SABANCI UNIVERSITY

Probabilistic Graphical Models for Brain Computer Interfaces

by

Jaime F. Delgado Saa

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Electronics Engineering

Sabanci University

February 2014

"Further conceive, I beg, that a stone, while continuing in motion, should be capable of thinking and knowing, that it is endeavoring, as far as it can, to continue to move. Such a stone, being conscious merely of its own endeavor and not at all indifferent, would believe itself to be completely free, and would think that it continued in motion solely because of its own wish. This is that human freedom, which all boast that they possess, and which consists solely in the fact, that men are conscious of their own desire, but are ignorant of the causes whereby that desire has been determined."

Baruch Spinoza

PROBABILISTIC GRAPHICAL MODELS FOR BRAIN COMPUTER INTERFACES

by Jaime F. Delgado Saa
Electronics Engineering, Ph.D. thesis, 2014

Thesis supervisor: Assoc. Prof. Müjdat Çetin

Keywords: Brain rhythms, brain computer interfaces, probabilistic graphical models, time-frequency representations, linear classifiers, event related potentials, sensorimotor rhythms, electroencephalogram, electrocorticogram.

ABSTRACT

Brain computer interfaces (BCI) are systems that aim to establish a new communication path for subjects who suffer from motor disabilities, allowing interaction with the environment through computer systems. BCIs make use of a diverse group of physiological phenomena recorded using electrodes placed on the scalp (Electroencephalography, EEG) or electrodes placed directly over the brain cortex (Electrocorticography, ECoG). One commonly used phenomenon is the activity observed in specific areas of the brain in response to external events, called Event Related Potentials (ERP). Among those, a type of response called P300 is the most used phenomenon. The P300 has found application in spellers that make use of the brain's response to the presentation of a sequence of visual stimuli. Another commonly used phenomenon is the synchronization or de-synchronization of brain rhythms during the execution or imagination of a motor task, which can be used to differentiate between two or more subject intentions. In the most basic scenario, a BCI system calculates the differences in the power of the EEG rhythms during execution of different tasks. Based on those differences, the BCI decides which task has been executed (e.g., motor imagination of left or right hand). Current approaches are mainly based on machine learning techniques that learn the distribution of the power values of the brain signals for each of the possible classes.

In this thesis, making use of EEG and ECoG recording methods, we propose the use of probabilistic graphical models for brain computer interfaces. In the case of ERPs, in particular P300-based spellers, we propose the incorporation of language models at the level of words to increase significantly the performance of the spelling system. The proposed framework allows also the incorporation of different methods that take into account language models based on n-grams, all of this in an integrated structure whose parameters can be efficiently learned. In the context of execution or imagination of motor tasks, we propose techniques that take into account the temporal structure of

the signals. Stochastic processes that model temporal dynamics of the brain signals in different frequency bands such as non-parametric Bayesian hidden Markov models are proposed in order to solve the problem of selection of the number of brain states during the execution of motor tasks as well as the selection of the number of components used to model the distribution of the brain signals. Following up on the same line of thought, hidden conditional random fields are proposed for classification of synchronous motor tasks. The combination of hidden states with the discriminative power of conditional random fields is shown to increase the classification performance of imaginary motor movements. In the context of asynchronous BCIs, we propose a method based on latent dynamic conditional random fields that is capable of modeling the internal temporal dynamics related to the generation of the brain signals, and external brain dynamics related to the execution of different mental tasks. Finally, in the context of asynchronous BCIs a model based on discriminative graphical models is presented for continuous classification of finger movements from ECoG data. We show that the incorporation of temporal dynamics of the brain signals in the classification stages increases significantly the classification accuracy of different mental states which can lead to a more effective interaction between the subject and the environment.

PROBABILISTIC GRAPHICAL MODELS FOR BRAIN COMPUTER INTERFACES

by Jaime F. Delgado Saa
Electronics Engineering, Ph.D. thesis, 2014

Thesis supervisor: Assoc. Prof. Müjdat Çetin

Anahtar Kelimeler: Beyin ritimleri, beyin bilgisayar arayüzleri, olasılıksal grafiksel modeller, zaman-frekans gösterimleri, doğrusal sınıflandırıcılar, olaya ilişkin potansiyeller, duyu-motor ritimleri, elektroensefalografi, elektrokortikografi

ABSTRACT

Beyin bilgisayar arayüzleri (BBA), motor hareketi yeteneğini kaybetmiş kişiler için yeni bir iletişim yolu kurmayı amaçlayan ve bu kişilerin bilgisayar sistemleri üzerinden çevreyle iletişim kurmalarına olanak sağlayan sistemlerdir. BBAlar kafa derisi üzerine takılan (elektroensefalografi, EEG) veya direk olarak beyin korteksine yerleştirilen elektrotlar (elektrokortikografi, ECoG) ile kaydedilen çeşitli fiziksel olgu gruplarını kullanırlar. En yaygın kullanılan olgu, beynin özel bölümlerinde dış olaylara cevap olarak oluşan Olaya İlişkin Potansiyeller (OİP) aktivitesidir. Bunlar içinde, bir cevap tipi olan P300 en çok kullanılan olgudur. P300ün kullanım alanı bir görsel uyaran dizisinin sunulmasına göre beyinde oluşan cevabı kullanan heceleyici uygulamalarıdır. Bir başka sıkça kullanılan olgu ise, kişinin iki veya daha çok sayıdaki isteğini ayırt etmek için kullanılabilecek, hayali motor hareketlerinin gerçekleştirilmesi sırasında oluşan beyin ritimlerinin eş zamanlaması veya eş zamanlama bozulumudur. En basit senaryoda, bir BBA sistemi farklı görevlerin yapılması sırasındaki EEG ritimlerindeki güç farklılıklarını hesaplar. Bu farklılıklara göre, BBA hangi görevin (ör. sağ veya sol el hayali motor hareketi) yapıldığına karar verir. Güncel yaklaşımlar, her olası sınıf için beyin sinyallerindeki güç değerlerinin dağılımını öğrenen makine öğrenmesi tekniklerine dayanmaktadır.

Bu tezde, EEG ve ECoG kayıt yöntemleri kullanılarak beyin bilgisayar arayüzleri için olasılıksal grafiksel model kullanımını öneriyoruz. OİP sırasında, özellikle P300 tabanlı heceleyicilerde, sistemin performansını belirgin olarak arttırmak için dil modellerini kelime seviyesinde birleştirmeyi öneriyoruz. Önerdiğimiz sistem, parametreleri etkin bir şekilde öğrenilebilen bütünlenmiş bir yapı içinde n-gram tabanlı dil modellerini hesaba katan farklı yöntemlerin de birleştirilmesine izin veriyor. Hayali veya gerçek motor hareketi görevinin gerçekleştirilmesi bağlamında, sinyalin zamansal yapısını dikkate alan teknikler öneriyoruz. parametresiz Bayes saklı Markov modelleri gibi farklı frekans bantlarındaki beyin sinyallerinin zamansal dinamiklerini modelleyen stokastik işlemler,

motor görevlerinin gerçekleştirilmesi sırasındaki beyin durumlarının sayısının ve beyin sinyallerinin dağılımının modellenmesinde kullanılan öğe sayısının seçimi sorununu çözmek için sunuluyor. Aynı düşünce şekliyle, eş zamanlamalı motor görevlerinin sınıflandırılması için saklı şartlı rastgele alanlar öneriliyor. Saklı durumlar ile şartlı rastgele alanlarının ayrımsal gücünün birleşiminin, hayali motor hareketlerinde sınıflandırıcı performansını arttırdığı görülüyor. Eş zamanlı olmayan BBAlar bağlamında, beyin sinyallerinin üretimi ile bağlantılı içsel zamansal dinamiklerini ve farklı ansal görevlerin gerçekleştirilmesine bağlı olarak dışsal beyin dinamiklerini modelleme yeteneğine sahip gizli dinamik koşullu rastgele alanlar tabanlı bir yöntem öneriyoruz. Son olarak, eş zamanlı olmayan BBAlar bağlamında, ECoG verisinden parmak hareketlerinin devamlı olarak sınıflandırılması için ayrımsal bir grafiksel model sunuluyor. Sınıflandırma aşamalarında beyin sinyallerinin zamansal dinamiklerinin birleşiminin, farklı ansal durumların sınıflandırılma performansını belirgin bir şekille arttırarak kişi-çevre etkileşiminin daha etkin olmasının sağlanabileceğini gösteriyoruz.

Acknowledgements

[1] Although the emotions that bring the finalization of the PhD could may me bias towards saying that everything during the last four years has been nothing but sunny days and smiles, I must say that it was not easy and not nice always. But yet, here I am, which suggests that in the middle of everything the good things were more than the bad things, and that was indeed the case. I had the chance to work in something that I am deeply interested so I had a lot of fun. It was enjoyable and constructive but this would not be possible without the collaboration and support of a group of people to whom I would like to dedicate the following lines.

My first words of gratitude go to my supervisor Müjdat Çetin who guided me during all this process, providing vision, inspiration and encouraging me. His insights were of great importance to give form to this thesis. Müjdat, besides being an excellent human being also possess a strong academic knowledge and a great capacity to see through the problems with clarity and more importantly with scientific strictness and honesty. I had also the opportunity to enjoy many discussions with him. In most cases we agreed, in a few others we did not but there was something to learn from him at the end of each discussion. For all this I thank Müjdat and I am looking forward to continue working with him.

Also, I would like to thank to Prof Hakan Erdoğan from Sabanci University for the innumerable discussions and insights on probabilistic graphical models. I also thank to Prof Hanks Frenk from Sabanci University and his wife Jikke Frenk for the long discussions over coffee and for their advices. Thanks to Professor Berrin Yanikoğlu from Sabanci University and Professor Zumray Dokur from Istanbul Technical University for their particiapation in the Thesis committee, their suggestions helped to improve the final form of this thesis.

Thanks to Jonathan Wolpaw at the Neural Injury and repair Laboratory in Wadsworth Center for giving me the opportunity to be the part of his prestigious lab. Thanks to Dennis McFarland, also from Wadsworth Center, for the long discussions about everything. It was a period during which I learn a lot. Our conversations included topics from linear regression and colorful plots to split brain experiments. Thanks to Gerwin Schalk, the director of Schalk Lab, for providing me the opportunity to join his prestigious lab, his encouragement and the interesting discussions, also for enabling access to valuable data without which the final chapter of this thesis would be incomplete.

Thanks to Adriana de Pesters, researcher at Schalk Lab in Wadsworth Center for the many interesting academic and philosophical discussions. Her unmatched passion for her research, the desire for understanding and very particular views on life catch my

[1] This work was partially supported by the Scientific and Technological Research Council of Turkey under Grant 111E056 and by Sabanc University under Grant IACF-11-00889.

attention and interest, making her subject of my admiration. I thank Adriana for the valuable suggestions and corrections which were very important to bring this document to its final form.

Thanks to Hugo Gmez, his wife Assel Saparova and their daughter Camila Gmez whom during my first years of PhD in Istanbul, far from home (more than 10000 Km) in a culture very different from the one I was raised in, made me part of their family. Hugo was also a source of fresh ideas, active part of many discussions and an example to be followed as scientist and human being. In my mind, I have also the memories of the nights when Hugo used to play guitar in the lab while I intended to follow hime with the maracas to the rhythm of "Chan Chan" from Buenavista Social Club.

To my family for the unconditional support and particularly to my father Carlos Emilio Delagdo Angulo for the long conversations over the phone, his rational insight on everything, for the good advice, for sharing his wisdom with me, for being constant, for the encouragement, for his friendship, and his support during all these years.

Thanks to all my friends for sharing many experiences and good moments enriched by our cultural differences. In particular, I would like to thank to Mireia Pérez, Marta López and Markéta Bílská for all the good memories in Istanbul, to Atia Shafique for her friendship, support and patience in difficult times, to Saygin Topkaya and Umut Sen from VPA laboratory for their help, to Lacides Ripoll, Oscar Serrano and Juan Carlos Villamizar for their support despite of the long distance and to Pandian Chelliah and Rupak Roy for the interesting conversations. Last but not least, thanks to the Universidad del Norte in Colombia for its support and in particular, to Beatriz de Torres for being an active part of this process.

Agradecimientos

[2] Aunque la emoción que trae la terminación del Doctorado me inclina a decir que cada día fue como un día de verano lleno de sonrisas y alegría, la verdad es que no fue así. Sin embargo, aquí estoy, lo que sugiere que en medio de todo, las cosas buenas superan a las no tan buenas, que es de hecho el caso. Durante el doctorado hice lo que quería hacer, el tema de esta tesis es de mi interés personal, así que me divertí bastante. Sin embargo, los resultados obtenidos no habrían sido posibles sin la colaboración de muchas personas a quienes les dedico las siguientes líneas.

Mis primeras palabras de gratitud van dirigidas a mi director Müjdat Çetin, quien me guió durante todo este proceso, proporcionando visión, inspiración y soporte. Su colaboración y consejo fueron primordiales en el proceso de dar forma a esta tesis. Müjdat es una persona con un amplio conocimiento, con capacidad de ver a través de los problemas con claridad, con rigor científico y con honestidad. Tuve la oportunidad de discutir sobre muchos temas con él, en muchos casos estuvimos de acuerdo en otros tantos no, pero en todo caso las conversaciones siempre me dejaron algo que aprender. Por todo esto un sincero agradecimiento a él,

También quiero agradecer al profesor Hakan Erdoğan de Sabanci Universitesi por las innumerables discusiones e ideas en el tema de modelos probabilísticos gráficos. Agradezco también al profesor Hanks Frenk y su esposa Jikke Frenk por las largas conversaciones al calor del café y por sus buenos consejos. Agradezco a la profesora Berrin Yanikoğlu de Sabanci Universitesi y a la profesora Zumray Dokur de la Universidad Tecnológica de Estambul por su participación como miembros del jurado de defensa de la tesis, sus sugerencias ayudaron en el mejoramiento del presente documento.

Gracias a Jonathan Wolpaw, director del Neural Injury and Repair laboratory en Wadsworth Center por darme la oportunidad de unirme a su grupo de investigadores. Agradezco a Dennis Mcfarland, también en Wadsworth Center, por las largas discusiones sobre básicamente todo, fue un periodo en el que aprendí mucho. Nuestras conversaciones iban desde gráficas coloridas y regresión lineal hasta experimentos relacionados a la separación de los hemisferios cerebrales. Agradezco también a Gerwin Schalk, director de Schalk Lab, por darme la oportunidad de unirme a su grupo de investigadores, por las interesantes discusiones y por proveer acceso a datos de incalculable valor, sin los cuales el capítulo final de esta tesis estaría incompleto.

Agradezco a Adriana de Pesters, miembro del equipo investigador en Schalk Lab, por las interesantes discusiones académicas y filosóficas. Su incomparable pasión por su área de investigación, su deseo por entender, y su particular forma de ver la vida, atraparon mi interés y la hicieron sujeto de mi admiración. Agradezco también a Adriana las

[2] Este trabajo ha sido parcialmente financiado por el Concejo Científico y Tecnológico de Turquía bajo el proyecto 111E056 y por Sabancı University bajo el proyecto IACF-11-00889

sugerencias y correcciones, que fueron de gran valor y contribuyeron enormemente a la forma final de este documento.

A Hugo Gómez, su esposa Assel Saparova y su hija Camila Gómez, quienes durante mi primer año de Doctorado en Estambul, a más de 10000 Km de distancia de mi hogar, en una cultura completamente diferente a la cultura en que fui criado, me hicieron parte de su familia. Hugo fue también fuente de nuevas ideas, muchas discusiones, apoyo y un ejemplo a seguir como científico y ser humano. Tengo frescas en la memoria las noches en el laboratorio en las que Hugo tocaba la guitarra mientras yo intentaba seguirle el ritmo con maracas al son de "Chan Chan" de Buenavista social Club.

Un agradecimiento muy especial va a toda mi familia por su apoyo incondicional. En particular quiero agradecer a mi padre Carlos Emilio Delgado Angulo por su incansable apoyo, por las largas y reconfortantes conversaciones telefónicas durante estos cuatro años y medio, por recordarme siempre que la vida hay que disfrutarla, por su amistad y por continuar compartiendo su sabiduría conmigo.

A todos mis amigos por las muchas experiencias, enriquecidas por las diferencias culturales. En particular agradezco a Mireia Pérez, Mata López y Marketa Bílská por los buenos momentos en Estambul, a Atia Shafique por su amistad y su paciencia en los momentos difíciles, a Saygin Topkaya y Umut Sen del VPA Lab por su colaboración desinteresada. A Lacides Ripoll, Oscar Serrano y Juan Carlos Villamizar por el apoyo y ánimo ofrecido durante estos cuatro aos y medio a pesar de la distancia. A Pandian Chelliah and Rupak Roy por las buenas e interesantes conversaciones. Finalmente agradezco a la Universidad del Norte por el apoyo y en particular a Beatriz de Torres por ser parte activa de este proceso.

... to my daughter, Nicolle.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Translating thoughts into computer commands have been for a long time material of science fiction movies. Brain Computer Interfaces (BCIs) have opened a door to make this possible. The main goal of a BCI is to provide a new communication path that allows people with severe disabilities to communicate with their environment. This non-muscular communication path is based on the analysis of brain signals during the execution of specific mental tasks. Recently, applications for healthy subjects in the fields of multimedia and gaming have started to incorporate these technologies as well [2, 3]. A BCI system involves a basic set of blocks: acquisition, pre-processing, classification and feedback. For acquisition of the signals related to brain activity different methods which can be grouped as invasive and non-invasive have been employed. Invasive technologies such as electro-corticography (ECoG) require implantation of electrodes in the brain cortex making the process risky for the subject, as well as expensive, but at the same time providing a higher signal to noise ratio (SNR) than other techniques. Non-invasive methods such as functional magnetic resonance imaging (fMRI), magneto-encephalography (MEG) and positron emission tomography (PET), require the use of complex and expensive equipment that may not be appropriate for practical BCI applications given that the equipment is confined to specific locations in a controlled environment. In contrast, techniques such as electroencephalography (EEG) and near infrared spectroscopy (NIRS) are non-invasive, portable and relatively inexpensive when compared to the alternatives mentioned above, which makes them suitable for practical BCI applications. The price to pay for these advantages includes lower SNR and poor spatial resolution. Pre-processing stages involve the use of signal processing techniques with the main purpose of enhancing the SNR. Here, two main tasks are executed: feature extraction and feature selection. The former aims to extract characteristics of the signal that provide information that is useful for discrimination of mental activities. Feature selection has the objective of selecting the most prominent features to avoid a well-known problem called curse of dimensionality that affect machine learning methods that are used in the classification stage. The classification stage involves the use of

machine learning techniques such as Linear Discriminant Analysis (LDA), Artificial Neural Networks (ANN), Support Vector Machines (SVM), among many others, where the main idea is to use previously acquired data to train a model which can then be used to discriminate among new inputs. The feedback stage is used to present to the subject the decision that the system has made and at the same time for controlling the actions through external devices, according to the mental activity recognized by the system. Two types of signals have commonly been exploited in BCI research. The first is a potential known as P300. P300 is an event related potential involving the response generated by the brain to low-probability visual or auditory stimuli that the subject is interested in. Such responses can experimentally be generated using the oddball paradigm [4]. In this paradigm low probability stimuli are mixed with high probability stimuli. The subject is requested (for example) to count each time that an uncommon stimulus appears. This set up is expected to generate a P300 response in the subject's brain. This phenomenon has been subject of exhaustive research (see [5] for a review of BCI systems that use P300) . Different laboratories around the world have opted for the P300 for the development of BCI systems that enable a subject to spell letters in a computer [6, 7, 4, 8]. Various classifiers including stepwise linear discriminant analysis (SWLDA), support vector machines (SVMs), etc. have been used in P300-based spellers with similar levels of success. In most work, each letter is classified independently of the other letters. However, in the context of typing words from a language, the letters are of course not independent, and just like in speech recognition, their dependence could be exploited. This observation led to recent interest in the use of language models in P300-based spellers [9, 10, 11] producing significant performance improvements. Most of this work involves building and using conditional probabilities of letters given previous letters in the typed sequence. Another observation one can make is that many BCI tasks involve typing from a limited dictionary. This observation motivates the use of even higher level, e.g., word-level language models in P300-based spellers.

The second type of signals commonly used for BCI is the sensorimotor rhythms. These rhythms are characterized by the increase or decrease of power with the execution of motor tasks, in different frequency bands. The classification of these rhythms involve the use of features measuring the power in different frequency bands of the brain signals during the execution of different tasks. This is commonly done by means of static classifiers, i.e., classifiers that do not involve dynamic models of the temporal structure of the inference task. (see [12] for a review of classifiers in BCI based in sensorimotor rhythms). However, the observed changes in time of the power of the brain signals in specific frequency bands [13], support the idea that dynamics of the signals contain information that can be used to discriminate between different type of classes. Preliminary work has been presented in line with these thoughts. The work in [14] makes use of Hidden Markov Models (HMM) for modeling the EEG signal during the imagination of movements, using as features the well-known Hjorth parameters, which provide information about the power, frequency and frequency rate of change in the EEG. In this approach, the HMM is used

for modeling different states in the EEG signal. As expected, a system that includes temporal information overperforms the classical approach based on a static classifier [14]. The disadvantage of this method is that it requires an extensive number of training samples and that the number of states should be defined based on experience or making use of cross-validation methods. Given that the number of samples for training is an issue in BCI, the proposed method by Obermaier does not make use of autoregressive parameters (AR) which have proven to be a powerful tool for modeling the EEG signal [15]. The main reason for not considering the AR parameters is that orders of 6 or greater ($\rho \geq 6$) are needed to represent the EEG signal reasonably accurately [15] (see [16] for an interesting discussion of this topic) and the numbers of features obtained from the EEG signals become large, $\rho \times N_e$, where N_e is the number of electrodes. The work in [17] presents a solution to the problem of high dimensionality of the set of features, when AR parameters are used. In this approach, the AR parameters for each electrode are obtained each 0.5 seconds using the last second of data. The parameters are concatenated producing $N_e \times \rho$ features for each one-second window of data. Then, the dimension of this set is reduced using principal component analysis. The resulting feature is applied to the HMM model. The number of states that produces the best accuracy is obtained testing the training model over a validation set. The results show that this approach overperforms the HMM method based on Hjorth parameters solving the problem of high dimensionality in the feature set. In [18] a two-layer HMM is proposed. In this approach signals from electrodes over the motor cortex region are modeled separately. That means that for each electrode and each class a different HMM is trained. A second layer of HMM uses the log-likelihood of the signal in each HMM in the first layer as input. The EEG signal features used involve time domain parameters [19] of the EEG, which can be understood as a generalization of the Hjorth parameters. Results presented in [18] show that this approach is comparable with the state-of-the-art. Furthermore, this method provides a physiological interpretation because it is observed that the states in the HMM are related to the event-related synchronization/desynchronization (ERS/ERD), well-known phenomena in motor task execution. Other works involve extensions of HMM, including e.g., the so called Input-Output HMM (IOHMM) [20]. This approach provides better performance in asynchronous BCI systems, when compared to HMM, which can be attributed to the discriminative properties of IOHMM and the fact that only one model with the ability to discriminate between different classes is trained. This is in contrast with HMM where for each class, a model must be learned. Recently, other works that involve discriminative models have been presented. [21] proposes a modified conditional random field (CRF) for synchronous BCI system. This work shows the advantages of a discriminative model over generative models in BCI. However, although this is a dynamic model, the structure proposed by [21] associates "states" with each of the possible classes in a synchronous scenario (a three class problem is presented). As a consequence of this, the temporal structure is not exploited.

The discussion above motivates several lines of inquiry about possible improvements

of BCI systems. In the case of the P300-based spellers, a probabilistic method that incorporates a word-level language model into the process of inferring on the typed letter sequence based on EEG data is currently missing in the literature. Given that many potential users of BCI technology are likely to be interested in communication through a limited dictionary, we expect such strong language models to be of great value in increasing the information transfer rate of P300-based spellers.

In the case of the sensorimotor rhythms our perspective is one of modeling and exploiting the dynamics of the signals. In this work we propose the use of several probabilistic graphical models that aim to address certain limitations of existing, mostly HMM-based, methods. Another aspect of the dynamic structure, not considered explicitly in past work on BCIs is the existence of two types of dynamics: the intrinsic dynamics of brain states through the process of execution of a specific mental task and the extrinsic dynamics of different mental tasks. This is another one of the new perspectives developed in this thesis.

1.1 Overview of Contributions

Here we describe briefly the contributions of this thesis:

- We propose a novel discriminative P300 framework that models the variables of a P300 speller system and makes use of a language model at the level of words, allowing the system to fit language characteristics particular to each BCI task or subject..
- A non-parametric HMM is proposed in the context of synchronous BCIs as a solution to the problem of selection of of the number of hidden states and the selection of the number of components needed to model the probability density functions of the data. This data driven method leads to better results than conventional techniques based on cross-validation. This is the first use of nonparametric Bayesian methods in the context of BCI.
- A latent discriminative model with hidden variables is proposed for classification in synchronous BCI systems based on sensorimotor rhythms. Here we make use of the temporal dynamics of the brains signals and exploit the advantages of discriminative models. The results show significant improvements in classification accuracy of motor tasks.
- We propose a discriminative graphical model based approach for classification in asynchronous BCIs. This approach exploits both the intrinsic dynamics of brain states during the execution of a particular mental task and the extrinsic dynamics across different mental tasks.

- We propose asynchronous classification of the independent movement of fingers from electrocorticography (ECoG) data, making use of classifiers based on conditional random fields. The proposed model provides ideas on how to include information about the relationships between the movement patterns of different fingers as well. This opens the door to the exploration of spatial relationships in brain signals during the execution of different tasks.

1.2 Thesis Organization

1.2.1 Chapter 2: Background

We begin with an overview of the definitions and methods used commonly in the BCI community. A summary of recording methods, pre-processing tools and classification methods is presented. At the end of this chapter an introduction to graphical models and motivation for their use in EEG signal processing is presented.

1.2.2 Chapter 3: A Word-level Language Modeling Framework for the P300 Speller

In this chapter we propose a discriminative graphical model for classification of P300 potentials in an application that allows people with motor limitations to spell letters in a computer. This approach overcomes many of the problems in traditional spellers by integrating all the variables of a BCI system into a single model. The model also includes a language model that is used as a prior on the words spelled by the subject. Through experiments with EEG, we provide evidence of the superiority of the proposed model as compared to conventional methods.

1.2.3 Chapter 4: Generative Graphical Models for Synchronous BCIs

In this chapter a type of BCIs that makes use of brain signals related to imagination of motor activity is studied. We propose the use of generative methods for modeling the temporal structure of the signals by defining different states in the ongoing EEG signal during the imagination of motor tasks. A nonparametric Bayesian method based on hierarchical Dirichlet processes is proposed to overcome the problem of model order parameter (number of hidden states and number of Gaussian mixture components). The results demonstrate that the modeling of the temporal structure of the signal provides an increased classification performance.

1.2.4 Chapter 5: A Latent Discriminative Graphical Model for Synchronous BCIs

In Chapter 5, a discriminative model based on conditional random fields with hidden states is proposed. This method overcomes some of the limitations of generative models by directly modeling the conditional distribution of the labels given the data. Hidden states are used to model the dynamics of the EEG signals during the execution of imaginary motor tasks. The results show that this method provides a significant improvement in the classification of motor tasks in synchronous BCIs.

1.2.5 Chapter 6: Discriminative Methods for Asynchronous BCI

We continue by exploiting the dynamics of the EEG signals in a type of BCI where the tasks are executed in asynchronous form, i.e. the subject decides, without waiting for cues, when to start or end a specific mental task. In this chapter, we propose a method that exploits the dynamics of the EEG signals together with dynamics of the task executed by the subject. This particular classification problem is more challenging than in the synchronous case because the algorithm has to determine the start and ending of each specific mental tasks. In addition to the motor tasks used in previous chapters, mental activity related to cognitive states are used as mechanisms of control. The proposed method is compared to the state-of-the-art methods in asynchronous classification in BCI showing significant performance improvements.. The experiments involve the use of publicly available data as well as data recorded in our laboratory from subjects without experience with BCI, to generate a more challenging scenario. The results evidence the robustness of our method.

1.2.6 Chapter 7: Asynchronous classification of Finger Movements using ECoG

In Chapter 7, we present an application of graphical models for decoding the movements of fingers using signals recorded directly from the brain cortex. We propose a model for asynchronous classification of the ECoG signals to determine the movement or rest of each finger as well as a model for the classification of which finger is in movement. Experimental results evidence the capability of the presented model for continuous decoding of movements. Furthermore, this model opens the door to a future incorporation of spatial features together with temporal features of the brain signals with the potential of creating a more integrative model that explains spatio-temporal dynamics in the brain during the execution of motor tasks.

1.2.7 Chapter 8: Contributions and Future Work

In this chapter, we conclude by surveying the contributions of this thesis and indicating possible directions for future work, motivated by the limitations and advantages of the proposed methods.

Chapter 2

Background

In this chapter, an overview of the basic concepts in BCI is given. Also methods for pre-processing and classification are presented. The chapter ends with an overview of probabilistic graphical models.

2.1 Neurophysiological Signals and Recording Methods

Neurophysiology is a branch of physiology and neuroscience that studies the function of the nervous system (NS). One important tool for the study of the function of the NS is electrophysiology; the study of the electrical properties of the cells or tissues. The cellular electrical phenomena observed in biological structures are explained by the flow of ions from the exterior of the cell to the interior of the cell and vice versa giving origin to currents and voltages that can be measured by electrodes placed in the interior of the cell (intracellular recordings) or at the exterior of the cells (extracellular recordings). The recording of extracellular electrical activity can be made on many scales, giving rise to different types of recording methods. In the NS, single neuron recordings are possible when the diameter of the electrode placed in the brain is in the order of micrometers (about 1 micrometer). Electrodes in the order of millimeters placed on the surface of the cortex measure the response of groups of many neurons, this type of recording is known as Electrocorticography (ECoG). If the electrodes are placed over the scalp, it is possible to measure the electrical activity of cells in wide regions of the brain. This noninvasive type of recording is known as Electroencephalography (EEG). Recording methods such as Magnetoencephalography (MEG) and Magnetic Resonance (MRI) among others, are currently used to measure brain activity. However, their applications to BCIs systems are limited in practice given the difficulty of access to such technologies both in terms of cost and of portability. In this thesis, the recording methods used are EEG and ECoG.

