# Factorial Hidden Markov Models for Speech Recognition: Preliminary Experiments

*Beth Logan   Pedro J. Moreno*

**Cambridge Research Laboratory**
Technical Report Series

**CRL 97/7**
September 1997

**Cambridge Research Laboratory**

The Cambridge Research Laboratory was founded in 1987 to advance the state of the art in both core computing and human-computer interaction, and to use the knowledge so gained to support the Company's corporate objectives. We believe this is best accomplished through interconnected pursuits in technology creation, advanced systems engineering, and business development. We are actively investigating scalable computing; mobile computing; vision-based human and scene sensing; speech interaction; computer-animated synthetic persona; intelligent information appliances; and the capture, coding, storage, indexing, retrieval, decoding, and rendering of multimedia data. We recognize and embrace a technology creation model which is characterized by three major phases:

**Freedom**: The life blood of the Laboratory comes from the observations and imaginations of our research staff. It is here that challenging research problems are uncovered (through discussions with customers, through interactions with others in the Corporation, through other professional interactions, through reading, and the like) or that new ideas are born. For any such problem or idea, this phase culminates in the nucleation of a project team around a well articulated central research question and the outlining of a research plan.

**Focus**: Once a team is formed, we aggressively pursue the creation of new technology based on the plan. This may involve direct collaboration with other technical professionals inside and outside the Corporation. This phase culminates in the demonstrable creation of new technology which may take any of a number of forms - a journal article, a technical talk, a working prototype, a patent application, or some combination of these. The research team is typically augmented with other resident professionals—engineering and business development—who work as integral members of the core team to prepare preliminary plans for how best to leverage this new knowledge, either through internal transfer of technology or through other means.

**Follow-through**: We actively pursue taking the best technologies to the marketplace. For those opportunities which are not immediately transferred internally and where the team has identified a significant opportunity, the business development and engineering staff will lead early-stage commercial development, often in conjunction with members of the research staff. While the value to the Corporation of taking these new ideas to the market is clear, it also has a significant positive impact on our future research work by providing the means to understand intimately the problems and opportunities in the market and to more fully exercise our ideas and concepts in real-world settings.

Throughout this process, communicating our understanding is a critical part of what we do, and participating in the larger technical community—through the publication of refereed journal articles and the presentation of our ideas at conferences–is essential. Our technical report series supports and facilitates broad and early dissemination of our work. We welcome your feedback on its effectiveness.

Robert A. Iannucci, Ph.D.
Director

# Factorial Hidden Markov Models for Speech Recognition: Preliminary Experiments

Beth Logan[1] Pedro J. Moreno

September 1997

**Abstract**

During the last decade the field of speech recognition has used the theory of hidden Markov models (HMMs) with great success. At the same time there is now a wide perception in the speech research community that new ideas are needed to continue improvements in performance. This report represents a small contribution in this effort. We explore an alternative acoustic modeling approach based on Factorial Hidden Markov Models (FHMMs). These are presented as possible extensions to HMMs. We show results for phonetic classification experiments using the phonetically balanced TIMIT database which compare the performance of FHMMs with HMMs and parallel HMMs.

[1] Beth Logan is a PhD student at the University of Cambridge, United Kingdom. This work was done during a summer internship.

CRL Technical reports are available on the CRL's web page at
http://www.crl.research.digital.com.

Digital Equipment Corporation
Cambridge Research Laboratory
One Kendall Square, Building 700
Cambridge, Massachusetts 02139 USA

# Contents

# List of Figures



# List of Tables

# 1 Introduction

In recent years 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 this success, there has recently been some interest in exploring possible extensions to HMMs. These include factorial HMMs [Ghahramani and Jordan, 1996] and coupled HMMs [Brand, 1997] among others. In this report we explore factorial HMMs. These were first introduced by Ghahramani [Ghahramani and Jordan, 1996] and 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 [Russell and Norvig, 1995] domain. In this report we choose to approach them as extensions to HMMs.

The report is organized as follows. We start by describing the basic theory of HMMs and then follow by presenting FHMMs as extensions of these. We continue by presenting an extension to the traditional HMM Baum-Welch learning algorithm applied to FHMMs. We describe then several experiments designed to compare their performance with traditional HMMs. We end this report with our conclusions and suggestions for future work.

