HMMs and Coupled HMMs for Multi-channel EEG Classification

Shi Zhong and Joydeep Ghosh
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712-1084, USA
szhong,ghosh@ece.utexas.edu

Abstract— A variety of Coupled HMMs (CHMMs) have recently been proposed as extensions of HMM to better characterize multiple interdependent sequences. This paper introduces a novel distance coupled HMM. It then compares the performance of several HMM and CHMM models for a multi-channel EEG classification problem. The results show that, of all approaches examined, the multivariate HMM that has low computational complexity surprisingly outperforms all other models.

I. INTRODUCTION

Classification of EEG is an important part of EEG-based brain-computer interfaces. An overview of EEG-based brain-computer interface systems is presented by Pfurtscheller and Neuper [1]. They summarize several approaches, such as linear discrimination analysis (LDA), artificial neural networks (ANN) and HMMs, for classifying features extracted from raw EEG data. When a neural network is used for EEG analysis, it is often modified to exploit time information. For example, Haselsteiner and Pfurtscheller [2] use a time-delayed neural network and collect features using an adaptive autoregressive (AR) model.

HMMs have been heavily researched and used for the past several decades, especially in the speech recognition area [3], and successfully applied to a wide variety of applications, including EEG classification [4], [5]. Works on EEG classification usually apply HMMs to the time-changing feature vectors extracted by an AR model or by some other digital signal processing techniques. Huang *et al* [4] use the mean frequency features, calculated from FFT spectrum, for detecting the arousal state changes. Obermaier, Guger, and Pfurtscheller [5] compare LDA and HMMs on bandpass-filtered feature vectors and experiment with the structure parameters of HMMs.

Penny and Roberts [6] conclude, based on experiments on synthetic data, that HMMs are capable of detecting nonstationary changes and are thus 'perfect' for EEG analysis. They point out that operating HMMs on AR coefficients is fundamentally flawed because the windowing procedure used in AR models may lead to incorrect estimates of state and state transitions in an HMM model.

In this paper, we use HMMs to model the (scaled) raw EEG data instead of the extracted features. This approach avoids the need for expert knowledge to construct a feature extractor. After all, why bother to construct features if the raw data can be well modeled? Our experimental results strongly support the feasibility of this approach.

Furthermore, we want to model multiple EEG channels simultaneously since EEG data often come as correlated time series from multiple electrodes on the scalp. There are many ways to do this. Using one HMM with multivariate gaussian observations is the most straightforward approach. Using one univariate HMM for each channel and then combining these HMMs is another one. Recently, CHMM models have been proposed to better model multiple interacting time series processes [7], [8] and they seem to work better than HMMs. Also some generalized HMM models have been suggested to enrich the HMM model for specific applications [9], [10]. In this paper, we propose a new CHMM formulation, named distance coupled HMM (DCHMM), that is related to the mixed memory Markov models [11]. We examine some of these sophisticated models on the EEG classification problem and compare their performance against the simple aforementioned HMM models.

The organization of this paper is as follows. Section II reviews HMMs and discusses several CHMM formulations. Section III details our DCHMM formulation with the extended forward-backward procedure and training algorithm. Section IV presents our experimental results on an EEG classification problem. Finally, section V concludes this paper.

II. HMM AND CHMM MODELS

A. Hidden Markov Models

The standard HMM model uses a discrete hidden state at time t to summarize all the information before t and thus the observation at any time depends only on the current hidden state. The hidden state time sequence in an HMM is a Markov chain. In this paper we use first order HMMs with gaussian observations, in which the observation distribution is normal at any state and the state sequence is a first order Markov chain. Such an HMM unrolled over several time slices is shown in Fig. 1.