(A) EEG (B) ECoG (C) MEG (D) MRI

FIGURE 2.1: Different recording methods for neurophysiological signals.

2.1.1 Electroencephalography

An EEG signal is a measure of currents that flow during synaptic excitations of the dendrites of many pyramidal neurons in the cerebral cortex. When neurons are activated, the synaptic currents are produced within dendrites. This current generates an electric field over the scalp measurable by EEG systems [22]. Differences of electric potentials are caused by summed post-synaptic graded potentials from pyramidal cells that create electrical dipoles between the body of the neuron (soma) and apical dendrites, which branch from neurons. The current in the brain is generated mostly by pumping the positive ions of sodium, potassium, calcium and the negative ion of chlorine, through the neuron membranes in the direction governed by the membrane potential [22]. The signals can be recorded over the scalp. However, different layers in the human head (scalp, skull, etc.) produce attenuation and sources of noise either within the brain or over the scalp (external noise) reduce the SNR.

The EEG signals provide information about neurological disorders and other abnormalities as well as physiological phenomena related to the functioning of the body which makes them useful for diagnostics.

2.1.2 Electrocorticography

The two main problems observed in EEG are the SNR and the spatial resolution. The low SNR of EEG recordings is due to the attenuation of the amplitude of the synaptic excitation of the dendrites as the signal travels across the skull. In order to avoid those issues, the electrodes can be placed in direct contact with the brain cortex which at the same time allows to reduce the separation of the electrodes from centimeters (in EEG) to millimeters. There is no fundamental difference between EEG and ECoG and for this reason ECoG is also named Intra-cranial EEG (iEEG). The technique is invasive, requiring surgery for the placement of the electrodes, and the amount of time that the

electrodes can remain in contact with the brain cortex is limited. For all these reasons the use of ECoG is limited to cases in which the patient needs surgery as it is the case in patients with epilepsy where ECoG is used to identify the areas of the brain from where the seizures originate. Despite its disadvantages, ECoG stands as a potential alternative for BCI in patients with serious motor limitations such as Amyotrophic Lateral Sclerosis (ALS) as recent work has shown [23, 24].

2.2 Brain Rhythms Used in BCIs

2.2.1 Slow Cortical Potentials

Slow Cortical Potentials (SCP) are positive and negative polarizations of the electroencephalogram that originate from the depolarization of the apical dendritic tree in the upper cortical layers. The SCP constitutes a threshold regulation mechanism for local excitatory mobilization or inhibition of cortical networks. Humans can learn to voluntarily generate these potentials after training, using immediate feedback and positive reinforcement. These shifts produced in the EEG signal at very low frequencies can be used as control signals for a BCI system [25].

2.2.2 P300

P300 is a positive deflection in the EEG time locked to auditory or visual stimuli. It is typically seen when participants are required to attend to rare target stimuli, within a stream of frequent standard stimuli [26]. P300 is generally observed in central and parietal regions, and it is understood as a correlate of an extinction process in short-term memory when new stimuli require an update of representations [26]. This potential is well known in the BCI community, and numerous pieces of work have been presented, predominantly applied to spelling systems [27, 28, 29, 30].

In a typical P300 spelling session, the subject sits up right in front of a screen observing a matrix of letters as shown in Figure 2.2. The task involves focusing attention on a specific letter of the matrix and counting the number of times that the character is intensified. The matrix is divided in rows and columns. Rather than highlighting the letters individually, the system intensifies columns or rows. It is expected that the intensification of the letter to which the subject focuses his/her attention will lead to the generation of an event-related response , namely the P300 response. Therefore, the presence of P300 detected after the intensification of any row or column implies that the target letter is in that row or column. The letter can be decoded by intercepting the row and column that contains P300s in the matrix of letters.

FIGURE 2.2: P300 Speller Matrix

2.2.3 Steady State Visual Evoked Potentials

Evoked potentials can be recorded in the occipital region over the electrode positions O1, O2, Oz (according to the international 10-20 standard montage shown in Figure 2.3) when subjects are exposed to repetitive visual stimuli. The subjects focus their gaze on flickering targets and evoked potentials become steady-state, with the higher intensity of the response occurring at the fundamental frequency of the stimulus and at second and third harmonics [31]. Parameters of the evoked potential as amplitude and phase depend on stimulus frequency and contrast [26]. The frequency resolution of SSVEP is about 0.2Hz and the bandwidth in which it can be detected reliably is between 6Hz and 24Hz [26]. The SSVEP phenomena can be used in BCIs by asking the subject to focus on one among different stimuli presented on a screen. The classification of the target observed by the subject is related to the estimation of the fundamental frequency in the spectrum of the recorded brain signals.

2.2.4 Sensorimotor Rhythms

Sensorimotor rhythms (SMRs) include the so called μ-rhythm with frequencies around 10Hz, often mixed with a β component around 20Hz. It is easily recorded over the motor cortex, preferably over the electrode positions C3 and C4 according to the international 10-20 standard montage (see Figure 2.3). The power of sensorimotor rhythms can decrease with the movement or preparation of movement and can increase in the post movement period. Furthermore, imagination of movements (motor imagery) can also generate a decrease of μ-rhythm power [26]. This phenomenon is known in the BCI literature as event-related de-synchronization / synchronization (ERD/ERS) [32] and is relevant for BCI given that the target population of users suffer from motor disabilities. The modulation of the SMR can be used as input for a BCI system. Subjects are instructed to execute the imagination of left or right hand movements, which produces a de-synchronization in the contralateral region in the brain (i.e., left hand motor imagination/movement produces de-synchronization of μ-rhythm in the right hemisphere

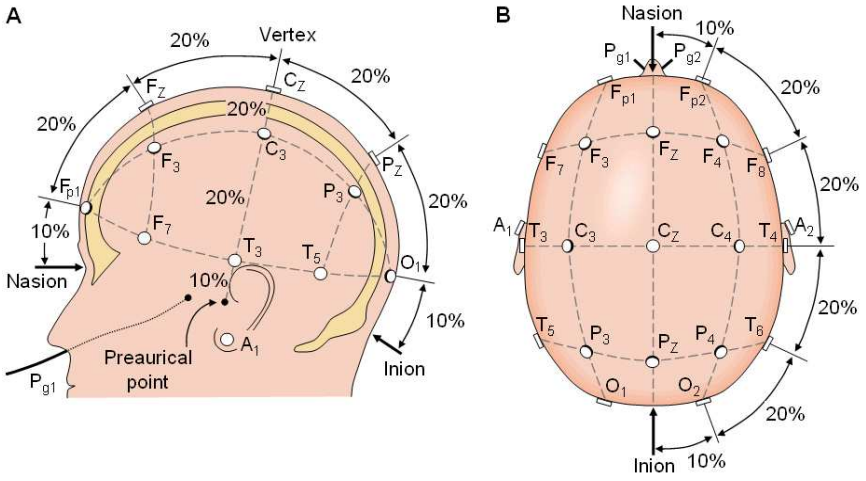


FIGURE 2.3: Standard 10-20 EEG montage.

and vice versa). After subsequent application of machine learning algorithms, feedback can be provided to the subject to reinforce the execution of the mental task. Several BCI systems use this kind of rhythms with good results, measured by classification accuracy [33, 14, 7, 34, 35, 18, 36].

2.3 Pre-processing Methods

2.3.1 Electrode Reference Methods

EEG recordings measure the voltages at electrode locations with respect to a reference that is usually placed on the mastoid, the left ears or the linked ears. However, a phenomenon known as volume conductor may affect the signal at different electrodes. The volume conductor phenomenon occurs when brain waves of a region propagates through the skin affecting the recording in distant locations. Re-referencing methods can be applied to the EEG recordings to minimize this effect. Commonly used techniques include Common Average Reference, Bipolar Reference and long Laplacian filters.

2.3.1.1 Common average reference

Common Average Reference (CAR) re-references the EEG by averaging the signals in all electrodes and subtracting this mean from each electrode. Assuming that $E = \{e_1, e_2, ..., e_N\}$ is the variable that represents the EEG recording, with each component e_i representing the signal recorded at each electrode, the re-referenced EEG recording E_r is obtained by:

$$E_r = E - \frac{1}{N}\sum_{i=1}^{N} e_i \tag{2.1}$$

This simple method can reduce the effects of the volume conductor phenomenon as well as artifacts that are common to all the electrodes. The main disadvantage is that an artifact with high amplitude that is only present in one electrode can distort the signals in all electrodes after CAR is applied.

2.3.1.2 Bipolar reference

Bipolar reference is a re-referencing method that measures the potential between two electrodes. This method is usually used when the area of interest in the scalp is known (for instance, electrodes in the anterior and posterior position to C3 according to the standard 10-20). This produces a more localized measure of the potential and eliminates the need of a global reference electrode on the mastoid or the ears. Given $E = \{e_1, e_2, ..., e_N\}$ representing the EEG recording, the bipolar re-referencing $e_{i,j}$ of electrodes e_i and e_j is obtained by:

$$e_{i,j} = e_i - e_j. \tag{2.2}$$

2.3.1.3 Laplacian Reference

A long Laplacian reference provides a measure of the local potential between one electrode and all neighbor electrodes that are separated from it by the equal distance in the scalp. The re-referenced potential on electrode e_i, P_{e_i}, using as reference the subset G of n_k electrodes $\in E = \{e_1, e_2, , e_N\}$, is determined by:

$$P_{e_i} = e_i - \frac{1}{n_k}\sum_{j \in G} e_j \tag{2.3}$$

2.3.2 Artifact Reduction

EEG recordings are affected by many sources of noise. Muscular activity (Electromyographic signals, EMG), heartbeat (Electrocardiographic signals, ECG) and potentials between the cornea and the retina (Electro-oculographic signals, EOG) are the most common causes of artifacts in the EEG signal. The removal of EOG signals is of great interest given that the magnitude of these potentials is several orders of magnitude larger than the EEG signals. In order to reduce the interference of EOG signals in the EEG recordings, linear regression methods can be employed [37]. In this approach, EOG

signals are recorded in parallel with the EEG signals. The signal recorded by the EEG electrodes is modeled as the summation of the actual underlying EEG signal and the noise, represented by a linear combination of the EOG signals interfering into the EEG electrodes [37]:

$$w(n) = s(n) + u(n).b \tag{2.4}$$

where n represents the discrete time index, $w(n)$ and $s(n)$ represent the noisy and the actual EEG signals at M electrodes, and $u(n)$ represents the EOG signal at N electrodes. Representing $w(n)$, $s(n)$, and $u(n)$ at a particular time point as row vectors of appropriate dimensions, b is an *unknown* matrix of size $N \times M$ representing the set of coefficients that explain how the EOG signals have propagated by volume conduction to each of the points on the scalp where the EEG measurements are made. The problem is to recover $s(n)$ from measurements of $w(n)$ and $u(n)$. Given that the EOG signals are large in magnitude compared to the EEG signals, the interference of EEG in the EOG recordings $u(n)$ can be neglected [37]. Knowing b, the original EEG signal can be found by $s(n) = w(n) - u(n).b$. Multiplying the signal $w(n)$ by $u(n)^T$ and taking expectation, we obtain:

$$E[u(n)^T w(n)] = E[u(n)^T s(n)] + E[u(n)^T u(n) b] \tag{2.5}$$

Under the assumption that there is no correlation between the EEG signals $s(n)$ and the EOG signals $u(n)$, an expression for estimating the coefficient matrix b is found:

$$\hat{b} = E[u(n)^T u(n)]^{-1} E[u(n)^T w(n)] \tag{2.6}$$

The correlation matrices above can be learned and $\hat{b}$ can be computed using the EOG and EEG measurements.

2.3.3 Frequency Band Separation

During the execution of different mental tasks, the characteristics of the brain signals change. These changes are strongly related to the increase or decrease of the power of the signals in different frequency bands. These changes are in general common among humans, and specific classification of these rhythms have been established in the literature. Delta waves fall in the frequency range of 0.5Hz - 4Hz. These waves are associated with deep sleep, although they can also be present in waking states. The low frequencies involved in Delta waves make them easy to be confused with artifacts caused by activity of muscular groups of the neck and jaw [22]. Theta waves, in the

range of 4Hz - 7.5Hz have been associated with access to unconscious material, creative inspiration and deep meditation. The Alpha waves in the range of 8Hz - 13Hz are usually found over the occipital region of the brain. The apparition of this rhythm is related to states of concentration or relaxed awareness without any attention. As it will be shown in the following sections, this rhythm plays an important role in BCI systems because the execution of motor activity modifies the alpha rhythm amplitude. Beta waves are associated with active thinking, active attention and focus on the outside world. This rhythm contains frequencies in the range of 14Hz - 26Hz. Beta waves are found in the frontal and central regions. In the central regions, Beta waves can be blocked by motor activity and tactile stimulation [22]. Gamma waves contain frequencies in the range of 25Hz 100Hz with typical values around 40Hz. This frequency band has been historically ignored because scalp EEG recordings display a very low SNR at frequencies above 30Hz, but with the development of ECoG recordings, Gamma waves have become of great interest for the neuro-scientific community [38, 39, 23]. Although there is no consensus on the meaning of the gamma waves, they are believed to play an important role in conscious perception. [40]. A subdivision of Gamma waves (high Gamma) is made to describe brain waves with frequencies in the range of 60Hz 200Hz. The nature of the high Gamma is believed to be related to the firing rate of populations of neurons in the brain cortex [39, 41, 38]. Depending on the BCI task, one might apply filters to extract one of more of these components defined in different frequency bands.

2.3.4 Spatial Filtering

2.3.4.1 Common spatial patterns

Common spatial patterns (CSP) are spatial filters that are well suited to discriminate mental states characterized by ERS/ERD phenomena [42]. Given the bandpass filtered, labeled EEG signals $s(n) \in R^M$ from a training set for classes C_1 and C_2, it is possible to estimate the $M \times M$ sample spatial covariance matrices Σ_{C_1} and Σ_{C_2} of the EEG signals. CSP performs simultaneous diagonalization of Σ_{C_1} and Σ_{C_2} in such a way that the eigenvalues of the diagonalized matrices sum to 1, that is:

$$V^T \Sigma_{C_1} V = D \text{ and } V^T (\Sigma_{C_1} + \Sigma_{C_2}) V = I, \tag{2.7}$$

where V is the matrix of generalized eigenvectors, D is a diagonal matrix of eigenvalues and I is the identity matrix. Hence the EEG signal $s(n)$ at each time point can be transformed from the electrode space to the CSP space through $s(n)V$. It is possible to focus on the j-th CSP component by using the filter V_j (j-th column of V) and the resulting projected signals $s(n)V_j$. If the signal is from class 1, the variance of the

projected signal would be $V_j^T \Sigma_{C_1} V_j = d_j$ (d_j is the corresponding eigenvalue for the eigenvector V_j). Likewise, for signals from class 2, the variance of the projected signal would be $1 - d_j$. In the case that the number of classes is two, it is possible to use CSP components that emphasize the contrast between the classes. As observed, the filters V_j that provide the best contrast between the two classes are those with large eigenvalues and low eigenvalues, producing large variance for class 1 and low variance for class 2, and vice versa. Then, choosing those particular components corresponding to high and low eigenvalues only, the spatial filtered signal is obtained as follows:

$$c(n) = s(n)W, \tag{2.8}$$

where W is a matrix whose columns are composed of a subset of the eigenvectors V_j, in particular those with relatively large and small eigenvalues.

2.4 Feature Extraction

2.4.1 Autoregressive Parameters

Autoregressive models (AR) are Markov processes. The basic idea in AR modeling is that the current value of a time series can be predicted from the p previous values of the signal. This can be expressed as

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + n_t \tag{2.9}$$

where a_i represent the coefficients of the model, p is the order of the model and n_t is the input of the system of noise function. The parameters of the model define the characteristics of the temporal signal y_t. A stationary AR model is such that its inversion exists.

The AR models are widely used in BCI [15, 17, 43, 44]. However, the assumption of stationarity does not hold in the case of brain signals. Therefore, the parameters of the model should be updated continuously. One way to do this is calculating several AR models in short windows overlapped in time, assuming that short segments of the EEG signals are stationary. The sequence of coefficients of different sets of AR models can correspond to changes in the statistics of the brain signals, which can be used for classification.

2.4.2 Spectro-Temporal Features

In Section 2.3.3 it was explained that the power in different frequency bands of the EEG signal carries information that could be used to discriminate between different mental states. Different methods are used to extract the temporal variation of power within specific frequency bands.

The most common approach consists of filtering the signal in the frequency of interest and then estimating its envelope. The estimation of the envelope can be done in different ways, however one of the most used methods is the Hilbert Transform.

Hilbert Transform Approach. Given the filtered EEG signal $x(t)$, its envelope can be calculated using the magnitude of the analytic signal $s(t)$, obtained by:

$$s(t) = x(t) + j\widehat{x}(t) \tag{2.10}$$

where $\widehat{x}(t)$ is the Hilbert transform of the EEG signal x(t)

$$\widehat{x}(t) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x(\tau)}{t-\tau}d\tau \tag{2.11}$$

that is, the Hilbert transform of $x(t)$ is the response of a filter with impulse response $\frac{1}{\pi t}$. Note that the integral in Equation 2.11 is improper therefore, the Hilbert transform is defined as the Cauchy Principal Value of the integral in Equation 2.11. With reference to 2.10, the analytic signal $s(t)$ can be expressed as:

$$s(t) = a(t)e^{j\theta(t)} \tag{2.12}$$

where $a(t)$ is the magnitude of $s(t)$ and $\theta(t)$ is the angle. Note that $x(t)$ is the real part of $s(t)$, that is

$$x(t) = a(t)cos(\theta(t)) \tag{2.13}$$

meaning that $x(t)$ can be represented as an amplitude modulated signal with envelope $a(t)$. Note that it is assumed that the frequency content of the envelope $a(t)$ and $\cos(\theta(t))$ are disjoint.

Short-Time Fourier Transform. The Fourier Transform (FT) $X(\omega)$ of a signal $x(t)$ provides a representation of a signal in the frequency domain and is given by:

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt \tag{2.14}$$

where ω is the angular frequency. However, given that the brain signal is non-stationary, its spectral representation changes with time. The Short-Time Fourier Transform (STFT) provides a representation of the signal in frequency and time by calculating the FT in windowed segments of the signal $x(t)$ at positions t providing an estimate of what frequencies exist in the signal and where in time those frequencies appear. Therefore the frequency content of the signal $x(t)$ at time τ is given by

$$X(\omega, \tau) = \int_{-\infty}^{\infty} x(t)g(t-\tau)e^{-j\omega t}dt \tag{2.15}$$

This time-frequency map can be used to extract the temporal changes of the signal in different frequency bands. Special attention should be given to the window function g. If the length of the window is not a multiple of the period of any of the components of $x(t)$, spurious responses may appear at many frequencies. This issue can be minimized by selecting a window function that attenuates the values of the signal that are separated from the center (i.e., a Hamming window). The main disadvantage of the STFT is that there is a trade off between the resolution in frequency and time domains. A high resolution in frequency implies the use of long windows, which produces low resolution in the time domain. A high resolution in time requires the use of short length windows which reduces the resolution in frequency.

Wavelets Transform. The main problem of the STFT is due to the constant length of the windows which generate problems at different frequencies. Note that for a specific length of the window g in Equation 2.15 several cycles of high frequency components can be observed (good frequency resolution, bad temporal resolution) while for low frequency components few cycles would be observed (bad frequency resolution, good temporal resolution). This issue can be solved by using a multi-resolution representation, that is, representing the signal at different scales. This can be achieved by using windows with different sizes at different frequencies. The Wavelet Transform (WT) of a signal $x(t)$ can be expressed as:

$$X(s, \tau) = \int_{-\infty}^{\infty} x(t)\phi_{s,\tau}(t)dt \tag{2.16}$$

Note that this implies that the signal $x(t)$ can be represented as a linear combination of the basis $\phi_{s,\tau}(t)$, and s and τ are parameters that define operations of dilation and translation of the analytic function ϕ according to

$$\phi_{s,\tau}(t) = s^{-1/2}\phi(\frac{t-\tau}{s}), s, \tau \in R, s > 0. \tag{2.17}$$

Note that the wavelets $\phi_{s,\tau}(t)$ that are a stretched version of the analytic wavelet $\phi(t)$ are able to produce good frequency representation for components with low frequency

while compressed versions of $\phi(t)$ provide a good time resolution for components with high frequency.

The WT has been used in BCI with good results. Applications include BCIs that make use of sensorimotor rhythms and P300 potentials [45, 46, 47].

2.4.3 Measures of Connectivity Across Brain Regions

Relationships between different regions in the brain have been of interest for the BCI community. According to the neurophysiological theory, the execution of movements involves activation of different structures in the brain and possible communication between them. [48] shows, in an experiment involving internally paced and externally paced finger extensions, that movement-related activation is predominant in the contralateral area and in the primary motor cortex, and that functional coupling occurs between primary sensorimotor cortex in both hemispheres and between primary sensorimotor cortex and the mesial premotor areas. However, a phenomenon known as volume conductor [49] can affect the measure of the coherence presented by [48] because EEG signals obtained in a specific scalp position can spread through the scalp, which leads to possible false identification of functional coupling. In order to solve this problem different measures of functional coupling have been proposed. Using a multichannel AR approach, [50] proposes the so-called directed transfer function (DTF). In this approach, the transfer function matrix composed by the AR coefficients of the multichannel model, properly normalized, is used as estimator of the propagation direction of the flow of information between brain regions. That is, the value of the AR coefficients can be used to determine how much information a signal provides about any of the other signals in the model. Given that this transfer function matrix is not symmetric, information about the direction of propagation of the signals is obtained. This approach solves the problem of volume conductor because a copy of the signal at a point A that appears at point B by volume conductor will not contain extra information about signal in A. Here, it is assumed that no time-lag is observed because of volume conduction. [51] proposes the use of partial coherence $x-y/z$ defined as a linear association between processes X and Y taking into account and removing the linear effect of the process Z. [51] proposes that this method is a reliable measure of the interhemispheric human EEG coherence. Results show that increase in the interhemispheric communication is present in the beta band during execution of movements. [52] proposed that the classical measures of coherence will lead to an erroneous determination of connectivity in the brain because of the volume conductor. However, given that the interference of one EEG channel with neighbors is assumed to have zero time-lags, the use of the imaginary part of the coherence is insensitive to false connectivity arising from volume conductor. Results show, as in previous works, that connectivity is observed during movements in frequencies corresponding to the beta band. Other approaches such as full frequency DTF (ffDTF) [53], Short - time DTF (SDTF)

[54] and partial directed coherence (PDC) [55] have been proposed for determining the connectivity in brain regions during execution of real or imaginary movement.

A different approach for measuring possible communication between brain regions is presented in [56, 57]. Lachaux proposes that frequency synchrony between two sites can be determined by a quantity named the phase-locking value (PLV). This quantity provides a measure of the instantaneous phase difference for two signals x and y as described in Equation 2.18

$$PLV(f,t) = \int_{t-\delta/2}^{t+\delta/2} \exp(j(\phi_y(f,\tau) - \phi_x(f,\tau)))d\tau \tag{2.18}$$

where $\phi(f,t)$ is the phase of the signal for a frequency f as a function of time. If the signals are in phase during the interval δ, the PLV is equal to one; when the differences are large, the PLV approaches to zero. Lachaux proposes that the determination of the instantaneous phase should be done by first filtering the signal in a narrow band around the frequency of interest f and second, by convolving the signal with a complex Gabor wavelet centered at f [56]. However [58] proposed a method based on the calculation of the Hilbert transform of the signal. In a comparative study by [59], it was shown that the differences between these two methods are minor and can be considered equivalent, but the Hilbert Transform based method is less costly in terms of computational resources, which is important for real-time applications. PLV approaches have been used for classification of EEG signals during execution of mental tasks in several works [60, 61, 62, 63, 64] using static classifiers. Although the definition of PLV implies the use of narrow-band frequencies, [63, 64] report better accuracies using a wide frequency band (8Hz - 30Hz). It is also interesting to note that effects of the volume conduction have not been taken into consideration in the cited works. [62] proposed that given that the EEG signals are composed of the superimposition of different signals, blind source separation methods are necessary to avoid false synchrony detection. The proposed method, temporal decorrelation source separation (TDSEP) [65], makes use of the time structure of signals and uses the fact that signals are assumed to be time-lagged. Klaus et al. [62] show that this approach allows to observe appreciable changes in PLV measures during self-paced finger movements.

2.5 Feature Selection

Feature selection algorithms define a way to add or remove features, mostly in a sequential manner. In the case of forward sequential forward selection (SFS), the initial set is empty and new features are added if they provide an increase in the value of a predefined cost function. Sequential Backward Selection (SBS) starts with a full set of features which are removed sequentially if an improvement is obtained in the predefined cost function

when the feature is removed. In this section we describe a more general approach for feature selection called Sequential Forward Selection.

Given a set of features $\mathcal{F} = \{f_1, f_2, ...f_D\}$, the Sequential Floating Forward Selection algorithm (SFFS) [66, 67, 68] finds a new set $\overline{\mathcal{F}}_k = \{\overline{f}_1, \overline{f}_2, ...\overline{f}_k\}$ such that $k \leq D$. Ideally the new set of features $\overline{\mathcal{F}}$ increases the performance of the system or produces the same performance with a reduced number of features, which reduces the computational cost. The selection of the feature subset $\overline{\mathcal{F}}_i$ from set $\mathcal{F}$ is performed according to an objective function $J(\overline{\mathcal{F}}_i)$, where if $J(\overline{\mathcal{F}}_i) > J(\overline{\mathcal{F}}_j)$ the subset $\overline{\mathcal{F}}_i$ performs better than subset $\overline{\mathcal{F}}_j$. SFFS sequentially adds a new feature from the original set to the output set according to the objective function. On each iteration the effect of removing each one of the features previously selected is evaluated, and if one feature is found to reduce the accuracy it is removed. The basic Sequential Feature Selection (SFS) algorithm is obtained if the effect of removing each feature is not considered. SFS method has the disadvantage of producing a monotonic growth of the feature vector which impacts the computational cost in the subsequent stage of classification.

2.6 Classification Methods

2.6.1 Linear Discriminant Analysis

In the case of binary classification, a discriminant function defines a hyperplane that separates the elements that belong to class 1 from elements that belong to class 2. The discriminant linear function can be expressed as:

$$g(\mathbf{x}) = \theta\mathbf{x} + \theta_0 \tag{2.19}$$

where θ is a weight vector and θ_0 is a bias. Note that in a classification problem, the samples of $\mathbf{x}$ that make $g(\mathbf{x} > 0)$ are assigned to class 1 while samples of $\mathbf{x}$ that make $g(\mathbf{x} < 0)$ are assigned to class 2. This idea can be extended to multiple classes by defining c discriminant functions, where c is the number of classes.

Different linear methods share the same structure, with differences in how the weights are calculated. Of particular interest is the Fisher's LDA which is the benchmark for determining the optimal separating hyperplane [69]. Fisher's LDA aims at finding a set of weights θ that maximize the ratio:

$$J(\theta) = \frac{\theta^T S_B \theta}{\theta^T S_w \theta} \tag{2.20}$$

where S_B is the scatter matrix between classes and S_w is the scatter matrix within a class. S_B and S_w are defined as:

$$S_B = \sum_c (\mu_c - \overline{\mathbf{x}})(\mu_c - \overline{\mathbf{x}})^T \tag{2.21}$$

$$S_W = \sum_c \sum_{i \in c} (\overline{\mathbf{x}}_i - \mu_c)(\overline{\mathbf{x}}_i - \mu_c)^T \tag{2.22}$$

Note that maximizing $J(\theta)$ can be understood as projecting the data $\mathbf{x}$ to a new space where the class means are well separated and the variances of each class are minimized.

2.6.2 Logistic Regression

Logistic regression is a discriminative classifier. Assuming a training set of pairs of examples and labels $(\mathbf{x}, y)$ with $\mathbf{x} \in R^d$, the logistic regression assumes that the $logp(y/\mathbf{x})$ is a linear function of the examples $\mathbf{x}$ plus a normalization factor [70]:

$$p(y/\mathbf{x}) = \frac{1}{Z} e^{\{\theta_y + \sum_{i=1}^{d} \theta_{i,y} x_i\}} \tag{2.23}$$

where Z is calculated according to:

$$Z = \sum_{y'} e^{\{\theta_y + \sum_{i=1}^{d} \theta_{i,y} x_i\}} \tag{2.24}$$

Learning the parameters θ of the classifier can be done by means of the maximum likelihood, i.e. finding the set of parameters θ that maximize the logarithm of the expression in Equation 2.23. For this, optimization methods based on gradient ascent can be used.

The logistic regression possesses advantages over other classifiers. First, as a discriminative method it is better suited for classification as it models directly the conditional probability of the labels given the data, avoiding the need to establish the distribution of the data. Note that in this method, the data $\mathbf{x}$ are not taken as a random variable. Second, Logistic regression is less sensitive to unbalanced data than other classifiers such as the naive Bayes classifier. Third, logistic regression will find optimal parameters θ disregarding the correlation between the samples. This is beneficial in cases where $\mathbf{x}$ has a large dimension. Note that other methods like LDA will have difficulty finding the optimal parameters if the set of features are highly correlated. Finally, the probabilities provided by this method are well calibrated. This last feature is in contrast with other classifiers used in the BCI community, such as the Step Wise LDA which outputs scores and

not probabilities, thus preventing to define confidence intervals about the selection of a specific class.[71].