# 2 Factorial Hidden Markov Models

Factorial HMMs were first described in [Ghahramani and Jordan, 1996]. 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.

To this end, we have made an effort to explain FHMMs as extensions of HMMs, making connections between these two techniques when possible. We assume the reader is somewhat familiar with HMM theory.

## 2.1 Model Description

The description requires us to first briefly introduce hidden Markov models. These models are the dominant technology used for speech recognition. Tractable, well understood training and testing algorithms exist to estimate the model parameters and evaluate the likelihood of alternative speech utterances. Their main strength lies in their ability to capture the dynamic information in the speech signal. They are able to model dynamic patterns, *i.e.*, patterns of variable length. This is important because for example the same phoneme when uttered by the same speaker can vary in length.

Hidden Markov models are probabilistic models which describe a sequence of acoustic observation vectors $Y = \{Y_t : t = 1, \dots, T\}$. The random process generating the observation is modeled as being in one of $K$ states. The states are not observable hence the "hidden" nature of the model. Each state can be thought of as representing particular speech patterns or regions.

The parameters of the HMM are the probability density functions (pdf) describing the statistics of the acoustic vectors being produced or generated by each of the states, and the transition probabilities modeling the likelihoods of evolving from one state to another. For a first order HMM, this transition probability depends only on the current state.

The probability that an observation $Y$ is generated given the model is expressed as follows

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

Here:

$$\begin{aligned}
Y &= \text{a sequence of } N \text{ dimensional vector observations } \{Y_t, t = 1, \dots, T\} \\
S &= \text{a sequence of states } \{S_t, t = 1, \dots, T\} \\
P(S_t|S_{t-1}) &= \text{transition probability from state } S_{t-1} \text{ to state } S_t \\
\Pi(S_1) &= \text{the probability of being in state } S_1 \text{ at time } t = 1 \\
p(Y_t|S_t) &= \text{pdf of the observation vector } Y_t \text{ given the state } S_t \\
&\quad\ \text{typically modeled as a mixture of Gaussians} \\
K &= \text{the number of states in the model} \\
\lambda &= \text{the model parameters} = \{K, \{P(S_t|S_{t-1})\}, \{p(Y_t|S_t)\}\}
\end{aligned}$$

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 [Russell and Norvig, 1995] as shown in Figure 2. This alternative representation shows the evolution of the state sequence with time since each node represents the state at each time slice. This context switch to dynamic belief networks shows the manner in which HMMs can be generalized to 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

$$S_t = S_t^{(1)}, \dots, S_t^{(M)} \tag{2}$$

Here the superscript is the layer index with $M$ being the number of layers. The layer nature of the model arises by restricting transitions between the states in different layers. Were we to allow unrestricted transitions between states in different layers we would simply have a regular HMM with a $K^M$x$K^M$ transition matrix. Intermediate architectures in which some limited transitions between states in different layers are allowed have also been presented in [Brand, 1997].