A standard HMM is usually denoted as a triplet $\lambda = (\pi, A, B)$. $\pi = \{\pi_i\}$ (where $\sum_i \pi_i = 1$) is the prior probability distribution of hidden states. $A = \{a_{ij}\}$ (where $\sum_j a_{ij} = 1$) is the transition probability distribution between hidden states. For discrete observation case, the observation distribution is $B = \{b_j(k)\}$ (where $\sum_k b_j(k) = 1$). For the continuous observation case, the observation distribution is usually modeled by a mixture

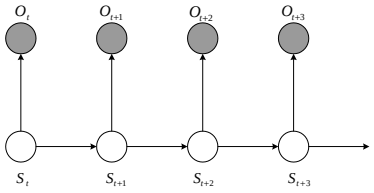

Fig. 1. A first order HMM model. The empty circles $\{s_t\}$ are hidden states and the shaded ones $\{o_t\}$ observation nodes.

of gaussians

$$b_j(o) = \sum_l c_{jl}\mathcal{N}[o, \mu_{jl}, U_{jl}], \quad \text{and} \ \sum_l c_{jl} = 1 \qquad (1)$$

where o is the observation vector being modeled, c_{jl} the mixture weight, μ_{jl} the mean vector of the l-th gaussian mixture for state j, U_{jl} the covariance matrix of the l-th mixture for state j, and $\mathcal{N}$ is the gaussian density function. When modeling EEG time series, we set the length of observation vector to be the number of channels. We call the HMM with scalar observations *univariate HMM* and the HMM with multivariate gaussian observations *multivariate HMM*.

Three basic problems of interest for HMMs are: evaluating the likelihood $P(o|\lambda)$ of an observation sequence o given the HMM λ; finding the most likely hidden state sequence S corresponding to an observation o given the model λ, and learning the parameters of a model λ given a set of observations O. The evaluation and learning of an HMM both exploit an efficient forward-backward inference procedure [12]. The inference is exact for standard HMM but can often only be approximate for more complex models discussed in next section.

B. Coupled HMMs

Various extended HMM models have been used to solve coupled sequence data analysis problems, such as complex human action recognition [13], traffic modeling [14] and biosignal analysis [8]. These new models aim to enhance the capabilities of standard HMM model by using more complex architectures, while still being able to utilize the established methodologies (e.g. EM algorithm) for standard HMM models. Several typical examples from recent literature are CHMMs [7], event-coupled HMMs [15], factorial HMMs (FHMMs) [10] and input-output HMMs (IOHMMs) [9], as shown in Fig. 2.

Fig. 2(b) is a specific type of coupled HMMs, developed by Kristjansson, Frey, and Huang [15], for modeling a class of loosely coupled time series where only the onsets of events are coupled in time. Bengio and Frasconi [9] develop the IOHMMs (Fig. 2(d)) to address the input-output sequence pair modeling problem. While the IOHMMs may be viewed as a superset of CHMMs (in which the hidden states from the previous time slice are treated as the inputs at current time slice), the inputs in IOHMM and the hidden states in CHMM are inherently

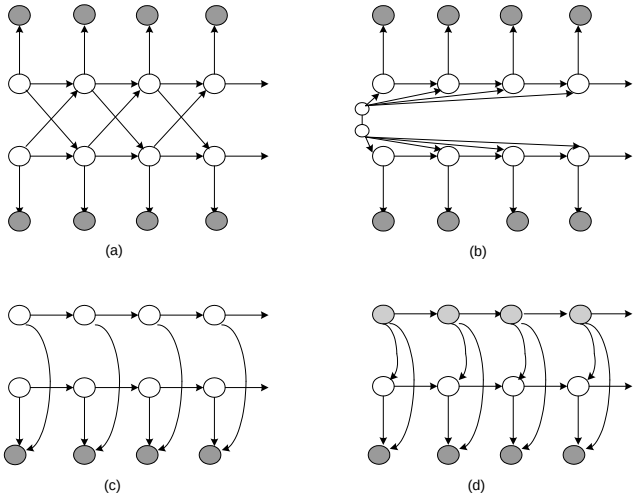

Fig. 2. Various new HMM architectures. The empty circles are the hidden states and the shaded ones the observation nodes (except for (d) where the lightly shaded ones are the input nodes). (a) Standard coupled HMMs; (b) Event-coupled HMMs; (c) Factorial HMMs; (d) Input-Output HMM.

different. A certain independence assumption of inputs does not apply to the hidden states and the EM algorithm used in [9] cannot be used for general CHMMs. The FHMM shown in Fig. 2(c) enriches the representation power of hidden states by putting in multiple hidden state chains for one HMM. It makes model training difficult or even impossible when the number of hidden state chains is large. Approximate inferences have to be used.