2.6.3 Support Vector Machines

The description of support vector machines (SVMs) presented here is based on [72]. The reader can refer to [72] for more details. In the separable case it is possible to find a hyperplane that separates positive samples from negatives. Given training data in the form x_i, y_i $i = 1...l$,where $y \in -1, 1$ and $x_i \in R^d$, the samples that fall in the hyperplane are found by:

$$W \cdot x + b = 0 \tag{2.25}$$

where W is a normal vector to the hyperplane and $b\frac{b}{\|W\|}$ is the distance from the hyperplane to the origin. The assignment of each sample to one of the two classes is made by:

$$y_i(x_i \cdot W + b) - 1 \geq 0 \tag{2.26}$$

The Lagrangian can be built by introducing a multiplier for each of the constrains in the Equation 2.26 which produces

$$L_p = \frac{1}{2}\|W\|^2 + \sum_{i=1}^{l} y_i(x_i \cdot W + b) + \sum_{i=1}^{l} \alpha_i \tag{2.27}$$

The problem of maximizing L_p can be reformulated as the minimization of L_p under the constraints that the gradient of L_p with respect to W and b vanishes and that the multipliers are positive ($\alpha_i \geq 0$). Two conditions are obtained:

$$W = \sum_i \alpha_i y_i x_i \tag{2.28}$$

$$\sum_i \alpha_i y_i = 0 \tag{2.29}$$

The dual is obtained as:

$$L_D = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \tag{2.30}$$

The solution obtained involves one multiplier for each sample point. The $\alpha_i > 0$ are called "Support Vector" and fall in the Hyperplanes H1 or H2; for the other sample points the multipliers are equal to zero. The support vectors are thus critical elements of the training set, because they determine the separating hyperplane.

2.7 Probabilistic Graphical Models

Graphical models provide a framework to encode the conditional probabilistic dependence between random variables. This framework contains, as special cases, many probabilistic methods such as Hidden Markov Model, Conditional Random Fields, etc. Also, many algorithms for inference, such as the forward-backward algorithm, can be seen as a particular case of more general methods. In this thesis, different approaches to BCI based on graphical models are presented. Graphical models provide a framework in which the different variables of a system and their relations can be represented in a principled probabilistic fashion. This framework also gives the opportunity to include knowledge of different areas in the structure of the graph, which could be translated in a better modeling of the process in hand.

A graphical model can be defined as $G = (V, E)$, where V represents a set of nodes or variables and E a set of edges connecting the nodes i and j. We start by describing undirected graphical models.

2.7.1 Undirected Graphs

In Undirected Graphical Models (UGM), given a set of random variables Y (nodes in a graph), the probability distribution over Y is represented as the product of factors of the form $\prod_{i=1}^{N} \Psi_i(Y_i)$, where N is the number of factors and Y_i is a subset of Y that affects the value of the factor Ψ_i. The factor $\Psi_i(Y_i)$ is a non-negative scalar that can be understood as the measure of compatibility between the variables in the subset Y_i. This means that it is easier to represent the probability density function (pdf) P of the random variables Y given that each factor Ψ_i depends of a subset of Y. The pdf can then be written as:

$$p(Y) = \frac{1}{Z} \prod_{i=1}^{N} \Psi_i(Y_i) \tag{2.31}$$

The factors are called local potentials or compatibility functions [70]. In Equation 2.31 Z is defined as:

$$Z = \sum_{Y} \prod_{i=1}^{N} \Psi_i(Y_i) \tag{2.32}$$

Z is a normalization term which makes $p(Y)$ a valid probability density function.

2.7.1.1 Log-linear Models

Given a set of examples x and a possible label for them y, a log linear model representing the probability of the labels given the examples is given by:

$$p(y/x, \theta) = \frac{1}{Z}\Psi(y, x) \tag{2.33}$$

where the local potentials Ψ are of the form

$$\Psi(y, x) = e^{\sum_i \theta_i f_i(x,y)} \tag{2.34}$$

where f_i is called a feature function. Note that the model assumes that the $logp(y/x, \theta)$ is a linear function of the feature functions. The linear function $\theta_i f_i(x, y)$ can take values from $[-\infty, \infty]$, the exponential makes those values range from $[0, \infty]$ and the normalization term Z limits the range to $[0, 1]$ making $p(y/x, \theta)$ a valid probability density function. The feature functions are real-valued functions depending both on the data and the labels. Each feature function is associated to a weight. The value of the weights capture information that defines the affinity of the attributes of the function with each label. This kind of Undirected discriminative models are of interest because the use of exponential families eases the calculation of log-likelihood functions. Note that although the exponential family may not be the best selection for certain types of data, in a graphical model with several nodes the conditional probability density function may take complex forms based on the expression in Equation 2.31. Conditional random fields are a particular case of log-linear models that will be used extensively in this thesis (see [71] for a more detailed description of log-linear models).

2.7.2 Directed Graphs

In undirected graphical models, the factors $\Psi_i(Y_i)$ do not necessarily represent probability density functions over the subset Y_i. In directed graphical models, the dependence between variables is indicated by directed edges (represented by arrows). Given a directed graph $G = (V, E)$, a node j is parent of a node k if there is a directed edge from j to k. The node k is then called a child of i. The probability density function $p(Y)$ in a directed graphical model can be facatorized as:

$$p(Y) = \prod_{j=1}^{M} p(Y_j/pa(Y_j)) \tag{2.35}$$

where M is the number of nodes in the graph and $pa(Y_j)$ represents the set of variables parent of Y_j. Note that $pa(Y_j)$ could be empty. A directed graphical model can be understood as an undirected graph in which $Z = 1$ and the factors Ψ represent local conditional probability distributions

Chapter 3

A Word-level Language Modeling Framework for the P300 Speller

One of the most widely studied BCI applications is spelling. This application targets patients with ALS who lost motor ability and rely on external mechanisms to communicate with other people and interact with the environment. The P300 speller, originally proposed in [4], makes use of the signals generated in the brain when a surprising event takes place (see Section 2.2.2). A set of characters are presented to the subject, arranged in rows and columns. The rows and columns are intensified randomly while the subject focuses his/her attention to the letter that he/she wants to communicate. A P300 potential is expected to be generated when the letter in which the subject is focusing his attention is intensified. The system then analyzes the signals in segments up to 800 ms following the intensification of a row or column and classifies the signal as P300 or non-P300. The row-column pair that has the higher probability to contain P300s defines the letter to be declared by the system.

The P300 speller system can be divided into three principal blocks: recording, pre-processing and classification. Most of the current P300-based BCI approaches make use of EEG recordings, although studies based on ECoG recordings have been presented [73, 74]. Basic approaches for pre-processing of the EEG signals for the P300 speller involve filtering of the signal in different frequency bands, and averaging across trials. However, given the low SNR of the EEG recordings, more advanced techniques that involve blind source separation have also been proposed (see [75] for a comparison of Blind Source Separation methods for P300 detection). For classification, linear methods such as Linear Discriminant Analysis (LDA), Step Wise LDA (SWLDA), Fisher LDA and linear kernel Support Vector Machines (SVM) have been used. (See [69] for a comparison of classification methods for P300). Although not a new idea [76], recently there has been a growing interest in the incorporation of language models in the P300 speller with the intention to increase the performance. This approach makes use of language

statistics as a prior, preventing the system from declaring sequences of letters that are unlikely in the language. Following this idea, a dictionary-based method was proposed in [77]. This speller auto-completes the words based on prior information obtained from a newspaper corpus. This method effectively increases the performance of the P300 speller by reducing the number of letters that the subject must type. However, it assumes that the first letters of the word are decoded correctly and in case of error the whole word will be decoded incorrectly. In [78] a solution to this problem is presented. This method classifies the EEG signal and outputs a word that is compared to the words in a custom dictionary. The word in the dictionary that is closest to the classifier output is then declared. This method assumes that the maximum number of misclassified letters in a word is 50% in an attempt to reduce the number of possible matches in the dictionary. The dictionary employed used a small subset of 942 words, all words with a length of four letters, which is restrictive. Recently, a Natural Language Processing (NLP) approach has been presented in [9]. In this approach each letter receives a score based on the output of the SWLDA. The scores are transformed into probabilities by assuming a Gaussian distribution. These probabilities are combined with language priors based on frequency statistics of the language. These statistics were simplified by assuming a second order Hidden Markov Model (HMM) structure in order to calculate 3-grams. One limiting factor in this work is that greedy decisions are made about letters, this means that once a letter is declared it is not possible to modify it based on new information obtained by new letters spelled by the subject. Given the dependence between letters in a word, if an error is made by the classifier, the error will propagate in the remaining letters of the word. Following the work in [9], in [10] a work using models of different orders shows that a 4-gram model provides the best results. Also in [10] is proposed that after the first letters of a word have been predicted, is possible to decrease the number of times that a letters should be flashed without compromising the performance. In this method as in [9], once a letter is declared, it is fixed. This issue was resolved in [11] where two generative methods are combined. A Bayesian LDA classifier is used to detect P300 vs Non-P300 and the estimated letters and their probabilities are the input of a second order HMM (3-grams). This allows to make inference in different ways. In particular, the use of the forward-backward algorithm for inference allows the method proposed in [11] to correct previously declared letters if current information (posterior spelled letters) support that change.

Motivated by the work described above we propose a discriminative graphical model framework for the P300 speller. The proposed model integrates all the elements of the BCI system from the input brain signals to the spelled word in a probabilistic framework. The language is modeled at the level of words which as we present in the results, has a better effect in the performance compared to n-grams. Also in cases where the subject has serious limitations the communication of a reduced set of works is necessary. In such cases a dictionary based method provides a mechanism for efficient communication. Furthermore, classification and language model are integrated in a single model and the

structure of the graph allows efficient inference methods making the system suitable for online applications. Also, the discriminative nature of the model avoids the need for imposing a particular probabilistic density distribution over the brain signals. Finally, the proposed method does not make greedy decisions about the letters spelled. This means that it has the ability to correct previously declared letters if current information, given by a current letter decoded, provides evidence that an error has occurred in previous classification outputs.

3.1 Proposed method

In this work, we propose a discriminative graphical model framework that is capable of including language model at the level of words rather than letters. Including the modeling of the language at the level of words has the potential to allow the system to predict the target word. More importantly, it imposes a prior over the sequence of spelled letters, providing the system with the ability to correct errors and to reduce the number of trials needed to achieve acceptable performance for communication.

We choose discriminative methods over generative ones because the formers are more suited for classification. While both generative and discriminative models describe distributions over a set of variables (i.e x and y) each of them are subject to practical considerations that provide advantages and disadvantages [70]. In particular, a generative model that models the joint probability density function $p(x, y)$ can be used to model the posterior probability $p(y/x)$ using the Bayes's rule, which is in also the goal of the discriminative model. However, to obtain $p(y/x)$ by means of Bayes's rule, one needs the distribution $p(x)$ [70]. The true probabilistic density function of the data is rarely known, giving an advantage to discriminative models in classification tasks. In the particular case of BCI applications, the use of generative models implies the modeling of the distribution of brain signals. Characteristics such as non-linearity and non-stationary make this a difficult task.

3.1.1 Overview of the Proposed Graphical Model

The proposed model is shown in Figure 3.1. The model represents a hierarchical structure where different aspects of the P300 speller system are integrated. The bottom layer (first layer) represents the EEG signal. The variables $x_{i,j}$ represent the EEG signal recorded during the intensification of each row and column (a total of twelve variables for each spelled letter). The index i is used to identify the number of the letter being spelled and the index j represents a row or column (values of j from 1 to 6 represent columns in the spelling matrix and values of j from 7 to 12 represent rows). The second layer contains a set of twelve variables $c_{i,j}$ indicating the presence or absence of the P300 potential in

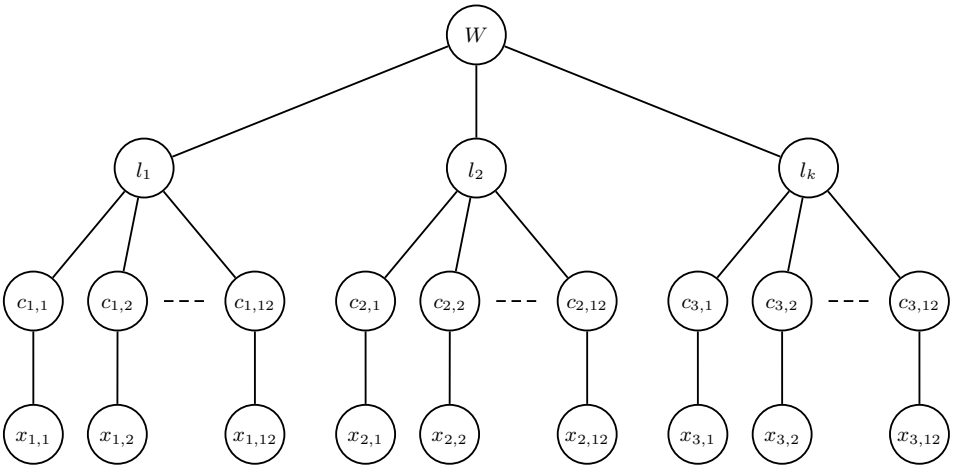


FIGURE 3.1: Proposed graphical model framework for the P300 speller

each row and column. Each $c_{i,j}$ is a binary variable taking values form the set $C \in \{0, 1\}$. The sub-graph formed by the variables $c_{i,j}$ and $x_{i,j} \in R^d$ and the edges between them encodes conditional dependence that can be expressed as:

$$c_{i,j}\big|x_{i,j} \sim F_{1,j}(\theta_1) \tag{3.1}$$

where $F_{1,j}$ is a probability density function with parameters θ_1. Note that the structure of the graph implies that for each row and column a different set of parameters should be learned. This feature of the model is beneficial if there is evidence that the characteristics of the P300 change according to the spatial position of the row or column that is highlighted. However, the parameters can be shared between nodes $c_{i,j}$, which is equivalent to learning a single set of parameters making use of all available data and then using those parameters in the model of the conditional dependence of each pair of variables $c_{i,j}$ and $x_{i,j}$.

The third layer contains variables l_i representing the letter being spelled. The variables l_i are related to the variables $c_{i,j}$ in the same fashion that is done in traditional P300 speller systems: the presence of a P300 potential in a particular pair of row-column encodes one letter. However, given that the detection of P300 potentials is not perfect (false detection or miss-detection of P300 potentials), a probabilistic approach is taken:

$$l_i\big|c_{i,j} \sim F_2(\theta_2) \tag{3.2}$$

where F_2 is a probability density function with parameters θ_2.

The fourth layer contains the variable w which represents valid words in the English language. This variable is used as a prior on the language. The conditional dependence between w and l_i can be expressed as:

$$w|l_i \sim F_3(\theta_3) \tag{3.3}$$

where F_3 is a probability density function with parameters θ_3. Note that at the level of the variable w the system predicts the target word based on the current number of letters spelled. At the level of the variables l_i the variable w imposes a prior on the sequence of letters which has the potential to reduce the error rate by forcing the sequence of letters to be a valid sequence in the language. Furthermore, the system does not make greedy assignments which implies that when a new letter is spelled by the subject, this information can be used to update the belief about the previously spelled letters.

3.1.2 Detailed Description of the Proposed Model.

Following the basic ideas from undirected graphical models (see Section 2.7.1), the pairwise distributions of all the variables in the model ($w, \mathbf{l} = \{l_1, ..., l_k\}, \mathbf{c} = \{c_{1,1:12}, ..., c_{k,1:12}\}$) given the observations ($\mathbf{x} = \{x_{1,1:12}, ..., x_{k,1:12}\}$) can be written as a product of factors over all the nodes and edges in the graph:

$$P(w, \boldsymbol{l}, \boldsymbol{c}/\boldsymbol{x}) = \frac{1}{Z}\Psi_4(w)\prod_i \left\{ \Psi_3(i, w, l_i) \prod_{j=1}^{12} \{\Psi_2(j, l_i, c_{i,j})\Psi_1(j, c_{i,j}, x_{i,j})\} \right\} \tag{3.4}$$

where Ψ_4 is the potential function for node w, Ψ_3 is the potential function for the edge connecting node w with node l_i, Ψ_2 is the potential function for the edge connecting node l_i with node $c_{i,j}$, Ψ_1 is the potential function for node $c_{i,j}$ and Z is the partition function, which is a normalization factor. The potential functions in Equation 3.4 are defined as follows:

$$\Psi_4(w) = e^{\theta_4 f_4(w)} \tag{3.5}$$

$$\Psi_3(i, w, l_i) = e^{\theta_{3i} f_3(i,w,l_i)} \tag{3.6}$$

$$\Psi_2(j, l_i, c_{i,j}) = e^{\theta_{2j} f_2(j,l_i,c_{i,j})} \tag{3.7}$$

$$\Psi_1(j, c_{i,j}, x_{i,j}) = e^{\sum_{m=1}^{d} \theta_{1j,m} f_{1m}(j,c_{i,j},x_{i,jm})} \tag{3.8}$$

where d is the dimensionality of the data. The parameter θ_4 is a vector of weights of length equal to the number of states of the node w (number of words in the dictionary used). The product $\theta_4 f_4(w)$ models a prior for the probability of a word in the language with the feature function $f_4(w) = \mathbf{1}_{\{w=w'\}}$. The notation $\mathbf{1}_{\{w=w'\}}$ denotes an indicator function of w that takes a value of 1 when $w = w'$ and 0 for any other value of w.

The product $\theta_{3i} f_3(i, w, l_i)$ models a prior for the probability of a letter l_i appearing in the position i of a word. The feature function $f_3(i, w, l_i) = \mathbf{1}_{\{w(i)=l_i, l_i=l'\}}$.

The product $\theta_{2j} f_2(j, l_i, c_{i,j})$ measures the compatibility between the variable $c_{i,j}$ and the variable l_i with the feature function $f_2(j, l_i, c_{i,j}) = \mathbf{1}_{\{C(l_i,j)=c_{i,j}, c_{i,j}=c'\}}$ where C is a code-book that maps the intersections of rows and columns in the spelling matrix to letters. For instance, the entry for A in the code-book assuming the spelling matrix in Figure 2.2 is $C(A, 1:12) = \{100000100000\}$.

The product $\theta_{1j,m} f_{1m}(j, c_{i,j}, x_{i,j_m})$ is a measure of the compatibility of the component m of the EEG signals $x_{i,j} \in R^d$ with the variable $c_{i,j}$ (presence or absence of P300). The feature function $f_{1m}(j, c_{i,j}, x_{i,j_m}) = x_{i,j_m} \mathbf{1}_{\{C_{i,j}=c'\}}$.

Learning in the model corresponds to finding the set of parameters $\Theta = \{\theta_4, \theta_3, \theta_2, \theta_1\}$ that maximize the log-likelihood of the conditional probability density function described in Equation 3.4. Given that the structure of the model does not involve loops, inference in the model can be made using the belief propagation algorithm which can efficiently provide the marginals of interest:

$$P(w/\boldsymbol{l}, \boldsymbol{c}, \boldsymbol{x}) = \sum_{\boldsymbol{l}} \sum_{\boldsymbol{c}} P(w, \boldsymbol{l}, \boldsymbol{c}/\boldsymbol{x}) \tag{3.9}$$

$$P(\boldsymbol{l}/w, \boldsymbol{c}, \boldsymbol{x}) = \sum_{\boldsymbol{w}} \sum_{\boldsymbol{c}} P(w, \boldsymbol{l}, \boldsymbol{c}/\boldsymbol{x}) \tag{3.10}$$

Such marginal can be used respectively to declare the word or letter that the subject intents to communicate. Finally, a word is declared according to:

$$\overline{w} = \arg\max_{w} P(w/\boldsymbol{l}, \boldsymbol{c}, \boldsymbol{x}) \tag{3.11}$$

and in the same fashion letters are declared according to:

$$\overline{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} P(\boldsymbol{l}/w, \boldsymbol{c}, \boldsymbol{x}) \tag{3.12}$$

3.2 Description of Experiments and Methods

Problem and Data Set Description.In a typical P300 spelling session, the subject sits up right in front of a screen observing a matrix of letters as shown in Figure 2.2. The task involves focusing attention on a specific letter of the matrix and counting the number of times that the character is intensified (see Section 2.2.2 for more details). The EEG signals were recorded using a cap (Electrode cap International Inc) embedded with 64 electrodes according to the 10-20 standard. All electrodes were referenced to the right earlobe and grounded to the right mastoid. All aspects of the data collection and experimental control were controlled by the BCI2000 system [79]. From the total set of electrodes a subset of 16 electrodes in positions F3, Fz, F4, FCz, C3, Cz, C4, CPz, P3, Pz, P4, PO7, PO8, O1, O2, Oz were selected, motivated by the study presented in [27]. The classification problem is to declare one letter out of 26 possible letters in the alphabet. In total each subject spelled 32 letters (9 words). The data was divided in training session and testing session. Training and testing sessions were recorded in different days on the same subjects.

In this work, sessions are divided into trials. For each trial, the subject is requested to spell a word. This means that the beginning and the end of the trial are known a priory. The classification performance measure used is accuracy and is based in correctly predicted number of letters rather than the number of corrected words.

Signal Pre-processing Segments of 600ms following the intensification of each row or column where calculated. For each segment and each electrode, the EEG signal was initially de-trended by removing a linear fit from the signal. The de-trended signals where then filtered between 0.5Hz and 8Hz using a zero-phase IIR filter of order 8 and decimated according to the high cutoff frequency of the filters. Signals from all electrodes were concatenated and used as the inputs for the classifier during training. For testing, the segments were averaged across repetitions (up to 15 repetitions) and fed to the classifier which allows to determine the performance as a function of the number of repetitions.

Model Selection

Referring to the Section 3.1.2, the parameters $\theta_4, \theta_{3i}, \theta_{2j}$ in Equations 3.5, 3.6, 3.7 respectively are independent of the brain signals and can be learned independently. The language-dependent parameters θ_4 are learned by calculating the relative frequency of each word in the language. The parameters θ_3 are learned by calculating the relative frequency of each letter appearing in the i-th position of all words. These statistics can be learned from a text corpus. However, the structure of the model allows to select a dictionary based on the specific application of the BCI system. This means that the number of words in the dictionary can be adjusted to satisfy particular requirements of the application. In this work, the statistics about the language were calculated using

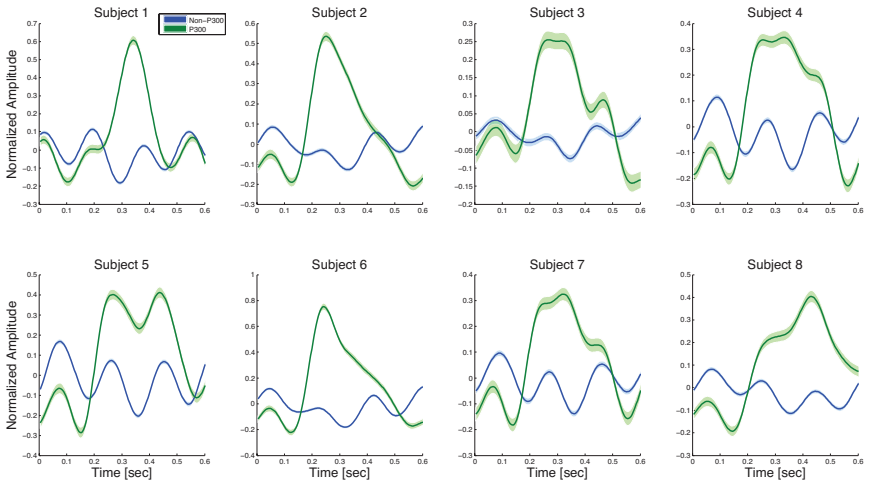


FIGURE 3.2: Mean and mean error of the normalized P300 and Non-P300 signal amplitude

the Corpus of Contemporary American English [80] which contains 450.000.000 words. The dictionary was then built using the 5000 words most frequently used in the English language. Note that further reducing the number of words has a positive impact on the performance of the BCI system. The node w on Figure 3.1 contains 5000 states, one for each word in the dictionary. As described previously, the parameters θ_2 represent a set of decoding vectors that map rows and columns to a letter in the spelling matrix. Note that the parameters θ_2 do not depend on i, the position of a letter in a word.

Once the parameters $\theta_4, \theta_3, \theta_2$ are fixed, the parameters θ_1 remain to be learned. In order to obtain a robust estimation of the parameters θ_1, the parameters are shared across nodes. This assumes that the generation of the P300 is independent of the position of the letter in the matrix. Therefore, the parameters $\theta_{1j,m}$ are the same for any value of j. For learning, we use an non-linear optimization method based on the BroydenFletcherGoldfarbShanno (BFGS) algorithm.

3.3 Results

The mean and mean error of the P300 and non-P300 trials are shown in Figure 3.2, where the electrode Cz has been used out of the 16 electrodes available. Note that for all subjects, the change in signal amplitude in the trials that contain P300 is clear. However, the exact time at which the maximum amplitude peak is observed is subject dependent. The separability between P300 and Non-P300 responses can be determined by calculating the correlation between each time point of the signal across trials with the labels. Given that more than one electrode is available, the correlations are better displayed by the topographical distribution of their squared value r^2. The r^2's distributions are displayed

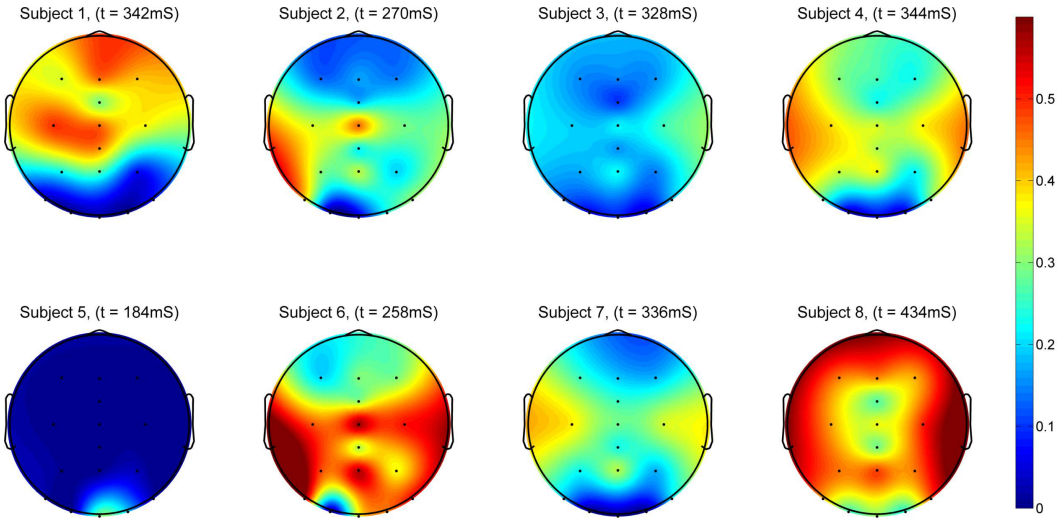


FIGURE 3.3: Topographical r^2 values for all subjects.

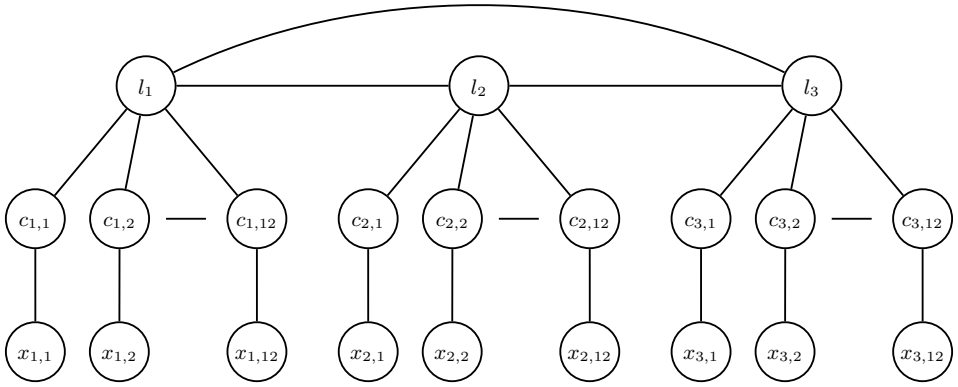


FIGURE 3.4: Example of a 3-gram Model for a 3 letters word.

in the Figure 3.3 at the time point when the maximum value across electrodes is achieved. In this plot all the trials related to one letter in the training session were averaged (15 repetitions per letter). A value of r^2 equal to 1 implies perfect classification. Note also that the central position (Cz) is in average the position with higher r^2. However, electrodes in the occipital region also contain useful information for classification. In general, based on the information obtained by the r^2's values a subject-dependent electrode location can be done. A high resolution for the topographic distribution is obtained by interpolating the data in a Cartesian grid.

