used in speech recognition. Recently other schemes have been proposed. One such scheme is semi-tied covariance matrices (STC) [1]. Systems employing STC generally yield better performance than standard diagonal covariance HMMs without dramatically increasing the number of model parameters. An alternative approach to improve intra-frame correlation modelling is to use schemes based on extensions to factor analysis [2, 3] or linear discriminant analysis [4]. The approach adopted in this paper is based on factor analysis. A separate factor analyser has been previously used for each of the component covariance matrices [2]. This gives a large number of model parameters due to the individual loading matrix attached to every component in the system. To reduce the number of model parameters, the loading matrix can be shared among several observation noise components as in shared factor analysis (SFA) [3]. However, SFA still assumes that the factors are distributed according to a standard normal distribution. A factor analysis model with a mixture of Gaussians, or more generally an HMM, generating the factors should provide a more flexible model.

A-V.I. Rosti is funded by an EPSRC studentship and Tampere Graduate School in Information Science and Engineering. He received additional support from the Nokia Foundation and the Finnish Cultural Foundation. This work made use of equipment kindly supplied by IBM under an SUR award.

Factor analysed HMMs (FAHMMs) use a mixture of Gaussians HMM as the state vector (an ordered set of factors) generating model and a shared factor analyser is used to generate the observations. Thus, an FAHMM assumes the state vectors are generated by a standard diagonal covariance mixture of Gaussians HMM, similar to a dynamic version of independent factor analysis (IFA) [5] without the independent factor (state-vector element) assumption, combined with SFA. A variety of model configurations will be examined. Some of these configurations correspond to standard systems. This paper presents the theory of FAHMMs. First, the generative model and the parameter estimation schemes for a FAHMM are described. Implementation and complexity issues in training and recognition are then considered. Finally, preliminary experiments on a large vocabulary continuous speech recognition task are described.

2. FACTOR ANALYSED HIDDEN MARKOV MODELS

A factor analysed hidden Markov model can be viewed as a state-space model with k dimensional state vectors, $\boldsymbol{x}_t$, and p dimensional observation vectors, $\boldsymbol{o}_t$. The state vectors are assumed to be generated by a standard mixture of Gaussians HMM with parameters $\mathcal{M}^{hmm} = \{a_{ij}, c_{jn}^{(x)}, \boldsymbol{\mu}_{jn}^{(x)}, \boldsymbol{\Sigma}_{jn}^{(x)}\}$ where a_{ij} denotes the probability of moving from state i to state j, $j \in (1, N_s)$ is a state indicator and $n \in (1, M^{(x)})$ is a state-space component indicator. The observations are generated by a general factor analysis model. The generative model of FAHMM can be represented as follows

$$\boldsymbol{x}_t \sim \mathcal{M}^{hmm} \tag{1}$$

$$\boldsymbol{o}_t = \boldsymbol{C}\boldsymbol{x}_t + \boldsymbol{v} \tag{2}$$

where the observation noise $\boldsymbol{v}$ can be distributed according to a mixture of Gaussians with parameters $c_{jm}^{(o)}$, $\boldsymbol{\mu}_{jm}^{(o)}$ and $\boldsymbol{\Sigma}_{jm}^{(o)}$ with $m \in (1, M^{(o)})$ as an observation noise component indicator and the p by k observation matrix, $\boldsymbol{C}$, (traditionally the loading matrix) can be arbitrarily shared in the state level or the model level. The observation noise covariance matrices are assumed to be diagonal. The state-space dimensionality can be chosen individually depending on the sharing of the observation parameters. By choosing $k = 0$ the model reduces to a standard HMM allowing arbitrary combinations of HMMs and FAHMMs to be employed.

Figure 1 shows the dynamic Bayesian network (DBN) that illustrates the independence assumptions in the model. In the figure, square nodes denote discrete and round nodes continuous variables. The nodes representing observable variables are shaded. The discrete HMM states are denoted by q_t and the standard HMM state conditional independence assumption applies for all other

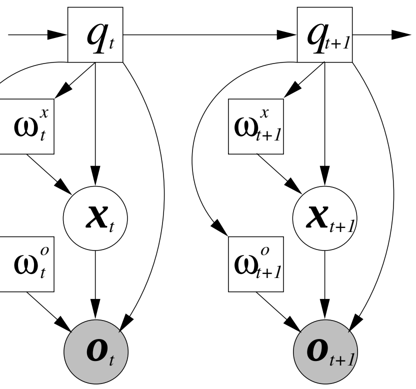

Fig. 1. DBN representing a factor analysed HMM.

variables. The mixture indicator variables ω_t^x and ω_t^o depend on the current state as does the state distribution and observation parameters. The state vectors, HMM states and both the mixture components are all hidden. Thus, expectation maximisation algorithm (EM) may be used in training the model parameters.