This paper focuses on the CHMMs wherein the state of one model at time t depends on the states of all models (including itself) at time $t-1$. Fig 2(a) shows two HMM chains coupled together. For C chains coupled together, the state transition probability is

$$P(S_t^{(c)}|S_{t-1}^{(1)}, S_{t-1}^{(2)}, ..., S_{t-1}^{(C)}) \qquad (2)$$

instead of $P(S_t^{(c)}|S_{t-1}^{(c)})$ as in a standard HMM model. Here $S_t^{(c)}$ is the hidden state of model c at time t. It is easy to see that the number of free parameters in the transition probability matrix is N^C (if the number of hidden states is N for every chain), which is exponential in the number of HMMs coupled together. This is not a desirable feature as it hinders accurate parameter learning.

There have been several variations of the standard CHMM for which the model size and inference problems are more tractable. Coupled HMMs proposed by Brand [7] is one of them. In his paper, Brand substitutes the joint conditional probability by the product of all marginal conditional probabilities, i. e.

$$P(S_t^{(c)}|S_{t-1}^{(1)}, S_{t-1}^{(2)}, ..., S_{t-1}^{(C)}) = \prod_{c'} P(S_t^{(c)}|S_{t-1}^{(c')}) \qquad (3)$$

This formulation is erroneous since the right hand side is not a properly defined probability density (does not

sum up to one). Rezek and Roberts [8] use a decoupled forward variable for each HMM chain in a CHMM, that is an approximation of the true forward variables. The computational complexity is reduced but still exponential in the number of HMM chains.

Kwon and Murphy [14] use CHMM to model freeway traffic. They cast CHMM in a more general framework called dynamic Bayesian network (DBN) in which the approximate inference can be done using Boyen-Koller (BK) algorithm [16]. Murphy and Weiss [17] examine a factored version of the BK algorithm that has a complexity of $O(TCN^{F+1})$, where T is the length of sequence and F the maximum fan-in of any node.

Saul and Jordan [11] reduce the number of parameters in Eq. (2) by representing it as a linear combination of conditional marginals. That results in a model they call mixed memory Markov model, in which the joint transition probability is

$$P(S_t^{(c)}|S_{t-1}^{(1)}, S_{t-1}^{(2)}, ..., S_{t-1}^{(C)}) = \sum_{c'} \theta_{c'c} P(S_t^{(c)}|S_{t-1}^{(c')}) \quad (4)$$

They then develop EM algorithm, by introducing a missing variable for c' in the equation, for training the model.

The DCHMM we propose in next section has the same representation as Eq. (4) but is motivated by a specific distance-coupled application. We develop a different training algorithm for the DCHMM model.

III. DCHMM FORMULATION

A. Motivation

In DCHMM, special parameters are used to directly characterize the coupling strengths. DCHMM replaces the joint conditional probability by a linear combination of marginals as in Eq. (4), and uses the combination weights $\theta_{c'c}$ to represent the coupling strengths between two objects, c' and c. This formulation was motivated by a Raytheon project involving signals emitted by interacting, mobile objects, wherein the degree of coupling depends on the (time varying) distance between two objects, with longer distances implying weaker coupling in some monotonic fashion. It retains the power of standard CHMM (capable of modeling interactions), using a much reduced parameter space.

The elements of DCHMM are the same as standard HMMs except we add one parameter—the coupling coefficients $\Theta = \{\theta_{c'c}\}$, with $1 \le c', c \le C$ and $\sum_{c'=1}^{C} \theta_{c'c} = 1$. Thus, the proposed DCHMM model can be characterized by a quadruplet $\lambda = (\pi, A, B, \Theta)$, where Θ is the new interaction parameter in the DCHMM formulation.