The proposed method is compared to a letter-level language model-based BCI approach using 3-grams for modeling letter sequences. The classification of P300 potentials is left unchanged and the modeling of the language is made based on a 3-gram method that makes use of language statistics of the sequence of 3 letters in the language. The graphical model for the 3-gram methods is presented in the Figure 3.4. Also for reference, results based on a common method used in the context of P300, the Step Wise LDA (SWLDA), are also shown. Classification results are presented in Figure 3.5. Note

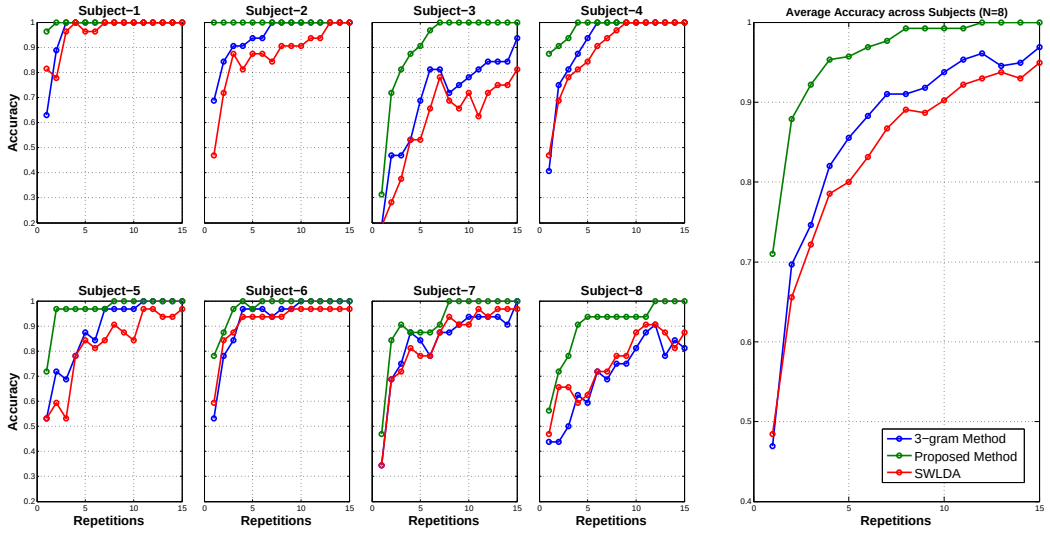


FIGURE 3.5: Comparison of performances between different classifiers

that for all subjects the proposed method provides better performance, obtaining in average an Accuracy of 99% in all subjects with as few as eight repetitions. In order to verify the results correct classification accuracy, a statistical test was performed. A repeated measures ANOVA on the performance results reveals significant difference ($F(2, 14), \epsilon = 0.74, p = 0.0019$, see Table 3.1 for details of the statistical analysis [81]) between the three methods compared). Using a post hoc Tukey-Kramer test, the proposed method performs significantly better ($p < 0.01$) than the 3-gram based method and than SWLDA, even so the 3-gram based method (as expected) performs significantly better ($p < 0.01$) than SWLDA which does not make use of language model.

It is important to note that the results observed in terms of performance are expected. The use of the language model at the level of words as presented here has the advantage that it is known that the target words are in the dictionary, and the proposed method exploits that. If the word spelled is not in the dictionary then the word-level model may not exhibit a performance gain over an n-gram model. This would be a case in which the prior model does not reflect the actual distribution of the random phenomenon to be estimated. The n-gram model could suffer from a similar issue if the conditional probabilities it uses do not accurately represent those in the test data. Nevertheless, in terms of our word-level model, although at the level of the variable w in Figure 3.1 only words existing in the dictionary can be predicted, at the level of the variables l_i words that do not exist in the dictionary can be obtained if there is strong evidence for them from the brain signals.

3.4 Conclusion

The goal of this work was to assess the efficiency of the integration of language models at the level of words rather that comparing the proposed method to the state-of-the-art BCIs based on P300 potentials. The SWLDA has been used for comparison given that it is a common method for classification in BCIs. Although the SWLDA-based approach does not involve language models, the proposed method addresses some of the limitations that SWLDA presents.

SWLDA makes use of the training data to learn a discriminant function between P300 responses and no response. During testing, using the brain signal features, a score is calculated determining the similarity of the observed signal with a P300 response. However, the classification of the spelled letter is only indirectly associated with the capacity of the classifier to detect P300 responses. Each row and column are associated to a vector of brain signal features and a score is assigned to each row and column. Predicting the spelled letter then corresponds to choosing the row and column with the highest score in spite of the possibility that the magnitude of the score of each row and column is low enough to assume that no P300 response has been detected. Note that this is the same mechanism used in most of the P300 BCIs systems described in the literature [69]. Furthermore it is difficult to determine a confidence interval for the scores obtained with this method given that this is not a probabilistic approach.

The proposed method attempts to solve several of these problems by integrating in a probabilistic model all variables in the P300 system. This framework would allow one to build up any language model in a consistent way and can be used to model language characteristics beyond the relative frequency of letters as proposed in [82]. The 3-gram method here presented makes use of a probabilistic framework as well and is presented here for comparison given the popularity of the language models based on n-grams [9, 11, 10]. Note that methods based on n-grams found in the literature need to incorporate separated modules for classification of the P300 and language modeling. The proposed method models in an integrated way the interactions between variables in the P300 speller system, from the level of the brain signals to the level of words. Its construction allows the parameters of the model to be learned independently, which reduces the complexity of the learning process. The top layer includes features of the language, the second layer includes information inherent to the construction of the speller (equivalence between column/rows and letters) and the bottom layer maps the brain signals into two states (i.e. P300 response vs. non-P300 response). The main differences with other methods that include language modeling by means of n-grams (i.e. 3-grams) are 1) the proposed method models directly the word spelled by the subject, using all the information available during the spelling of a word. This means that greedy decisions on the spelled letters are not made and the probability of each letter is used to determine the whole word. 2) The language can be adapted to the particular situations of the

Sphericity	F	ϵ	Degrees of freedom	$p - value$
Sphericity Assumed	13.98	1	(2,14)	< 0.0001
Greenhouse & Geisser	13.98	0.74	(1.480,10.359)	0.0019
Huynh and Feldt (Lecoutre)	13.98	0.89	(1.782,12.475)	0.0055

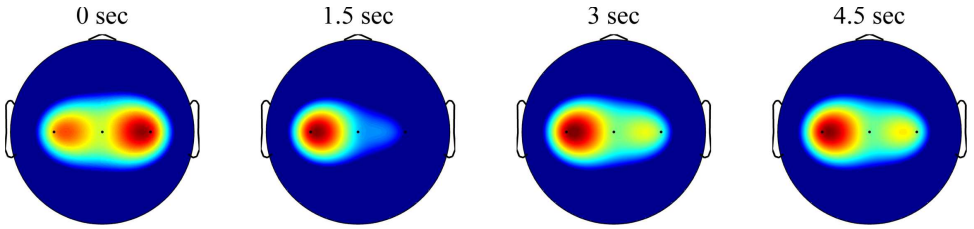
TABLE 3.1: Repeated measures ANOVA statistical tests from comparison of the proposed method

subject by limiting the size of the number of words in the language. As a result the performance of the system is increased, which would allow us to reduce the number of repetitions needed to achieve a level of practical usability of the system. Furthermore, given that the structure of the model has been carefully designed to avoid loops, efficient inference algorithms such as belief propagation can be used without compromising the use of the model in online applications. It is worth noting that the features used in the proposed model are not fixed: different approaches can be incorporated by redefining the feature function $f_{1m}(j, c_{i,j}, x_{i,j_m})$ in Equation 3.8 to include any state-of-the-art signal processing method.

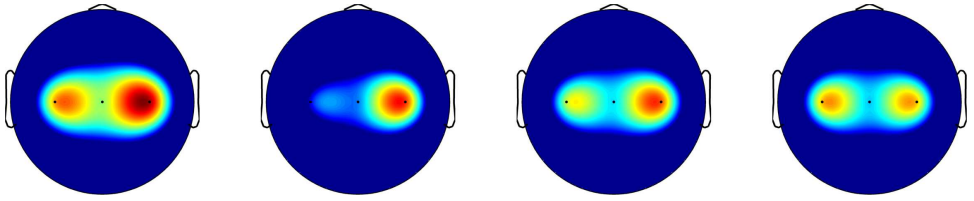
Chapter 4

Generative Graphical Models for Synchronous BCIs

A Brain Computer Interface (BCI) is a system that provides an alternative communication way for people who suffered a disease or an accident that compromises their ability to perform motor tasks. Also, applications for healthy subjects in areas of multimedia and gaming started to incorporate these technologies in the last years[2]. BCIs make use of the brain signals to control external devices that help the subject to communicate and interact with the environment. Current approaches to BCIs are based on the comparison of the values of power of the EEG during the execution of imaginary motor tasks. However, the well-known phenomena of Event Related Synchronization and Event Related De-synchronization [13] provide more information that can be employed to improve the performance of the BCIs. This information is related not only to the difference of power of the signals but their change in time for different frequency bands 4.1. For this, algorithms that take into account the changes of the signal on time as Hidden Markov Models (HMM) [14, 18] have been used in combination with features that describe the temporal behavior of the EEG signals [83, 19]. Although the Time-Frequency analysis of the EEG signals have shown good results in previous works [84, 85, 86, 87, 88] in BCI applications, a combination of the time-frequency power distribution of the signals and algorithms that take into account the changes in the distribution have not been reported. One possible reason for this is that the selection of the parameters of the models (states, Gaussian mixtures, etc. in HMM) along with the selection of the frequency bands becomes problematic. In this work we propose the use of the Time-Frequency distribution of the power of the signals, using Autoregressive Models for calculation of the Power Spectral Density (PSD) and HMM for classification of two different motor tasks. The problem of selection of parameters of the model is handled in three different ways: 1) Prior information based on the description of the task is used to select the number of states and Gaussian mixtures, 2) Parameters are selected based on exhaustive search using cross-validation on the number of states and the number of Gaussian mixtures,



(A) Class one: Left hand



(B) Class two: right hand

FIGURE 4.1: Scalp topographical distribution of the power during the execution of two different imaginary motor tasks.

and 3) A Non-parametric Bayesian HMM for BCI applications is proposed, inferring the parameters directly from the training data.

4.1 HMM Approach

A HMM is a finite automaton which contains a set of discrete states H emitting a feature vector x_t at each time point t. Given that these kind of models are generative, it is necessary to determine the joint probability over observations and labels, which requires all possible observation sequences to be enumerated. In order to make the inference problem tractable, conditional independence is assumed, meaning that the future states are independent from the past states given that the current state is known. The structure of the state sequence is described by

$$h_t|h_{t-1} \sim \pi_{h_{t-1}} \tag{4.1}$$

Where $\pi_{h_{t-1}}$ is a state dependent transition distribution for the state h_{t-1}. This also implies that given h_t, the probability for the emission x_t is independent of observations and states at different time points

$$x_t|h_t \sim F(\theta_{h_t}) \tag{4.2}$$

Where θ_{h_t} are the emission parameters related to the state h_t.

The learning problem is defined as a search for the parameters that maximize the log-likelihood of the observation. These parameters are the state transition matrix π, for which each entry $\pi_{i,j}$ represents the probability to pass from state i to state j, the

vector of initial probabilities π^0 so that $h_1 \sim \pi^0$ and the parameters θ of the emission distribution F for each state. In this work, the emission distribution for each state is modeled using Gaussian mixtures.

The joint probability of the observations and states is given by:

$$P(h_{1:n}, x_{1:n}) = \pi^0(z_t)p(x_1/h_1)\prod_{t=2}^{n} p(x_t/h_t)p(h_t/h_{t-1}) \tag{4.3}$$

The joint distribution of the HMM is represented by the graph in the Figure 4.2. The hidden states h represent different states during the execution of specific mental tasks. Such states are related to the power levels of the signal. The assumption of these different states during the execution of the task is justified by the time course of the EEG power in specific frequency bands. Figure 4.1 shows the average spatial distribution of the scalp EEG power in the alpha band for one subject, using bipolar references over electrodes C3, Cz and C4 (see subsection 2.3.1.2) at different time points for two different classes. This reveals the dynamics (sequence of states) of the signal during the execution of different motor tasks. The sequence of these states is modeled by a HMM for each class using training data. Inference in the model is done using the *forward-backward algorithm* [89] which provides an efficient scheme for passing messages in the graph making possible the calculation of marginals for problems of filtering $p(h_t/x_1, ..., x_t)$, prediction $p(h_{t+\tau}/x_1, ..., x_t)$ for $\tau > 0$ and smoothing $p(h_t/x_1, ..., x_\tau)$ for $\tau > t$.

In the BCI problem studied in this chapter, the decision about the task executed by the subject is made at the end of the trial. Therefore, the problem to be solved is one of filtering. Model parameters (π, θ) are learned from training data using the Expectation-Maximization algorithm [90]. During the expectation step the expected value of the likelihood of the model is calculated using the *forward-backward* algorithm. During the maximization step, gradient search is used to maximize the expected value of the likelihood.

The selection of the number of hidden states and Gaussian mixtures is made in two ways: 1.) Based on the description of the task, the number of hidden states is set to three, representing start, execution and ending of the task. The number of Gaussian mixtures was fixed to two as a reasonable tradeoff between the expressive power of the model and its simplicity. 2) Three-fold Cross-validation using the training data, looking for a combination of hidden states and number of Gaussian mixtures that maximizes a cost function, that in this work is selected to be accuracy.

Once the number of hidden states and the number of Gaussian mixtures is fixed, for each class, one model is trained using all the training data available. During testing, the label assigned to each sequence is determined by calculation of the likelihood of the data on each model. The model with higher likelihood determines the assigned label.

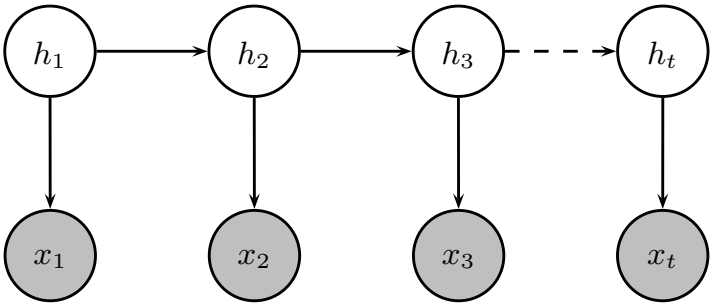


FIGURE 4.2: Graphical model representation of a HMM

4.2 Bayesian Nonparametric HMM Approach

The motivation for using HMM-like models is to extract temporal information from the brain signals. The underlying assumption is that during the execution of the task the rhythms produced by the brain go through a sequence of states. We link these states to changes in the power of the brain signals in specific frequency bands. However, the number of states is not known a priori. Furthermore, during each one of these states the distributions of the brain signals are complex and therefore, are better modeled by multimodal distributions. Again, the number of components of each multimodal distribution is unknown. The Hierarchical Dirichlet Processes HMM (HDP-HMM) with Dirichlet Processes Gaussian Mixtures present a solution to those problems faced with conventional HMM approaches, by solving the problem of selection of the model order parameters (number of hidden states and number of mixtures).

4.2.1 The HDP-HMM and the Sticky HDP-HMM

In [91] a nonparametric Bayesian approach to HMM in which the HDP defines a prior distribution on transition matrices over countably infinite state spaces is presented. The HDP-HMM leads to data-driven learning algorithms which infer posterior distributions over the number of states. One serious limitation of the HDP-HMM presented by [91] is that the model has a tendency to produce unrealistic rapid dynamics. In [92] a modified HDP-HMM so called sticky-HDP-HMM is proposed, augmenting the original HDP-HMM by including a parameter for self-transition bias and placing a separate prior on this parameter.

A DP denoted by $\mathrm{DP}(\gamma, H)$ is a distribution over countably infinite random measures:

$$G_0(\theta) = \sum_{k=1}^{\infty} \beta_k \delta(\theta - \theta_k) \qquad \theta \sim H \tag{4.4}$$

on a parameter space Θ. The weights β_k are sampled via a stick-breaking construction, so that $\beta \sim \text{GEM}(\gamma)$ where GEM$(\cdot)$ denotes the stick breaking construction:

$$\beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l) \qquad \beta'_k \sim Beta(1, \gamma) \tag{4.5}$$

The HDP proposed by [91] extends the DP to cases in which groups of data are produced by unique, generative processes [92]. The HDP places a global Dirichlet process prior DP(α,G_0) on Θ and draws group specific distributions:

$$G_j(\theta) = \sum_{k=1}^{\infty} \pi_{jk} \delta(\theta - \theta_{jk}) \qquad \pi_j \sim DP(\alpha, \beta) \tag{4.6}$$

where π_j represents the transition distribution for a specific state, h_t denotes the state of the Markov chain at time t. Figure 4.3 shows the Graphical model for the Sticky HDP-HMM. Given the properties in the HMM the state h_{t-1} indexes the group to which the observation x_t is assigned and the current state h_t defines the parameter θ_{h_t} used to generate the observation x_t. The modification proposed by [92] allows to incorporate prior information that slow, smoothly varying dynamics are more likely. Unlike the original HDP-HMM, this modified version does not allow state sequences with unrealistic fast dynamics to have large posterior probability. To that end, the work in [92] proposes to sample transition distributions as follows:

$$\pi_j \sim DP\left(\alpha + \kappa, \frac{\alpha\beta + \kappa\delta_j}{\alpha + \kappa}\right) \tag{4.7}$$

This modification for κ values greater than zero, has the effect of increasing the prior probability of self transitions $E[\pi_{jj}]$. When κ is equal to zero the original HDP-HMM is obtained.

In many applications the distribution of the data associated with each hidden state is complex and is better modeled with a multimodal distribution. This motivates the use of HMM in which each hidden state is associated with a mixture of Gaussian distributions, which poses the problem of selection of the number of mixtures. Just like the model order problem associated with the number of states, this problem can be solved by using DP as well. The Sticky HDP-HMM is extended in [92] by defining a DP mixture of Gaussians, including a new variable s_t (see Figure 4.4) which indexes the Gaussian mixture component of the h_t^{th} hidden state. Therefore for the variables in Figure 4.4, we have:

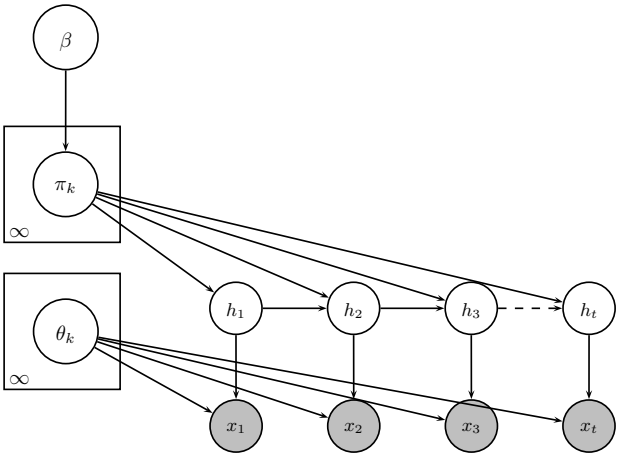


FIGURE 4.3: Sticky HDP-HMM Graph

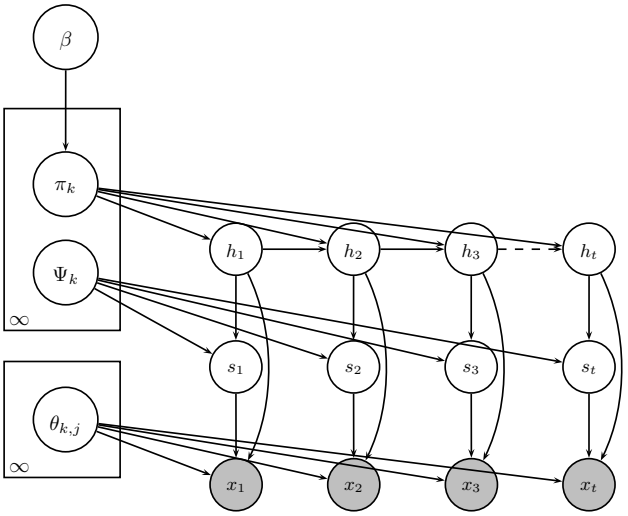


FIGURE 4.4: Sticky HDP-HMM Graph with DP Gaussian Mixtures

$$s_t \Big| \{\psi_j\}_{j=1}^{\infty}, h_t \sim \psi_{h_t} \tag{4.8}$$

$$x_t \Big| \{\theta\}_{k,j=1}^{\infty}, s_t \sim F(\theta_{h_t,s_t}) \tag{4.9}$$

We model the brain signals by fitting one HDP-HMM with DP Gaussian Mixtures model for each class. For the Gaussian Mixtures, a Gaussian prior was placed on the mean while an Inverse-Wishart prior was placed on the variance of the components of each mixture following [92]. Classification is performed in the same fashion that for the HMM approach

in Section 4.1. Using test data, calculation of the log-likelihood of the data given the model is performed. The model that is more likely to produce the data determines the declared class.

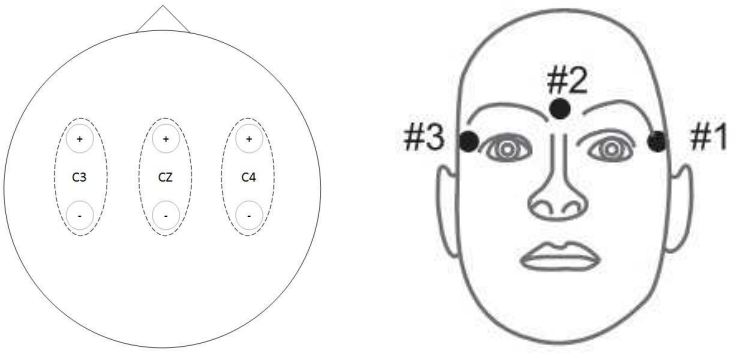
4.3 Description of Experiments and Methods

Problem and Data Set Description. In typical BCI applications based on the imagination of motor activity, the subject is requested to execute imaginary motor tasks following a visual cue. It is known that the imagination of motor activities produces synchronization and/or desynchronization of the electrical signals recorded over the motor cortex and that this process has an asymmetrical spatial distribution during the imagination of the motor task (e.g., imagination of movement of a particular leg produces changes in the power of electrical signals in the contra-lateral region of the brain, see Figure 4.1). Given data obtained from an initial session in which the subject is requested to execute different motor tasks, a model is trained. Then, given some new (test) data, the task is to run an inference algorithm to perform classification of the imaginary motor task.

In this work, Data Set 2b of BCI competition IV [93], which consists of bipolar EEG recordings over scalp positions for electrodes C3, Cz and C4 (see Figure 4.5a) in 9 subjects, was used. The cue-based BCI paradigm involved two classes, represented by the imagination of the movement of the left hand and the right hand, respectively. The time scheme of the sessions is depicted in Figure 4.6. At the beginning of each trial, a fixation cross and a warning tone are presented. Three seconds later, a cue (indicating left or right movement) is presented and the subject is requested to perform the imaginary movement of the corresponding hand. The data set contains five sessions, three for training and the remaining two for testing. Some of these sessions involved feedback, indicating to the subject how well the imagination of the motor task has been executed, and others did not. In our work we have used the sessions with feedback because previous work have shown that temporal behavior of the EEG signals could be modified due to the feedback influence [94].

The data was recorded in 3 different sessions in different days. The training session contains 160 trails, 80 trials for each class (imaginary right hand movement, imaginary left hand movement). The testing data consist on 2 sessions, each one with 160 trials equally divided between the two classes.

Artifact Reduction. Linear regression was used in order to reduce the interference of EOG signals in the EEG recordings. EOG data recorded at $N = 3$ channels at electrode locations shown in Figure 4.5b provide a measure of the eye movements executed by the subjects. In this approach, the signal recorded by the EEG electrodes is modeled as the summation of the actual underlying EEG signal and the noise, represented by a linear



(A) EEG channels (B) EOG channels

FIGURE 4.5: Electrode positioning for the BCI competition IV data set 2b.

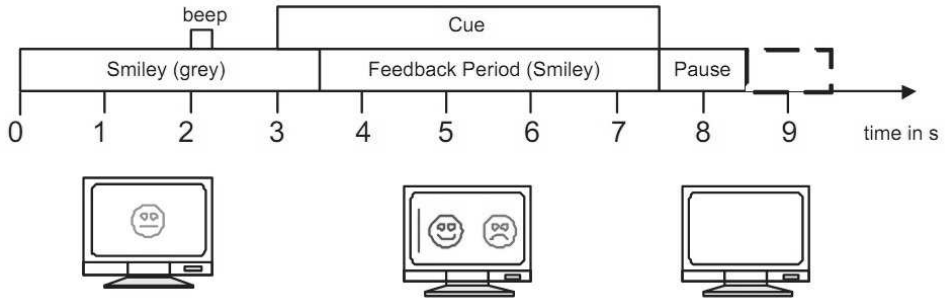


FIGURE 4.6: Time scheme for the experimental procedure.

combination of the EOG signals interfering into the EEG electrodes. The coefficients that explain how the EOG signal affects the EEG signal are calculated using the procedure described in Section 2.3.2.

Before the beginning of the motor task classification sessions, subjects are requested to execute different ocular movements enabling the estimation of the artifact-free EEG signal $s(n)$ based on data $w(n)$ and $u(n)$ recorded by EEG and EOG electrodes respectively (see Figure 4.5). As a complementary step, the obtained signals are band-pass filtered in the frequency band of interest for real/imaginary motor activity (8Hz - 35Hz), eliminating other sources of artifact in the low and high frequency range of the EEG recordings.

Spatial Filter. Spatial filters can be incorporated into the preprocessing stage by means of Common Spatial Pattern (CSP). Given a two-class classification problem related to motor task, CSP provides a method to maximize the power of the signal during the execution of the class 1, while minimizing it during the execution of class 2 and vice versa. Given the artifact-free signal $s(n)$ the spatial filtered signal $c(n)$ is obtained by:

$$c(n) = s(n)W, \tag{4.10}$$

EEG rhythm	Frequency band (Hz)
Alpha	08-13
Sigma	11-15
Low Beta	18-23
High Beta	21-26
Low Gamma	25-35

TABLE 4.1: Selected frequency bands used as features.

where W is a matrix for which each column represents a set of subject-specific spatial patterns. The procedure for calculation of the spatial filters with CSP is described in Section 2.3.4.1. The signal $s(n)$ composed of three electrodes is linearly transformed into 2 CSP-filtered EEG signals $c(n)$ to be used in subsequent preprocessing stages.

Spectral Power Estimation. The power spectrum of the signal is computed by parametric methods involving the calculation of autoregressive (AR) models of the signal. Burg's method for AR model estimation was used because it provides better stability than the Yule-Walker method, by minimizing the error in backward and in forward direction [95]. The power spectrum of the EEG signal is estimated as the frequency response of the auto-regressive model:

$$c_i(n) = \sum_{k=1}^{p} a_k c_i(n-k) + g(n). \tag{4.11}$$

where the sub index i represents each of the CSP components, n represents the discrete time index, p is the model order, a_k is the k^{th} coefficient of the model and $g(n)$ is the system input or noise function. The system function in the z-domain can be expressed as:

$$H_i(z) = \frac{C_i(z)}{G(z)} = \left(1 - \sum_{k=1}^{p} a_k z^{-k}\right)^{-1} \tag{4.12}$$

The AR spectrum can be obtained by evaluating $H_i(z)$ on the unit circle where $z = \exp(j\omega)$ [96]. For estimating the AR parameters a one-second sliding window was used over the spatial filtered signals $c(n)$. For each signal segment of one second, the model is estimated and the frequency response is obtained. The overlap of the segments was fixed to 90% of the window length. This produces a time-frequency map for each signal. From this time-frequency representation, the features to be used in the classifiers (HMM and Sticky HDP-HMM) are selected based on physiological information (see Section 2.3.3) of the frequency bands related to execution/imagination of motor tasks. Table 4.1 shows the selected frequency bands used in this work. The features are calculated by taking the average power across frequency at the indicated frequency bands. The frequency resolution used was 1 Hz.

4.4 Results

Figure 4.7 shows an example of the artifact reduction stage. Note that the EOG interference on the EEG recordings is appreciable. However, this effect can be reduced by linear regression. The bottom plot shows a noisy segment of EEG. The red area corresponds to the portion of the EEG recordings that were used to learn the coefficients b (see Equation 2.4) that explain how the EOG signals propagate through the scalp by volume conduction affecting the EEG recordings. The segment in green shows a considerable reduction of the artifacts.

The CSPs can be projected over the scalp to visualize how the signals from different regions in the scalp contribute to form the common spatial patters. These plots are presented in Figure 4.8. Given that the positions of the electrodes in this data set basically represent left, right and central areas over the motor cortex it is easy to visualize the operation of the filters. Using as reference the subject B01, it can be observed that while one of the filters gives more importance to electrode C4 (right) the other filter gives more importance to the electrode C3 (left). This is consistent with the fact that the representation of the hand is contralateral, i.e imagination of the movement of the left hand produces ERD in the right hemisphere while imagination of movement in the right hand produces ERD in the left hemisphere. According to the spatial filters in the Figure 4.8 the electrode Cz does not provide significant information for classification in most of the subjects. It is worth noting that the symmetry between the filters seems to be an indicative of good performance and is consistent with the description of ERD phenomena discussed above.

Classification results are summarized in Table 4.2. Following the methodology used in the competition, we use the kappa values [97] as the metric for comparing different methods:

$$\kappa = \frac{C \times P_{cc} - 1}{C - 1} \tag{4.13}$$

where C is the number of classes and P_{cc} is the probability of correct classification[1]. Relatively larger kappa values indicate better performance.

The average performance of the Sticky HDP-HMM is higher than the results obtained by the winners of the BCI competition. Note that the value reported by the BCI competition corresponds to the maximum value obtained along the execution of the task. In the case of the Sticky HDP-HMM and the HMM-like models presented in the Table 4.2 the time of maximum performance is known to be at the end of the trial. The HMM with fixed parameters (HMM-FP column in Table 4.2) makes use of 3 hidden states representing beginning, execution and end of the motor task and Gaussian mixtures of two components were allowed per each hidden state. The HMM-CV refers to the

[1]Equation (4.13) takes this simple form given that the same number of samples for each class is available for each subject in each session

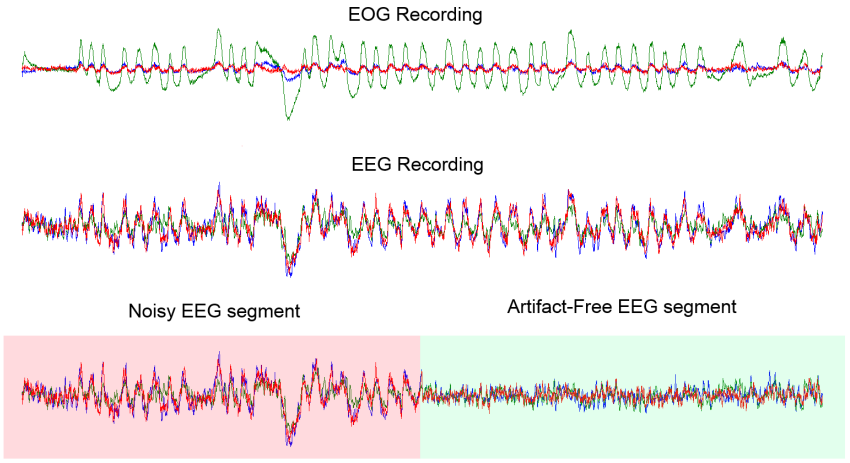


FIGURE 4.7: EOG artifact removal

Subject	Chin.	Gan	Coyle	HMM-FP	HMM-CV	Sticky HDP-HMM
B01	0.40	0.42	0.19	0.38	0.43	**0.57**
B02	0.21	0.21	0.12	0.10	0.16	0.14
B03	0.22	0.14	0.12	0.03	0.08	0.13
B04	0.95	0.94	0.77	0.93	0.94	0.92
B05	0.86	0.71	0.57	0.81	0.86	0.83
B06	0.61	0.62	0.49	0.61	0.66	**0.81**
B07	0.56	0.61	0.38	0.59	0.63	0.57
B08	0.85	0.84	0.85	0.79	0.80	0.81
B09	0.74	0.78	0.61	0.70	0.71	**0.79**
Average	0.60	0.58	0.46	0.55	0.59	**0.62**

TABLE 4.2: Comparison of the proposed Sticky HDP-HMM approach with the top three methods in BCI competition IV as well as with HMM. HMM-FP corresponds to a HMM with parameters fixed a priori (3 hidden states, Gaussian Mixtures of 2 components per hidden state). HMM-CV corresponds to HMM with parameters selected by 3 Folds-Crossvalidation. HMM-FP, HMM-CV and Sticky HDP-HMM use the same set of features. The metric used is Kappa Cohen's.

