# Coupled Hidden Markov Model for Electrocorticographic Signal Classification

Rui Zhao
ECSE Department
Rensselaer Polytechnic Institute
Troy, NY, USA, 12180
zhaor@rpi.edu

Gerwin Schalk
Neural Injury and Repair Lab
Wadsworth Ctr, NYS Dept of Health
Albany, NY, USA, 12201
schalk@wadsworth.org

Qiang Ji
ECSE Department
Rensselaer Polytechnic Institute
Troy, NY, USA, 12180
jiq@rpi.edu

***Abstract*—This paper investigates the spatial and temporal dynamics in multi-channel electrocorticographic (ECoG) time series signals using Coupled Hidden Markov Model (CHMM). The signals are recorded in a hand motion control task, when the subject uses a joystick to move a cursor appearing on the screen to hit a virtual target. We detect signal onset using two heuristic schemes based on the experiment process. We apply CHMM to capture the spatial and temporal dynamics between two different channels within fixed length of duration, where each channel is modelled by HMM. The interdependence between two channels are modelled by transitions between hidden states of different individual HMM. There are eight possible directions that the target may appear. We learn eight sets of parameters using EM algorithm to characterize the signal patterns for each possible direction of movement. Given the test signals, the set of learned parameters which produces highest probability likelihood decides the class label. The effectiveness of the model is measured by classification accuracy. The results indicate that CHMM outperforms conventional HMM in most of the cases and is significantly better than first order autoregressive model.**

## I. Introduction

Brain computer interface (BCI) is a communication technique that enables people to interact with the outside world using brain signals without performing body movement. With better prediction of people's intention of movement, BCI can not only help improve the motor capability of people with handicap but also enhance the performance of normal people. The analysis of brain signals is crucial for BCI technique.

There are different sensor modalities of signal recordings used in BCI. (Figure 1). The most commonly used one is electroencephalography (EEG), which is recorded from the scalp. Since it is non-invasive, EEG is flexible and easy to obtain. However, it suffers from problem like low signal noise ratio. During the past several decades, an invasive modality electrocorticographic (ECoG) is gaining scientific interest in many animal studies. Due to the surgery requirement and health condition risk, up to date, the majority of human ECoG recordings are obtained from epilepsy patients who accepted craniotomy. For solely clinical purpose, electrodes are placed on the brain surface subdurally or epidurally in order to monitor and localize seizure foci. The implantation will be removed in periods of several days to 1-2 weeks. Despite the surgery risk, the major advantage of ECoG is the fine resolution in both space and time with substantially high signal noise ratio compared to EEG (Schalk and Leuthardt [1]).

One important task is to differentiate ECoG signals under different physical conditions or in response to different exterior stimulus. This can be treated as a time series classification problem. There are a variety of classifiers that can be applied to BCI signals. Lotte *et al.* [2] provided a comprehensive survey on existed approaches to classify EEG signals. Similar categorization can be applied to ECoG signals classification. Due to limited availability of data, not all the classification techniques are applied to ECoG signals. Depending on whether the temporal relation is considered during the classification, we have static classifier, e.g., support vector machines (SVM), and dynamic classifier, e.g., hidden Markov model (HMM). We choose dynamic classifier in order to model the temporal interactions among different channels of signals.

HMM has been studied intensively since 1970s and widely used in time series modelling and analysis especially in speech recognition and synthesis (Rabiner [3]). HMM characterizes the system evolution by introducing hidden state variable, which governs the transition of the system among different status. The observed quantity is considered as random sample drawn from certain pre-determined probabilistic distribution. Obermaier *et al.* [4] applied HMM to classify EEG signal on a motor imagery task. Zhong and Ghosh [5] compared several variants of coupled HMM for EEG data classification. Suk and Lee [6] constructed a two-layer HMM to differentiate EEG signals. However, the classification was limited to motor imagery task. Recently, ECoG signals are used for more sophisticated motor task classification such as finger and hand movement. Onaran *et al.* [7] employed a hybrid approach combining SVM and HMM to classify the movement of individual finger. Wang *et al.* [8] took a two step approach to decode the onset and moving direction, where the onset prediction is completed