2.1. Training FAHMMs

From the DBN in Figure 1, the joint likelihood of an utterance $\boldsymbol{O} = \boldsymbol{o}_1, \ldots, \boldsymbol{o}_T$, state vector sequence $\boldsymbol{X} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ and HMM state sequence $Q = q_1, \ldots, q_T$ is given by

$$p(\boldsymbol{O}, \boldsymbol{X}, Q) = \prod_{t=2}^{T} P(q_t|q_{t-1}) \prod_{t=1}^{T} p(\boldsymbol{x}_t|q_t) p(\boldsymbol{o}_t|\boldsymbol{x}_t, q_t) \tag{3}$$

where $P(q_t|q_{t-1})$ is the normal HMM transition probability often denoted by a_{ij}, $p(\boldsymbol{x}_t|q_t)$ is the HMM state conditional output likelihood associated with state q_t and the observation likelihood can be obtained from the generative model as

$$p(\boldsymbol{o}_t|\boldsymbol{x}_t, j) = \sum_{m=1}^{M^{(o)}} c_{jm}^{(o)} \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{C}_j \boldsymbol{x}_t + \boldsymbol{\mu}_{jm}^{(o)}, \boldsymbol{\Sigma}_{jm}^{(o)}) \tag{4}$$

when $q_t = j$ and a separate observation matrix, $\boldsymbol{C}_j$, for every state is used. Using the above joint likelihood, an auxiliary function characteristic to standard EM learning schemes can be formed. Statistics for the HMM state posterior probabilities and state vector posterior likelihoods are estimated using the old set of model parameters, $\mathcal{M}$, in the E step. This is based on factoring the posterior likelihood of the state vector and HMM state sequence, $p(\boldsymbol{X}, Q|\boldsymbol{O}, \mathcal{M}) = p(\boldsymbol{X}|\boldsymbol{O}, Q, \mathcal{M}) P(Q|\boldsymbol{O}, \mathcal{M})$, which has previously been employed in learning the parameters of IFA [5]. The maximum likelihood estimates of the model parameters, $\hat{\mathcal{M}}$, are obtained in the M step using the sufficient statistics from the E step and applying standard optimisation techniques. The E and M steps are applied iteratively setting the new parameters obtained in the M step as the old parameters for the E step in the next iteration until the change in the log-likelihood for the training data becomes small. A detailed derivation of the EM algorithm for FAHMMs is presented in [6].

2.2. Sufficient Statistics

The posterior probability of being in state j at time t, $\gamma_j(t) = P(j|\boldsymbol{O}, \mathcal{M})$, being in state j at time t and in state i at time $t-1$, $\xi_{ij}(t) = P(i, j|\boldsymbol{O}, \mathcal{M})$, and being in state j and in state-space mixture n at time t, $\gamma_{jn}^{(x)}(t) = P(j, n|\boldsymbol{O}, \mathcal{M})$, can be obtained using the traditional forward backward algorithm for mixture of Gaussians HMM [7] with the following posterior likelihood of an observation given the state j and state-space component n

$$p(\boldsymbol{o}_t|j, n, \mathcal{M}) = \sum_{m=1}^{M^{(o)}} c_{jm}^{(o)} \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{C}_j \boldsymbol{\mu}_{jn}^{(x)} + \boldsymbol{\mu}_{jm}^{(o)}, \boldsymbol{C}_j \boldsymbol{\Sigma}_{jn}^{(x)} \boldsymbol{C}_j' + \boldsymbol{\Sigma}_{jm}^{(o)}) \tag{5}$$

where a prime, $(\cdot)'$, denotes transpose. The observation posterior likelihood for the state j can be obtained by marginalising the above likelihood over the state-space components. The same likelihood is also used in recognition algorithms which otherwise are exactly the same as for HMMs.