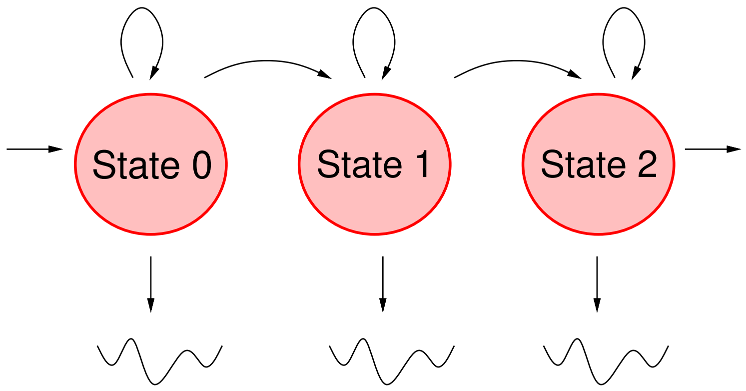


Figure 1: Topological representation of a Hidden Markov Model

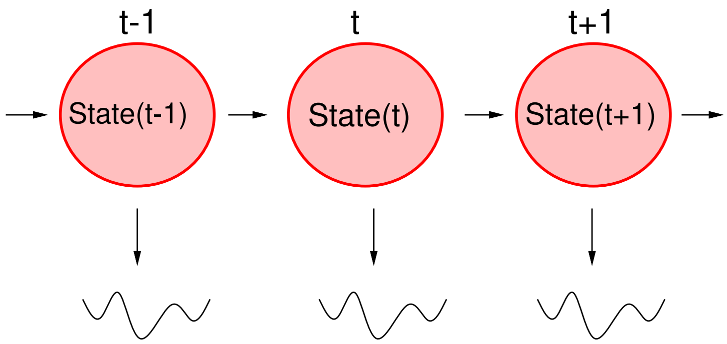


Figure 2: Dynamic Belief Network representation of a Hidden Markov Model

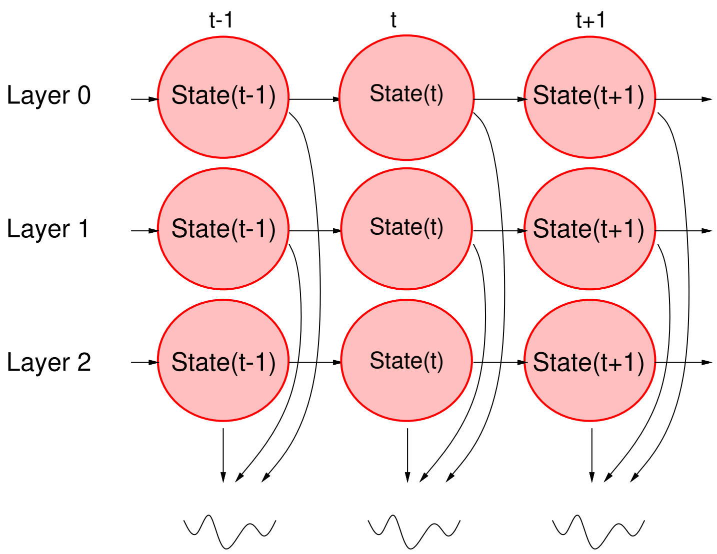


Figure 3: Factorial Hidden Markov Model

By dividing the states into layers we form a system that can model several processes with independent dynamics which are loosely coupled. 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. In our formulation it is assumed for simplicity that in each layer, the state variable can take on one of $K$ distinct values at each time (rather than assuming that the number of possible states within each layers is different). Thus we have a system that requires $M$ $K$x$K$ transition matrices. It should be noted that this system could still be represented as a regular HMM with a $K^M$x$K^M$ transition matrix with zeros representing illegal transitions.

For example, consider a 2-layer system with 3 states per layer. Let the transition matrices for layer 0 and layer 1 be $A_0$ and $A_1$ respectively.

$$A_0 = \begin{pmatrix} a_0 & b_0 & c_0 \\ 0 & d_0 & e_0 \\ 0 & 0 & 1 \end{pmatrix} \quad A_1 = \begin{pmatrix} a_1 & b_1 & c_1 \\ 0 & d_1 & e_1 \\ 0 & 0 & 1 \end{pmatrix}$$

The transition matrix for the equivalent basic HMM system is built by creating a Cartesian product of the two original matrices $A_0$ and $A_1$

$$\begin{pmatrix} a_0\,a_1 & a_0\,b_1 & a_0\,c_1 & b_0\,a_1 & b_0\,b_1 & b_0\,c_1 & c_0\,a_1 & c_0\,b_1 & c_0\,c_1 \\ 0 & a_0\,d_1 & a_0\,e_1 & 0 & b_0\,d_1 & b_1\,e_1 & 0 & c_0\,d_1 & c_0\,e_1 \\ 0 & 0 & a_0 & 0 & 0 & b_0 & 0 & 0 & c_0 \\ 0 & 0 & 0 & d_0\,a_1 & d_0\,b_1 & d_0\,c_1 & e_0\,a_1 & e_0\,b_1 & e_0\,c_1 \\ 0 & 0 & 0 & 0 & d_0\,d_1 & d_0\,e_1 & 0 & e_0\,d_1 & e_0\,e_1 \\ 0 & 0 & 0 & 0 & 0 & d_0 & 0 & 0 & e_0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_1 & b_1 & c_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & d_1 & e_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

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 we note in section 2.3 it is preferable to use the $M$ $K$x$K$ transition matrices over the equivalent $K^M$x$K^M$ representation simply on computational grounds.