HMM approach with the number of hidden states and Gaussian mixtures learned by cross-validation. HMM-CV method uses Three-fold cross-validation for selection of the number of states and number of components of the Gaussian mixtures.

4.5 Conclusion

In this chapter a HMM-based approach has been presented for classification of sensorimotor rhythms in synchronous BCIs. First, preprocessing methods including artifact reduction by linear regression, spatial filtering by common spatial patterns and spectral power estimation by autoregressive modeling have been performed on the data. The HMM method aims to model the dynamics of the EEG signal during the execution of

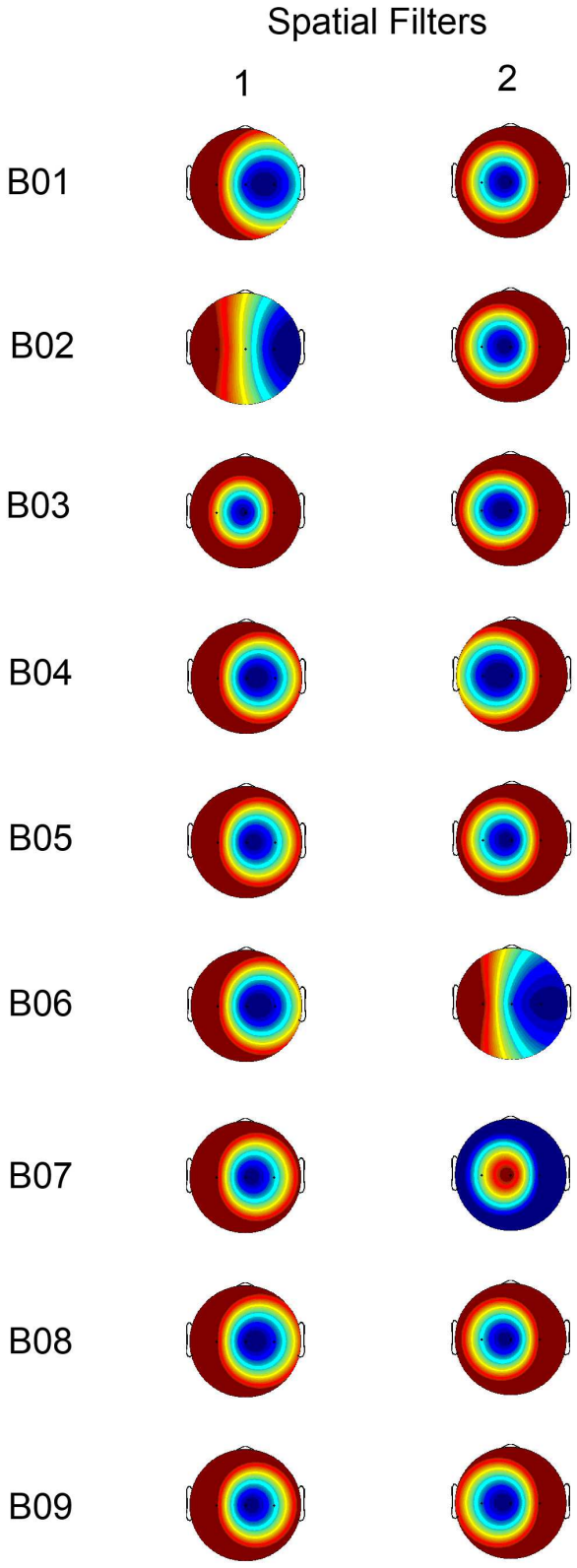


FIGURE 4.8: Topographical projection of the spatial filters.

the motor task proving that this information can increase the performance of the system. The main disadvantage of classic HMM methods is that the number of hidden states and

the parameters of the distribution of the data have to be set *a priori*. The HDP-HMM approach provides a solution for the selection of the number of hidden states and the number of components of a mixture of Gaussians used to model the distribution of the data. This is achieved by inferring a posterior distribution over the number of hidden states by means of Hierarchical Dirichlet Process, which allows an infinite number of states, which translates in a data driven method. Also Dirichlet processes are used to model the number of components of the mixture of Gaussians. The results show that these methods provide good classification results. In particular and as expected, the HDP-HMM approach with self bias transition (Sticky HDP-HMM) provides the best results, above the performance obtained by the winners of the BCI competition IV. Despite the good performance of the proposed methods, the modeling of the data as a mixture of Gaussians may be restrictive in the case of brain signals. Furthermore, the model requires to build independent models for each class, which may limit the ability of the system for discrimination. It is then necessary to extend this work to models that directly model the posterior probability of the classes given the observed data, integrating in a single model the ability to discriminate between different classes. Those kind of discriminative models are studied in the remainder of this thesis.

Chapter 5

A Latent Discriminative Graphical Model for Synchronous BCIs

In Chapter 2 it was shown that HMM-based classifiers provide good results by taking into consideration the dynamics of the signal and modeling different states that can be linked to the phenomena of Event Related Synchronization (ERS) and Event Related Desynchronization (ERD). The HDP-HMM model provides an elegant and efficient solution for the selection of the number of hidden states and number of Gaussian mixtures in HMM using Dirichlet processes. Nevertheless, if the EEG signal is modeled by HMM-based methods, the distribution of the data must be estimated and conditional independence assumptions of the data given the underlying states should be incorporated in order to make the inference problem tractable. A remedy for this problem is the use of Conditional Random Fields (CRFs) [98]. Although this is a discriminative model that does not require the estimation of the distribution of the data, there is one more issue in the case of BCI applications, where, unlike the analysis of sleep EEG signals based on CRFs as proposed in [99], the sequence of states is unknown. A modified CRF method has been proposed for BCI in [21] where the classes are associated with states in the CRF. However this method does not make use of intermediate states to model EEG signals related to each mental state, which have been proved to increase the performance, as presented in Chapter 4 and reported in other pieces of work [14, 18, 17]. This motivates the use of hidden states in CRF. Gunawardana et al. [100] have proposed a hidden-state CRF with application to phone classification which has been generalized by Sugiura et al. in [101], to the so-called hierarchical hidden state CRF (HHCRF). Sugiura et al. have presented an application of HHCRF in EEG signal segmentation in an asynchronous BCI application exhibiting advantages when compared to the generative counterpart, the Hierarchical HMM. However, the model proposed in [101] is based on a complicated structure making the parameter estimation and state sequence approximation computationally expensive.

Pieces of work [102] have proposed a Hidden Conditional Random Field (HCRF) model that uses hidden variables to model the latent structure of the input domain and defines a joint conditional distribution over the class labels and the hidden variables given the observations. Contrary to the work in [100], the HCRF model defined by Quattoni et al. does not fix the sufficient statistics used in the potential function of the CRF and does not assume Gaussianity of the data, which leads to a more flexible model selection process.

Motivated by the work in [102], we present an HCRF-based approach for classification of imaginary motor tasks in a synchronous BCI scenario, where the labels do not change with time, making it unnecessary to define a top layer with different states as required by the HHCRF approach of [101].

5.1 Hidden Conditional Random Fields for BCI

In the task of labeling sequence data, one of the most widely used tools is the hidden Markov model [89], a finite automaton which contains discrete-valued states Q emitting a data vector X at each time point, the distribution of the data at each time point depends on the current state. Given that models of this kind are generative, they require computation of the joint probability density function of the observed data samples over multiple time points. In order to make the inference problem tractable, assumptions about independence of the data at each time point conditioned on the states should be made. Such assumptions are violated in many practical scenarios. CRFs are discriminative models that overcome these issues [98], avoiding the need to explicitly model the data distribution as well as the need for the independence assumptions. For *linear - chain* CRFs, Lafferty et al. [98] define the probability of a particular label sequence $\overline{\mathbf{y}}$ given an observation sequence $\mathbf{x}$ to be of the form:

$$P_\theta(\overline{\mathbf{y}}|\mathbf{x}) \propto \exp\{\sum_{l\in L_1}\sum_{j=1}^{m} f_{1,l}(\overline{y}_{j-1},\overline{y}_j,\mathbf{x},j)\theta_{1,l} + \sum_{l\in L_2}\sum_{j=1}^{m} f_{2,l}(\overline{y}_j,\mathbf{x},j)\theta_{2,l}\} \tag{5.1}$$

where j represent the discrete time index, m is the length of the sequence $\mathbf{x}$, $f_{1,l}$ and $f_{2,l}$ are the *CRF-features*[1] related to the edges and nodes of the graph, respectively, and are given and fixed. L_1 and L_2 denote the sets of indices for the *CRF-features*. The parameters $\theta_{1,l}$ and $\theta_{2,l}$ must be estimated based on training data. For a more detailed description of CRFs the reader is referred to [98].

This approach overcomes the problems stated above for HMMs. However, CRFs focus on assigning a label for each observation (e.g., each time point in a sequence), and they neither capture hidden states nor directly provide a way to estimate the conditional

[1]These are simply called features in the CRF literature. However to distinguish them from features to be extracted from EEG signal, we call them CRF-features.

probability of a class label for an entire sequence. In the BCI problem, which is of interest in this paper, labels are not available for temporal segments of (training) EEG data recorded during the execution of a motor task, and the central problem of interest is to assign a class label for an entire sequence. As a result, it would be necessary to use a model that facilitates classifying an entire sequence, and that involves hidden states. Such a model has been proposed in [102], and is called the hidden conditional random field (HCRF). HCRFs are able to capture intermediate structures through hidden states, combined with the power of discriminative models provided by CRFs. Furthermore, unlike CRFs, they also provide a way to estimate the conditional probability of a class label for an entire sequence. A HCRF is constructed as follows. The task is to infer the class y from the data $\mathbf{x}$, where y is an element of the set Y of possible labels for the entire data and $\mathbf{x}$ is the set of vectors of temporal EEG features $\mathbf{x} = \{x_1, x_2, ..., x_m\}$. The subindex m represents the number temporal observations. The training data consists of a set of labeled samples $(\mathbf{x_i}, y_i)$, for $i = 1, ..., n$ where $y_i \in Y$ and $\mathbf{x_i} = \{x_{i,1}, x_{i,2}, ..., x_{i,m}\}$. For any $\mathbf{x_i}$, a vector of latent variables $\mathbf{h} = \{h_1, h_2, ..., h_m\}$ is assumed, providing the state sequence of the data. Each possible value for h_j is a member of a finite set H of possible hidden states. The joint probability of the labels and the states given the data is described by:

$$P(y, \mathbf{h}|\mathbf{x}, \theta) = \frac{\exp(\Psi(y, \mathbf{h}, \mathbf{x}; \theta)}{\sum_{y', \mathbf{h}} \exp(\Psi(y', \mathbf{h}, \mathbf{x}; \theta)} \tag{5.2}$$

where θ are the parameters of the models and $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$ is a potential function $\in R$. The conditional probability of the labels given the data can be found by marginalizing out h:

$$P(y|\mathbf{x}, \theta) = \sum_{\mathbf{h}} P(y, \mathbf{h} \mid \mathbf{x}, \theta) = \frac{\sum_{\mathbf{h}} exp(\Psi(y, \mathbf{h}, \mathbf{x}; \theta)}{\sum_{y', \mathbf{h}} exp(\Psi(y', \mathbf{h}, \mathbf{x}; \theta)} \tag{5.3}$$

Following [102], the estimation of parameter values, using the training data, can be performed by maximizing the following objective function:

$$L(\theta) = \sum_i \log P(y_i|\mathbf{x_i}, \theta) - \frac{1}{2\sigma^2} \|\theta\|^2, \tag{5.4}$$

where the first term in (5.4) is the log-likelihood of the data. The second term is the log of a Gaussian prior with variance σ^2. Given this objective function, various nonlinear optimization algorithms can be used to search for the optimal parameter values $\theta^* = \arg\max_\theta L(\theta)$. In our work, we use a quasi-Newton algorithm using Hessian updates based on the Broyden–Fletcher–Goldfarb–Shanno (BFGS) formula. Given a new test example $\mathbf{x}$ and parameter values θ^* induced from the training set, the label for the example is taken to be $\arg\max_{y \in Y} P(y|\mathbf{x}, \theta^*)$

HCRFs use undirected graphical structures, with the graph defined by $G = (V, E)$ where the V denotes to the vertices in the graph and E denotes the edges. Based on this, the potential function $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$ is defined as:

$$\Psi(y, \mathbf{h}, \mathbf{x}; \theta) = \sum_{j=1}^{m} \sum_{l \in L_1} f_{1,l}(j, y, h_j, \mathbf{x})\theta_{1,l} + \sum_{(j,k) \in E} \sum_{l \in L_2} f_{2,l}(j, k, y, h_j, h_k, \mathbf{x})\theta_{2,l} \quad (5.5)$$

where $f_{1,l}$ and $f_{2,l}$ are the *HCRF-features* related to the nodes and edges of the graph, respectively, and are given and fixed. L_1 and L_2 denote the sets of indices for the *HCRF-features*. It is important to note that $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$ is decomposed into a series of potentially local functions of the hidden variables. This property is the key for efficient inference over such models. If the set of hidden states form a tree-structured graph, then exact methods for inference and parameter estimation can be used. In particular, the belief propagation algorithm [103] can be used to compute the marginal distributions of hidden states given the data, which can in turn be used in the solution of the classification problem defined above [102]. If the graph G contains cycles, approximate methods such as loopy belief-propagation can be used for approximate inference.

Figure 5.1 shows an HCRF graphical model. The graphical structure of this model encodes which variables are involved in each of the functions defining the *HCRF-features* in $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$ in Equation (5.5). For example the chain structure of the hidden variables in the particular graphical model in Figure 5.1 implies that the only hidden variables appearing in the edge *HCRF-features* $f_{2,l}$ in Equation (5.5) are those with adjacent indices, i.e., with $|j - k| = 1$. Likewise, in the case in which the possible edges indicated by dashed lines in Figure 5.1 are missing, the node *HCRF-features* in Equation (5.5) for the graph in Figure 1 would take the form $f_{1,l}(j, y, h_j, x_j)$. Furthermore, since y and x_j are not directly connected, but connected through h_j; $f_{1,l}$ would further decompose into two functions, one expressing the compatibility between y and h_j, and the other between h_j and x_j. Hence the graphical model contains information directly related to the decomposition of the potential function $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$, which in turn specifies how the posterior probability of the labels in Equation 5.3 is expressed in terms of local functions.

5.2 Description of Experiments and Methods

In order be able to compare the results obtained in Chapter 2 with the proposed method based on HCRF presented in this chapter, the preprocessing steps used for the approach based on HMM and Sticky HDP-HMM have been left unchanged. The reader is referred to Section 4.3 for details of the preprocessing stage.

EEG feature vectors obtained using the auto-regressive power spectrum as described in 4.3, constitute the data $\mathbf{x}$ to be fed to the HCRF-based inference algorithm to be

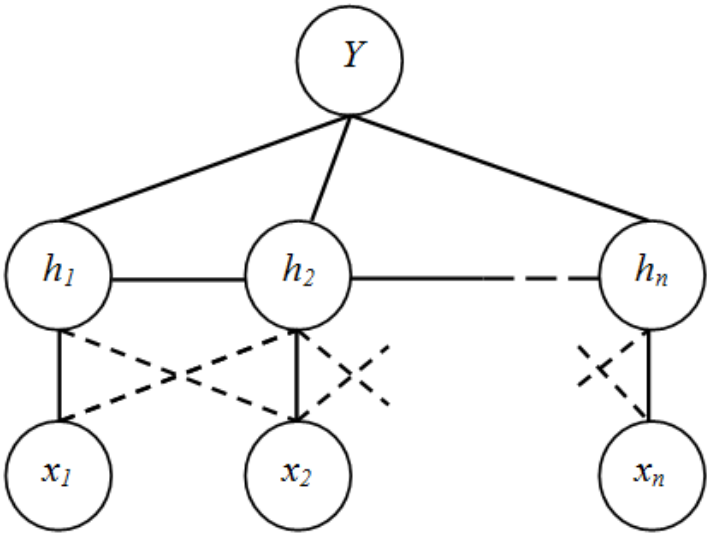


FIGURE 5.1: An HCRF graphical model. Dashed lines indicate the possibility of including long range dependencies between the data and the hidden states.

labeled. Since we use 5 frequency bands and two CSP components, the component x_j of the vector $\mathbf{x}$ at time point j is ten dimensional.

The particular HCRF model used in our work is a special case of the general form appearing in Equation (5.5). In particular, we use a model represented by the graphical structure in Figure 5.1, without the presence of the long range dependencies indicated by the dashed lines. This leads to decoupling and a number of simplifications in the potential function $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$ of Equation (5.5). First, since y and $\mathbf{x}$ are only connected through $\mathbf{h}$, the node potential function decomposes into two terms, one relating y and $\mathbf{h}$, and the other one relating $\mathbf{h}$ and $\mathbf{x}$. Second, since long range dependencies are not present, only x_j (rather than the past and future values present in the input sequence $\mathbf{x}$) is involved in the potential function for h_j. Third, the edge potential function involves cliques formed by consecutive nodes h_j and h_k (where $|j - k| = 1$) and the label y. Putting all of this together, we obtain the following potential function used in our work:

$$\Psi(y, \mathbf{h}, \mathbf{x}; \theta) = \sum_j f_{1,1}(x_j) \cdot \theta_h[h_j] + \sum_j f_{1,2}(y, h_j)\theta_y[y, h_j] + \sum_{(j,k)\in E} f_{2,1}(y, h_j, h_k)\theta_e[y, h_j, h_k] \tag{5.6}$$

where the node HCRF-feature function $f_{1,1} = x_j$. The dot product $f_{1,1}(x_j) \cdot \theta_h[h_j]$ measures the compatibility between the current EEG feature and the state h_j, where $\theta_h[h_j]$ are the weights associated with h_j. The HCRF-feature function $f_{1,2}$ is a binary function with the product $f_{1,2}(y, h_j)\theta_y[y, h_j]$ measuring the compatibility between the current state h_j and the motor task (label) y. The edge HCRF-feature function $f_{2,1}$ is also a binary function with the product $f_{2,1}(y, h_j, h_k)\theta_e[y, h_j, h_k]$ measuring the compatibility between the state transition from h_j to h_k and the motor task y. As the potential

Subject	Cross-val. Accuracy (%)	Number of states
B01	83	2
B02	68	3
B03	50	3
B04	99	2
B05	95	2
B06	85	2
B07	90	2
B08	87	3
B09	87	2

TABLE 5.1: Cross-validation accuracy in training data and the number of states in the HCRF model that maximizes the performance for each subject.

function in (5.6) can be written in the same form as (5.5) and the graphical structure modeling the hidden state transitions is a chain, algorithms such as belief propagation can be used for inference [102, 104].

One important issue in the BCI problem treated here is that the number of different brain states encountered during the execution or imagination of motor tasks is not obvious. In order to find the number of states that explain the signal well, a Four-Folds cross-validation is performed using the training data, with possible values of 2, 3 and 4 for the number of distinct states.[2] From this set of models, with different numbers of hidden states, the model which provides the best classification accuracy after the cross validation process, over the training data, is selected. Once the model is selected, classification is performed by assigning the label y for a test sequence $\mathbf{x}$ as follows:

$$\hat{y} = \arg\max_{y \in Y} P(y/\mathbf{x}; \theta^*). \tag{5.7}$$

5.3 Results

We evaluate the performance of the HCRF-based approach presented above on BCI Competition IV data set 2b. The number of states in the HCRF model was selected using a Four-Folds cross-validation on the training data. Table 5.1 shows the final selection of the number of states in the HCRF model for each subject. The selected model for each subject was used to classify the data in the test sessions identified in the data set as B0X04E and B0X05E, with X indicating the respective subject.

We compare the results of our approach to the top three results in the competition for this data set. In addition, we also present a comparison with the HMM-based approaches (HMM-CV and Sticky HDP-HMM) presented in Chapter 4 and with a CRF-based

[2]The value of 1 was not considered because it is physically inconsistent with phenomena involving changes (synchronization and desynchronization) in the EEG signal.

Subject	Chin.	Gan	Coyle	HMM-CV	Sticky HDP-HMM	CRF	HCRF
B01	0.40	0.42	0.19	0.43	0.57	0.49	0.60
B02	0.21	0.21	0.12	0.16	0.14	0.23	0.32
B03	0.22	0.14	0.12	0.08	0.13	-0.03	0.06
B04	0.95	0.94	0.77	0.94	0.92	0.94	0.97
B05	0.86	0.71	0.57	0.86	0.83	0.73	0.87
B06	0.61	0.62	0.49	0.66	0.81	0.73	0.78
B07	0.56	0.61	0.38	0.63	0.57	0.46	0.63
B08	0.85	0.84	0.85	0.80	0.81	0.70	0.88
B09	0.74	0.78	0.61	0.71	0.79	0.48	0.81
Average	0.60	0.58	0.46	0.59	0.62	0.53	**0.66**

TABLE 5.2: Comparison of the proposed HCRF-based approach with the top three methods in BCI competition IV as well as with HMM and CRF based techniques in terms of classification accuracy (kappa values).

	Shahid *et al*			HCRF		
Subject	04E	05E	Max kappa	04E	05E	Max kappa
B01	0.64	0.44	0.64	0.70	0.51	0.70
B02	0.33	0.25	0.33	0.33	0.38	0.38
B03	0.29	0.15	0.29	0.11	0.00	0.11
B04	0.96	0.89	0.96	1.00	0.94	1.00
B05	0.60	0.68	0.68	0.88	0.86	0.88
B06	0.64	0.73	0.73	0.74	0.85	0.85
B07	0.43	0.57	0.57	0.56	0.75	0.75
B08	0.69	0.94	0.94	0.79	0.96	0.96
B09	0.81	0.68	0.81	0.84	0.78	0.84
Average	0.60	0.59	0.66	**0.66**	**0.67**	**0.72**

TABLE 5.3: Comparison between the Bispectrum + LDA approach and the proposed HCRF-based approach. 04E and 05E denote two distinct sessions in the test data. Max kappa refers to picking the best kappa value for each subject across the two sessions (following the analysis in [1]).

method (using the same features employed for the HCRF model). All methods used for comparison in Table 5.2 use spatial filters (CSP) in the pre-processing stage. Furthermore a comparison with a recently published method based on the bispectrum of the EEG signal [1] is presented Table 5.3.

The results of our experiments are shown in Table 5.2. We observe that the method proposed in this paper provides higher kappa values than the top algorithms in the BCI competition, the HMM-based classifiers and the CRF-based classifier. The proposed method outperforms all three algorithms from the BCI competition in 8 out of 9 subjects and produces an average kappa value of 0.66 compared to 0.60 for the winner of the competition.

All the methods from BCI competition IV we have compared against, except Coyle et al.'s method, use EOG artifact removal. In order to ensure fairness in our comparisons with Coyle et al.'s method, we have repeated our experiments without EOG artifact

removal. In this case, our HCRF-based approach has produced an average kappa value of 0.65. Note that our average kappa value with EOG artifact reduction was 0.66. Thus, our approach without EOG artifact reduction still performs significantly better than that of Coyle et al.

The time course of the kappa value produced by our approach for each subject in each evaluation session is shown in Figure 5.2. Given the structure of the model as depicted in Figure 5.1, the HCRF model does not provide output for each sample point. Then, the plots in Figure 5.2 are obtained by simulating an on-line experiment where data from the beginning of a trial to the current time point are used. In this way, the model calculates the likelihood of the sequence for each class and provides an output for each sample point. The discussion on the time course of the kappa values also helps us contrast HCRFs with CRFs for synchronous BCI problems. If we plotted similar time courses for the CRF-based method whose results were presented in Table 5.2 in comparison with our HCRF-based approach, we would observe that the time course is constant. CRFs are sequential labeling models able to model the extrinsic dynamics of the labels given the data. However, there is no label dynamics in a synchronous BCI paradigm, that is, during a trial no transitions among class labels occur. This will be learned by the CRF model generating a strong bias to remain in the same label during the trial. Then, an error in the label based on information at the beginning of a trial will propagate in time to the end of the trial. This explains why CRFs are not well-suited for synchronous BCI applications and is also the reason for their rather poor performance, presented in Table 5.2. A solution for this is proposed by [21] where the transitions are not modeled directly. However, as the signals (or EEG features) obtained during each trial are assumed to belong to the same state (which is also the label in this case), temporal intrinsic dynamics of data for each class are not exploited, contrary to which is actually achieved by the HCRF-based approach proposed. A comparison between HCRF and CRF-based models using one particular type of feature and one particular classification methodology have been done. While the features were not choosen to favor one model versus the other, we acknowledge that other choices (e.g., as in [21]) might lead to different performance results for the CRF-based model.

We also compare our HCRF-based approach to the recent work in [1] where a high order statistic method involving the bispectrum of the EEG signal, together with linear discriminant analysis was used for classification of motor imaginary tasks. The results are presented in Table 5.3 following the methodology employed in [1]. These results demonstrate that our proposed HCRF-based approach outperforms the method in [1] on the BCI competition data set.

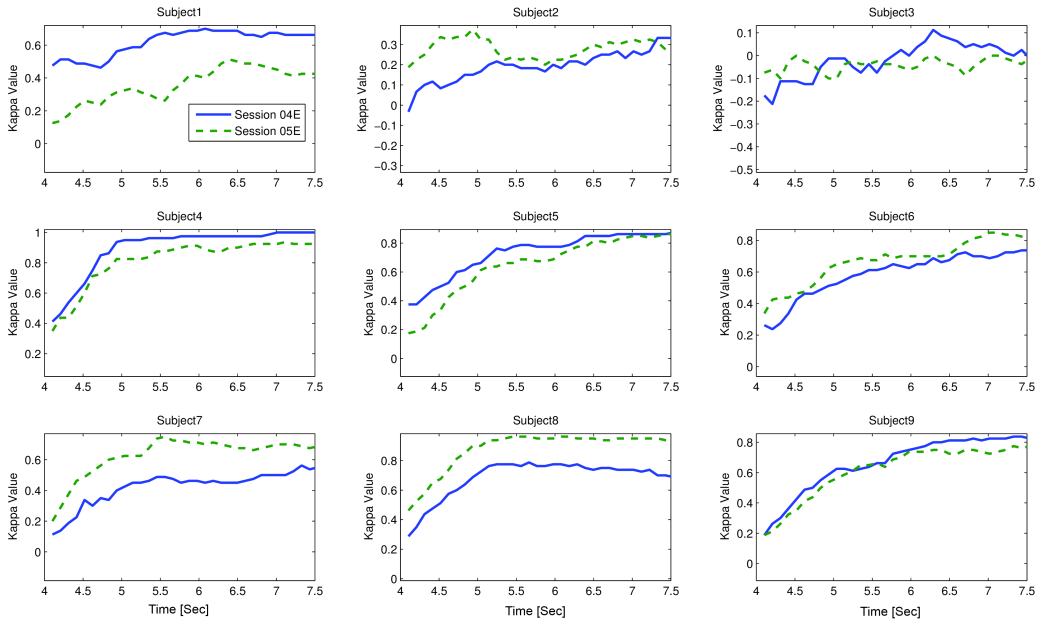


FIGURE 5.2: Time course of the kappa values for the proposed method in evaluation sessions 04E and 05E.

5.4 Conclusion

A new method for classification of imaginary motor tasks based on HCRFs is proposed. The autoregressive modeling of the CSP components, followed by the computation of the power spectrum and the selection of the frequency bands according to neurophysiological information, produces the feature vector that is fed to the HCRF-based classifier. Although subject-dependent selection of the frequency bands could lead to higher accuracy, we have opted here for common frequency bands for all subjects making the approach more general, which, given the performance obtained, shows the robustness of this method. Furthermore, the discriminative nature of the model proposed makes it unnecessary to model the distribution of the data or make assumptions about independence. Experimental results demonstrate the improvements in the classification accuracy provided by this approach over other methods. In addition, this method is based on modeling of the temporal changes of the EEG signal and the analysis of the state sequences could provide insights into the physical phenomena underlying the execution of the imaginary motor tasks.

Chapter 6

Discriminative Methods for Asynchronous BCI

In the case of non-invasive BCI systems based on electroencephalographic (EEG) signals, two types of BCI systems are used: synchronous and asynchronous. In a synchronous BCI approach, as discussed in Chapters 4 and Chapter 5, the subject receives cues that indicate when the mental task should be executed. Although this approach can be appropriate for laboratory research, it is not useful for most real life applications in which the subject will need to control the interface continuously without cues or temporal constraints for the execution of the mental task. A BCI system operating in this manner is called asynchronous. Most of the existing pieces of work on asynchronous systems make use of windowed EEG signals (or features of the EEG signals) and static classifiers (e.g., LDA, Gaussian classifiers, neural networks) [105, 106, 107, 108, 109, 110, 111, 112, 113]. In those approaches the difference of power in the EEG signals in different frequency bands is used to determine the subject's intention. Other studies involve the detection of transitions between tasks by detection of abrupt changes in the estimated power densities of the EEG signals [114, 112]. This so-called mental task transition detector offers increased performance in the classification accuracy of EEG signals [114, 112]. However, the temporal structure of the EEG signal which has been shown to increase the performance of the synchronous BCI systems[14, 18, 17, 36, 115, 116] is not exploited.