The joint posterior probability of being in state j and in observation noise component m, $\gamma_{jm}^{(o)}(t) = P(j, m|\boldsymbol{O}, \mathcal{M})$, is also required for the observation parameter estimation. This can be obtained as follows

$$\gamma_{jm}^{(o)}(t) = \sum_{n=1}^{M^{(x)}} P(m|j, n, \boldsymbol{o}_t, \mathcal{M}) \gamma_{jn}^{(x)}(t) \tag{6}$$

where the posterior probability of being in observation noise component m given the state j and state-space component n can be written as

$$P(m|j, n, \boldsymbol{o}_t, \mathcal{M}) = \frac{c_{jm}^{(o)} \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{C}_j \boldsymbol{\mu}_{jn}^{(x)} + \boldsymbol{\mu}_{jm}^{(o)}, \boldsymbol{C}_j \boldsymbol{\Sigma}_{jn}^{(x)} \boldsymbol{C}_j' + \boldsymbol{\Sigma}_{jm}^{(o)})}{\sum_{l=1}^{M^{(o)}} c_{jl}^{(o)} \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{C}_j \boldsymbol{\mu}_{jn}^{(x)} + \boldsymbol{\mu}_{jl}^{(o)}, \boldsymbol{C}_j \boldsymbol{\Sigma}_{jn}^{(x)} \boldsymbol{C}_j' + \boldsymbol{\Sigma}_{jl}^{(o)})} \tag{7}$$

The estimation of the HMM parameters requires different state vector posterior statistics than the estimation of the observation parameters. It can be shown that the state vector posterior distribution given the state and both the mixture components is a Gaussian. Thus, only first and second order statistics are needed and they can be written as

$$\hat{\boldsymbol{x}}_{jmn}(t) = \boldsymbol{\mu}_{jn}^{(x)} + \boldsymbol{K}_{jmn}(\boldsymbol{o}_t - \boldsymbol{C}_j - \boldsymbol{\mu}_{jm}^{(o)}) \tag{8}$$

$$\hat{\boldsymbol{R}}_{jmn}(t) = \boldsymbol{\Sigma}_{jn}^{(x)} - \boldsymbol{K}_{jmn} \boldsymbol{C}_j \boldsymbol{\Sigma}_{jn}^{(x)} + \hat{\boldsymbol{x}}_{jmn}(t) \hat{\boldsymbol{x}}_{jmn}'(t) \tag{9}$$

where

$$\boldsymbol{K}_{jmn} = \boldsymbol{\Sigma}_{jn}^{(x)} \boldsymbol{C}_j' (\boldsymbol{C}_j \boldsymbol{\Sigma}_{jn}^{(x)} \boldsymbol{C}_j' + \boldsymbol{\Sigma}_{jm}^{(o)})^{-1} \tag{10}$$

The sufficient statistics $\hat{\boldsymbol{x}}_{jm}^{(o)}(t)$, $\hat{\boldsymbol{R}}_{jm}^{(o)}(t)$, $\hat{\boldsymbol{x}}_{jn}^{(x)}(t)$ and $\hat{\boldsymbol{R}}_{jn}^{(x)}(t)$ can be obtained by marginalising the above statistics using the component prior probabilities $c_{jm}^{(o)}$ and $c_{jn}^{(x)}$.

2.3. Re-estimation Formulae

The re-estimation formulae for the HMM parameters are similar to the standard formulae described in [7]. However, the observations $\boldsymbol{o}_t$ are replaced by the posterior mean estimates $\hat{\boldsymbol{x}}_{jn}^{(x)}(t)$ and the second moments $\boldsymbol{o}_t\boldsymbol{o}_t'$ by $\hat{\boldsymbol{R}}_{jn}^{(x)}(t)$. Using the above notation, the formulae can be written as

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} \xi_{ij}(t)}{\sum_{t=2}^{T} \gamma_i(t-1)} \tag{11}$$

$$\hat{c}_{jn}^{(x)} = \frac{\sum_{t=1}^{T} \gamma_{jn}^{(x)}(t)}{\sum_{t=1}^{T} \gamma_j(t)} \tag{12}$$

$$\hat{\boldsymbol{\mu}}_{jn}^{(x)} = \frac{\sum_{t=1}^{T} \gamma_{jn}^{(x)}(t)\hat{\boldsymbol{x}}_{jn}^{(x)}(t)}{\sum_{t=1}^{T} \gamma_{jn}^{(x)}(t)} \tag{13}$$

$$\hat{\boldsymbol{\Sigma}}_{jn}^{(x)} = \operatorname{diag}\Big(\frac{\sum_{t=1}^{T} \gamma_{jn}^{(x)}(t)\hat{\boldsymbol{R}}_{jn}^{(x)}(t)}{\sum_{t=1}^{T} \gamma_{jn}^{(x)}(t)} - \hat{\boldsymbol{\mu}}_{jn}^{(x)}\hat{\boldsymbol{\mu}}_{jn}^{(x)\prime}\Big) \tag{14}$$

