DISCRIMINANT INITIALIZATION FOR FACTOR ANALYZED HMM TRAINING

Fabrice Lefevre and Jean-Luc Gauvain*

Spoken Language Processing Group
LIMSI-CNRS, France
{lefevre,gauvain}@limsi.fr

ABSTRACT

Factor analysis has been recently used to model the covariance of the feature vector in speech recognition systems. Maximum likelihood estimation of the parameters of factor analyzed HMMs (FAHMMs) is usually done via the EM algorithm, meaning that initial estimates of the model parameters is a key issue. In this paper we report on experiments showing some evidence that the use of a discriminative criterion to initialize the FAHMM maximum likelihood parameter estimation can be effective. The proposed approach relies on the estimation of a discriminant linear transformation to provide initial values for the factor loading matrices, as well as appropriate initializations for the other model parameters. Speech recognition experiments were carried out on the *Wall Street Journal* LVCSR task with a 65k vocabulary. Contrastive results are reported with various model sizes using discriminant and non discriminant initialization.

1. INTRODUCTION

Over the last few years there has been renewed interest in improving covariance modeling in HMM-based automatic speech recognition (ASR) systems [1, 3, 10, 8]. Although desirable, the use of full covariance Gaussians increases dramatically the number of parameters and complicates parameter estimation. Hence, Gaussians with diagonal covariances matrices are commonly used in HMMs. Factor analysis provides an intermediate modeling strategy which allows full covariances with fewer parameters to be derived. A generative model of speech based on a statistical signal filtering scheme is used. In this model, the assumption is made that the observations result from a linear transformation of a lower dimension hidden random vector. Factor analyzed models have been recently generalized in the context of HMM states [8]. The Factor Analyzed HMM (FAHMM) is a linear Gaussian model based on a piecewise constant state evolution process. The state vectors are generated by a standard diagonal covariance Gaussian mixture HMMs. As shown in [8], FAHMM provides a general framework encompassing many standard covariance modeling schemes such as shared factor analysis (SFA) [10], independent factor analysis (IFA) [1] or semi-tied covariance (STC) models [3]. With FAHMMs various levels of tying can be applied resulting in various degrees of complexity for the statistical components.

An optimal use of FAHMMs relies on many parameter settings. As for conventional HMMs, the size of the observation vectors and the number of Gaussians per state need to be chosen. In addition, for FAHMMs the size of the state space and the number of Gaussians associated to it also have to be decided. The model parameters are usually obtained using an EM procedure. Good initial values of these parameters are crucial to ensure a proper convergence. The approach evaluated in this work addresses this point through the introduction of a discriminant criterion in the selection of the state space dimensions. We propose to derive the state space dimensions from an Heteroscedastic Linear Discriminant (HLDA) transformation [5].

Speech recognition experiments were carried out on the *Wall Street Journal* large vocabulary continuous speech recognition task. A comparison is made between diagonal covariance HMMs, full covariance HMMs, and FAHMMs with non discriminant and discriminant initialization.

2. FACTOR ANALYZED HMMS

FAHMM is a dynamic state space generalization of a multiple component factor analysis system. The generative model used in FAHMM for each time step t and a given state j is described by the following equations (using the same notations as [8])

$$o_t = C_j x_t + v_t \tag{1}$$

$$x_t \sim \sum_k c_{jk}^{(x)} \mathcal{N}(x_t; \mu_{jk}^{(x)}, \Sigma_{jk}^{(x)})) \tag{2}$$

$$v_t \sim \sum_l c_{jl}^{(o)} \mathcal{N}(v_t; \mu_{jl}^{(o)}, \Sigma_{jl}^{(o)})) \tag{3}$$

where is o_t an n-dimensional observation vector, x_t is a p-dimensional state vector and v_t is an n-dimensional observation noise vector. All the covariance matrices being diagonal, the covariance structure is captured by the matrix C_j known as the factor loading matrix. The distribution of an

*F. Lefevre is also with the Human-Machine Dialog Team at LIA-University of Avignon

observation vector o_t for a given state j, and for a given state space component k and an observation noise component l, is obtained by integrating over the state vector x_t. The resulting distribution is Gaussian with the following mean vector and covariance matrix

$$\mu_{jkl} = C_j \mu_{jk}^{(x)} + \mu_{jl}^{(o)} \quad (4)$$

$$\Sigma_{jkl} = C_j \Sigma_{jk}^{(x)} C_j' + \Sigma_{jk}^{(o)}. \quad (5)$$

The conditional observation density of an FAHMM state can be viewed as a $M^{(o)}M^{(x)}$ component full covariance matrix GMM with mean vectors and covariance matrices given by equation (4) and (5), and the marginal likelihood of an observation given only the state j is obtained by summing over all the two sets of Gaussians. This calculation requires inverting $M^{(o)}M^{(x)}$ full $n \times n$ covariance matrices.