In an asynchronous scenario, the subjects execute different mental tasks without cues, which means that it is unknown when the subjects start the execution of a specific task. In this case one of the problems is the labeling of sequential data. Statistical models such as hidden Markov models (HMM) and conditional random fields (CRF) have been used with success in other fields such as gesture recognition and natural language processing [117, 104, 118, 100]. Given that CRFs can in principle be used to model the dynamics of sequential data, they are attractive for asynchronous BCI applications. However, although CRFs can model the extrinsic dynamics of the data (or features),

which in asynchronous BCI corresponds to dynamics across different tasks, they lack the ability to model intrinsic dynamics, i.e., the temporal evolution in the course of execution of a particular task. Physiological theory indicates that different states in the human brain emerge during the execution of mental tasks and those states are observed in the EEG signal through the well-known phenomena of event related synchronization and de-synchronization (ERS/ERD) [13]. Several studies have attempted to capture that structure through various random process models, as we describe below.

The so called Hidden Conditional Random Fields (HCRF) has been used for synchronous BCI in [116]. This method takes into consideration the dynamics of the signal during the execution of one task. However, it assigns a unique class label to an entire segment of EEG signals. This approach is not very attractive for asynchronous BCI because it does not provide a straightforward mechanism to model both the intrinsic and the extrinsic dynamics of the signal. Of particular interest is a method capable of modeling the intrinsic structure, proposed by Sugiura et al. [101]. This method is based on hierarchical hidden CRFs (HHCRFs), which generalize the HCRF model of [100]. Sugiura et al. apply HHCRF to EEG signal segmentation in an asynchronous BCI application and demonstrate the performance improvements it provides over the generative counterpart, the hierarchical HMM [119, 120]. Sugiura et al.'s work shares certain aspects of our perspective. In particular, similar to our work, it also involves a discriminative model for asynchronous BCI. However, their model is focused on building the hierarchy of various state variables and leads to a rather complicated structure requiring an extra level involving indicator variables. We feel the nature of the asynchronous BCI problem can be effectively captured by the simpler discriminative model presented in our work. We experimentally demonstrate the advantages offered by our model over that proposed by Sugiura et al. in this chapter. Another algorithm used for classification of temporal patterns is presented in a recent work by Cano et al. [121]. This algorithm is based on neural networks and fuzzy theory, the S-dFasArt. Cano et al. show that the S-dFasArt algorithm provides an improvement in the classification rate of spontaneous mental activity using dataset V of the BCI competition III.

A method that provides the combined advantages of CRF with the use of hidden states, has been proposed by Morency et al. for gesture recognition [122]. The so called latent dynamic CRF (LDCRF) allows modeling extrinsic dynamics of the sequential data as well as the intrinsic dynamics within each class by means of hidden states. This permits modeling different states during the execution of a specific mental task and at the same time modeling the transitions between different mental tasks. Given these features, LDCRF can be applied directly to sequential data avoiding the need for windowing the signal. In this chapter two methods for asynchronous BCI, one based on CRF and another based on LDCRF, are presented. For CRF the nodes in the model are used for representation of the mental task executed by the user. For LDCRF, hidden variables are incorporated and represent different states that take place during the execution of

a specific task. Nodes in a second layer of the graph represent different mental tasks. Surface Laplacian filters are used to obtain the signals over centro-parietal electrode positions and power spectral densities of the signals in specific frequency bands are used as features. Feature selection is performed by sequential floating forward selection (SFFS) producing an optimal set of features used as input to the CRF-based and LDCRF-based classifiers.

6.1 Conditional Random Fields

As we discussed in Chapter 5, CRFs are discriminative graphical models. Lafferty et al. [98] define the probability of a particular label sequence $\mathbf{y} = \{y_1, y_2, ..., y_m\}$ given an observation sequence $\mathbf{x} = \{x_1, x_2, ..., x_m\}$ with $x_j \in \mathcal{R}^d$ to be of the form:

$$P_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\{\sum_{l \in L_1} \sum_{j=1}^{m} f_{1,l}(y_{j-1}, y_j, \mathbf{x}, j)\theta_{1,l} + \sum_{l \in L_2} \sum_{j=1}^{m} f_{2,l}(y_j, \mathbf{x}, j)\theta_{2,l}\} \quad (6.1)$$

where $f_{1,l}$ and $f_{2,l}$ are feature functions related to the edges and nodes of the graph, respectively, and are given and fixed. L_1 and L_2 are the set of indices for the feature functions related to the edges and nodes respectively (see Figure 6.1.). The feature functions are real-valued and express sufficient statistics describing their arguments and their relationships.

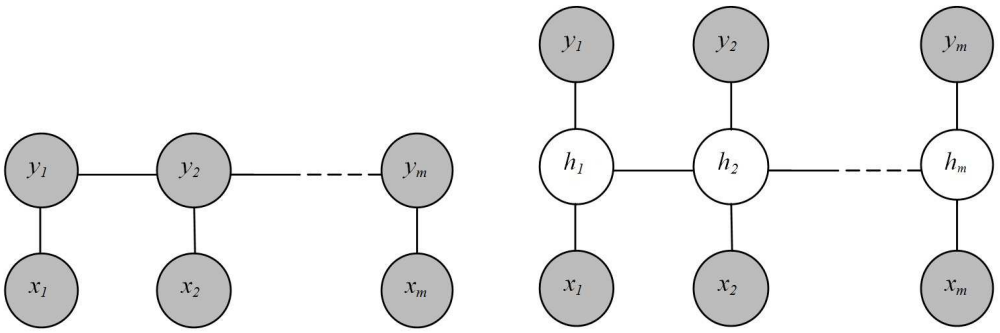
The conditional probability expressed in (6.1) can be simplified by writing:

$$P_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\{\sum_{l \in L} \sum_{j=1}^{m} f_l(y_{j-1}, y_j, \mathbf{x}, j)\theta_l\} \quad (6.2)$$

where L is a set of indices for the feature functions, each $f_l(y_{j-1}, y_j, \mathbf{x}, j)$ is either a state (node) function of a transition (edge) function and $Z(\mathbf{x})$ is a normalization factor .

In an asynchronous BCI scenario with reference to Figure 6.1(a) , the observation sequence $\mathbf{x}$ corresponds to EEG features and each element y_j of the label sequence $\mathbf{y}$ correspond to the imagined mental/motor task (relax, right finger movement, left finger movement, mathematical mental operation, etc.) at time point j. Then the feature functions provide sufficient statistics for classification of motor tasks.

Parameter estimation in CRFs for a linear chain (considered here for BCI signals) can be performed through a maximum likelihood approach [98]. Given independent identically distributed (i.i.d.) training data $D = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{N}$, where $\mathbf{x}_{(i)} = \{x_1^{(i)}, x_2^{(i)}, ...x_m^{(i)}\}$ is a sequence of inputs and each $\mathbf{y}_{(i)} = \{y_1^{(i)}, y_2^{(i)}, ...y_m^{(i)}\}$ is a sequence of mental/motor task labels, the conditional log - likelihood of the training data can be expressed as follows::



(A) (B)

FIGURE 6.1: (a) CRF model (b) LDCRF model. Shaded nodes represent observed variables in the training set. Although only one link between x_j and hidden nodes h is shown in the graph for simplicity, long range dependencies are also possible in these models.

$$l(\theta) = \sum_{i=1}^{N} \log P(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}) - \frac{\theta^2}{2\sigma^2} \tag{6.3}$$

where the regularization term $\frac{\theta^2}{2\sigma^2}$ is the log of a Gaussian prior with variance σ^2, that is $P(\theta) = \exp(\frac{1}{2\sigma^2}\|\theta\|^2)$. By substituting (6.2) into (6.3) and including a regularization term as a measure to avoid over fitting [98] the following expression is obtained:

$$\tilde{l}(\theta) = \sum_{i=1}^{N}\sum_{j=1}^{m}\sum_{l\in L} f_l(y_{j-1}^{(i)}, y_j^{(i)}, \mathbf{x}^{(i)}, j)\theta_l - \sum_{i=1}^{N} \log Z(\mathbf{x}^{(\mathbf{i})}) - \sum_{l\in L} \frac{\theta_l^2}{2\sigma^2}. \tag{6.4}$$

The parameters θ_l which maximize the regularized conditional log-likelihood above can be found by iterative optimization methods. In our work, we use a quasi-Newton algorithm using Hessian updates based on the Broyden–Fletcher–Goldfarb–Shanno (BFGS) formula.

6.2 Latent Dynamics Conditional Random Fields

CRFs allow modeling transitions between classes, capturing the extrinsic dynamics of the EEG features, but lack the ability to represent internal states for each class which can be used to increase the differentiability between classes. A model that incorporates the ability to capture both extrinsic and intrinsic dynamics is the Latent Dynamics CRF (LDCRF) proposed by Morency et al. [122]. LDCRF offers several advantages over previous discriminative models such as CRFs and hidden conditional random fields (HCRF) [104] combining their strengths. As in CRF, LDCRF models the transitions between classes; and as in HCRF, includes hidden states allowing to model within class dynamics. These characteristics allow the LDCRF model to be directly applied for labeling unsegmented sequences.

In the application of LDCRF models to BCI, the task is to learn a mapping between a sequence of EEG features $\mathbf{x} = \{x_1, x_2, ...x_m\}$ obtained during the imagination of motor activity and a sequence of labels $\mathbf{y} = \{y_1, y_2, ...y_m\}$ for the imaginary task executed; where each y_j is a class label for the j^{th} element of the sequence $\mathbf{x}$ and is a member of the set $\mathcal{Y}$ of possible class labels. LDCRFs also contain a vector of substructures $\mathbf{h} = \{h_1, h_2, ...h_m\}$ which form a set of hidden variables in the model, because they are not observed in the training examples, and represent different mental states in the brain during the execution of each of the imaginary tasks.

Morency et al. define the latent conditional model:

$$P(\mathbf{y}|\mathbf{x}, \theta) = \sum_{\mathbf{h}} P(\mathbf{y}|\mathbf{h}, \mathbf{x}, \theta) P(\mathbf{h}|\mathbf{x}, \theta). \tag{6.5}$$

where θ are the parameters of the model. In order to keep training and inference tractable, Morency et al. restrict the model to have disjoint sets of hidden states associated with each class. Then, the set of all possible states $\mathcal{H}$ is the union of all $\mathcal{H}_y$ sets, where $\mathcal{H}_y$ refers to the class-specific set of hidden states for class y. Under this assumption, the conditional probability in (6.5) can be written as:

$$P(\mathbf{y}|\mathbf{x}, \theta) = \sum_{\mathbf{h}:\forall h_j \in \mathcal{H}_y} P(\mathbf{h}|\mathbf{x}, \theta). \tag{6.6}$$

This is because the assumption of disjoint hidden states produces $P(\mathbf{y}|\mathbf{h}, \mathbf{x}, \theta) = 0$ for $h_j \notin \mathcal{H}_y$ and $P(\mathbf{y}|\mathbf{h}, \mathbf{x}, \theta) = 1$ for $h_j \in \mathcal{H}_y$. Using the usual conditional random field formulation:

$$P(\mathbf{h}|\mathbf{x}, \theta) = \frac{1}{Z(\mathbf{x}, \theta)} \exp\{\sum_{l} \mathbf{F}_l(\mathbf{h}, \mathbf{x})\theta_l\}. \tag{6.7}$$

with $\mathbf{F}_l$ defined as:

$$\mathbf{F}_l(\mathbf{h}, \mathbf{x}) = \sum_{j=1}^{m} f_l(h_{j-1}, h_j, \mathbf{x}, j) \tag{6.8}$$

Each feature function $f_l(h_{j-1}, h_j, \mathbf{x}, j)$ as in the case of CRF is either a transition function or a state function.

Learning the parameters in the LDCRF model can be achieved in the same way as for CRF, finding the optimal parameters θ^* that maximize the objective function in equation 6.3.

The feature functions in the LDCRF model correspond to transition and state feature functions. Note that transitions can be among hidden states within the same class (hence intrinsic) or among hidden states of different classes (hence extrinsic). Accordingly,

weights associated with the hidden states in the same subset $\mathcal{H}_y$ model the intrinsic dynamics while weights associated with hidden variables from different sets model the extrinsic dynamics. The number of transition functions in the model is given by the square of the cardinality of the set $\mathcal{H}$.

The number of state feature functions will be equal to the dimension of $\mathbf{x}$ times the number of possible hidden states $|\mathcal{H}|$. Figure 6.1(b) shows a diagram for the LDCRF model where the input sequence $\mathbf{x}$ corresponds to EEG features and the labels y_j represent the mental task executed. Given that $\mathbf{x}$ and $\mathbf{y}$ are observed in the training set, they are represented by shaded nodes in the graph of Figure 6.1(b).

6.3 Description of Experiments and Methods

6.3.1 Preprocessing

Problem and Dataset Description. Dataset V of the BCI competition III was used in this work. The dataset contains data from three normal subjects during four non-feedback sessions. The subject is requested to execute one out of three mental tasks: 1) Imagination of repetitive left hand movements, 2) Imagination of repetitive right hand movements, and 3) Generation of words beginning with the same random letter. The subject executes a mental task during fifteen seconds and then switches randomly to another task at the operator's request. For each subject, four sessions of four minutes length are available. The first three sessions are used for training and the fourth session is used for testing. The data available provide pre-computed features, obtained as follows. EEG signals are spatially filtered using a surface Laplacian filter and the power spectral density of these signals is calculated every 62.5 ms using the last second of data. The power spectral density was calculated between 8Hz - 30Hz with a resolution of 2Hz over centro-parietal electrodes C3, Cz, C4, CP1, CP2, P3, Pz, and P4. As a result, the pre-computed feature vector for each temporal window is a 96-dimensional vector (8 channels $\times$ 12 frequency components). Additionally we have performed experiments to create a second dataset at the Signal Processing and Information Systems (SPIS) Laboratory using the same setup described above. This dataset contains data from 4 subjects who are naive to BCIs, meaning that none of them has previous experience on BCI applications. The methods described in subsequent sections are applied equally to all subjects included in the two datasets, but the results for each dataset are presented separately.

Feature Extraction. Using the vector of pre-computed features, the average power across frequency in Alpha (8Hz - 12Hz), Sigma (12Hz - 16Hz) and Beta (18Hz - 26Hz) bands were computed for each of the eight electrodes. Figure 6.2 shows the topographic power distribution in the selected bands, for each subject. The topographic distribution

is obtained by interpolating the data in a Cartesian grid and shows for each class and each frequency band the logarithm of the average power during the execution of each mental task, using all data available for each class in the training set. Differences in the amplitude of the signal provide information about the type of CRF-features and LDCRF-features that could be used, as will be discussed later. The frequency bands alpha, sigma and beta, were selected because these rhythms are related to the well-known phenomena of ERS/ERD observed during the execution of mental tasks. This provides a new feature vector with 24 features, based on which we perform automatic feature selection for maximizing classification performance.

Feature Selection. The selection of features is done by means of Sequential Floating Feature Selection algorithm (SFFS). SFFS is describe in Section 2.5. In this work, classification accuracy is used as the cost function to be maximized. A three-fold cross-validation process is implemented, dividing the data in three sets, using each time two sets for training and one for testing. This process is repeated each time that a feature is added or removed to from the set of selected features.

6.3.2 Model Selection and Classification

CRF model. For the case of linear-chain CRF, given a new input sequence $\mathbf{x}$, the most likely labeling $\mathbf{y}^* = \arg\max_y p(\mathbf{y}|\mathbf{x})$ can be efficiently and exactly calculated by variants of dynamic programming algorithms for HMM, as described in [98]. The particular form we use for the conditional probability of the labels given the data is given by:

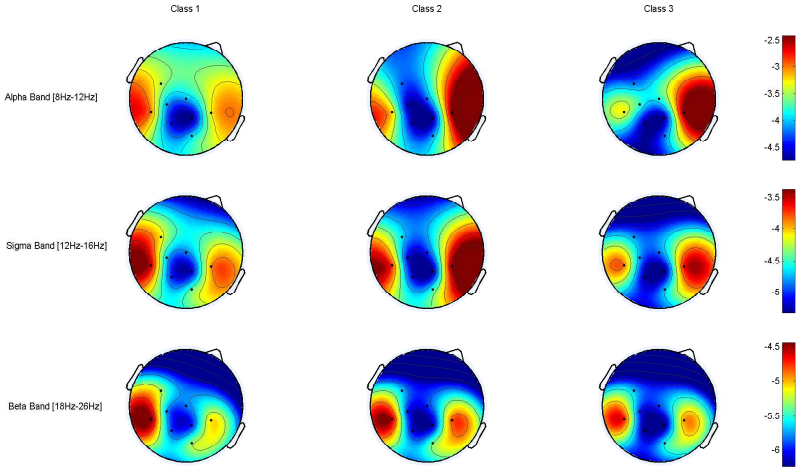
$$P_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\{\sum_{j=1}^{m} f_{1,1}(y_{j-1}, y_j) \cdot \theta_{1,1} + \sum_{j=1}^{m} f_{2,1}(x_j) \cdot \theta_{2,1}[y]\} \tag{6.9}$$

The dot product $f_{1,1}(y_{j-1}, y_j) \cdot \theta_{1,1}$ measures the compatibility of a transition from a particular motor task at $j-1$ to the same or another motor task at j. Each element of the *edge weight* vector $\theta_{1,1}$ contains a weight for a particular pair of labels. The feature function $f_{1,1}(y_{j-1}, y_j)$ is an indicator vector, with a value of 1 for the entry corresponding to the particular set of values (y_{j-1}, y_j), and 0 for all the other entries. The second term, which involves $f_{2,1}(x_j) \cdot \theta_{2,l}[y]$ with $f_{2,1}(x_j) = x_j$, measures the compatibility between the current EEG feature x_j and the label y_j.

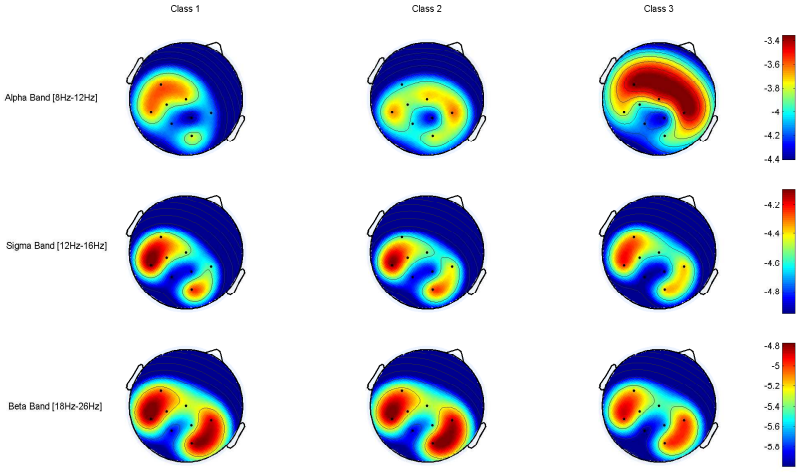
The class-dependent structure of the features as shown by the topographic distributions in Figure 6.2 suggest that the node compatibility function chosen in this manner has the potential of being useful in classification.

LDCRF model.

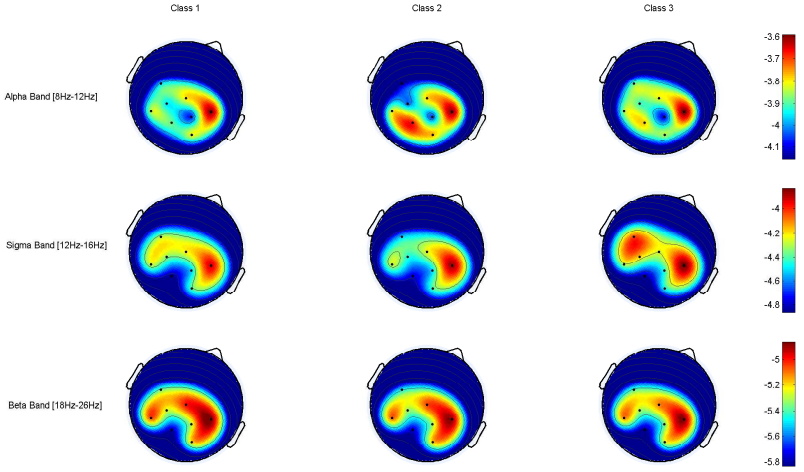
The topographic power distributions shown in Figure 6.2 highlight the differences in power distribution when different motor tasks are executed. However, the temporal



(A) Subject 1.



(B) Subject 2.



(C) Subject 3.

FIGURE 6.2: Average topographic distribution of power in different frequency bands.

variations of power during the execution of a particular task can also be observed. Figure

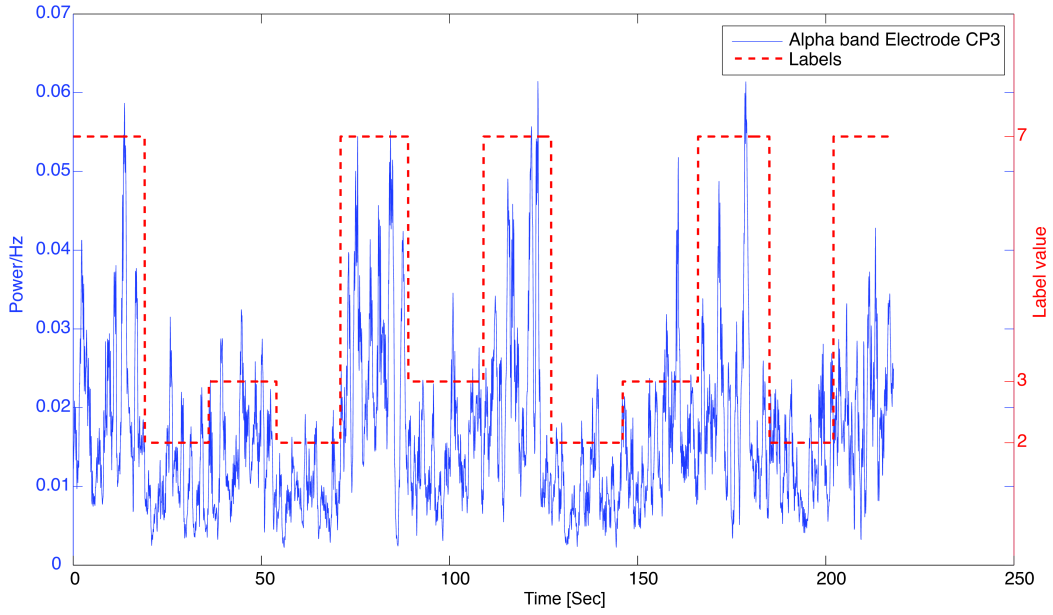


FIGURE 6.3: Example of EEG dynamics for different classes. Differences between classes and also intra-class differences are observed. The signal corresponds to alpha band in electrode CP3.

6.3 aims to display both of these phenomena. For the case of motor tasks, phenomena such as ERD and ERS explain the within class temporal variations. As observed in Figure 6.3, the magnitude of the signal is class-dependent but variations of the power during execution of the same task are also evident. The LDCRF model has the potential to fit and explain such data well, because LDCRFs are able to model extrinsic and intrinsic dynamics of the signal. Based on this, the feature functions are selected to obtain information about those dynamics. The conditional distribution of the labels given the data can be written as:

$$P(\mathbf{y}|\mathbf{x},\theta) = \sum_{\mathbf{h}:\forall h_j \in H_y} \frac{1}{Z(\mathbf{h},\mathbf{x})} \exp\{\sum_{j=1}^{m} f_1(h_{j-1},h_j)\cdot\theta_1 + \sum_{j=1}^{m} f_2(x_j)\cdot\theta[h_j]\}. \quad (6.10)$$

were the dot product $f_1(h_{j-1},h_j)\cdot\theta_1$ measures the compatibility of the state transitions, where states could correspond to the same or different classes. Each element of the *edge weight* vector θ_1 contains a weight for a particular pairs of hidden states. The feature function $f_1(h_{j-1},h_j)$ is an indicator vector, with a value of 1 for the entry corresponding to the particular set of values (h_{j-1},h_j), and 0 for all the other entries. It is worth noting, that this feature function models the intrinsic dynamics by means of the weights associated with pairs of hidden states in the same subset $\mathcal{H}_y$ and extrinsic dynamics by means of the weights associated with hidden states in different subsets. The second term, which involves the dot product $f_2(x_j)\cdot\theta[h_j]$ with $f_2(x_j)=x_j$ measures the compatibility of the current EEG feature $x_j$with the hidden state h_j.

TABLE 6.1: Cross validation results in training data for the proposed CRF and LDCRF based methods. BCI competition dataset.

Subject	CRF(%)	LDCRF (%).	Hidden states (LDCRF).
B01	89.34	91.55	2
B02	78.08	83.89	2
B03	59.73	59.30	3

For testing, given a new test sequence $\mathbf{x}$, the most probable sequence $\mathbf{y}^*$ that maximizes the conditional model [122] should be estimated:

$$\mathbf{y}^* = \arg\max_y \sum_{\mathbf{h}:\forall h_i \in \mathcal{H}_y} P(\mathbf{h}|\mathbf{x}, \theta^*) \tag{6.11}$$

To estimate the label y_j^* of the element x_j of the sequence $\mathbf{x}$, the marginal probabilities $P(h_j = a|\mathbf{x}, \theta^*)$ are evaluated for all possible hidden states $a \in \mathcal{H}$. Then the probabilities of hidden states corresponding to each distinct label are summed up, and the label corresponding to the maximum probability hidden state set is chosen. That is, given that it is assumed that the states are not shared across classes, the set of states with the higher global probability define the label to be declared. The marginal probabilities mentioned above can be calculated by belief propagation [122, 103].

In our experiments, we use three different models with 2,3, and 4 states per class. For each model SFFS is employed to select the optimal set of features and the accuracies in the three-fold cross - validation process in the training data are compared. The model which provides the best accuracy is selected and used for labeling the test data.

6.4 Results

Table 6.1 shows the classification accuracies on the BCI competition dataset, obtained by cross-validation in the training set using CRF and LDCRF, as well as the number of states in the LDCRF model, that provides the best results. Table 6.2 shows the selected electrodes and frequency bands using SFFS for each subject using CRF and LDCRF. The input feature vector is formed by concatenation of the power of the signals in each of the selected frequency bands for each electrode in Table 6.2. Experimental results on test data are shown in Table 6.3. The proposed CRF and LDCRF-based methods are compared to the top result in the BCI competition [112], to the HHMM and HHCRF-based methods presented in [101], to a method proposed by Lin et al. that makes use of neural networks based on particle swarm optimization [113], to the recently proposed S-dFasArt method of Cano et al. [121], and to the popular Linear discriminant analysis (LDA) classifier using SFFS as feature selection method. Results evidence the

superiority of the proposed methods. LDCRF performs better than CRF, which can be explained by the use of hidden variables that allow modeling, besides extrinsic dynamics, the intrinsic dynamics of the signal during the execution of a particular task. In order to visualize the statistical significance of the results, a statistical test has been performed and the results are presented in Table 6.4.

In the evaluation above, we compared our approach against the winner of the BCI competition, and hence we have demonstrated that our approach offers better classification performance than all methods considered by the competition organizers. There were a number of other methods submitted to the BCI competition and not considered by the competition organizers as they did not follow the requirements for evaluation. It may be important to note a number of interesting observations related to these methods. In particular from the left-out methods the one with the highest performance was proposed by John Q. Gan et al. This method included post processing stages following a linear classifier. The post processing stage smoothes the output of the classifier, that is, previous values of the output were used to define the current output under the assumption that rapid changes are not observed during the execution of the mental tasks. This method obtains an average accuracy of 80.97%. The proposed CRF and LDCRF methods yield better performance in terms of accuracy. Furthermore, they do not need any post-processing based on smoothing of the output of the classifier (See Figure 6.4). The proposed models are able to learn from training data that rapid changes in the executed task are unlikely. However, if such transitions do appear in the training data, they will be automatically taken into consideration in the learning phase. We believe this is a principled approach to learning and exploiting the dynamics of transitions among tasks in an asynchronous BCI system. Other studies [123, 124] have proposed other post-processing operations that involve parameters such as the dwell time and the refractory period. The dwell time defines the minimum time that the detector should be above the threshold value before declaring a positive output. The refractory period defines a time interval in which the detector is suppressed once a positive output is declared. Such ideas can be implemented as post-processing operations within the context of the algorithms presented in our paper.

Results on the dataset collected in our laboratory (the SPIS dataset) are presented in Table 6.5. The methods proposed have been compared to LDA, which has in common with other methods presented in Table 6.3 that it does not take the dynamics of the EEG signal into consideration for classification. Note that this dataset is more challenging as observed in the reduction of the performance on the all methods considered. However, the CRF and LDCRF methods provide higher performance.

TABLE 6.2: Frequency bands for each electrode selected by SFFS for the LDCRF and the CRF based methods.

		LDCRF			CRF		
	Chn	Frequency Band			Frequency Band		
Subj		Alpha	Sigma	Beta	Alpha	Sigma	Beta
	C3	✓	-	✓	-	✓	✓
B01	CP1	✓	-	-	✓	✓	-
	P3	-	✓	-	-	-	-
	C3	✓	-	-	✓	-	-
	Cz	-	-	-	-	-	✓
B02	C4	✓	✓	✓	✓	-	✓
	CP1	-	✓	-	-	-	-
	P4	-	-	✓	-	-	✓
B03	C3	-	✓	-	✓	✓	-
	CP2	-	-	✓	-	-	-
	Pz	-	-	-	-	-	✓

TABLE 6.3: Correct classification percentages achieved by various methods on a 3-class asynchronous BCI task.