The new observation matrix has to be estimated row by row as in SFA [3]. The scheme adopted in this paper follows closely the maximum likelihood linear regression transform matrix optimisation [8]. The lth row vector $\hat{\boldsymbol{c}}_l$ of the new observation matrix, $\hat{\boldsymbol{C}}_j$, can be written as

$$\hat{\boldsymbol{c}}_l = \boldsymbol{k}_l'\boldsymbol{G}_l^{-1} \tag{15}$$

where the k by k matrices $\boldsymbol{G}_l$ and the k dimensional column vectors $\boldsymbol{k}_l$ are defined as follows

$$\boldsymbol{G}_l = \sum_{m=1}^{M^{(o)}} \frac{1}{\sigma_{jml}^{(o)2}} \sum_{t=1}^{T} \gamma_{jm}^{(o)}(t)\hat{\boldsymbol{R}}_{jm}^{(o)}(t) \tag{16}$$

$$\boldsymbol{k}_l = \sum_{m=1}^{M^{(o)}} \frac{1}{\sigma_{jml}^{(o)2}} \sum_{t=1}^{T} \gamma_{jm}^{(o)}(t)(o_{tl} - \mu_{jml}^{(o)})\hat{\boldsymbol{x}}_{jm}^{(o)}(t) \tag{17}$$

where $\sigma_{jml}^{(o)2}$ is the lth diagonal element of the observation covariance matrix $\boldsymbol{\Sigma}_{jm}^{(o)}$, o_{tl} and $\mu_{jml}^{(o)}$ are the lth elements of the current observation and the observation noise mean vectors, respectively.

The observation noise parameters are updated using the following formulae

$$\hat{c}_{jm}^{(o)} = \frac{\sum_{t=1}^{T} \gamma_{jm}^{(o)}(t)}{\sum_{t} \gamma_j(t)} \tag{18}$$

$$\hat{\boldsymbol{\mu}}_{jm}^{(o)} = \frac{\sum_{t=1}^{T} \gamma_{jm}^{(o)}(t)(\boldsymbol{o}_t - \hat{\boldsymbol{C}}_j\hat{\boldsymbol{x}}_{jm}^{(o)}(t))}{\sum_{t=1}^{T} \gamma_{jm}^{(o)}(t)} \tag{19}$$

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{jm}^{(o)} = \frac{1}{\sum_{t=1}^{T} \gamma_{jm}^{(o)}(t)} \sum_{t=1}^{T} \gamma_{jm}^{(o)}(t)\operatorname{diag}\Big(&\boldsymbol{o}_t\boldsymbol{o}_t' \\ &- \Big[\, \hat{\boldsymbol{C}}_j \;\; \hat{\boldsymbol{\mu}}_{jm}^{(o)} \,\Big]\Big[\, \boldsymbol{o}_t\hat{\boldsymbol{x}}_{jm}^{(o)\prime}(t) \;\; \boldsymbol{o}_t \,\Big]' \\ &- \Big[\, \boldsymbol{o}_t\hat{\boldsymbol{x}}_{jm}^{(o)\prime}(t) \;\; \boldsymbol{o}_t \,\Big]\Big[\, \hat{\boldsymbol{C}}_j \;\; \hat{\boldsymbol{\mu}}_{jm}^{(o)} \,\Big]' \\ &+ \Big[\, \hat{\boldsymbol{C}}_j \;\; \hat{\boldsymbol{\mu}}_{jm}^{(o)} \,\Big]\begin{bmatrix} \hat{\boldsymbol{R}}_{jm}^{(o)}(t) & \hat{\boldsymbol{x}}_{jm}^{(o)}(t) \\ \hat{\boldsymbol{x}}_{jm}^{(o)\prime}(t) & 1 \end{bmatrix}\Big[\, \hat{\boldsymbol{C}}_j \;\; \hat{\boldsymbol{\mu}}_{jm}^{(o)} \,\Big]'\Big)\end{aligned} \tag{20}$$