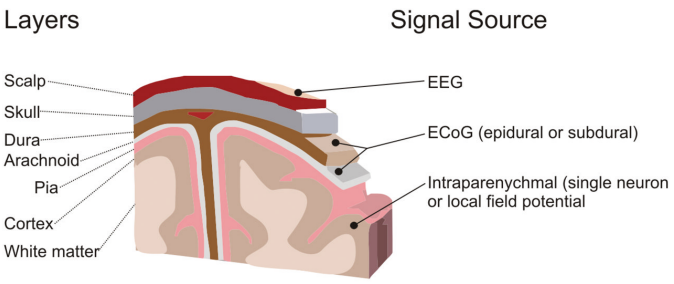


Fig. 1. Different sensor modalities used in BCI signal recording (Schalk and Leuthardt [1] with permission to use by author).

by SVM and multi-channel ECoG signals are modelled by STVDBN. But the model is complicated and requires an accurate onset detection. In this paper, we propose a model that keeps a balance between complexity and effectiveness. We focus on the classification of hand movement direction and heuristically detect onset by taking advantage of the controlled experiment process.

The main contribution of our work is applying the coupled HMM to the new application condition of ECoG signal classification on sophisticated hand movement. Experiment results suggest the superiority of proposed model to conventional HMM. The rest of paper is organized as follows. In section 2, we introduce the dynamic models and methods. Then we describe the experiment process and discuss the results in section 3. The last section summarizes the work and points out future direction.

## II. Models and Methods

### A. Hidden Markov Model

HMM is a probabilistic graphical model that describes stochastic evolution of a set of random variables over time. Figure 2 is an illustration of the graph structure. The shaded nodes represent observed quantity, which can be multivariate random variables. Each observed node has one and only one parent node, which is hidden, from the same time frame. We use $X^t$ to represent discrete latent variables and $Y^t$ for continuous observed variables for the rest of paper. Intuitively,

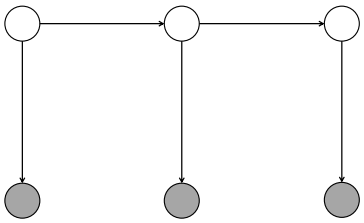

Fig. 2. Graphical topology of standard HMM. Shaded nodes are observed and unshaded nodes are hidden.

the dynamics of observed quantity are characterized by the transition among the virtual states across time. There are two basic assumptions for HMM. First, the transition among hidden nodes forms a first order Markov process. Second, all the conditional probability distribution (CPD) are time-invariant. Based on the assumptions, the joint probability distribution of length $T$ data $\mathbf{Y} = \{Y^1, ..., Y^T\}$ and a set of realization of hidden nodes $\mathbf{X} = \{X^1, ..., X^T\}$ can be written as

$$P(\mathbf{X}, \mathbf{Y}) = P(X^1)\prod_{t=2}^{T} P(X^t|X^{t-1})\prod_{t=1}^{T} P(Y^t|X^t) \quad (1)$$

To fully determine the joint distribution (1), we only need to define $P(X^1), P(X^t|X^{t-1})$ and $P(Y^t|X^t)$. In this paper, we parametrize the three CPDs as follows.

1) Initial distribution: $\pi = \{\pi_i\}$.

$$P(X^1 = i) = \pi_i, 1 \le i \le N \quad (2)$$

where $N$ is the number of hidden states. To make a valid probability distribution, we constrain $\sum_{i=1}^{N} \pi_i = 1, \pi_i \ge 0$

2) Transition probability distribution: $A = \{a_i^j\}$

$$P(X^{t+1} = j|X^t = i) = a_i^j, 1 \le i, j \le N \quad (3)$$

Similarly, to make a valid probability distribution, we constrain $\sum_{j=1}^{N} a_i^j = 1, \forall i, a_i^j \ge 0$.