A detailed presentation of the EM reestimation formulae and general training setup for the FAHMM models can be found in [8]. In our system, these reestimation formulae are used with the exception that during the EM training a constant segmentation is used (*Viterbi training*). To further decrease the computational demand during model training, the two level algorithm has been adopted (a fast inner loop speeds up convergence).

Initialization of the model parameters is an important issue when using the EM algorithm as it can improve the possibility of reaching a good solution. For FAHMMs, a sensible starting point is to convert a standard HMM (with single Gaussian components) to an equivalent FAHMM by using the static cepstrum features as the state space dimensions (see [8]). This initialization assumes that the state space vector is highly correlated with the static cepstrum which may not be correct. For this reason, we propose to use an HLDA projection for model initialization.

3. DISCRIMINANT INITIALIZATION

Recent work in acoustic modeling [5, 3, 9] has led to the widespread adoption of HLDA techniques in state-of-the-art ASR systems. The objective of discriminant feature transformation is to find a projected feature space of low dimensionality while keeping most of the discriminant information. HLDA is an ML method for estimating a linear projection of n-dimensional feature vectors onto a p-dimensional sub-space. As for LDA, the resulting sub-space is supposed to show increased separation between the considered classes, with each class modeled by a Gaussian distribution. HLDA generalizes LDA by removing the restriction of a common within class covariance matrix, resulting in better feature projections when the classes are heteroscedastic.

Practically, the linear transform A is partitioned into two matrices: A_p transforming the original feature space to the reduced projected subspace of dimension p and A_{n-p} to the rejected subspace. The assumption is that the means and variances of the rejected n-p dimensions are represented by the corresponding dimensions of the transformed global mean and variances in the transformed feature space. Optimization of this HLDA objective function can be done by means of numerical methods (such as conjugate gradient) or using an ML optimization procedure performed row by row [3]. In the second approach (used in our experiments), each row of the transform matrix is updated sequentially, using the cofactor vector of the row and the current projected model parameters. When HLDA is applied to HMMs for speech recognition, good results are generally observed by using the HMM tied states as the HLDA classes, each represented by a full covariance Gaussian [9]. Several initializations are possible for the projection matrix, if an identity matrix is the simplest, we observed slightly better results using the Fisher ratio or an LDA solution (the latter is used in the experiments reported in this work).

Common use of the HLDA technique consists in reducing optimally a large feature space to a more discriminant one and then to build the speech models in the new space. In this work, the HLDA transformation is used to define a discriminant state space for the FAHMM. To do so, the state space is defined by the useful dimensions of the HLDA projection.

By combining the HLDA sub-space definition $x_t = A_p o_t$ with the FAHMM model given in equation (1), we see that a pseudo-inverse of A_p (a rectangular $p \times n$ matrix) can be good solution to initialize the factor loading matrices. For this work we used the Moore-Penrose matrix inverse [2], $C_j = A_p^+$.

Once the factor loading matrix is defined, the state space and observation noise Gaussian mixtures have to be initialized. They are obtained directly from the HLDA transformed space. This can be done either by training the models on the HLDA transformed feature vectors or by transforming the observation space parameters.

With the state vector distribution defined as in Equation (2), the state distribution parameters in the first method *(training)* is obtained by EM training in the HLDA-projected observation sub-space which gives for every state j

$$\mu_{jk}^{(x)} = A_p \mu_{jk} \quad (6)$$

$$\Sigma_{jk}^{(x)} = \mathrm{diag}(A_p W_{jk} A_p^T) \quad (7)$$

where μ_{jk} and W_{jk} are the mean vector and grand covariance matrix of the original data associated to state i of the k'th Gaussian. In the second method *(projection)*, the mean vectors still follow Equation (6) but the covariances become

$$\Sigma_{jk}^{(x)} = \mathrm{diag}(A_p \Sigma_{jk} A_p^T) \quad (8)$$

where μ_{jk} and Σ_{jk} are obtained by EM training on the observation space.