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 our work, we have used two different ways of combining the information from the layers. The first method assumes that the observation is distributed according to a Gaussian pdf with a common covariance and the mean being a linear combination of the state means. This formulation was originally proposed by Ghahramani [Ghahramani and Jordan, 1996] and is shown in Equation 3. We refer to this model as a "linear" factorial HMM.

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

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.

The second combination method 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. Each layer of the FHMM models a stream of the observation vector. The idea of streams has already been proposed in the speech research community. Recognition engines like SPHINX [Lee et al., 1990] and HTK [Young et al., 1993] allow similar formulations in their HMM systems. The difference between our formulation and their's is that a "streamed" FHMM allows more decoupling in the streams' dynamics.

The equation for the observation probabilities in our streamed case is

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

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

$$\begin{aligned} M_0 &= \left(\ \mathbf{I}_D \mid \mathbf{0}_D\ \right) \\ M_1 &= \left(\ \mathbf{0}_D \mid \mathbf{I}_D\ \right) \end{aligned}$$

Here $\mathbf{I}_D$ is the $D$x$D$ identity matrix and $D$ is the dimensionality of each of the streams. We will discuss later in more detail the motivation for this alternative formulation.

Notice that here we use a single covariance although extending this formulation to use a different covariance for each stream or for each state within the stream is straightforward.

## 2.2 Estimation of Parameters

The model parameters are the means of the states in each layer, the transition probabilities between states in each layer, the prior probabilities of each state and the covariance. All these parameters can be estimated using the Expectation Maximization (EM) algorithm [Dempster et al., 1977]. Due to our slightly different formulation of the acoustic probability, the algorithm we present here is different but equivalent to that presented in [Ghahramani and Jordan, 1996].

The basic workings of the algorithm are well known. Model parameters are initialized and then reestimated to maximize a so-called auxiliary function. The algorithm guarantees to increase the likelihood of the observations given the model on each iteration. Only convergence to a local maximum is guaranteed.

We first discuss reestimation of the model parameters by maximization of the auxiliary function.

The auxiliary function to be maximized is

$$\phi(\lambda, \lambda') = \sum_S p_\lambda(S|Y) \ln p_{\lambda'}(S, Y) \tag{5}$$

In this and subsequent equations the prime denotes the reestimated or new model parameters.

Substituting Equations 1 and 2 into 5 we have

$$\phi(\lambda, \lambda') = \sum_S p_\lambda(S|Y) \left[ \ln \Pi(S_1)' + \sum_{t=2}^{T} \sum_{m=1}^{M} \ln P(S_t^{(m)}|S_{t-1}^{(m)})' + \sum_{t=1}^{T} \ln p(Y_t|S_t)' \right]$$

Here $P(S_t^{(m)}|S_{t-1}^{(m)})$ is the transition probability between state $S_{t-1}^{(m)}$ and $S_t^{(m)}$. This equation can be separated into components which depend only on each set of parameters to be reestimated.

$$\phi(\lambda, \lambda') = \phi_a(\lambda, \lambda') + \phi_b(\lambda, \lambda') + \phi_c(\lambda, \lambda')$$

Here $\phi_a(\lambda, \lambda')$ is the part of $\phi(\lambda, \lambda')$ which depends on the prior probabilities, $\phi_b(\lambda, \lambda')$ is the part which depends only on the transition probabilities and $\phi_c(\lambda, \lambda')$ is the part which depends only on the means and covariance.

We present here formulas for the single observation case. Extension to multiple observations is straightforward.

### 2.2.1 Reestimation of the Means

The means are reestimated by maximizing $\phi_c(\lambda, \lambda')$. For linear FHMMs (means combined using Equation 3) the auxiliary function becomes (ignoring the term in $\phi_c(\lambda, \lambda')$ containing only the covariance)

$$\phi_c(\lambda, \lambda') = \sum_S p_\lambda(S|Y) \sum_{t=1}^{T} \left[ -\frac{1}{2} \left( Y_t - \sum_{m=1}^{M} \mu^{(m|S_t)'} \right)^t C^{-1} \left( Y_t - \sum_{m=1}^{M} \mu^{(m|S_t)'} \right) \right] \quad (6)$$

To reestimate the $i$th mean of the $n$th layer we take the derivative of Equation 6 with respect to $\mu_i^{(n)}$ and set it equal to zero. This leads to the following equation

$$0 = \sum_{t=1}^{T} \sum_{S_t, S_t^{(n)} \equiv i} P(S_t|Y, \lambda) \left( Y_t - \sum_{m=1}^{M} \mu^{(m|S_t)'} \right)$$

where $P(S_t|Y, \lambda)$ is the posterior probability of meta-state $S_t$ given the observations and the model.