3) Emission probability distribution: $B = \{\mu_i, \Lambda_i\}$

$$P(Y^t|X^t = i) = \mathcal{N}(\mu_i, \Lambda_i),\ 1 \le i \le N \quad (4)$$

where $\mu_i$ and $\Lambda_i$ are the mean and covariance matrix of Gaussian distribution.

### B. Coupled Hidden Markov Model

CHMM is an extension to conventional HMM. It has been applied to speech recognition (Nefian *et al.* [9]) and activity recognition (Brand *et al.* [10]), where different information sources or processes interact with each other. In our application, we are interested in modelling the interaction between multiple channels of signals. We aggregate multiple HMMs together by allowing transition between hidden nodes from each HMM. Figure 3 is the graph structure of CHMM with two channels. CHMM is more expressive than HMM

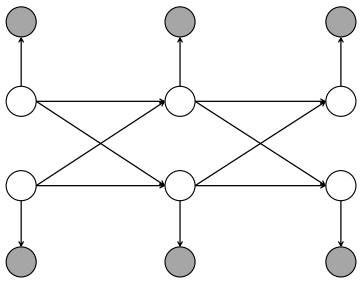

Fig. 3. Graphical topology of two-channel CHMM. Shaded nodes are observed and unshaded nodes are hidden.

because it allows different number of states to be applied for each individual HMM. The parametrization of CHMM is a natural extension from HMM. Let $\mathbf{Y}_c = \{Y_c^1, ..., Y_c^T\},\ \mathbf{X}_c = \{X_c^1, ..., X_c^T\},\ c = 1, ..., C$ be the $c^{\text{th}}$ HMM with $C$ channels in total. The way we defining the CPDs is similar to HMM.

1) Initial distribution:

$$P(X_c^1 = n_c) = \pi_{n_c}, 1 \le n_c \le N_c \quad (5)$$

where $N_c$ is the number of hidden states of $c^{\text{th}}$ HMM and $\sum_{n_c=1}^{N_c} \pi_{n_c} = 1, \pi_{n_c} \ge 0$.

2) Transition distribution: each hidden node will have $C$ parent nodes from previous time frame. We define

$$P(X_k^{t+1} = n_k|X_1^t = n_1, ..., X_C^t = n_C) = a_{n_1...n_C}^{n_k} \ge 0 \quad (6)$$

where $\sum_{n_k=1}^{N_k} a_{n_1...n_C}^{n_k} = 1$.

3) Emission distribution: we use unimodal Gaussian distribution, which is the same as HMM.

$$P(Y_c^t|X_c^t = n_c) = \mathcal{N}(\mu_{n_c}, \Lambda_{n_c}) \quad (7)$$

where $1 \le n_c \le N_c$, $\mu_{n_c}$ and $\Lambda_{n_c}$ are the mean and covariance matrix.

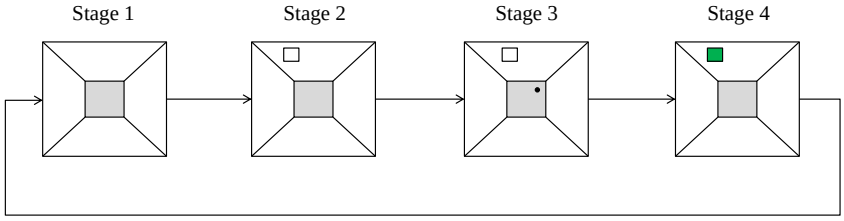


Fig. 4. Timeline of single experiment trial. The big square box is a visualization of the screen. Stage 1 is inter-trial period when nothing is shown on the screen. It lasts 1 second. Stage 2 is pre-trial pause period when there is a box appearing on the screen. The subject is supposed to pay attention but not move joystick yet. Stage 3 is trial period when a cursor appears on the screen and the subject can move joystick immediately. The maximum duration of trial period is 2 seconds. Stage 4 is reward period. The box will be highlighted when target is hit correctly.