B. Forward-backward procedure

The forward-backward procedure is an essential part of the HMM inference problem. In DCHMM formulation, the exact forward-backward procedure needs exponentially large number of forward and backward variables. We have to resort to approximate inference when the number of coupled HMM chains is large.

For C HMMs coupled together, the extended forward and backward variables should be defined jointly across C HMMs as

$$\alpha_t(j_1, ..., j_C) = P(o_1, ..., o_t, S_{t,j_1}, ..., S_{t,j_C}|\lambda)$$

and

$$\beta_t(j_1, ..., j_C) = P(o_{t+1}, ..., o_T|S_{t,j_1}, ..., S_{t,j_C}, \lambda)$$

It is easy to check that both variables cannot be simply decoupled. The computational complexity for forward-backward procedure would be $O(TN^{2C})$, which is not practical. Note that the time complexity for a single-chain standard HMM is just $O(TN^2)$. Therefore we use a slightly modified forward variable that can be calculated in time $O(TCN^2)$ for each HMM chain. This modified forward variable is calculated inductively as follows:

1. Initialization :
   $\alpha_1^{(c)}(j) = \pi_j^{(c)} \cdot b_j^{(c)}(o_1^{(c)})$
2. Induction :
   $\alpha_t^{(c)}(j) = b_j^{(c)}(o_t) \sum_{c'} \theta_{c'c} \sum_i \left(\alpha_{t-1}^{(c')}(i) \cdot a_{ij}^{(c',c)}\right),\ t > 1$
3. Termination :
   $P(O|\lambda) = \prod_c \left(\sum_j \alpha_T^{(c)}(j)\right)$

This can be seen as a factored version of the exact forward procedure. Experiments show that $P(O|\lambda)$ calculated this way is close to true $P(O|\lambda)$ and the training algorithm based on this new forward variable produces reasonably good models. Later in the paper, we refer to *exact* inference as the exact forward procedure and *factored* inference as the modified forward procedure.

C. Learning a DCHMM

The EM algorithm can be derived for learning a DCHMM, as shown by Saul and Jordan [11]. But the algorithm finally amounts to using statistics that are similar to forward variables but need factored approximation. We take a different approach from EM and train DCHMM using an iterative algorithm based on a self-mapping transformation τ described by Baum [12]. The transformation is motivated by the optimality condition of standard Lagrange multiplier method and leads to an iterative reestimation procedure. Convergence of the procedure is guaranteed by the following theorem.

Theorem 1: [18] Let $\mathcal{P}$ be a homogeneous polynomial

$$\mathcal{P}(z_1, ..., z_n) = \sum_{\mu_1, \mu_2, ..., \mu_n} c_{\mu_1, \mu_2, ..., \mu_n} z_1^{\mu_1} z_2^{\mu_2} \cdots z_n^{\mu_n} \quad (5)$$

where $c_{\mu_1, \mu_2, ..., \mu_n} \ge 0$ and $\mu_1 + \cdots + \mu_n = d$. Then

$$\tau : z_i \to \frac{z_i \partial P/\partial z_i}{\sum_j z_j \partial \mathcal{P}/\partial z_j} \quad (6)$$

maps $D : z_i \ge 0, \sum z_i = 1$ into itself and satisfies $\mathcal{P}(\tau(z_i)) \ge \mathcal{P}(z_i)$. In fact, strict inequality holds unless z_i is a critical point of $\mathcal{P}$ in D.