This equation is clearly not solvable for $\mu_i^{(n)'}$. However, if the process is repeated for all the means of all the layers, $K$x$M$ equations will be generated for the $K$x$M$ means. These can be solved using matrix algebra, although in practice efficient matrix inversion techniques capable of handling ill-conditioned matrices are needed.

If the streamed method is used to combine the means then the equations become somewhat more decoupled. The auxiliary function is now

$$\phi_c(\lambda, \lambda') = \sum_S p_\lambda(S|Y) \sum_{t=1}^{T} \left[ -\frac{1}{2} \sum_{m=1}^{M} \left( M_m\, Y_t - \mu^{(m|S_t)'} \right)^t C^{-1} \left( M_m\, Y_t - \mu^{(m|S_t)'} \right) \right] \quad (7)$$

Solving for $\mu_i^{(n)'}$ we have

$$\mu_i^{(n)'} = \frac{\sum_{t=1}^{T} \sum_{S_t, S_t^{(n)} \equiv i} P(S_t|Y, \lambda) M_m\, Y_t}{\sum_{t=1}^{T} \sum_{S_t, S_t^{(n)} \equiv i} P(S_t|Y, \lambda)}$$

### 2.2.2 Reestimating the Covariance

The covariance is reestimated by maximizing $\phi_c(\lambda, \lambda')$ with respect to $C$. In the linear case, the reestimation formula is

$$C = \frac{1}{T} \sum_{t=1}^{T} \sum_{S_t} P(S_t|Y, \lambda) \left( Y_t - \sum_{m=1}^{M} \mu^{(m)} \right) \left( Y_t - \sum_{m=1}^{M} \mu^{(m)} \right)^t$$

For the streamed case, notice that the reestimation formula is very similar to the usual covariance reestimation formula for HMMs.

$$C = \frac{1}{T * M} \sum_{t=1}^{T} \sum_{S_t} \sum_{m=1}^{M} P(S_t|Y, \lambda)(M_m\, Y_t - \mu^{(m)})(M_m\, Y_t - \mu^{(m)})^t$$

If it is desired that each stream has a separate covariance, then the reestimation formula reduces to the usual HMM covariance reestimation formula with the observation being the part of the feature vector for that stream.

### 2.2.3 Reestimating the Transition Probabilities

The transition probabilities are reestimated by maximizing $\phi_b(\lambda, \lambda')$. We have

$$\phi_b(\lambda, \lambda') = \sum_S p_\lambda(S|Y) \sum_{t=2}^{T} \sum_{m=1}^{M} \ln p(S_t^{(m)}|S_{t-1}^{(m)\prime}) \tag{8}$$

Now let $a_{ij}^{(n)}$ be the transition probability from state $i$ to state $j$ in layer $n$. Maximizing Equation 8 with respect to $a_{ij}^{(n)}$ gives the following reestimation formula

$$a_{xy}^{n\;\prime} = \frac{\sum_{t=2}^{T} \sum_{S_{t-1} S_t, S_{t-1}^{(m)} \equiv x, S_t^{(m)} \equiv y} P(S_t|S_{t-1}, Y)}{\sum_{t=2}^{T} \sum_{S_{t-1} S_t, S_{t-1}^{(m)} \equiv x} P(S_t|S_{t-1}, Y)}$$

## 2.3 Calculation of the Posterior Probabilities

The reestimation formulas require the calculation of $P(S_t|Y, \lambda)$ and $P(S_t|S_{t-1}, \lambda)$, which we will refer for notational simplicity as $P(S_t|Y)$ and $P(S_t|S_{t-1})$ respectively.

Direct computation of these using Equations 1 and 2 would require $O(2T(K^M)^T)$ calculations which is intractable. This can be reduced to $O(TK^{2M})$ by use of the so-called Forward-Backward or Baum-Welsh algorithm [Rabiner, 1989].