### C. Autoregressive Model

For comparison purpose, we also implement the autoregressive (AR) model, which is widely used for time series analysis. AR model assumes that the random process is stationary and current observation can be expressed as linear weighted sum of previous observations. In our experiment, we use first order AR model to characterize the signals. To be specific, let $Y^t \in \mathbb{R}^d$ for all discrete time step $t$, then

$$Y^{t+1} = AY^t + \epsilon,\ \epsilon \sim \mathcal{N}(0, I) \tag{8}$$

where $A \in \mathbb{R}^{d\times d}$ is the regression matrix and $\epsilon$ is Gaussian white noise with identity covariance matrix $I$.

### D. Learning and Inference

The goal of learning in HMM and CHMM is the parameters associated with the model. The algorithms we used are based on expectation maximization (EM). For HMM, the E-step computes the expectation of sufficient statistics of parameters conditioned on observation. Then M-step computes the maximum likelihood estimate of parameters using the conditional expectation obtained in E-step. The algorithm iteratively updates parameters with guarantee of increasing likelihood at each iteration. The same algorithm can be extended to CHMM by modifying the condition expectation computed in E-step.
The computation of E-step requires the posterior distribution of a hidden state $X^t$ given the observation of $\mathbf{Y}$. The inference is completed efficiently using forward-backward algorithm (Rabiner [3]), which takes the advantage of the chain structure. The computation complexity is linear to the length of sequence. To be specific, we define

$$\alpha_t(i) \triangleq P(Y^1, ..., Y^{t-1}, X^t = i)P(Y^t|X^t = i)$$
$$\beta_t(i) \triangleq P(Y^{t+1}, ...Y^T|X^t = i)$$

Given the model structure shown in Figure 2. We have the following recursion.

$$\alpha_t(i) = P(Y^t|X^t = i)\sum_j P(X^t = i|X^{t-1} = j)\alpha_{t-1}(j) \tag{9}$$

$$\beta_{t-1}(i) = \sum_j P(Y^t|X^t = j)P(X^t = j|X^{t-1} = i)\beta_t(j) \tag{10}$$

with initialization $\alpha_1(i) = P(Y^1|X^1 = i)P(X^1 = i),\ \beta_T(i) = 1,\ \forall i \in \mathcal{I}$, where $T$ is maximum time step and $\mathcal{I}$ is some discrete index set. Therefore we can compute $\alpha, \beta$ in two rounds of recursion and then use them to compute the posterior distribution required in E-step.
For classification purpose, we need to compute likelihood of observation, which is given by $P(\mathbf{Y}) = \sum_{X^T} \alpha(X^T)$. Therefore during testing, given the learned parameters, we only need to run the forward recursion (9) once for each testing sequence.
For AR, the learning and inference is integrated as one task which is estimating regression matrix $A$. We compute $A$ by minimizing the mean square error between actual value and regression value of data.

$$\hat{A}_{i*} = \arg\min_{A_{i*}} \sum_{t=2}^{T}\sum_{m=1}^{M}(y_i^{m,t} - A_{i*}Y^{m,t-1})^2 + \lambda||A_{i*}||^2 \tag{11}$$

where $A_{i*}$ is the $i^{\text{th}}$ row of $A$, $M$ is the number of training sequences from one class, $y_i^{m,t}$ is the $i^{\text{th}}$ dimension of $m^{\text{th}}$ sequence at time step $t$. In general, each sequence does not have to be the same length, i.e. $T$ can be different for each sequence. $\lambda$ is regularization coefficient which can be selected by validation during training.