2.4. Implementation Issues

The estimation of the sufficient statistics in the EM algorithm for FAHMMs requires inverting $M^{(o)}M^{(x)}$ full p by p covariance matrices of the form $\boldsymbol{C}_j\boldsymbol{\Sigma}_{jn}^{(x)}\boldsymbol{C}_j' + \boldsymbol{\Sigma}_{jm}^{(o)}$. The inverses are also needed in the recognition. If the amount of memory is not an issue, the inverses and the corresponding determinants can be computed prior to starting off with the training and recognition. A more memory efficient implementation requires the computation of the inverses and determinants on the fly. It should be noted that these can be efficiently calculated using the following equality for matrix inverses

$$\begin{aligned}&(\boldsymbol{C}_j\boldsymbol{\Sigma}_{jn}^{(x)}\boldsymbol{C}_j' + \boldsymbol{\Sigma}_{jm}^{(o)})^{-1} = \\ &\quad \boldsymbol{\Sigma}_{jm}^{(o)-1} - \boldsymbol{\Sigma}_{jm}^{(o)-1}\boldsymbol{C}_j(\boldsymbol{C}_j'\boldsymbol{\Sigma}_{jm}^{(o)-1}\boldsymbol{C}_j + \boldsymbol{\Sigma}_{jn}^{(x)-1})^{-1}\boldsymbol{C}_j'\boldsymbol{\Sigma}_{jm}^{(o)-1}\end{aligned} \tag{21}$$

where the inverses of the diagonal covariance matrices $\boldsymbol{\Sigma}_{jm}^{(o)}$ and $\boldsymbol{\Sigma}_{jn}^{(x)}$ are trivial to compute and the full matrix $\boldsymbol{C}_j'\boldsymbol{\Sigma}_{jm}^{(o)-1}\boldsymbol{C}_j + \boldsymbol{\Sigma}_{jn}^{(x)-1}$ to be inverted is only a k by k matrix. It is dramatically faster than inverting a full p by p matrix if $k \ll p$. The determinants needed in the likelihood calculations can be obtained using the following equality

$$\begin{aligned}&|\boldsymbol{C}_j\boldsymbol{\Sigma}_{jn}^{(x)}\boldsymbol{C}_j' + \boldsymbol{\Sigma}_{jm}^{(o)}| = \\ &\quad |\boldsymbol{\Sigma}_{jm}^{(o)}||\boldsymbol{\Sigma}_{jn}^{(x)}||\boldsymbol{C}_j'\boldsymbol{\Sigma}_{jm}^{(o)-1}\boldsymbol{C}_j + \boldsymbol{\Sigma}_{jn}^{(x)-1}|\end{aligned} \tag{22}$$

where again the determinants of the diagonal covariance matrices are trivial to compute and provided Cholesky decomposition was used to invert the full k by k matrix, its determinant is obtained as a by-product.

The number of Gaussian components in a large vocabulary speech recogniser is often huge. The new estimate of a particular component may not be reliable if the number of observation vectors assigned to the component is small. For this reason some of the above full covariance matrices may become singular. Since the first term, $\boldsymbol{C}_j\boldsymbol{\Sigma}_{jn}^{(x)}\boldsymbol{C}_j'$, is generally singular a single zero observation variance element can make the full matrix non-invertible. Thus, flooring of the observation variance elements is adopted [6]. The flooring can be performed by using a fraction of the global variance computed from all the observations as a minimum value for the corresponding element. This is the way flooring is often done in HMM based systems [7].

Table 1. Order of number of free parameters using $M^{(x)}$ state-space components, $M^{(o)}$ observation noise components and no sharing of individual FAHMM parameters.

System	Free Parameters
HMM ($M^{(x)} = 0$)	$2M^{(o)}p$
FAHMM ($M^{(x)} > 0$)	$2(M^{(x)} - 1)k + pk + 2M^{(o)}p$

Table 1 describes the order of number of free parameters in a standard diagonal covariance HMM and an FAHMM with a separate factor analyser per state. HMMs can be viewed as a special case of FAHMMs when the state-space distributions are discarded, $M^{(x)} = 0$. Alternatively, by discarding the observation noise and using same state and observation space dimensionalities, $k = p$, the model reduces to an STC [1] for which a different training scheme has to be applied. An FAHMM with only one state-space component can be viewed as SFA [3] whereas a single observation noise component FAHMM is a dynamic version of IFA [5].

3. RESULTS

The baseline used for the experiments was a gender independent decision tree clustered tied state cross-word triphone mixture of

Gaussians HMM system. The setup was the same as the Minitrain 1998 Hub5 HTK system [9]. The number of distinct states was 3091 including three silence states. The 18 hour Minitrain set containing 398 conversation sides of Switchboard-1 corpus defined by BBN was used as the acoustic training data. The test set used was the subset of the 1997 Hub5 evaluation set used in [9]. It contains 10 conversation sides of Switchboard-2 data and 10 of Call Home English. The state output distributions of the baseline models were mixed up to 12 components running 4 re-estimation iterations prior to every mixture splitting as described in [7]. The test set word error rates (WER) for different mixture configurations of the baseline system are shown in Table 2 with the number of free parameters. The performance began degrading after more than 12 components were used giving WER of 47.8% for a 14 component system.

