FACTORIAL HMMS FOR ACOUSTIC MODELING

Beth Logan[*] *and Pedro Moreno*

Cambridge Research Laboratories
Digital Equipment Corporation
One Kendall Square, Building 700, 2nd Floor
Cambridge, Massachusetts 02139
United States

ABSTRACT

Despite the success of hidden Markov models (HMMs) and other techniques for speech recognition, there remains a wide perception in the speech research community that new ideas are needed to continue improvements in performance. This paper represents a contribution to this effort.

We describe preliminary experiments using an alternative modeling approach known as factorial hidden Markov models (FHMMs). We present these models as extensions of HMMs and detail a modification to the original formulation which seems to allow a more natural fit to speech. We present experimental results on the phonetically balanced TIMIT database comparing the performance of FHMMs with HMMs. We also study alternative feature represetations that might be more suited to FHMMs.

1. INTRODUCTION

Over the last decade hidden Markov models have become the dominant technology in speech recognition. HMMs provide a very useful paradigm to model the dynamics of speech signals. They provide a solid mathematical formulation for the problem of learning HMM parameters from speech observations. Furthermore, efficient and fast algorithms exist for the problem of computing the most likely model given a sequence of observations.

Due to their success, there has recently been some interest in exploring possible extensions to HMMs. These include factorial HMMs [5] and coupled HMMs [2]. In this paper we explore factorial HMMs. These were first introduced by Ghahramani [5]. They attempt to extend HMMs by allowing the modeling of several stochastic random processes loosely coupled. Factorial HMMs can be seen as both an extension to HMMs or as a modeling technique in the Bayesian belief networks [10] domain. In our work we choose to approach them as extensions to HMMs. Further detail can be found in [9].

The paper is organized as follows. We start by describing the basic theory of HMMs and then follow by presenting FHMMs as extensions of these. A modification to the original formulation is then proposed which allows better modeling of speech. We describe then several experiments designed to compare the performance of FHMMs with traditional HMMs. We end this paper with our conclusions and suggestions for future work.

2. FACTORIAL HIDDEN MARKOV MODELS

Factorial hidden Markov models were first described by Ghahramani [5]. In his original work Ghahramani presents FHMMs and introduces several methods to efficiently learn their parameters. Our focus, however, is on studying the applicability of FHMMs to speech modeling. Our goal is to study FHMMs as a viable replacement for HMMs.

2.1. Model Description

Hidden Markov models are probabilistic models describing a sequence of observation acoustic vectors $Y = \{Y_t : t = 1, \ldots, T\}$. They are characterized by a hidden state sequence and an output probability which depends on the current state.

The probability density function (pdf) of Y given the model λ is

$$p(Y|\lambda) = \sum_S \Pi(S_1) p(Y_1|S_1) \prod_{t=2}^{T} P(S_t|S_{t-1}) p(Y_t|S_t). \quad (1)$$

Here S is a sequence of states $\{S_t, t = 1, \ldots, T\}$, $P(S_t|S_{t-1})$ is the transition probability from state S_{t-1} to state S_t, $\Pi(S_1)$ is the (prior) probability of being in state S_1 at time $t = 1$, and $p(Y_t|S_t)$ is the pdf of the observation vector Y_t given the state S_t. $p(Y_t|S_t)$ is typically modeled as a Gaussian mixture. We assume that the model has K states.

In the speech community a HMM is typically represented as shown in Figure 1. Here each state is shown explicitly and the arrows show allowable transitions between states. However a HMM can also be represented as a dynamic belief network [10] as shown in Figure 2. This alternative representation shows the evolution of the state sequence in time. Each node represents the state at each time slice. This context switch to dynamic belief networks allows many new modeling posibilities such as FHMMs.

The factorial HMM arises by forming a dynamic belief network composed of several 'layers'. This is shown in Figure 3. We see here that each layer has independent dynamics but that the observation vector depends upon the current state in each of the layers. This is achieved by allowing the state variable in Equation 1 to be composed of a collection of states. That is, we now have a 'meta-state' variable S_t which is composed of M states as follows

$$S_t = S_t^{(1)}, \ldots, S_t^{(M)}. \quad (2)$$

[*] Beth Logan is a PhD student at the University of Cambridge, United Kingdom. This work was performed during a summer internship at Cambridge Research Laboratories, Massachusetts.