We now derive the iterative optimization procedure for learning the parameters of DCHMM. For simplicity, we use P for $P(O|\lambda)$ and restrict the discussion to optimization of P with respect to A. Actually all parameters in $\lambda = (\pi, A, B, \Theta)$ are subject to similar stochastic constraints so the discussion with respect to A here can be easily duplicated for π, B and Θ. Let $\mathcal{L}$ be the Lagrangian of P with respect to the constraints associated with A, we have

$$\mathcal{L} = P + \sum_{i,c',c} \lambda_i^{(c',c)} \left(\sum_{j=1}^{N} a_{ij}^{(c',c)} - 1 \right) \tag{7}$$

where $\lambda_i^{(c',c)}$'s are undetermined Lagrange multipliers. It is easy to verify that P is locally maximized when

$$a_{ij}^{(c',c)} = \frac{a_{ij}^{(c',c)} \partial P / \partial a_{ij}^{(c',c)}}{\sum_k a_{ik}^{(c',c)} \partial P / \partial a_{ik}^{(c',c)}} \tag{8}$$

Similar arguments can be made for π, B and Θ parameters. The reestimation formula suggested by the above equation is exactly the transformation τ shown in Eq. (6).

The transformation τ can be applied to more general likelihood functions that are polynomials with positive coefficients (not necessarily homogeneous), according to a relaxation presented by Baum and Sell [19]. One advantage of this learning algorithm is that it can be applied to minimize more complex objective functions for which EM algorithm may be difficult to derive.

The difference from the standard HMM case for using this optimization algorithm is that we do not have a simple form of calculating the derivative of the likelihood function. While in the standard HMM training algorithm, these derivatives reduce to a form in which only the forward and backward variables are needed. In our case, we need to calculate the derivatives iteratively by rippling through time, just like the way we calculate forward variables. Fortunately, this does not add too much computational complexity since we can use similar forward procedures for the derivatives and only one pass through time is needed to calculate the forward variables and all the derivatives.

IV. EXPERIMENTAL RESULTS

A. EEG data

The real EEG data used was downloaded from UCI KDD archive website [20]. The data arose from a large study to examine EEG correlates of genetic predisposition to alcoholism. There are two groups of subjects in the study: alcoholic and control. Each subject is exposed to stimuli that are pictures of objects chosen from the 1980 Snodgrass and Vanderwart picture set. It contains measurements, sampled at 256 Hz for 1 second, from 64 electrodes placed on the scalp.

We extracted from the archive two EEG datasets that we call *EEG-1* dataset and *EEG-2* dataset, respectively. Each dataset contains measurements for two subjects—one alcoholic and one control subject. The *EEG-1* dataset contains 10 measurements from 2 electrodes (F4, P8) for each subject, i. e. there are two feature sequences (from electrodes F4 and P8) for each data sample. The data with scaled magnitude is plotted in Fig. 3. It can be seen that extracting discriminative features from the data is a highly nontrivial task. The *EEG-2* dataset contains 20 measurements from the same 2 electrodes for each subject, but 10 of them are from out-of-sample runs and thus chosen as test samples.

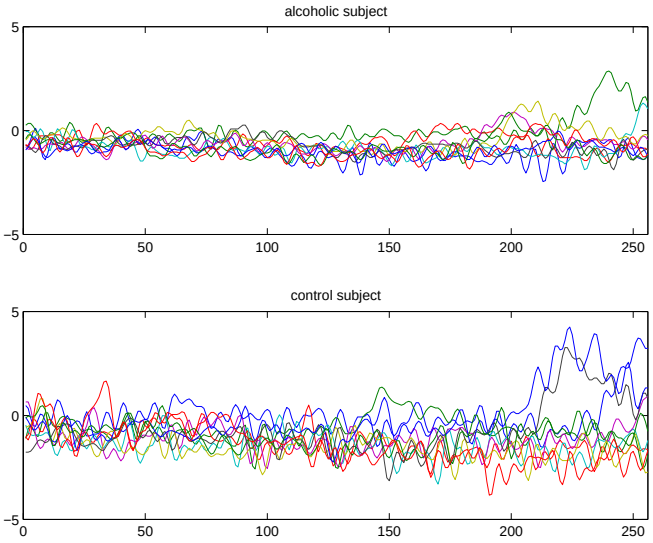

Fig. 3. EEG data samples for one alcoholic subject and one control subject.