Subject	B01	B02	B03	Average
Fisher LDA	71.97	62.53	49.68	61.39
Galan [112]	79.60	70.31	56.02	68.64
HHMM [101]	79.05	61.58	34.40	58.34
HHCRF [101]	94.58	70.17	32.11	65.62
IPSONN [113]	78.31	70.27	56.46	68.35
S-dFasArt [121]	87.21	82.26	58.72	76.07
CRF	92.95	89.63	61.81	**81.46**
LDCRF	95.63	89.75	72.36	**85.91**

6.5 Conclusion

In this work two statistical methods are proposed for use in modeling the dynamics of the EEG signal during the execution of mental tasks in an asynchronous BCI scenario. The preprocessing of the signals involve the use of global Laplacian filters and estimation of the spectral density of the segmented EEG signals using the last second of data. SFFS was used for selection of relevant features.

A CRF-based model and a LDCRF-based model were employed. The former method is able to model extrinsic dynamics of the EEG features. Those dynamics are related to the transitions from one mental task to the other in an asynchronous BCI system.

TABLE 6.4: One-sided paired-ttest results for the methods compared in Table 6.3.

	Vs CRF	Vs LDCRF
Subject	p-value	p-value
LDA	0.0036	0.0001
Galan [112]	0.0095	0.0001
HHMM [101]	0.0029	0.0040
HHCRF [101]	0.0467	0.0425
IPSONN [113]	0.0102	0.0001
S-dFasArt [121]	0.0044	0.0027

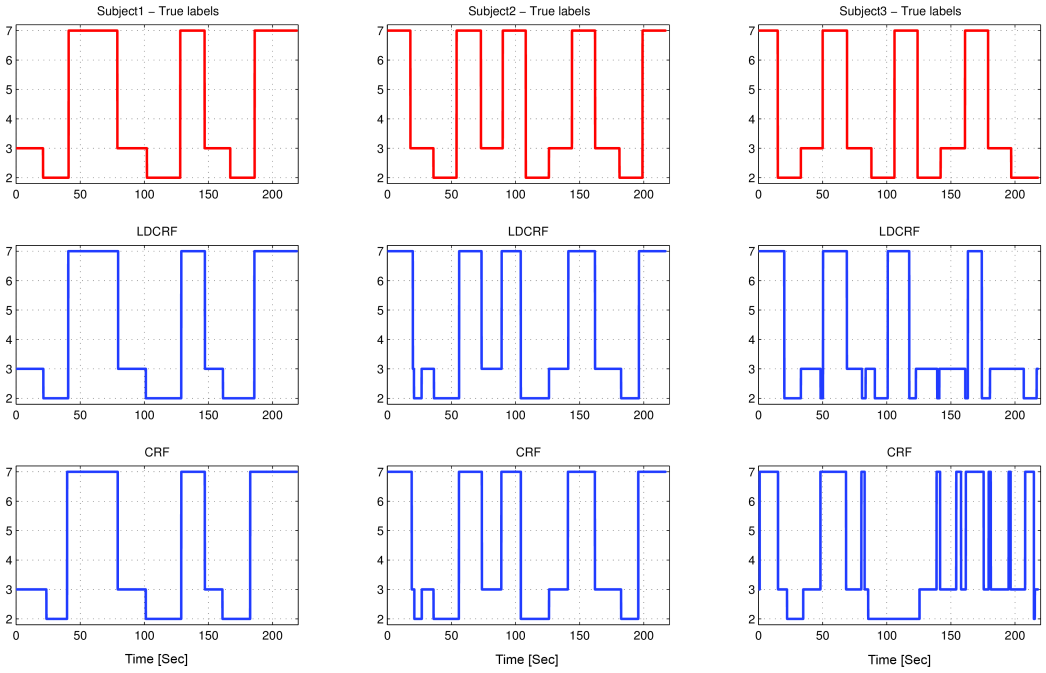


FIGURE 6.4: Classification output for the proposed methods,CRF and LDCRF on the test data. Labels 2,3 and 7 correspond to right hand imaginary, left hand imaginary and word association respectively.

LDCRF goes beyond that approach and models, in addition to the extrinsic dynamics, the internal structure of the signals. We assert that this structure is related to different mental states during the execution of a specific mental task (ERD / ERS for imaginary motor tasks). Our work provides the first application of LDCRFs in BCI. In particular, we demonstrate how this discriminative random field model used in other applications before can be utilized to capture the dynamics of the asynchronous BCI process. The superiority of the presented CRF-based and LDCRF-based methods is evidenced in the results presented using a publicly available dataset, a dataset recorded in our lab, and by comparison with recent works. Furthermore, it is worth of noting that the proposed methods do not need to make use of post-processing stages as they learn the dynamics of the data automatically. Another advantage of the proposed methods is that there is

TABLE 6.5: Comparison of the proposed methods with LDA method. SPIS dataset. (Values in %)

Method	B01	B02	B03	B04	Average
LDA	56.07	52.44	57.32	68.76	58.65
CRF	62.20	54.85	67.22	80.75	**67.47**
LDCRF	71.27	62.76	62.76	81.31	**69.53**

no need for windowing the EEG features thanks to the fact that the proposed methods inherently model the temporal structure of the signals, and carry temporal information through the state variables.

Chapter 7

Asynchronous Classification of Finger Movements using ECoG

Despite the disadvantages of electrocorticography (ECoG) discussed in Section 2.1.2, it has the advantage of providing high SNR and a good spatial resolution. Although the use of this technique is questionable in practical BCIs, in severe cases of ALS its use becomes justified.

The exceptional features of ECoG recordings allow us to obtain high spatial resolution in motor tasks. Particularly, in the case of EEG the tasks are limited to the actual or imaginary movement of arms, legs, feet, etc. This is possible because execution of the task involves many areas in the cortical representation of the body in the primary motor cortex. Therefore the imagination of the arm, for instance, will activate a vast area in the motor cortex, which then can be observed with EEG recordings. In ECoG, prediction of more specific movements is possible, to the extent that individual finger movements can be predicted or decoded [41]. One of the reasons that make this possible is that ECoG is capable of measuring a spectrum of frequencies beyond 40Hz, which is the practical limit usually observed for BCI. In particular, frequencies in the range of 60Hz to 200Hz (high gamma) reflect the neuronal activity of areas in the cortex that are related to the execution of particular tasks [125]. In [41] it was shown that signals in the Gamma band can be used to decode the movement of individual fingers, using a linear regression in brain signals recorded in the contralateral area of the hand executing the movement. Similar results were obtained in [126, 127, 128].

In this chapter, we present what is to our best knowledge the first attempt to classify the movements of individual fingers in an asynchronous scheme from ECoG signals. We begin with the analysis of the ECoG signals and investigate the relationship between motor movements and the brain signals for different frequency bands. We then propose the use of a probabilistic method based on graphical models for continuous classification of the finger movements.

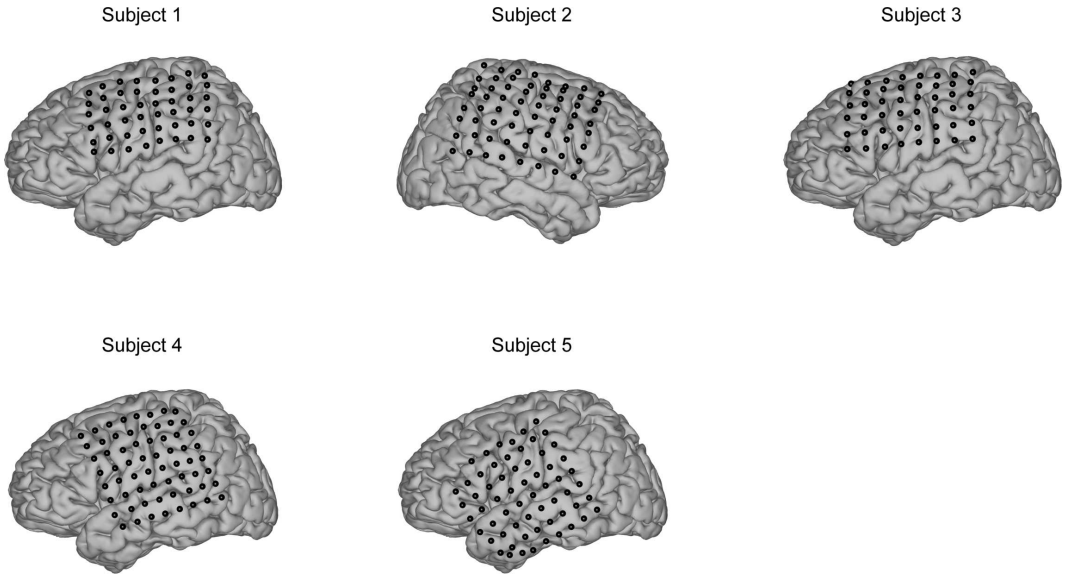


FIGURE 7.1: ECoG electrode grid placement for all subjects

7.1 Signal Analysis

Data set description. ECoG signals from 5 subjects were recorded during 10 minutes. The subjects sit up in front of a screen and are requested to move their fingers according to a stimulus presented on the screen indicating the name of the finger to move. Additionally, the subjects wear a data glove that is used to record the actual movement of the finger which is used as ground truth. ECoG signals were recorded using a grid of electrodes (the number of electrodes is different for each subject) ranging from 48 to 64 electrodes. The position of the electrodes for each subject is displayed in Figure 7.1 and the subjects were required to execute the movements in the hand contralateral to the placement of the grid of electrodes. The sampling frequency was set to 1000Hz for the ECoG recordings and 25Hz for the data glove.

Frequency Selection and feature extraction. We selected two frequency bands related to the execution of motor tasks. Alpha (8Hz-12Hz) and high gamma in a wide range of 65Hz to 200Hz. We use two band-pass Butterworth IIR filters of order 8 each one. Based on previous pieces of work, it is assumed that the envelope of the brain signals in these frequency bands contains information about the task being executed [41, 128]. We calculate the envelope by squaring the signals and then applying a low-pass filter of order 8 with a cutoff frequency of 4Hz. The value of 4Hz was selected because it is the maximum frequency at which the subject moves the finger and the envelope of the ECoG signals in the Gamma band is assumed to reflex the dynamics of the movement.

Electrode Selection

The motor tasks that the subjects execute are related to specific brain areas. In order to reduce the number of electrodes we calculate the average correlation between the brain

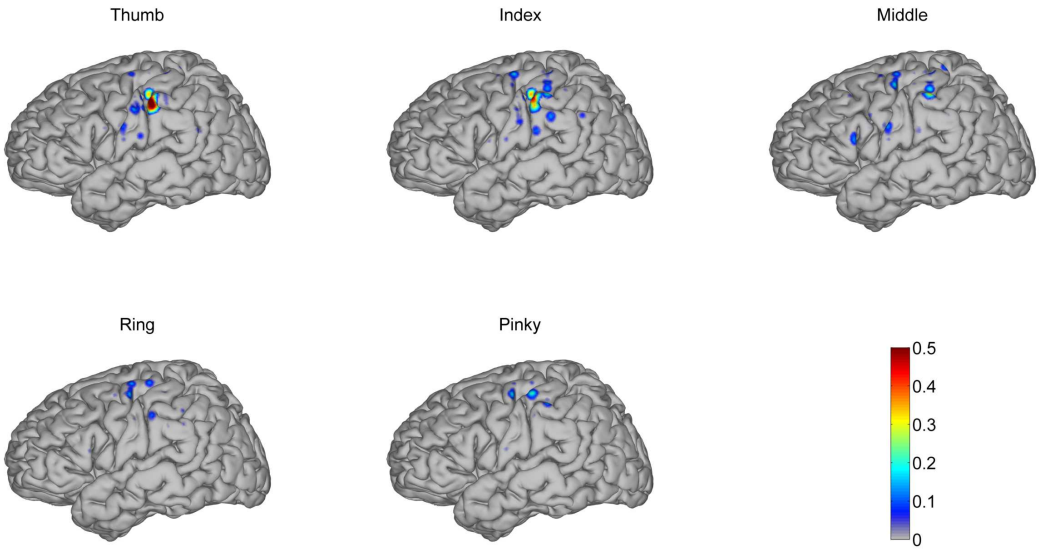


FIGURE 7.2: Distribution of correlations for the high Gamma (60Hz - 200Hz) for one subject during finger movements.

signals and the actual movement of each finger measured by the data glove. We rank the electrodes according to the magnitude of the correlation and select the first 20 electrodes according to this rank. Figure 7.2 shows the distribution of the correlations in the brain model for one subject. Note that the electrodes with higher correlation are in the area that correspond to the motor cortex. Also, we observe that for all subjects the finger whose actual movement is most highly correlated with the ECoG data is the thumb.

7.2 Classification Problems

Two classification problems can be defined: 1) Classification of movement versus rest in each finger (two-class problem for each finger), 2) Multiple class classification declaring which of the five fingers is moving or if all fingers are in rest (six-class problem).

7.2.1 Classification of Movement Versus Rest

7.2.1.1 Approach one: Independent chain-CRFs

In this approach, the main idea is to determine the periods of time during which one finger is moving and when it rests. For this, we build a Conditional Random Field (CRF) for modeling the activity of each finger.

Referring Figure 7.3, the brain signals features are represented by $\mathbf{x} = \{x_1, ..., x_n\}$ with $x_i \in R^d$, while the labels obtained from the data glove signals are represented by $\mathbf{y} = \{y_{1,1}, ..., y_{n,m}\}$, where n represents the time points and m is the nimber of fingers

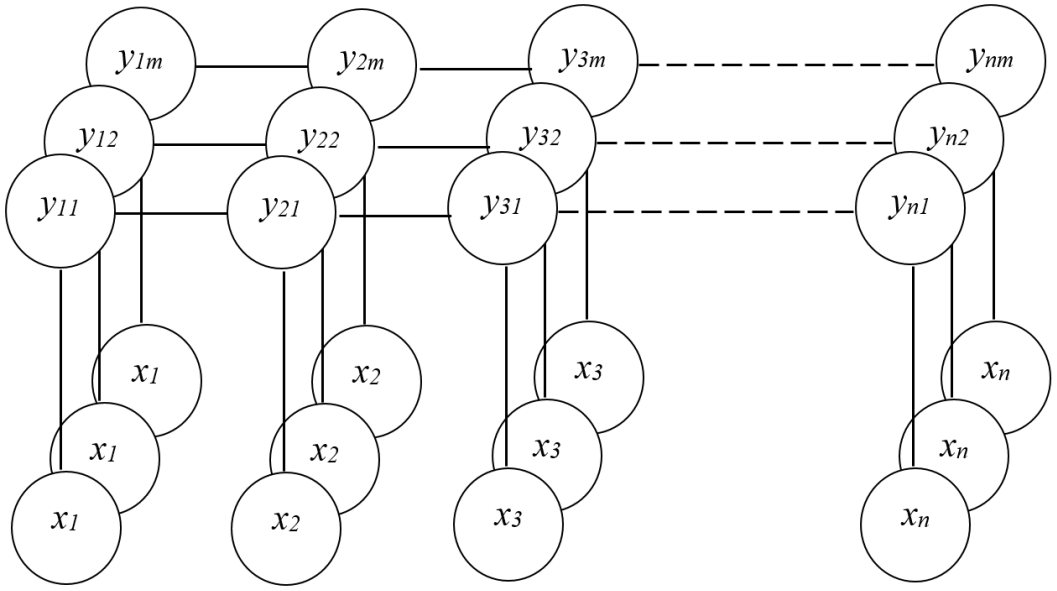


FIGURE 7.3: Graphical model for the independent chain-CRF

with $y \in [0, 1]$, with 0 and 1 representing resting and movement respectively for each finger. The conditional probability of the labels given the observed brain signals features is given by:

$$p_{chain}(\mathbf{y_m}/\mathbf{x}, \Theta) = \frac{1}{Z} \prod_{i=1}^{n} \Psi_i(y_{i,m}, x_i) \prod_{i=2}^{n} \Phi_i(y_{i,m}, y_{i-1,m}) \tag{7.1}$$

where Z is a normalization factor, $\Psi_i(y_{i,m}, x_i)$ is a node potential function and $\Phi_i(y_{i,m}, y_{i-1,m})$ is an edge potential function. The node potential functions are defined according to:

$$\Psi_i(y_{i,m}, x_i) = e^{\sum_{j=1}^{d} \theta_{Vj} f_{Vj}(x_i, y_{i,m})} \tag{7.2}$$

where $f_{Vj}(x_i) = \mathbf{1}_{y_{i,m}=y'} x_i$ and the sum of products $\sum_{j=1}^{d} \theta_{Vj} f_{Vj}(x_i, y_{i,m})$ is a measure of the compatibility between the brain signals x_i and the movement of the finger represented by the label $y_{i,m}$. The edge potential function is defined as:

$$\Phi_i(y_{i,m}, y_{i-1,m}) = e^{\theta_E f_E(y_{i,m}, y_{i-1,m})} \tag{7.3}$$

where $f_E(y_{i,m}, y_{i-1,m}) = \mathbf{1}_{y_{i,m}=y', y_{i-1,m}=y''}$. Therefore, the product $\theta_e f_e(y_{i,m}, y_{i-1,m})$ is the compatibility between the current state and the previous state, i.e it provides a prior on the transition probabilities between classes. This information is useful given that it is known that the subject does not execute rapid changes between movement and rest. In this way, information about the extrinsic dynamics of the brain signals is incorporated in the model.

7.2.1.2 Approach two: Grid-CRF

The second approach is an extension of the chain-CRF described in the previous section. In this case we allow connections between chain-CRFs as described in Figure 7.4, where a top view is shown for convenience. Note that although not visible in Figure 7.4, each component x_i is connected to $y_{i,j}$ where j indexes the finger that the chain models.

In this graph a new set of edges is added. These connections can in principle allow one to incorporate more prior information about the extrinsic dynamics of the movements. The new edges, for instance, can provide information that explains that only one finger moves at a specific time point. Furthermore, if more information can be obtained (i.e., correlation between the movements of the fingers), it can be incorporated in this model. We model the conditional probability distribution for the grid-CRF model as:

$$p_{grid}(\mathbf{y}/\mathbf{x},\Theta) = \frac{1}{Z}\prod_{i=1}^{n}\prod_{k=1}^{m}\Psi_i(y_{i,m},x_i)\prod_{i=2}^{n}\prod_{k=1}^{m}\Phi_i(y_{i,m},y_{i-1,m})\prod_{k=2}^{m}\prod_{i=2}^{n}\Gamma_{i,k}(y_{i,k},y_{i,k-1}) \tag{7.4}$$

where n is the number of nodes and m the number of fingers. The potential functions Ψ_i and Φ_i were defined in Equations 7.2 and 7.3 respectively. The edge potential function $\Gamma_{i,k}(y_{i,k},y_{i,k-1})$ is defined as

$$\Gamma_{i,k}(y_{i,k},y_{i,k-1}) = e^{\theta_{Ck}f_{Ck}(y_{i,k},y_{i,k-1})} \tag{7.5}$$

where the feature function $f_{Ck}(y_{i,k},y_{i,k-1}) = \mathbf{1}_{y_{i,k}=y',y_{i,k-1}=y''}$ and the product $\theta_{Ck}f_{Ck}(y_{i,k},y_{i,k-1})$ is a measure of the compatibility between the label assigned to the movement of different fingers at time point i. Note that these edges only connect models related to neighbor fingers.

7.2.2 Multi-class Classification

The multi-class classification problem is solved by using the same structure used in Section 7.2.1.1 for one independent chain-CRF. In this case, the labels take values from 0 to 5 and describe a six-class classification problem.

7.3 Classification Results

The classifiers were tested using a cross-validation scheme. The signals were divided into 5 segments. Each time, 4 segments are used for training while 1 segment is used for

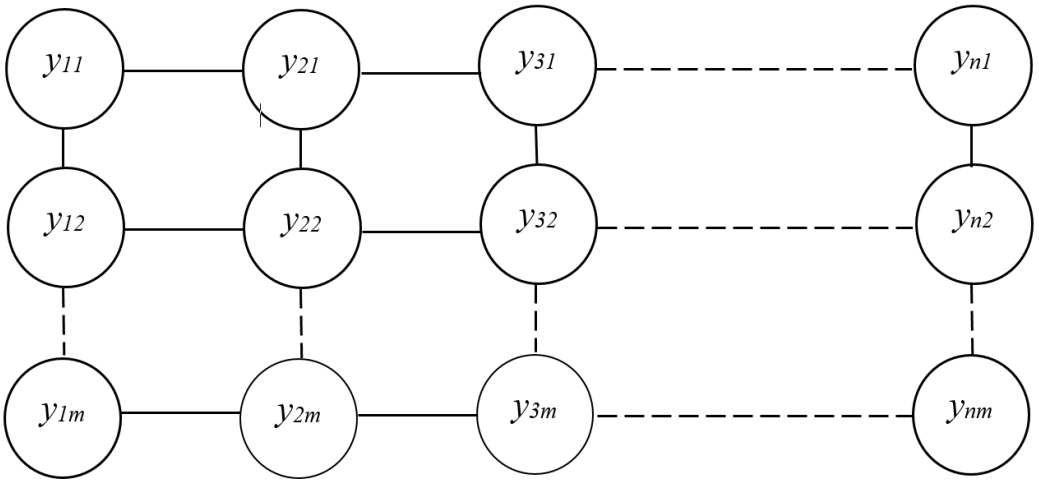


FIGURE 7.4: Graph for the grid-CRF Model

testing. We repeat this process until all segments have been used one time for testing. Also, a static classifier based on logistic regression was trained for comparison, using the same features that are applied to the method proposed. Finally, we used the Cohen's kappa coefficient as a metric to measure the performance of the classifiers. The use of this metric is justified because the labels are highly unbalanced, that is, the numbers of samples in the testing set for each class are not equal, which can lead to misinterpretations of the results.

7.3.1 Classification of Movement Versus Rest

The classification results are summarized in Figure 7.5. The results show that the grid-CRF provides the best results. From a theoretical point of view, these results are expected. Note that each one of the methods used for comparison is in fact a log-linear model. The logistic regression models the relationship between the signals and the labels. The chain-CRF goes beyond this by modeling the relationships between nodes. These nodes represent the variables at different time points and the edges between them model the dynamics of the labels (see Equation 7.3). As shown, this prior knowledge improves the classification results. The grid-CRF model adds another source of information by taking into consideration relationships between nodes that represent the activities in different fingers. This information is learned on the parameters θ_C and included in the model, further increasing the performance.

7.3.2 Multi-class classification

The results obtained for multi-class classification are summarized in Figure 7.6. Here, the multi-class chain-CRF is compared to logistic regression. The main difference between these models is that the chain-CRF takes into consideration the extrinsic dynamics by

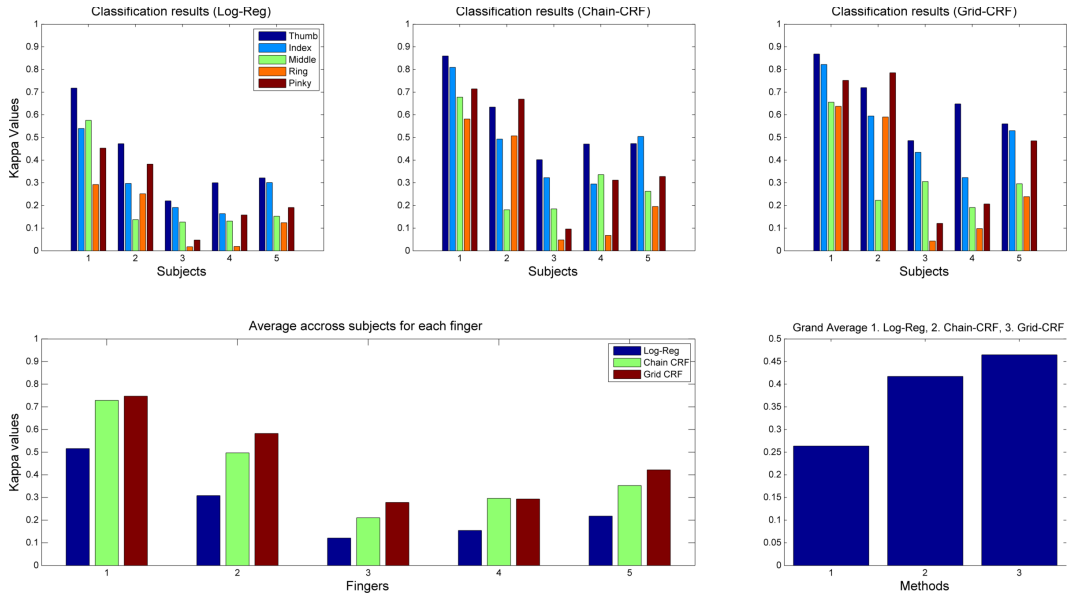


FIGURE 7.5: Summary of classification results for movement versus rest for each finger

means of the parameters θ_E. Including this information, as in the case of classification between movement and rest, results in better classification results. Also note that further improvement could be obtained with the logistic regression by imposing smoothness to the output of the classifier. Although this is a technique used in BCI, it requires the parameters of the smoothing filter to be set by the designer. The advantage of the proposed method is that this information is learned from the data, therefore if the data supports smoothness or not, or to what degree, is learned automatically.

7.4 conclusion

In this chapter we have proposed the use of graphical models for continuous classification of finger movements from ECoG recordings. The high spatial resolution of ECoG allows the determination of active regions during the execution of specific motor tasks, which is useful for preliminary electrode selection. We present two classification problems. The first involves the classification of movement versus rest in all fingers. The second is a multi-class problem in which the goal is to determine which finger moves or if all fingers are in rest (six-class problem). For the binary classification problem we propose the use of independent chain-CRFs and a more general CRF (grid-CRF). The grid-CRF provides best results in terms of performance given that it takes into consideration the temporal dynamics of the task as well as information about the movements of neighbor fingers. Note that the proposed structure allows one to include more information about the relationship between the processes to be modeled as well. In the second classification problem we use a chain-CRF with augmented states allowing the classification of the movement of each of the five fingers and rest periods. The results shows that including

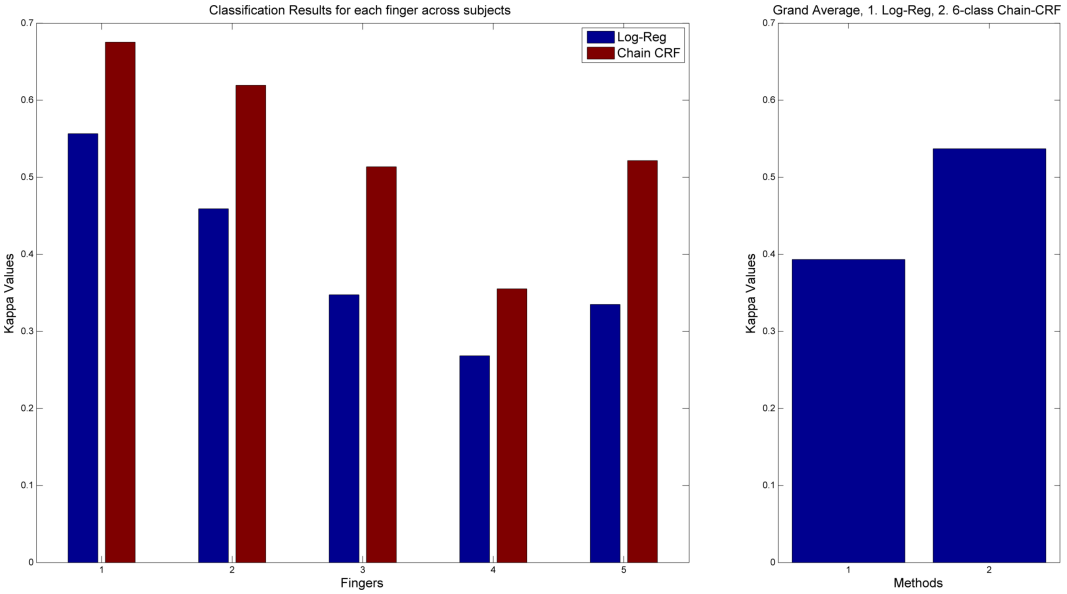


FIGURE 7.6: Summary of classification results for the multi-class problem

temporal information that capture the extrinsic dynamics of the brain signals, has a significant impact on the classification performance. Future work will be focus on the modeling of spatial relationships between electrodes during the execution of motor task. For this, measures of connectivity as presented in Section 2.4.3 will be incorporated in models with similar structure to the grid-CRF model proposed in this chapter.

Chapter 8

Contributions and Future Work

8.1 Summary of Contributions

As we have shown in this thesis, probabilistic graphical models have the capability to incorporate information that can help to improve the performance of BCI systems. In Chapter 3, we presented a discriminative graphical model for application on the P300 speller system. The proposed model improves the performance of the P300 speller by including a language model at the level of words. This makes it possible to decrease the time that the subject takes to spell a letter, and hence increases the transfer rate of the BCI. The proposed model integrates all the aspects of the BCI P300 speller in a single model, in a probabilistic fashion which allows to determine confidence intervals at any level in the system, from the brain signals to the probability of the words in a custom dictionary that can be modified to fit particularities of the subjects.

Then we turned our attention from P300 to sensorimotor rhythms and examined the potential of various graphical modeling formalisms in this setting. We began by proposing in Chapter 4 a non-parametric Bayesian approach for modeling the internal dynamics of the brain signals in the context of a synchronous imaginary motor task classification problem. The results show that considering the brain signals as a sequence of states improves the performance in the detection of imaginary motor movements. One main problem was to determine the number of states, a problem common to HMM applications. The proposed method overcomes this problem by making use of Hierarchical Dirichlet Processes, allowing to let the number of states to be determined according to the characteristics of the data. The idea of modeling the temporal structure of the data shows that the different states can be related to processes of Event Related Synchronization/Desynchronization of different brain rhythms.