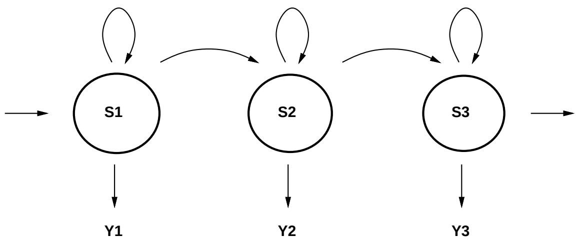

Figure 1: Topological representation of a Hidden Markov Model

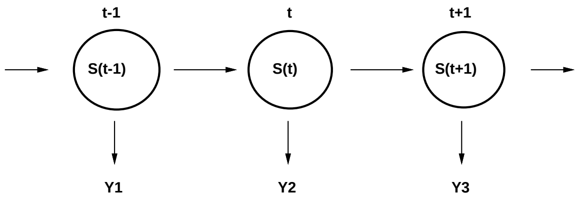

Figure 2: Dynamic Belief Network representation of a Hidden Markov Model

Here the superscript is the layer index with M being the number of layers. The layer nature of the model arises by only allowing transitions between states in the same layer. Were we to allow unrestricted transitions between states we would have a regular HMM with a K^MxK^M transition matrix. Intermediate architectures in which some limited transitions between states in different layers are allowed have also been presented in [2].

By dividing the states into layers we form a system that models several processes with loosely coupled dynamics. Each layer has similar dynamics to a basic hidden Markov model but the probability of an observation at each time depends upon the current state in all of the layers. For simplicity the number of possible states in each layer is K. Thus we have a system that requires M KxK transition matrices.

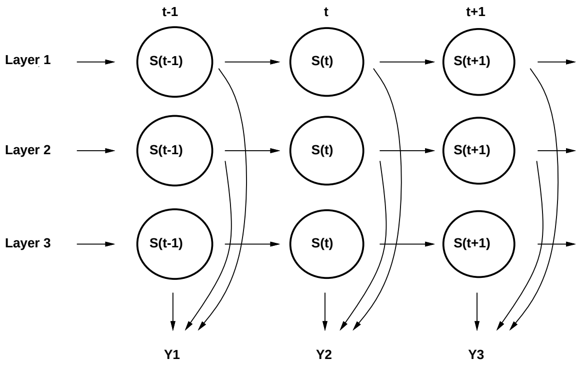

Figure 3: Dynamic Belief Network representation of a Factorial Hidden Markov Model

2.1.1. *Topological Equivalence to a Basic HMM*

Notice that a factorial HMM system could still be represented as a traditional HMM with a K^MxK^M transition matrix. For example, consider a two-layer system with three states per layer. Let the transition matrices for layer 1 and layer 2 be A_1 and A_2 respectively.

$$A_1 = \begin{pmatrix} a_1 & b_1 & c_1 \\ 0 & d_1 & e_1 \\ 0 & 0 & 1 \end{pmatrix} \quad A_2 = \begin{pmatrix} a_2 & b_2 & c_2 \\ 0 & d_2 & e_2 \\ 0 & 0 & 1 \end{pmatrix}$$

The transition matrix for the equivalent basic HMM system is built by creating a Cartesian product of the two original matrices A_1 and A_2

$$\begin{pmatrix} a_1a_2 & a_1b_2 & a_1c_2 & b_1a_2 & b_1b_2 & b_1c_2 & c_1a_2 & c_1b_2 & c_1c_2 \\ 0 & a_1d_2 & a_1e_2 & 0 & b_1d_2 & b_2e_2 & 0 & c_1d_2 & c_1e_2 \\ 0 & 0 & a_1 & 0 & 0 & b_1 & 0 & 0 & c_1 \\ 0 & 0 & 0 & d_1a_2 & d_1b_2 & d_1c_2 & e_1a_2 & e_1b_2 & e_1c_2 \\ 0 & 0 & 0 & 0 & d_1d_2 & d_1e_2 & 0 & e_1d_2 & e_1e_2 \\ 0 & 0 & 0 & 0 & 0 & d_1 & 0 & 0 & e_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_2 & b_2 & c_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & d_2 & e_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

resulting in a transition matrix with $K^M = 9$ states. As we can see an explosion in the number of states occurs. For this reason, as noted in [5], it is preferable to use the M KxK transition matrices over the equivalent K^MxK^M representation simply on computational grounds.