Our goal is to recognize the correct subject class (normal or alcoholic) when presented a test sample that contains two feature sequences. Recognition accuracy is measured as the percentage of all test samples that have been recognized correctly.

EEG signals are believed to be highly correlated with the sleep stages of brain cells. The number of sleep stages is about 5 or 6 according to Geva and Kerem [21]. We set the number of hidden states to 5 when using HMMs to model the EEG data.

B. Experimental setting

We want to mention a few details of the training of HMMs here. Juang, Levinson, and Sondhi [22] point out that using mixture of gaussians as the observation model of HMM sometimes results in singularity problems during training. They suggest solving the problem by re-training from a different initialization. This is the method we use for dealing with the singularity problem. Rabiner *et al* [23] find, through empirical study, that accurately estimating the means of gaussians is critical to learning good models for continuous HMMs. In his experiment, Smyth [24] uses a clever initialization scheme—using the k-means algorithm to locate the means. We adopt the same strategy. In our experiments, we scale all EEG values to be within $[-5, 5]$ to avoid severe mismatches between the data and the initial (random) models.

For the *EEG-2* dataset, we have 10 training samples and 10 out-of-sample test samples for each subject. For the *EEG-1* dataset, we use 5-fold cross validation since we don't have out-of-sample test samples. That is, the full dataset is partitioned into five equal-sized sets and each set in turn serves as test data with the rest as training data. The accuracy of each experiment is averaged over the five sets. We repeat each experiment ten times for each model to get the average and standard deviation of classification accuracy across the ten experiments.

We compare ten models in total, for the EEG classification task. For each type of model, we train two models, one for each subject class. Then for any test sample, we fit it to the two trained models and classify it as the one with higher (log-)likelihood value. The ten models are:
HMM-1: univariate HMMs trained on feature 1 (F4) sequences only.
HMM-2: univariate HMMs trained on feature 2 (P8) sequences only.
Combined HMM: models combining HMM-1 and HMM-2. For each test sample, we sum up the log-likelihood value of fitting feature 1 sequence to HMM-1 and that of fitting feature 2 sequence to HMM-2 and recognize the sample based on the summed log-likelihood.
Multivariate HMM: bivariate HMMs trained on feature 1 and feature 2 sequence pairs.
FHMM-exact: bivariate FHMMs with exact inference.
FHMM-approximate: bivariate FHMMs with approximate inference (using a structured variational approximation).
DCHMM-exact: DCHMMs with exact inference.
DCHMM-factored: DCHMMs with approximate inference.
BK-exact: the exact Boyen-Koller algorithm applied to a two-chain CHMM structure.
BK-factored: the factored Boyen-Koller algorithm.
For FHMM models, we use the software package that can be downloaded from http://www.cs.toronto.edu/~zoubin/. For BK models, we use BNT toolbox that is created by Murphy and can be downloaded from his website http://www.cs.berkeley.edu/~murphyk/.

C. Results and discussions

The classification accuracy results on *EEG-1* dataset are shown in Table-I and the results on *EEG-2* dataset in Table-II. The first column lists the ten models. The middle column presents the average accuracies with 95% confidence intervals ($\pm 1.96\sigma$, σ is the standard deviation). The last column gives results of t-tests of each method compared to the best one (multivariate HMM). A p-value of less than 0.05 means that the difference between two methods is statistically significant. When a method is compared to itself, the p-value would be 1.0, as shown in the tables.

All models except HMM-2 achieve reasonably good recognition accuracies, indicating that the raw EEG data can be successfully modeled by HMMs. The accuracies in Table-II are generally lower than those in Table-I since out-of-sample test samples are used for the experiments on *EEG-2* dataset.