## III. Experiment and Result

### A. Data collection and preprocessing

*1) Subjects:* Four subjects participated in this study. They were patients with electrodes placed subdurally on the surface of brain for solely clinical purpose of identifying epilepsy seizure foci prior to surgical resection. All the subjects had normal cognitive capability and were functionally independently. They all gave informed consent. The study was approved by the Institutional Review Board of Albany Medical College as well as by the Human Research Protections Office of the US Army Medical Research and Materiel Command.
There were two male subjects and two female subjects. The number of electrodes placed on each subject varies from 96 to 112. During the motion control experiment, the subject held

TABLE I. Subjects profile and number of segmented sequences used in training and testing

| Subject | Gender | Age | Training Sequences | Testing Sequences |
|---|---|---|---|---|
| A | Female | 29 | 206 | 203 |
| B | Male | 25 | 93 | 91 |
| C | Male | 25 | 186 | 180 |
| D | Female | 49 | 95 | 89 |

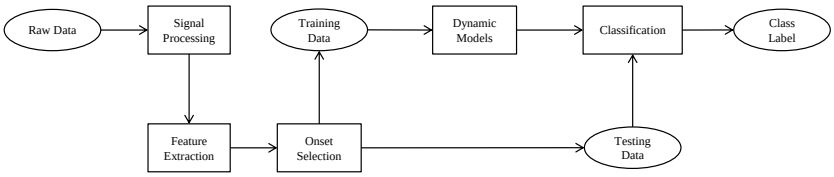


Fig. 5. General process of the experiment. Ellipse blocks are data items and square blocks are processing units.

a joystick to move a cursor appearing on the screen to hit virtual target. For each trial, the subject went through different stages from rest to moving according to the timeline specified by BCI2000 CursorTask protocol (Schalk and Mellinger [11]). Figure 4 illustrates the timeline of one trial.

*2) Signal processing:* The ECoG signal is sampled at 1200Hz for each channel. We first exclude channels with significant line noise and then apply common average filter to all channels. We then apply notch filter to further reduce harmonics of line noise. For feature extraction, we apply spectrum filter followed by Hilbert transform to extract amplitude of representative frequency band. In our experiment, we use high gamma band of range 70-170 Hz, which has been demonstrated with significant correlation to motor activity (Schalk [12]). Finally, we downsample the signal to 200 Hz.

### B. Channel selection and segmentation

Among all the electrodes, only a few of them covered the motor cortex area. Although we can blindly feed all the channels of signal into proposed dynamic models, the channels that recorded signal with irrelevant neural activities will be of little use. In our experiment, we select channels from the ones that are covering or close to the motor cortex area. For each channel we compare the average amplitude change between motion period and rest period. Finally, we select top two channels with the largest average amplitude change over all trials. Figure 6 provides two segments of feature points in selected channel for the same subject.
After choosing channels, we segment a fixed length sequence

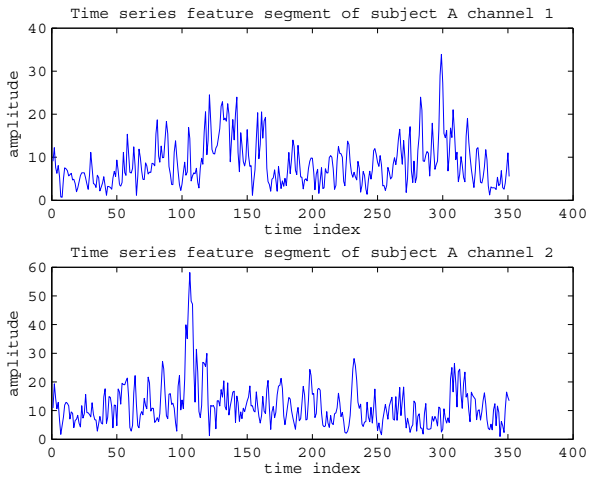


Fig. 6. Feature segment from two selected channels of subject A.

of 1 second duration, which is typical duration of joystick movement in our study. This gives us 200 data points for each sequence. For each subject, we choose half of the samples for training and the rest half for testing. Table I lists the number of training and testing samples we used in the experiment for each subject. Since the subject performed multiple trials in a consecutive manner, we need to determine which part of the signals correspond to the actual neural activity of motion control. The detection of exact signal onset is still open research problem. In our experiment, we heuristically determine the onset location by taking advantage of the experiment protocol. Considering the variation of response time for each subject in each single trial, we compare two schemes of onset selection. First, we decide onset as soon as the virtual target appears on the screen namely the beginning of stage 3 in Figure 4. Second, we decide onset as soon as the joystick position has been changed, which is recorded together with ECoG signals.