2.1.2. *Posterior Probability Formulation*

We now consider the probability of the observation given the meta-state. As mentioned, this probability depends on the current state in all the layers. In Ghahramani's original work, this probability was modeled by a Gaussian pdf with a common covariance and the mean being a linear combination of the state means. This pdf is given by Equation 3. We refer to this model as a 'linear' factorial HMM.

$$p(Y_t|S_t) \propto \tag{3}$$
$$\exp\left\{-\frac{1}{2}\left(Y_t - \sum_{m=1}^{M}\mu^{(m|S_t)}\right)^t C^{-1}\left(Y_t - \sum_{m=1}^{M}\mu^{(m|S_t)}\right)\right\}$$

Here $\mu^{(m|S_t)}$ is the mean of layer m given the meta-state S_t and C is the covariance. Other symbols are as previously defined.

A problem with this combination technique is that is it not extendible to the multiple Gaussian mixture. Neither is it a very natural fit to speech.

We propose a combination method that assumes that $p(Y_t|S_t)$ is the product of the (Gaussian) distributions of each layer. We refer to this technique as the 'streamed' method with each layer of the FHMM modeling a 'stream' of the observation vector. This method is extendible to multiple Gaussian mixtures. This pdf is defined by Equation 4 below.

$$p(Y_t|S_t) \propto \tag{4}$$
$$-\frac{1}{2}\prod_{m=1}^{M}\exp\left\{\left(M_mY_t - \mu^{(m|S_t)}\right)^t C^{-1}\left(M_mY_t - \mu^{(m|S_t)}\right)\right\}.$$

Here the matrix M_m partitions the observation vector into streams. For example in a two-layer system we have

$$M_0 = \left(\mathbf{I}_K \mid \mathbf{0}_K \right) \quad (5)$$

$$M_1 = \left(\mathbf{0}_K \mid \mathbf{I}_K \right). \quad (6)$$

Here $\mathbf{I}_K$ is the KxK identity matrix.

This formulation of the FHMM seems a more natural fit to speech feature vectors since these are often composed of several streams of sub-vectors. For example, a typical feature vector may consist of the cepstrum, delta cepstrum, second delta cepstrum, and sometimes even energy and its derivatives. If these different streams have somewhat decoupled dynamics then a factorial HMM could be a logical alternative to HMMs. Each distinct sub-vector stream could be modeled by each of the layers in the FHMM.

The idea of streams has already been proposed in the speech research community. Recognition engines like SPHINX [8] and HTK [11] allow similar formulations in their HMM systems. The difference between our formulation and theirs is that the streamed FHMM allows more decoupling of the streams' dynamics.

Notice that in Equation 4 we show a single covariance although extending this formulation to use a different covariance for each stream or each state in each stream is straightforward.

3. ESTIMATION OF PARAMETERS

The parameters of the FHMM are estimated using the Estimation-Maximization algorithm [3]. For further details refer to [5] and [9].

4. EXPERIMENTAL RESULTS

Our experiments tested a factorial HMM system on a phoneme classification task. We used the phonetically balanced TIMIT database [4]. Training was performed on the 'sx' and 'si' training sentences. These create a training set with 3696 utterances from 168 different speakers. 250 sentences from the test set were used for testing. The factorial HMM had two layers and three states in each layer. The standard Lee phonetic clustering [7] was used resulting in 48 phoneme models with these being further clustered during scoring to 39 models.