Given the statistics on the original observation space, the initial values for the observation noise parameters are obtained by subtracting the "loaded" state space parameters

(i.e. projected back to the observation space) from the observation prior parameters. A diagonal covariance prior is sufficient as no error compensation is made for the off-diagonal covariance coefficients. The observation and noise mixtures must have the same number of Gaussians ($M^{(o)}$) and the state space distribution must be single Gaussian ($M^{(x)} = 1$). Otherwise, the association between the observation and state space Gaussians would be undefined and complex to establish.

With the observation noise vector distribution defined as in Equation (3), the *training* initialization method leads to

$$\mu_{jl}^{(o)} = \mu_{jl} - \hat{C}\mu_{j1}^{(x)} \quad (9)$$

$$\Sigma_{jl}^{(o)} = \Sigma_{jl} - \mathrm{diag}(\hat{C}\Sigma_{j1}^{(x)}\hat{C}^T) \quad (10)$$

and the projected initialization modify the covariances such as

$$\Sigma_{jl}^{(o)} = \Sigma_{jl} - \mathrm{diag}(\hat{C}\mathrm{diag}(A_p\Sigma_{j1}A_p^T)\hat{C}^T) \quad (11)$$

In this context, special care must be taken to correctly floor the observation noise variances.

4. CORPUS AND SYSTEM DESCRIPTIONS

Experiments were carried out on a large vocabulary continuous speech recognition task using the *Wall Street Journal* (WSJ) corpus [7] and following the ARPA 1995 test conditions. The acoustic training data consist of about 100 hours of studio quality, read speech from 355 speakers (WSJ0 and WSJ1 corpora). The acoustic analysis derives cepstral parameters from a Mel frequency spectrum estimated on the 0-8kHz band every 10ms. Cepstral mean removal is applied to the cepstral coefficients. The 39-component acoustic feature vector consists of 12 cepstrum coefficients and the log energy, along with their first and second derivatives.

The speech recognizer uses continuous density HMMs with Gaussian mixture for acoustic modeling and *n*-gram statistics estimated on large text corpora for language modeling. Each context-dependent phone model is a tied-state left-to-right CD-HMM with Gaussian mixture observation densities where the tied states are obtained by means of a decision tree. The EM training is performed with a fixed segmentation (*Viterbi training*). Gender-dependent acoustic models are estimated using MAP adaptation of speaker-independent seed models [4]. The system has 4k context, position and gender-dependent phone models with 9k independent HMM states. The recognition vocabulary contains 65k words with 77k phone transcriptions. The 3-gram and 4-gram back-off language models result from the interpolation of models trained on different data sets (acoustic data transcriptions and newspapers texts). A pronunciation graph is associated with each word so as to allow for alternate pronunciations.

	$M^{(o)}$	*1*	*8*	*16*	*32*
Diagonal	η	78	624	1248	2496
	WER	13.8	8.8	8.5	8.0
	$M^{(o)}$	*1*	*2*	*4*	*8*
Full	η	819	1638	3276	6552
	WER	10.9	9.5	9.2	10.3

Table 1: Word error rates (%) and number of parameters (η) per state for the diagonal and full covariance HMM systems. $M^{(o)}$ is the number of Gaussians per state.

Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation. The initial hypotheses are used for acoustic model adaptation using the MLLR technique [6] on both model means and variances prior to word graph generation (MLLR is only applied to the model means when full matrix covariances are used). A 3-gram language model is used in the first two decoding steps. The final hypotheses are generated with a 4-gram language model and acoustic models adapted with the hypotheses of step 2.

5. EXPERIMENTAL RESULTS

To set a baseline, an HMM configuration with diagonal covariance matrices was tested. All reported results are with gender dependent models and unsupervised speaker adaptation. Table 1 gives the word error rates and number of free parameters per state (η) for 4 values of $M^{(o)}$ (the number Gaussian components per state). The WER decreases monotonically with the number of parameters, going down to 8% with 32 Gaussians per state.

A constrastive experiment was carried out with full covariance HMMs. The results obtained with 1 to 8 Gaussian components per state are given in the second part of the Table 1. The lowest WER is 9.2% with 4 Gaussians per state, i.e. about 1% above the best result with diagonal covariance matrices. This result is consistent with other reported results for conversational speech [11], and it justifies the search for an intermediate modeling scheme.