### C. Model training and classification

There are eight classes corresponding to eight possible directions of movement to hit the targets. For HMM and CHMM, we learn one set of parameters for each class during training, yielding eight models. During testing, for the same test sample, we compute the likelihood using each of the eight sets of parameters. Then the class label is determined the by the following criteria.

$$\hat{k} = \arg\max_k \log P(\mathbf{Y}|\theta_k) \tag{12}$$

where $k$ is the class label and $\theta_k$ is the set of parameters associated with class $k$. Although we only use high gamma band amplitude as feature for each channel, the model is completely general and can be applied to higher dimensional features. Figure 5 summarizes the overall experiment flow.
Since we compute maximum likelihood estimate of parameters iteratively using EM algorithm, which only converges to local optimum. Choosing initial parameters is crucial. We applied K-means algorithm to pre-cluster data at each time step, where the K value is chosen to maximize average classification accuracy in a 5-fold validation process. Given the cluster result, we initialize all the parameters as follows. 1) Initial distribution is initialized as the number counts of data points from the same cluster. 2) Transition distribution is initialized using the number counts of state transition between neighbouring data points. 3) Observed node mean and covariance are initialized using sample mean and sample covariance of data points from the same cluster respectively. The model implementation uses BNT toolbox (Murphy et al. [13]).
For AR, we compute eight regression matrices using training samples. During testing, we compute square error using the

TABLE II. CLASSIFICATION ACCURACY ON TEST SET FOR EACH SUBJECT

| Subject | A | | B | | C | | D | |
|---|---|---|---|---|---|---|---|---|
| Onset | Stimulus | Joystick | Stimulus | Joystick | Stimulus | Joystick | Stimulus | Joystick |
| CHMM | 0.1872 | **0.2167** | **0.2308** | 0.1868 | **0.2111** | **0.2278** | **0.2135** | **0.2360** |
| HMM | **0.1921** | 0.1872 | 0.1978 | **0.1978** | 0.1778 | 0.2111 | 0.1461 | 0.1685 |
| AR | 0.1478 | 0.1379 | 0.1099 | 0.1209 | 0.1667 | 0.1500 | 0.2022 | **0.2360** |

eight matrices and choose the one with least value.

$$\hat{k} = \arg\min_{k} \sum_{t=2}^{T} ||Y^t - A_k Y^{t-1}||^2 \tag{13}$$

where $A_k$ is the regression matrix of class $k$.
The number of hidden states varies from two to six. We reported result with highest classification accuracy on test set. The regularization parameter $\lambda$ is chosen by 5-fold validation during training between 0-10 with a granularity of 0.1.

### D. Results and Discussion

The experiments compare two different temporal alignment schemes and three models on four different subjects. Considering the physical difference between individuals, we perform classification for each subject individually. The classification accuracy on test set are listed in Table 2. The chance level for classification is 0.125. Given the same temporal alignment scheme, CHMM performs better than HMM for subject C and D. The average relative improvement for subject C and D are $14\%$ and $43\%$ respectively. For subject A, CHMM is better than HMM using joystick onset alignment by $16\%$, though CHMM is worse than HMM using stimulus onset alignment by $3\%$. Subject B has the opposite result to subject A. CHMM improves classification rate by $17\%$ with stimulus onset alignment and deteriorate by $6\%$ with joystick onset alignment. Overall, CHMM outperforms HMM most of the case. Even if not the case, the deterioration is less significant than the improvement.
In addition, for all subjects CHMM performs better than AR, which even produces results below chance level. This indicates that first order AR model is not sufficient to describe the temporal transition among multiple channels of ECoG signals. Comparing two different onset alignment schemes, joystick onset are in general more discriminative in determining the direction of movement. One possible explanation is we only use channels that cover motor cortex area, which mainly affect subject's motor activity. According to the experiment timeline, after the target appears on the screen, there will be one second planning time before the cursor appears. During this period of time, the subject's motor cortex area may or may not be involved in the response to visual stimulus. On the other hand, with joystick onset, it is more likely that motor cortex area becomes active. Therefore, for CHMM and HMM, joystick onset alignment tends to produce higher classification rate.