A baseline system was also implemented. This was a three-state left-to-right HMM system. Mixtures of Gaussians were used to model the posterior probabilities of the observation given the state. 8 mixture components were used per state.

We used cepstral and delta-cepstral features derived from 25.6ms long window frames. The dimension of the feature vector was 24 (12 cepstral and 12 delta cepstral features).

4.1. Linear Factorial HMMs

The first experiment investigated the performance of a linear factorial HMM. The results are shown in Table 1. For this experiment, the means and covariance were initialized using the mean and covariance of the pooled training data.

These results demonstrate that the linear factorial HMM models speech poorly. A major problem here is that there are not enough system parameters to form a good model. Adding more layers or states would increase the computational complexity exponentially while only providing small modeling advantages. We therefore turn our attention to the streamed FHMM.

Model	% Error
Baseline HMM	42.9
Linear FHMM	71.3

Table 1: Classification Results - Linear FHMM *vs* HMM

Model	Feature Vector	% Error
Baseline HMM	Cepstrum + Delta Cepstrum	42.9
Baseline HMM	Cepstrum	51.6
Baseline HMM	Delta Cepstrum	62.3
Streamed FHMM	Cepstrum + Delta Cepstrum	46.3

Table 2: Classification Results - Streamed FHMM *vs* HMM

4.2. Streamed Factorial HMMs

The parameters for each stream are initialized using regular HMMs trained on the features of the corresponding stream. Table 2 shows the results when one layer models the cepstrum and the other models the delta cepstrum. For completeness, the error rates of the HMMs trained on the cepstrum and delta cepstrum only are also shown. 8 mixture components per state were used in all the models.

We can see that while the streamed FHMM produces reasonable results it is not able to improve upon the basic HMM model.

A reason for this may be that there is only an advantage in using the FHMM if the layers model processes with different dynamics. The cepstrum and delta cepstrum are highly correlated hence it is to be expected that they would have similar dynamics.

We therefore tried feature vectors that we expected to be somewhat more decorrelated. It was hoped that perhaps the modeling assumptions of FHMMs might be more adequate and provide an edge over traditional HMMs.

4.3. Subband-based Speech Classification

Recently, researchers have considered modeling partial frequency bands by separate HMMs and combining the probabilities from these at a suitable level (e.g. the phoneme level) [1], [6]. The idea has its roots in models of human auditory perception. Figure 4 shows the sub-band model.

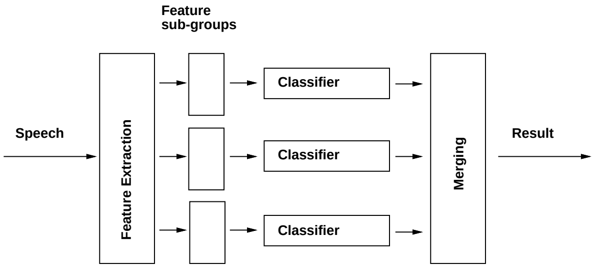

Figure 4: Feature Sub-band Classification Model

Model	Feature Vector	% Error
Baseline HMM	Upper + Lower band	46.9
Baseline HMM	Upper band	66.7
Baseline HMM	Lower band	59.5
Parallel HMM	Upper + Lower band	45.6
Streamed FHMM	Upper + Lower band	48.3

Table 3: Classification Results - Streamed FHMM

Examining this figure we can see there is clearly a great deal of scope for research when chosing the number of feature sub-groups and the merging technique. We do not consider these issues in our work. We have implemented a simple two-band version of the sub-band model using addition of the acoustic log likelihood at the phoneme level as the merging technique. We call this system a 'parallel' HMM.

The feature vectors for this system were derived as follows. A traditional mel-based log spectrum vector with 40 components was generated. The log spectrum was divided in two streams, the first one containing the lower 20 components and the second one containing the the upper 20 vector components. Each of the sub-vectors was rotated by a DCT matrix of dimension 20x12 generating two cepstral vectors each of dimension 12. Each of these streams of vectors was then mean normalized. Delta features for the resulting two streams were produced and appended to them.