In HMMs the usual method to calculate $P(S_t|Y)$ and $P(S_t|S_{t-1}, Y)$ is to define so-called Forward and Backward probabilities. The Forward Probability $\alpha_t(j)$ is defined as

$$\alpha_t(j) = P(Y_1, \ldots, Y_t, S_t = j|\lambda)$$

That is the probability of observing the first $t$ speech vectors and being in $j$th state at time $t$. Similarly the Backward Probability $\beta_t(j)$ is defined as

$$\beta_t(j) = P(Y_{t+1}, \ldots, Y_T|S_t = j, \lambda)$$

These probabilities can be calculated using simple recursion and they can be combined to give $P(S_t|Y)$ and $P(S_t|S_{t-1}, Y)$.

$$\begin{aligned} P(S_t = j|Y) &= \frac{\alpha_t(j)\, \beta_t(j)}{\sum_{S_t} \alpha_t(j)\beta_t(j)} \\ P(S_t = j|S_{t-1} = i, Y) &= \frac{\alpha_{t-1}(i)\, a_{ij}\, P(Y_t|S_t{=}j)\, \beta_t(j)}{\sum_{S_t} \alpha_t(j)\beta_t(j)} \end{aligned}$$

Unfortunately, in the factorial HMM case, the state $S_t$ is actually a meta-state. Therefore, to calculate the $\alpha_t$ and $\beta_t$ terms we would have to perform recursion over all the layers as well as all time. Ghahramani [Ghahramani and Jordan, 1996] presents a modified version of the Baum-Welch algorithm which does not depend on a $K^M$x$K^M$ transition matrix. Making use of the fact that each layer has independent dynamics, the calculations can be reduced to $O(TMK^{M+1})$. This is tractable for small $K$ and $M$. We present here Ghahramani's method with the equations to calculate $\alpha_t$ in slightly more detail. The equations for $\beta_t$ follow a similar pattern and are not presented here.

To calculate $\alpha_t$ we use the following recursion in space, *i.e.* for every time instant we perform a recursion across the layers

$$\alpha_t(i,j,\dots,z) = p_\lambda(Y_t|S_t)\alpha_t^{(0)} \tag{9}$$

$$\alpha_t^{(m-1)}(i,j,\dots,z) = \sum_{S_{t-1}^{(m)}} P(S_t^{(m)}|S_{t-1}^{(m)},\lambda)\alpha_t^{(m)} \tag{10}$$

$$\alpha_t^{(n)}(i,j,\dots,z) = P(S_{t-1}^{(i)},S_{t-1}^{(j)},\dots,S_t^{(n+1)},\dots,S_t^{(z)},Y_1,\dots,Y_{t-1}|\lambda) \tag{11}$$

$$\alpha_t^{(M)} = \alpha_{t-1}(i,j,\dots,z) \tag{12}$$

Here the indices of $\alpha_t(i,j,\dots,z)$ refer to the states in each layer. That is, the state at time $t$ in layer 0 takes value $i$, the state in layer 1 value $j$ and so on. To clarify these formulas, we briefly study the two-layer three-state case.

We initialize $\alpha_1(i,j)$ using the prior probabilities of states $i$ and $j$.

$$\alpha_1(i,j) = \Pi_i^{(0)}\Pi_j^{(1)} \qquad \forall i,j\epsilon\{0,1,2\}$$

Using Equation 12 we have

$$\alpha_2(i,j)^{(2)} = \alpha_1(i,j)$$

We now calculate $\alpha_2^{(1)}(i,j)$ and $\alpha_2^{(0)}(i,j)$ for all $i$ and $j$ using Equations 10 and 11.

$$\begin{aligned}
\alpha_2^{(1)}(i,j) &= \sum_{S_{t-1}^{(2)}} P(S_t^{(2)}=j|S_{t-1}^{(2)},\lambda)\alpha_t^{(2)}(i,S_{t-1}^{(2)}) \\
\alpha_2^{(0)}(i,j) &= \sum_{S_{t-1}^{(1)}} P(S_t^{(1)}=i|S_{t-1}^{(1)},\lambda)\alpha_t^{(1)}(S_{t-1}^{(1)},j)
\end{aligned}$$

Having calculated $\alpha_2^{(0)}(i,j)$ we use Equation 9 to calculate $\alpha_2(i,j)$, completing the recursion.