Table 2. Number of free parameters (nofp) and word error rates (wer%) for the baseline HMM systems with $M^{(o)}$ components.

$M^{(o)}$	1	2	3	5	7	10	12
nofp	78	156	234	390	546	780	936
wer%	56.0	53.4	51.6	49.4	48.4	47.6	47.1

The single component states in the initial baseline model set were modified to factor analysis models with a state-space dimensionality of 13. The separate observation matrices for every state were initialised to be 39 by 13 identity matrices, the first 13 elements of the baseline distributions were set as the HMM state distribution parameters and the resulting 26 elements were set as the observation noise parameters leaving their first 13 mean elements as zeroes and the first 13 variance elements as ones. First, the state distributions of FAHMMs were mixed up to 7 components leaving the number of observation components as one. The test set performance of the single observation component system began degrading after more than 7 state-space components were used giving WER of 48.2% for a 10 state-space component system. The observation noise distributions of the final models following state-space mixture splitting and 4 re-estimation iterations were mixed up until no further performance gain was achieved. Using more than 3 observation noise components degraded performance.

Table 3. Number of free parameters (nofp) and word error rates (wer%) for different state-space $M^{(x)}$ and observation noise $M^{(o)}$ component configurations of FAHMMs.

$M^{(o)}$	$M^{(x)}$	1	2	3	5	7
1	nofp	585	611	637	689	741
	wer%	48.0	47.8	48.0	47.8	47.6
2	nofp	663	689	715	767	
	wer%	47.5	47.4	47.4	47.6	
3	nofp	741	767	793		
	wer%	46.9	46.3	47.3		

The test set error rates for all the configurations described above are depicted in Table 3. The columns of the table correspond to the number of state-space components and the rows correspond to the number of observation noise components. The configurations with degrading performance are omitted leaving the corresponding table cells shaded. As discussed earlier, SFA can be viewed as an FAHMM with only one state-space component. Thus, the third column of the table refers to the SFA performance. The performance of an FAHMM with 3 observation noise and 2 state-space components exceeds the best performance of the baseline system with considerably fewer parameters. The results of these preliminary experiments in a rather small task suggest that there may be advantages in using subspace models. For these experiments no optimisation of the subspace dimensionality was performed.

4. CONCLUSIONS

A general form of acoustic model, the factor analysed HMM, is introduced in this paper. It combines the standard mixture of Gaussians continuous HMM with a shared and independent factor analysis models. The model should provide better intra-frame correlation modelling than the standard diagonal covariance matrix HMM which is a special case of the FAHMM. In addition, it allows a variety of linear subspaces to be investigated. HMM training and recognition algorithms are extended to apply for FAHMMs. Some preliminary experiments have been conducted and the results indicate that FAHMMs could prove to be useful. Future work will examine different structures including both the sharing and mixture configurations. Also, the optimal way to decide the state-space dimensionality and more sophisticated initialisation schemes will be examined.

5. REFERENCES

[1] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[2] L. Saul and M. Rahim, "Maximum likelihood and minimum classification error factor analysis for automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 115–125, 1999.

[3] R.A. Gopinath, B. Ramabhadran, and S. Dharanipragada, "Factor analysis invariant to linear transformations of data," in *Proc. ICSLP'98*, 1998, pp. 397–400.

[4] M.J.F. Gales, "Maximum likelihood multiple projection schemes for hidden Markov models," Tech. Rep. CUED/F-INFENG/TR.365, Cambridge University Engineering Department, 1999, Available via anonymous ftp from `ftp://svr-ftp.eng.cam.ac.uk/pub/reports/`.

[5] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.

[6] A-V.I. Rosti and M.J.F. Gales, "Generalised linear Gaussian models," Tech. Rep. CUED/F-INFENG/TR.420, Cambridge University Engineering Department, 2001, Available via anonymous ftp from `ftp://svr-ftp.eng.cam.ac.uk/pub/reports/`.

[7] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.0)*, Cambridge University, 2000.

[8] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[9] T. Hain, P.C. Woodland, T.R. Niesler, and E.W.D. Whittaker, "The 1998 HTK system for transcription of conversational telephone speech," in *Proc. ICASSP'99*, 1999, pp. 57–60.