Table 3 shows the results for experiments using the banded feature vectors. We present results for tests using the baseline HMMs, FHMMs, parallel HMMs and also for HMMs trained on only the lower or upper band and their delta coefficients.

The factorial HMM was initialized as follows. Each of the layers was trained first using traditional HMM techniques. These HMMs were the initial models used by the FHMM training algorithm.

Again we can see that there is no advantage in using the FHMM model.

5. DISCUSSION

Further work is needed to conclude if factorial HMMs are a good alternative to HMMs. Since the major advantage offered by these models appears to be their ability to model a process which is composed of independently evolving sub-processes, the choice of features is critical. If the features are indeed highly correlated factorial HMMs do not seem to offer compelling advantages. This fact is noted by Brand [2] who states that 'conventional HMMs excel for processes that evolve in lockstep; FHMMs are meant for processes that evolve independently'.

We postulate however along similar lines as [6] that there could be some advantage in using the FHMM framework to model speech and noise if these were uncorrelated. Alternatively if sub-band features were used the FHMM could provide more robust recognition in the case of corruption in one sub-band. Further work is needed in this area.

The most interesting research direction however would be to investigate the combination of traditional speech features with other information such as articulator positions or language models or lip tracking information. The FHMM framework provides an interesting alternative to combining several features without the need to collapse them into a single augmented feature vector.

It is important to notice that alternative formulations combining the information from each of the states in the meta-state are possible. In this paper we have described the linear FHMM and the streamed FHMM. Perhaps other alternatives could be explored.

Our conclusion, therefore, is that further research is needed to decide if algorithmic extensions to HMMs such as factorial HMMs or coupled HMMs offer a good alternative to traditional HMM techniques. The work in this paper only represents a very first effort in this direction.

6. CONCLUSIONS

We have presented factorial HMMs as possible extensions of hidden Markov models. These models were investigated in the context of phoneme classification as a possible replacement for traditional HMMs. We have also introduced and explored the concept of streamed factorial HMMs. Our experimental results proved inconclusive. In the experiments presented in this paper, factorial HMMs did not appear to offer any advantage over regular HMMs when traditional feature vectors were used. We postulate that this is because any modeling advantage offered by factorial HMMs will only become evident if less correlated features are used. We conclude the paper with suggestions for future work.

7. ACKNOWLEDGMENTS

We would like to thank Jim Rehg and Kevin Murphy for helpful discussions and the Speech Group at CRL for support.

8. REFERENCES

[1] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands, " *Proceedings International Conference on Spoken Language Processing*, 1996.

[2] M. Brand,, "Coupled hidden Markov models for modeling interacting processes", *MIT Media Lab Perceptual Computing/Learning and Common Sense Techincal Report 405 (Revised)*, June 1997.

[3] A. Dempster, N. Laird and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B*, 39:1-38, 1977.

[4] W. Fisher, G. Doddington and K. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status", *Proceedings of the DARPA Speech Recognition Workshop*, pp. 93-99, 1986.

[5] Z. Ghahramani and M. Jordan, "Factorial Hidden Markov Models", *Computational Cognitive Science Technical Report 9502 (Revised)*, July 1996.

[6] H. Hernansky, M. Pavel and S. Tibrewala, "Towards ASR on partially corrupted speech", *Proceedings International Conference on Spoken Language Processing*, 1996.

[7] K. Lee and H. Hon, "Speaker-independent phone recognition using hidden Markov models", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, No. 11, 1989.

[8] K. Lee, H. Hon and D. Reddy, “An overview of the SPHINX speech recognition system”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, No. 1, pp. 35-45, 1990.

[9] B. Logan and P. Moreno, “Factorial hidden Markov models for speech recognition: preliminary experiments”, *Cambridge Research Laboratories Technical Report 97/7*, 1997.

[10] S. Russell and P. Norvig, Artificial Intelligence A Modern Approach, Prentice Hall, 1995.

[11] S. Young, P. Woodland and W. Byrne, “HTK: Hidden Markov Model Toolkit V1.5”, Cambridge University Engineering Department Speech Group and Entropic Research Laboratories Inc. 1993.