TABLE I
CLASSIFICATION RESULTS ON *EEG-1* DATASET

Model	Accuracy (%)	p-value
HMM-1	80.5 ± 10.8	< 0.0001
HMM-2	70.5 ± 10.8	< 0.0001
Combined HMM	87.0 ± 6.9	0.024
Multivariate HMM	$\mathbf{90.5 \pm 5.6}$	**1.0**
FHMM-exact	87.0 ± 6.9	0.024
FHMM-approximate	$\mathbf{88.0 \pm 9.5}$	**0.18**
DCHMM-exact	$\mathbf{90.0 \pm 0.0}$	**0.58**
DCHMM-factored	80.0 ± 14.6	0.0006
BK-exact	84.0 ± 11.1	0.005
BK-factored	84.5 ± 12.6	0.015

TABLE II
CLASSIFICATION RESULTS ON *EEG-2* DATASET

Model	Accuracy (%)	p-value
HMM-1	73.5 ± 6.6	0.008
HMM-2	61.5 ± 6.6	< 0.0001
Combined HMM	$\mathbf{77.0 \pm 8.3}$	**0.43**
Multivariate HMM	$\mathbf{78.5 \pm 8.0}$	**1.0**
FHMM-exact	72.5 ± 9.5	0.008
FHMM-approximate	73.5 ± 10.4	0.03
DCHMM-exact	72.0 ± 8.3	0.0026
DCHMM-factored	66.0 ± 7.7	< 0.0001
BK-exact	70.0 ± 20.7	0.029
BK-factored	70.0 ± 20.7	0.029

Of all the multi-channel models, the last six (FHMMs, DCHMMs and BK algorithms) are the more complex ones, with more complex structure and more parameters than combined HMM and multivariate HMM. Comparing Table-I and II, we can see that the relative performance of these more complex models (compared to simpler ones, combined HMM and multivariate HMM) drops from dataset *EEG-1* to *EEG-2*. We think this is because the simpler models generalize better so they work relatively better on the out-of-sample test data.

The multivariate HMM emerges as the best approach on both datasets, as highlighted in the tables. It outperforms all the other models and is significantly better than all but the highlighted methods. The reason may be that since in EEG experiments the multiple electrodes are placed on the same head, one state space is enough and the correlations can be accurately captured by the covariance matrix.

HMM-1 and HMM-2 are among the worst models because they only use one channel for classification, i. e. use less information than other methods do. HMM-1 performs much better than HMM-2 on both datasets, which suggest that feature 1 might be more informative (or dis-

criminative) than feature 2. The combined HMM performs fairly well, boosting the accuracy of single channel HMMs dramatically by a simple combining strategy.

DCHMM-exact has superior performance on *EEG-1* but it is computationally infeasible to model more channels. DCHMM-factored does not perform well, probably due to inaccurate approximation of the forward variables. One of our future interests is how to improve the approximation. The FHMM models are similar to multivariate HMM except they use more complex hidden state structure and more hidden states. Their performance is in the middle class. The performance of BK algorithms is not so good, probably because they are not designed specifically for HMM/CHMM structures.

V. CONCLUSION AND FUTURE WORK

We have compared several HMM and CHMM approaches, including a new CHMM formulation of our own—DCHMM, on a multi-channel EEG classification problem. Results show that the simple multivariate HMM is superior in classification accuracy and low in computational complexity. Since the multivariate HMM models the interdependence of two sequences by an observation covariance matrix, the results suggest that the interactions between two EEG channels can be well modeled in the observation space. Modeling them in the state space (CHMM approaches) does not necessarily translate to better results, due to increasing model complexity and additional associated assumptions.

Future work can proceed in several directions:
Modeling more channels. In the future we can apply the HMM and CHMM approaches to simultaneously model more channels of the EEG data.
More EEG data. We plan to experiment with larger EEG datasets, for which we can use part of the data as validation set for model selection and thus to probably increase classification accuracy.
Better approximate inference for CHMM. CHMMs with exact inference give good results but are computationally infeasible. More work has to be done to reduce the computational complexity while retaining the modeling power.

ACKNOWLEDGEMENTS

We thank Henri Begleiter for providing the EEG data, Ghahramani for providing the FHMM matlab code, and Murphy for providing the BNT software package. This work is partially supported by the NSF grant ECS9900353, and a gift from Intel.

REFERENCES

[1] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proceedings of IEEE*, vol. 89, no. 7, pp. 1123–1134, July 2001.