# 3 Experimental results

Our experiments tested a factorial HMM system on a phoneme classification task. We used the phonetically balanced TIMIT database [Fisher et al., 1986]. Training was performed on the "sx" and "si" training sentences. These create a training set with 3696

| Model | % Error |
|---|---|
| Baseline HMM | 42.9 |
| Linear FHMM | 71.3 |

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

utterances from 168 different speakers. 250 sentences from the test set were used for testing. The factorial HMM had 2 layers and 3 states in each layer. The standard Lee phonetic clustering [Lee and Hon, 1989] 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 3-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).

## 3.1 Linear Factorial HMMs

The first experiment investigated the performance of the 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. The only way to introduce more system parameters would be to add more layers and/or states because there is no obvious way to incorporate mixtures of Gaussians into the linear FHMM framework.

We therefore turn our attention to the streamed FHMM.

## 3.2 Streamed Factorial HMMs

The reestimation formulas for streamed FHMMs can be easily extended to the multiple Gaussian mixture case. It also seems a more natural fit to speech feature vectors normally 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, we hypothesize 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.

In our experiments the parameters for each stream were 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 both HMMs and FHMMs.

| 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

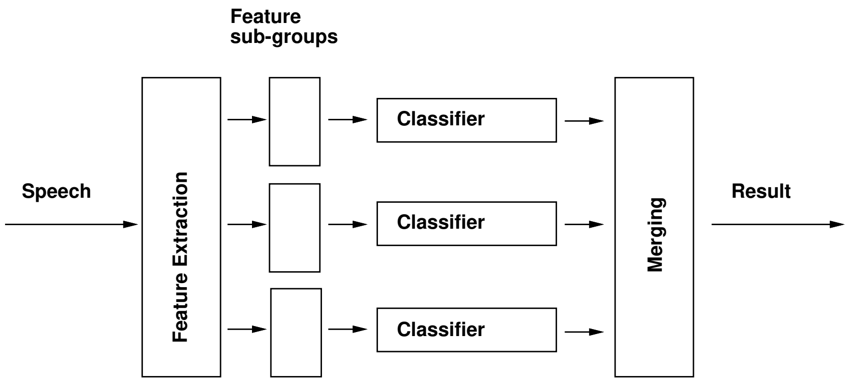


Figure 4: Sub-band Model

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.

## 3.3 Sub-band-based Speech Classification

Recently, researchers such as [Bourlard and Dupont, 1996], [Hermansky et al., 1996] and [Bourlard and Dupont, 1997], have considered modeling partial frequency bands by separate HMMs and combining the probabilities from these at a suitable level (e.g. the phoneme level). The idea has its roots in models of human auditory perception. Figure 4 shows the sub-band model.

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.

| 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

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 2 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.

# 4 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 [Brand, 1997] 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 [Hermansky et al., 1996] 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 report we have described the linear FHMM and the streamed FHMM. Perhaps other alternatives can be explored.

We believe, therefore, 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 report only represents a very first effort in this direction.

# 5 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 report, 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 report with suggestions for future work.

# 6 Acknowledgments

We would like to thank Jim Rehg and Kevin Murphy for helpful discussions and the Speech Group at CRL for support. We also thank Mark Tuttle at CRL for his help in formatting this document.

# References

[Bourlard and Dupont, 1996] 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*, Philadelphia, October 1996.

[Bourlard and Dupont, 1997] H. Bourlard and S. Dupont Subband-based speech recognition *Proceedings International Conference on Acoustics, Speech and Signal Processing*, Munich, 1997.

[Brand, 1997] 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.

[Dempster et al., 1977] 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.

[Fisher et al., 1986] 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.

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

[Hermansky et al., 1996] H. Hernansky, M. Pavel and S. Tibrewala Towards ASR on partially corrupted speech *Proceedings International Conference on Spoken Language Processing*, Philadelphia, October 1996.

[Lee and Hon, 1989] 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, Nov. 1989.

[Lee et al., 1990] 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, pp. 35-45, Jan. 1990.

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

[Rabiner, 1989] L. Rabiner A tutorial on hidden Markov models and selected applications in speech recognition *Proceedings of the IEEE*, vol. 77, pp. 257-285, Feb. 1989.

[Young et al., 1993] 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.

digital™

**Factorial Hidden Markov Models for Speech Recognition: Preliminary Experiments**

Beth Logan Pedro J. Moreno

**CRL 97/7**

September 1997