Table 2 contains the word error rates for various FAHMM setups along with the number of parameters per state. The dimensionality of the state space is set to 13 with a 39-dimensional observation space. For the regular initialization case, a set of HMMs with diagonal covariance matrix Gaussian mixtures is used to initialize the FAHMMs as proposed in [8]. One factor loading matrix is used per state and is thus shared by all its Gaussians. The two first rows correspond to the regular initialization with 6 state space components and from 1 to 6 observation space components. The lowest WER of 8.8% is obtained for $M^{(o)} = 6$. Although the number of free parameters per state stays low (1105) compared to the diagonal covariance models, the decoding time is significantly higher. This is why we cannot easily go beyond the 6×6 configuration.

$M^{(x)}$	$M^{(o)}$	*1*	*2*	*3*	*4*	*5*	*6*
		Regular intialization					
6	η	715	793	871	949	1027	1105
	WER	9.2	8.9	8.8	8.9	9.1	8.8
		HLDA init. (projection)					
	WER	9.0	8.5	8.5	8.4	8.6	8.3
		HLDA init. (training)					
	WER	9.2	8.9	8.6	8.8	8.5	8.3
3	η	637	715	793	871	949	1027
	WER	9.5	9.4	9.0	8.7	8.7	8.3
1	η	585	663	741	819	897	975
	WER	11.0	10.0	9.9	9.2	9.4	9.2

Table 2: Word error rates (%) and number of parameters per state (η) for three FAHMM configurations (all with $n = 39$ and $p = 13$: standard, HLDA projected, HLDA trained initializations.$M^{(x)}$ and $M^{(o)}$ are respectively the numbers of state space and observation space components.

The results for the discriminant initialization of FAHMMs are given in the next two entries, for both the *projection* and *training* schemes. When the number of noise components is increased, the performance improves faster with the *projection* method although both schemes lead to comparable performance for $M_{(o)} = 6$ (8.3%). With the discriminant initializations, the WER is reduced by 0.5% for the best configurations ($M^{(x)}$=6, $M^{(o)}$=6) compared to the regular initialization. With about 1k free parameters the WER is also lower than the WER obtained with the standard diagonal covariance models.

The lower part of the table gives additional results for the *training* initialization method using fewer state space components (3 and 1). With $M^{(x)} = 3$ the results are better than with $M^{(x)} = 6$ when compared at the corresponding number of parameters. However, the performance decreases with $M^{(x)} = 1$. These results tend to show that the balance between $M^{(x)}$ and $M^{(o)}$ is a sensitive key to reach good performance with FAHMMs.

6. CONCLUSIONS

Factor analyzed HMMs have been applied to an LVCSR task using about 100 hour training data from the LDC WSJ corpus. This system use a 65k 4-gram language model and unsupervised acoustic model adaptation. An method has been proposed to improve the state space initialization in FAHMM training by using a discriminant criterion (HLDA). In a set of experiments, we observed that FAHMMs with the proposed training method give slightly lower results than standard diagonal covariance HMMs (8.3% vs 8.0% WER) but improved results compared to full covariance HMMs (8.3% vs 9.2%) and regular FAHMMs (8.3% vs 8.8%). If the number of free parameters per state is fixed to a around 1k, FAHMMs with the proposed training also performed better than diagonal covariance models.

REFERENCES

[1] H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999.

[2] S. Campbell and C. Meyer. *Generalized Inverses of Linear Transformations*. Dover Publications, New-York, 1991.

[3] M. Gales. Semi-tied covariance matrices for hidden markov models. *IEEE Transactions on Speech and Audio Processing*, 7(3):272–281, 1999.

[4] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994.

[5] N Kumar and A Andreou. Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. *Speech Communication*, 26(4):283–297, 1998.

[6] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9:171–185, 1995.

[7] D. Paul and J. Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the ICSLP*, pages 899–902, Banff, 1992.

[8] A-V.I. Rosti and M.J.F. Gales. Factor analysed hidden markov models for speech recognition. *Computer Speech and Language*, 18(2):181–200, 2004.

[9] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen. Maximum likelihood discriminant feature spaces. In *Proceedings of the IEEE ICASSP*, Istanbul, 2000.

[10] L. Saul and M. Rahim. Maximum likelihood and minimum classification error factor analysis for automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 8(2):115–125, 2000.

[11] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig. The ibm 2004 conversational telephony system for rich transcription. In *Proceedings of the IEEE ICASSP*, volume I, pages 205–208, Philadelphia, 2005.