[2] Ernst Haselsteiner and Gert Pfurtscheller, "Using time-dependent neural networks for EEG classification," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 457–463, December 2000.

[3] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[4] Ruey S. Huang, Chung J. Kuo, Ling-Ling Tsai, and Oscal T.C. Chen, "EEG pattern recognition—arousal states detection and classification," in *Proceedings of IEEE International Conference on Neural Networks*, 1996, vol. 2, pp. 641–646.

[5] B. Obermaier, C. Guger, and G. Pfurtscheller, "HMM used for the offline classification of EEG data," *Biomedizinsche Technik*, June 1999.

[6] W. Penny and S. Roberts, "Gaussian observation hidden markov models for EEG analysis," Tech. Rep. TR-98-12, Imperial College, London, October 1998.

[7] Matthew Brand, "Coupled hidden Markov models for modeling interactive processes," Tech. Rep. 405, MIT Media Lab, 1997.

[8] I. Rezek and S. J. Roberts, "Estimation of coupled hidden Markov models with application to biosignal interaction modelling," in *Proc. IEEE Int. Conf. on Neural Network for Signal Processing*, 2000, vol. 2, pp. 804–813.

[9] Y. Bengio and P. Frasconi, "Input-Output HMMs for sequence processing," *IEEE Trans. Neural Networks*, vol. 7, no. 5, pp. 1231–1249, September 1996.

[10] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, pp. 245–275, 1997.

[11] Lawrence K. Saul and Michael I. Jordan, "Mixed memory Markov models: Decomposing complex stochastic processes as mixtures of simpler ones," *Machine Learning*, vol. 37, pp. 75–87, 1999.

[12] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1–8, 1969.

[13] Matthew Brand, Nuria Oliver, and Alex Pentland, "Coupled hidden Markov models for complex action recognition," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 1997, pp. 994–999.

[14] J. Kwon and K. Murphy, "Modeling freeway traffic with coupled HMMs," Tech. Rep., University of California at Berkeley, May 2000.

[15] T. T. Kristjansson, B. J. Frey, and T. Huang, "Event-coupled hidden Markov models," in *Proc. IEEE Int. Conf. on Multimedia and Exposition*, 2000, vol. 1, pp. 385–388.

[16] Xavier Boyen and Daphne Koller, "Approximate learning of dynamic models," in *Advances in Neural Information Processing Systems*, Denver, Colorado, 1998, Morgan Kaufmann.

[17] Kevin Murphy and Yair Weiss, "The factored frontier algorithm for approximate inference in DBNs," in *Proceedings of the 17th Conference on Uncertainty in Artifical Intelligence*, Seattle, Washington, August 2001.

[18] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology," *Bulletin AMS*, vol. 73, pp. 360–363, 1967.

[19] L. E. Baum and G. R. Sell, "Growth transformations for functions on manifolds," *Pacific Journal of Mathematics*, pp. 211–227, 1968.

[20] S. Hettish and S. D. Bay, "The UCI KDD archive [http://kdd.ics.uci.edu]," Irvine, CA: University of California, Department of Information and Computer Science, 1999.

[21] Amir B. Geva and Dan H. Kerem, "Brain state identification and forecasting of acute pathology using unsupervised fuzzy clustering of EEG temporal patterns," in *Fuzzy and Neuro-Fuzzy Systems in Medicine*, Horia-Nicolai Teodorescu, Abraham Kandel, and Lakhmi C. Jain, Eds., chapter 3, pp. 57–93. CRC Press, 1998.

[22] B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. Inform. Theory*, vol. 32, no. 2, pp. 307–309, 1986.

[23] L. R. Rabiner, B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Some properties of continuous hidden Markov model representations," *AT&T Technical Journal*, vol. 64, no. 6, pp. 1251–1269, 1985.

[24] Padhraic Smyth, "Clustering sequences with hidden Markov models," in *Advances in Neural Information Processing Systems*, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. 1997, vol. 9, pp. 648–654, MIT Press.