## IV. CONCLUSION AND FUTURE RESEARCH

In this paper, we employed CHMM on multi-channel ECoG signals to explicitly capture the temporal interactions among brain signals during multi-direction hand movement task. We choose two signal alignment schemes based on the experiment process. By comparing CHMM with conventional HMM and AR model, we demonstrate that CHMM is more expressive in characterizing the temporal dynamics of ECoG signals. The relative improvement of classification accuracy is noticeable. In the future, we plan to include more channels and multiple frequency bands as features and extend experiment to more subjects. We are also interested in designing models robust to the temporal variation between different trials.

## ACKNOWLEDGMENT

The authors would like to thank Peter Brunner for his help on signal processing and the anonymous reviewers for their helpful suggestions. This research is supported in part by the grant W911NF-08-1-0216 from the US Army Research Office.

## REFERENCES

[1] G. Schalk and E. C. Leuthardt, "Brain-computer interfaces using electrocorticographic signals," *Biomedical Engineering, IEEE Reviews in*, vol. 4, pp. 140–154, 2011.

[2] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, B. Arnaldi *et al.*, "A review of classification algorithms for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 4, 2007.

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

[4] B. Obermaier, C. Guger, C. Neuper, and G. Pfurtscheller, "Hidden markov models for online classification of single trial eeg data," *Pattern recognition letters*, vol. 22, no. 12, pp. 1299–1309, 2001.

[5] S. Zhong and J. Ghosh, "Hmms and coupled hmms for multi-channel eeg classification," in *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*, vol. 2. IEEE, 2002, pp. 1154–1159.

[6] H.-I. Suk and S.-W. Lee, "Two-layer hidden markov models for multi-class motor imagery classification," in *Brain Decoding: Pattern Recognition Challenges in Neuroimaging (WBD), 2010 First Workshop on*. IEEE, 2010, pp. 5–8.

[7] I. Onaran, N. F. Ince, A. E. Cetin, and A. Abosch, "A hybrid svm/hmm based system for the state detection of individual finger movements from multichannel ecog signals," in *Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference on*. IEEE, 2011, pp. 457–460.

[8] Z. Wang, A. Gunduz, P. Brunner, A. L. Ritaccio, Q. Ji, and G. Schalk, "Decoding onset and direction of movements using electrocorticographic (ecog) signals in humans," *Frontiers in Neuroengineering*, vol. 5, 2012.

[9] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled hmm for audio-visual speech recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 2. IEEE, 2002, pp. II–2013.

[10] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*. IEEE, 1997, pp. 994–999.

[11] G. Schalk and J. Mellinger, *Human-Computer Interaction: Practical Guide to Brain-Computer Interfacing with Bci2000: General-Purpose Software for Brain-Computer Interface Research, Data Acquisition, Stimulus Presentation, and Brain Monitoring*. Springer, 2010.

[12] G. Schalk, K. Miller, N. Anderson, J. Wilson, M. Smyth, J. Ojemann, D. Moran, J. Wolpaw, and E. Leuthardt, "Two-dimensional movement control using electrocorticographic signals in humans," *Journal of neural engineering*, vol. 5, no. 1, p. 75, 2008.

[13] K. Murphy *et al.*, "The bayes net toolbox for matlab," *Computing science and statistics*, vol. 33, no. 2, pp. 1024–1034, 2001.