Results in Chapter 4 raise questions about the classification problem. It is necessary to train one model for each class and calculate the likelihood of the data given the model

in order to declare a particular class. Also the generative nature of the models may not be well suited for the classification problem given that it models the brain signals assuming that the distribution of the data is known. In order to solve this issue we proposed the use of a Discriminative latent model, capable of modeling the dynamics of the brain signals over a discriminative framework integrating the modeling of the signals during different motor tasks in a single model. This is proposed in Chapter 5 in the context of synchronous BCI tasks using Hidden Conditional Random Fields (HCRF). The results show that this approach over performs classical approaches to BCI based on static classifiers as well as methods proposed recently.

The methods proposed in Chapters 4 and 5 are suited for synchronous BCI systems. In the case of asynchronous BCIs, the subjects do not receive cues that indicate the beginning or end of a particular task. We proposed in Chapter 6 a model suitable for asynchronous BCI that takes into account the intrinsic dynamics of the brain signals (as in the HCRF model) and also the extrinsic dynamics. The results obtained with the proposed Latent Dynamics CRF-based method show that this approach overperforms other methods based on Hierarchical HMM, Hierarchical CRF and static classifiers. Furthermore, the method provides a mechanism to take into consideration characteristics of the executed tasks. Assuming that the subject alternates between tasks in a smooth way (which is a realistic assumption), the model can take that this information into consideration if the data support it. This is in contrast with methods that use smoothing filters that should be tuned manually.

In order to solve some of the limitations of the EEG recordings related to low SNR and low spatial resolution, we moved into the use of another recording method. We used electro-corticographic signals in subjects during the execution of voluntary finger movements. In Chapter 7 we proposed the use of graphical models for continuous classification of finger movements. The proposed methods are based on CRFs. We proposed initially the modeling of the movement of the fingers independently using only spectro-temporal information and showed that in fact a significant improvement is obtained when compared to static methods. Also we proposed a model that that can incorporate statistical relationships of the states of movements of different fingers. This method, called in this thesis grid-CRF, obtains as expected a better performance. We note that although we only include a simple piece of prior information that only one finger moves at a time point for illustration,, different measures of interaction between different areas of the brain can be incorporated in this model as we will discuss in the next section..

8.2 Future Work

We end this thesis by suggesting future lines of work in relation with the results and observations in this thesis. The method proposed in Chapter 3 can be extended by modeling the language in higher levels. This does not only involve the use of phrases but also the incorporation of context in the model. Furthermore, a model that is able to learn from the particularities of the subject in terms of language could provide further improvements in the performance. In terms of the structure of the speller matrix, it is observed that many of the mistakes involve declaration of letters that are placed in rows or columns that are neighbors to the actual letter that the subject wants to spell. This information can be included in the system by changing the concept of row and columns by modeling directly each letter in the speller matrix. Within this approach, a map of probabilities for each letter can be obtained and the distribution of errors can be learned from training data. On the top of this structure, the language model as proposed in this thesis can be built. Finally, in Chapter 3 the model proposed assumes that the beginning and end of the words is known. In online decoding, the character "-" can be used to denote the end of each word. With this augmentation, the method proposed in Chapter 3 can be implemented.

In Chapter 4, we proposed the use of non-parametric Bayesian methods involving Hierarchical Dirichlet Processes as a solution for the selection of the number of hidden states. Future work can include a deeper analysis of the meaning of the states. Although it was found that the states are related to synchronization or de-synchronization of specific rhythms, more detailed analysis could reveal different processes happening at different locations and at different frequency bands.

In Chapters 5 and 6 discriminative models for synchronous and asynchronous BCI were proposed. These methods make use of hidden states and the optimal number of states is selected by cross-validation. Alternatively, similar to the use of HDPs in Chapter 4 in the context of generative models, nonparametric models could possibly be used in the context of discriminative models as well. Such a data driven approach might be likely to produce better results and improve the learning process in HCRF and LDCRF methods.

In Chapters 3, 4, 5, 6 and 7 the features involve a sequence of brain signals in different frequency bands. Different features are concatenated and used as input of the classifiers. These features are built by extracting power in different frequency bands. Given that different frequency bands are related to different kind of dynamics (e.g., rebound in alpha, de-synchronization in alpha, onset of gamma with initiation of motor task), an independent model for each frequency band could provide more insight about the complex dynamics that take place in the brain during the execution of a specific task. Note, for instance, that the number of states needed to model dynamics in alpha band does not necessarily have to match the number of states needed to model the dynamics in beta or gamma.

In general, independently of the application, the exploration of features that measure connectivity in different areas of the brain could lead to more descriptive models. In particular, in the case of motor movements or imagination of motor activity, it is well known that different parts of the brain interact (e.g., interactions between pre-motor cortex - motor cortex - somatosensory cortex). This kind of interactions can be incorporated in a graphical model framework together with the temporal structure of the brain signals in different frequency bands, generating spectro-temporal and spatial representations of the brain activity during specific tasks. To achieve this, ECoG recordings offer a great opportunity given their excellent spatial resolution and high signal-to-noise ratio. The literature offers various attempts to model such interactions. As was described in Section 2.4.3, measures based on methods such as the Directed Transfer Function (DTF), Partial Directed Coherence (PDC) and Phase Locking Value (PLV), among others could be incorporated in a graphical model. However, these measures have limitations related to the amount of data available for learning autoregressive parameters in the cases of DTF and PDC which poses a challenge for their use in single trial operation. In the case of PLV volume conduction poses a major problem as presented in [129] where it is shown that in EEG recordings the PLV measure is likely to be dominated by propagation of the brain signals between different sites. Therefore, new measures for effective modeling of the interaction between different regions are needed.

Bibliography

[1] S. Shahjahan and P. Girijesh, "Bispectrum - based feature extraction technique for devising a practical brain-computer interface.," *Journal of neural engineering*, vol. 8, p. 025014, Mar. 2011.

[2] A. Nijholt, B. Reuderink, and D. Oude Bos, "Turning Shortcomings into Challenges: Brain - Computer Interfaces for Games," in *Intelligent Technologies for Interactive Entertainment* (O. Akan, P. Bellavista, and Cao, eds.), vol. 9, pp. 153–168, Springer Berlin Heidelberg, 2009.

[3] D.-O. Bos, M. Duvinage, O. Oktay, J. Delgado Saa, H. Guruler, A. Istanbullu, M. van Vliet, B. van de Laar, M. Poel, L. Roijendijk, L. Tonin, A. Bahramisharif, and B. Reuderink, "Looking around with your brain in a virtual world," in *Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), 2011 IEEE Symposium on*, pp. 1–8, 2011.

[4] L. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalography and Clinical Neurophysiology*, vol. 70, no. 6, pp. 510 – 523, 1988.

[5] J. Elshout, "Review of Brain-Computer Interfaces based on the P300 evoked potential," Master's thesis, Universiteit Utrecht, 2009.

[6] J. R. Wolpaw, D. J. McFarland, T. M. Vaughan, and G. Schalk, "The Wadsworth Center brain-computer interface (BCI) research and development program.," *IEEE Trans Neural Syst Rehabil Eng*, vol. 11, pp. 204–7, 06/2003 2003.

[7] T. J. Sejnowski, G. Dornhege, J. d. R. Millán, T. Hinterberger, D. J. McFarland, and K.-R. Müller, *Toward Brain-Computer Interfacing*, ch. 2. Cambridge, Massachusetts: MIT Press, 2007.

[8] J. N. Mak, Y. Arbel, J. W. Minett, L. M. McCane, B. Yuksel, D. Ryan, D. Thompson, L. Bianchi, and D. Erdogmus, "Optimizing the P300-based braincomputer interface: current status, limitations and future directions," *Journal of Neural Engineering*, vol. 8, no. 2, p. 025003, 2011.

[9] W. Speier, C. Arnold, J. Lu, R. K. Taira, and N. Pouratian, "Natural language processing with dynamic classification improves p300 speller accuracy and bit rate," *Journal of Neural Engineering*, vol. 9, no. 1, p. 016004, 2012.

[10] U. Orhan, D. Erdogmus, B. Roark, S. Purwar, K. Hild, B. Oken, H. Nezamfar, and M. Fried-Oken, "Fusion with language models improves spelling accuracy for erp-based brain computer interface spellers," in *Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE*, pp. 5774–5777, 2011.

[11] C. Ulas and M. Cetin, "The first brain-computer interface utilizing a turkish language model," in *Signal Processing and Communications Applications Conference (SIU), 2013 21st*, pp. 1–4, 2013.

[12] F. Lotte, M. Congedo, A. Lcuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based braincomputer interfaces," *Journal of Neural Engineering*, vol. 4, no. 2, p. R1, 2007.

[13] L. da Silva and G. Pfurtscheller, *Basic concepts on EEG synchronization and desynchronization*, vol. 6, pp. 1–14. Elsevier, 1999.

[14] 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, pp. 1299–1309, Oct. 2001.

[15] A. Schlögl, *The Electroencephalogram and the Adaptive Autoregressive Models: Theory and Applications*. PhD thesis, vorgelegt an der Technischen Universitt Graz, Graz, April 2000.

[16] D. J. McFarland and J. R. Wolpaw, "Sensorimotor rhythm-based braincomputer interface (BCI): model order selection for autoregressive spectral analysis," *Journal of Neural Engineering*, vol. 5, no. 2, p. 155, 2008.

[17] A. O. Argunsah and M. Çetin, "AR-PCA-HMM Approach for Sensorimotor Task Classification in EEG-based Brain-Computer Interfaces," *2010 20th International Conference on Pattern Recognition*, pp. 113–116, Aug. 2010.

[18] H.-I. Suk and S.-W. Lee, "Two - Layer Hidden Markov Models for Multi-class Motor Imagery Classification," *First Workshop on Brain Decoding: Pattern Recognition Challenges in Neuroimaging*, pp. 5–8, Aug. 2010.

[19] C. Vidaurre, N. Krämer, B. Blankertz, and A. Schlögl, "Time Domain Parameters as a feature for EEG-based Brain-Computer Interfaces," *Neural Networks*, vol. 22, no. 9, pp. 1313 – 1319, 2009. Brain-Machine Interface.

[20] S. Bengio, "Hmm and iohmm modeling of eeg rhythms for asynchronous bci systems," in *In European Symposium on Artificial Neural Networks, ESANN*, 2004.

[21] H. B. Awwad Shiekh and J. Q. Gan, "Conditional random fields as classifiers for three-class motor-imagery brain-computer interfaces.," *Journal of neural engineering*, vol. 8, p. 025013, Mar. 2011.

[22] S. Sanei and J. Chambers, *EEG signal processing.* Prentece Hall, 2007.

[23] G. Schalk, "Can electrocorticography (ecog) support robust and powerful brain-computer interfaces?," *Frontiers in Neuroengineering*, vol. 3, no. 9, 2010.

[24] J. Wolpaw, E. Winter-Wolpaw, and G. Schalk, *BCIs That Use Electrocorticographic Activity.*, ch. 15. Oxford University Press, 2012.

[25] N. Birbaumer, A. Kubler, N. Ghanayim, T. Hinterberger, J. Perelmouter, J. Kaiser, I. Iversen, B. Kotchoubey, N. Neumann, and H. Flor, "The thought translation device (ttd) for completely paralyzed patients," *Rehabilitation Engineering, IEEE Transactions on*, vol. 8, pp. 190 –193, jun 2000.

[26] T. J. Sejnowski, G. Dornhege, J. d. R. Millán, T. Hinterberger, D. J. McFarland, and K.-R. Müller, *Toward Brain-Computer Interfacing*, ch. 1. Cambridge, Massachusetts: MIT Press, 2007.

[27] D. Krusienski, E. Sellers, D. McFarland, T. Vaughan, and J. Wolpaw, "Toward enhanced {P300} speller performance," *Journal of Neuroscience Methods*, vol. 167, no. 1, pp. 15 – 21, 2008. ¡ce:title¿Brain-Computer Interfaces (BCIs)¡/ce:title¿.

[28] E. Donchin, K. Spencer, and R. Wijesinghe, "The mental prosthesis: assessing the speed of a p300-based brain-computer interface," *Rehabilitation Engineering, IEEE Transactions on*, vol. 8, pp. 174 –179, jun 2000.

[29] E. W. Sellers, D. J. Krusienski, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "A p300 event-related potential brain-computer interface (bci): The effects of matrix size and inter stimulus interval on performance," *Biological Psychology*, vol. 73, no. 3, pp. 242 – 252, 2006.

[30] T. Vaughan, D. McFarland, G. Schalk, W. Sarnacki, D. Krusienski, E. Sellers, and J. Wolpaw, "The wadsworth bci research and development program: at home with bci," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 14, pp. 229 –233, june 2006.

[31] Y. Wang, R. Wang, X. Gao, B. Hong, and S. Gao, "A practical vep-based brain-computer interface," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 14, pp. 234 –240, june 2006.

[32] G. Pfurtscheller and F. Lopes da Silva, "Event - related EEG/MEG synchronization and desynchronization: basic principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842 – 1857, 1999.

[33] D. McFarland and J. Wolpaw, "Sensorimotor rhythm-based brain-computer interface (bci): feature selection by regression improves performance," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 13, no. 3, pp. 372–379, 2005.

[34] R. Leeb, F. Lee, C. Keinrath, R. Scherer, H. Bischof, and G. Pfurtscheller, "Brain-computer communication: motivation, aim, and impact of exploring a virtual apartment.," *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 15, pp. 473–82, Dec. 2007.

[35] J. d. R. Millan, "On the need for on-line learning in brain-computer interfaces," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, vol. 4, pp. 2877 – 2882 vol.4, july 2004.

[36] J. F. Delgado Saa and M. Çetin., "Modeling differences in the time - frequency presentation of EEG signals through HMMs for classification of imaginary motor tasks," *Sabanci University*, 2011.

[37] A. Schlögl, C. Keinrath, D. Zimmermann, R. Scherer, R. Leeb, and G. Pfurtscheller, "A fully automated correction method of EOG artifacts in EEG recordings.," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 118, pp. 98–104, Jan. 2007.

[38] J. R. Hughes, "Gamma, fast, and ultrafast waves of the brain: Their relationships with epilepsy and behavior," *Epilepsy and Behavior*, vol. 13, no. 1, pp. 25 – 31, 2008.

[39] Z. Wang, q. ji, K. J. Miller, and G. Schalk, "Prior knowledge improves decoding of finger flexion from electrocorticographic (ecog) signals," *Frontiers in Neuroscience*, vol. 5, no. 127, 2011.

[40] I. Gold, "Does 40-hz oscillation play a role in visual consciousness?," *Consciousness and Cognition*, vol. 8, no. 2, pp. 186 – 195, 1999.

[41] J. Kubánek, J. W. Miller, J. G. Ojemann, J. R. Wolpaw, and G. Schalk, "Decoding flexion of individual fingers using electrocorticographic signals in humans.," *J Neural Eng*, vol. 6, p. 066001, 12/2009 2009.

[42] T. J. Sejnowski, G. Dornhege, J. d. R. Millán, T. Hinterberger, D. J. McFarland, and K.-R. Müller, *Toward Brain-Computer Interfacing*, ch. 13. Cambridge, Massachusetts: MIT Press, 2007.

[43] H. Nai-Jen and R. Palaniappan, "Classification of mental tasks using fixed and adaptive autoregressive models of EEG signals," in *Engineering in Medicine and Biology Society, 2004. IEMBS '04. 26th Annual International Conference of the IEEE*, vol. 1, pp. 507–510, 2004.

[44] A. Schloegl, K. Lugger, and G. Pfurtscheller, "Using adaptive autoregressive parameters for a brain-computer-interface experiment," in *Engineering in Medicine and Biology Society, 1997. Proceedings of the 19th Annual International Conference of the IEEE*, vol. 4, pp. 1533–1535 vol.4, 1997.

[45] T. Bassani and J. Nievola, "Brain-Computer Interface Using Wavelet Transformation and Naive Bayes Classifier," in *Brain Inspired Cognitive Systems 2008* (A. Hussain, I. Aleksander, L. S. Smith, A. K. Barros, R. Chrisley, and V. Cutsuridis, eds.), vol. 657 of *Advances in Experimental Medicine and Biology*, pp. 147–165, Springer New York, 2010.

[46] M. Khalid, N. Rao, I. Rizwan-i Haque, S. Munir, and F. Tahir, "Towards a Brain Computer Interface using wavelet transform with averaged and time segmented adapted wavelets," in *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, pp. 1–4, 2009.

[47] K. Li, V. Narayan Raju, R. Sankar, Y. Arbel, and E. Donchin, "Advances and Challenges in Signal Analysis for Single Trial P300-BCI," in *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems* (D. Schmorrow and C. Fidopiastis, eds.), vol. 6780 of *Lecture Notes in Computer Science*, pp. 87–94, Springer Berlin Heidelberg, 2011.

[48] C. Gerloff, J. Richard, J. Hadley, A. E. Schulman, M. Honda, and M. Hallett, "Functional coupling and regional activation of human cortical motor areas during simple, internally paced and externally paced finger movements.," *Brain*, vol. 121, no. 8, pp. 1513–1531, 1998.

[49] J. Sarvas, "Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem," *Phys Med Biol*, vol. 32, no. 2, p. 11 22, 1987.

[50] M. Kaminski and K. Blinowska, "A new method of the description of the information flow in the brain structures," *Biological Cybernetics*, vol. 65, pp. 203–210, 1991. 10.1007/BF00198091.

[51] T. Mima, T. Matsuoka, and M. Hallett, "Functional coupling of human right and left cortical motor areas demonstrated with partial coherence analysis," *Neuroscience Letters*, vol. 287, no. 2, pp. 93 – 96, 2000.

[52] G. Nolte, O. Bai, L. Wheaton, Z. Mari, S. Vorbach, and M. Hallett, "Identifying true brain interaction from eeg data using the imaginary part of coherency," *Clinical Neurophysiology*, vol. 115, no. 10, pp. 2292 – 2307, 2004.

[53] A. Korzeniewska, "Determination of information flow direction among brain structures by a modified directed transfer function (dDTF) method," *Journal of Neuroscience Methods*, vol. 125, pp. 195–207, May 2003.

[54] J. Ginter, K. J. Blinowska, M. Kaminski, P. J. Durka, G. Pfurtscheller, and C. Neuper, "Propagation of EEG activity in the beta and gamma band during movement imagery in humans.," *Methods of information in medicine*, vol. 44, pp. 106–13, Jan. 2005.

[55] L. A. Baccalá and K. Sameshima, "Partial directed coherence: a new concept in neural structure determination.," *Biological cybernetics*, vol. 84, pp. 463–74, June 2001.

[56] J. P. Lachaux, E. Rodriguez, J. Martinerie, and F. J. Varela, "Measuring phase synchrony in brain signals.," *Human Brain Mapping*, vol. 8, pp. 194–208, jan 1999.

[57] J. P. Lachaux, E. Rodriguez, J. Martinerie, M. L. V. Quyen, A. Lutz, and F. J. Varela, "Studying Single-Trials of Phase Synchronous Activity in the Brain," *International Journal of Bifurcation and Chaos*, vol. 10, no. 10, pp. 2429–2439, 2000.

[58] P. Tass, M. G. Rosenblum, J. Weule, J. Kurths, A. Pikovsky, J. Volkmann, A. Schnitzler, and H. Freund, "Detection of n:m Phase Locking from Noisy Data : Application to Magnetoencephalography," *Physical Review Letters*, vol. 81, no. 15, pp. 3291–3294, 1998.

[59] M. Le Van Quyen, J. Foucher, J. Lachaux, E. Rodriguez, A. Lutz, J. Martinerie, and F. J. Varela, "Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony.," *Journal of neuroscience methods*, vol. 111, pp. 83–98, Oct. 2001.

[60] Y. Wang, B. Hong, X. Gao, and S. Gao, "Phase synchrony measurement in motor cortex for classifying single-trial eeg during motor imagery," in *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE*, pp. 75 –78, 30 2006-sept. 3 2006.

[61] E. Gysels and P. Celka, "Phase synchronization for the recognition of mental tasks in a brain-computer interface," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 12, pp. 406 – 415, dec. 2004.

[62] F. C. Meinecke, A. Ziehe, J. Kurths, and K.-R. Müller, "Measuring phase synchronization of superimposed signals," *Phys. Rev. Lett.*, vol. 94, p. 084102, Mar 2005.

[63] C. Brunner, R. Scherer, B. Graimann, G. Supp, and G. Pfurtscheller, "Online control of a brain-computer interface using phase synchronization," *Biomedical Engineering, IEEE Transactions on*, vol. 53, pp. 2501 –2506, dec. 2006.

[64] J. Hu, Z. Mu, and J. Wang, "Phase locking analysis of motor imagery in brain-computer interface," in *Proceedings of the 2008 International Conference on*

BioMedical Engineering and Informatics - Volume 02, (Washington, DC, USA), pp. 478–481, IEEE Computer Society, 2008.

[65] A. Ziehe and K.-R. Müller, "TDSEP – an efficient algorithm for blind separation using time structure," in *Proceedings of the 8th International Conference on Artificial Neural Networks, ICANN'98* (L. Niklasson, M. Bodén, and T. Ziemke, eds.), Perspectives in Neural Computing, (Berlin), pp. 675 – 680, Springer Verlag, 1998.

[66] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recogn. Lett.*, vol. 15, pp. 1119–1125, November 1994.

[67] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Pattern Recognition*, vol. 33, no. 1, pp. 25 – 41, 2000.

[68] J. Schenk, M. Kaiser, and G. Rigoll, "Selecting Features in On-Line Handwritten Whiteboard Note Recognition: SFS or SFFS?," in *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, pp. 1251 –1254, july 2009.

[69] D. J. Krusienski, E. W. Sellers, F. Cabestaing, S. Bayoudh, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "A comparison of classification techniques for the p300 speller," *Journal of Neural Engineering*, vol. 3, no. 4, p. 299, 2006.

[70] C. Sutton and A. McCallum, "An introduction to conditional random fields.," *Found. Trends Mach. Learn.*, vol. 4, no. 4, pp. 267–373, 2011.

[71] C. Elkan, "Log-linear models and conditional random fields," 2008.

[72] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[73] P. Brunner, A. L. Ritaccio, J. F. Emrich, H. Bischof, and G. Schalk, "Rapid communication with a ?p300? matrix speller using electrocorticographic signals (ecog)," *Frontiers in Neuroscience*, vol. 5, no. 5, 2011.

[74] D. Krusienski and J. Shih, "Spectral components of the p300 speller response in electrocorticography," in *Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference on*, pp. 282–285, 2011.

[75] Z. Cashero, *Comparison of EEG preprocessing methods to improve the performance of the P300 speller*. PhD thesis, Colorado State University, Feb. 2012.

[76] E. Donchin, K. Spencer, and R. Wijesinghe, "The mental prosthesis: assessing the speed of a p300-based brain-computer interface," *Rehabilitation Engineering, IEEE Transactions on*, vol. 8, no. 2, pp. 174–179, 2000.

[77] *Corpus-Driven Enhancement of a BCI Spelling Component*, Sept. 2007.

[78] S. Ahi, H. Kambara, and Y. Koike, "A dictionary-driven p300 speller with a modified interface," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 19, no. 1, pp. 6–14, 2011.

[79] G. Schalk, D. McFarland, T. Hinterberger, N. Birbaumer, and J. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 6, pp. 1034–1043, 2004.

[80] U. Brigham-Young, "Corpus of contemporary american english," 2012.

[81] D. C. Howell, *Statistical Methods for Psychology*. Belmont: Wadsworth Publishing Co Inc, international ed of 7th revised ed ed., 2009.

[82] S. M. M. Martens, J. M. Mooij, N. J. Hill, J. Farquhar, and B. Schölkopf, "A graphical model framework for decoding in the visual erp-based bci speller," *Neural Comput.*, vol. 23, pp. 160–182, Jan. 2011.

[83] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalography and Clinical Neurophysiology*, vol. 29, no. 3, pp. 306 – 310, 1970.

[84] R. Magjarevic, A. Yonas, A. S. Prihatmanto, and T. L. Mengko, "Time-Frequency Features Combination to Improve Single-Trial EEG Classification," in *World Congress on Medical Physics and Biomedical Engineering* (O. Dssel and W. C. Schlegel, eds.), vol. 25/4, pp. 805–808, Springer Berlin Heidelberg, 2010.

[85] R. T. Mina, A. Atiya, M. I. Owis, and Y. M. Kadah, "Brain-Computer Interface Based on Classification of Statistical and Power Spectral Density Features," *Biomedical Engineering*, pp. 2–5, 2006.

[86] R. Palaniappan, "Brain Computer Interface Design Using Band Powers Extracted During Mental Tasks," *Conference Proceedings. 2nd International IEEE EMBS Conference on Neural Engineering, 2005.*, pp. 321–324, 2005.

[87] Z. Mu, D. Xiao, and J. Hu, "Classification of Motor Imagery EEG Signals Based on TimeFrequency Analysis," *International Journal of Digital Content Technology and its Applications*, pp. 116–119, 2009.

[88] J. F. Delgado Saa and M. Sotaquir?, "Eeg signal classification using ar-power spectral features and linear discriminant analysis," *Proceedings of Latin American and Caribbean Consortium of Engineering Institutions, Arequipa - Per?*, 2010.

[89] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Readings in speech recognition* (A. Waibel and K.-F. Lee, eds.), pp. 267–296, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990.

[90] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 39, no. 1, pp. 1–38, 1977.

[91] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, vol. 101, pp. 1566–1581, Dec. 2006.

[92] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "Bayesian nonparametric methods for learning markov switching processes," *Signal Processing Magazine, IEEE*, vol. 27, pp. 43 –54, nov. 2010.

[93] G. Pfurstcheller *et al.*, "Data set 2b," in *BCI Competition 2008*, (Graz, Austria), 2008.

[94] C. Neuper, A. Schlögl, and G. Pfurstcheller, "Enhancement of left-right sensorimotor EEG differences during feedback-regulated motor imagery," *Clin. Neurophysiol*, vol. 16, no. 4, pp. 373–382, 1999.

[95] P. Stoica and R. L. . Moses, *Introduction to Spectral Analysis*. Prentece Hall, 1997.

[96] B. Jansen, J. Bourn, and J. Ward, "Autorregressive Estimation of Short Segment Spectra for Computarized EEG Analysis," *IEEE Transaction on Biomedical Engineering*, vol. BME-28, no. 8, pp. 630–637, 1981.

[97] T. J. Sejnowski, G. Dornhege, J. d. R. Millán, T. Hinterberger, D. J. McFarland, and K.-R. Müller, *Toward Brain-Computer Interfacing*, ch. 19. Cambridge, Massachusetts: MIT Press, 2007.

[98] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, (San Francisco, CA, USA), pp. 282–289, Morgan Kaufmann Publishers Inc., 2001.

[99] G. Luo and W. Min, "Subject-adaptive real-time sleep stage classification based on conditional random field.," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pp. 488–492, 2007.

[100] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *in Interspeech*, pp. 1117–1120, 2005.

[101] T. Sugiura, N. Goto, and A. Hayashi, "A discriminative model corresponding to hierarchical hmms," in *Proceedings of the 8th international conference on Intelligent data engineering and automated learning*, IDEAL'07, (Berlin, Heidelberg), pp. 375–384, Springer-Verlag, 2007.

[102] A. Quattoni, W. Sybor, L.-P. Morency, M. Collins, and T. Darrell, "Hidden Conditional Random Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1848–1852, 2007.

[103] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[104] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1521 – 1527, 2006.

[105] A. Satti, D. Coyle, and G. Prasad, "Continuous eeg classification for a self-paced bci," in *Neural Engineering, 2009. NER '09. 4th International IEEE/EMBS Conference on*, pp. 315–318, 2009.

[106] J. d. R. Millán, P. W. Ferrez, and A. Buttfield, "The idiap brain-computer interface: An asynchronous multi-class approach," in *Towards Brain-Computer Interfacing* (G. Dornhege, J. d. R. Millán, T. Hinterberger, D. McFarland, and K. R. Müller, eds.), The MIT Press, 0 2007.

[107] R. Leeb, D. Friedman, G. R. Müller-Putz, R. Scherer, M. Slater, and G. Pfurtscheller, "Self-paced (asynchronous) BCI control of a wheelchair in virtual environments: a case study with a tetraplegic," *Intell. Neuroscience*, vol. 2007, pp. 7:1–7:12, April 2007.

[108] J. Millan and J. Mourino, "Asynchronous BCI and local neural classifiers: an overview of the adaptive brain interface project," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, pp. 159 –161, june 2003.

[109] E. Sadeghian and M. Moradi, "Continuous detection of motor imagery in a four-class asynchronous bci," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pp. 3241 –3244, aug. 2007.

[110] C. Tsui and J. Gan, "Asynchronous BCI Control of a Robot Simulator with Supervised Online Training," in *Intelligent Data Engineering and Automated Learning - IDEAL 2007* (H. Yin, P. Tino, E. Corchado, W. Byrne, and X. Yao, eds.), vol. 4881 of *Lecture Notes in Computer Science*, pp. 125–134, Springer Berlin / Heidelberg, 2007.

[111] F. Velasco-Álvarez and R. Ron-Angevin, "Asynchronous brain-computer interface to navigate in virtual environments using one motor imagery," in *Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part I: Bio-Inspired Systems: Computational and Ambient Intelligence*, IWANN '09, (Berlin, Heidelberg), pp. 698–705, Springer-Verlag, 2009.

[112] F. Galán, F. Oliva, and J. Guárdia, "Using mental tasks transitions detection to improve spontaneous mental activity classification," *Medical and Biological Engineering and Computing*, vol. 45, pp. 603–609, 2007.

[113] Cheng-Jian Lin and Ming-Hua Hsieh, "Classification of mental task from EEG data using neural networks based on particle swarm optimization," *Neurocomputing*, vol. 72, no. 4?6, pp. 1121 – 1130, 2009.

[114] R. Aler, I. M. Galván, and J. M. Valls, "Transition detection for brain computer interface classification," in *Biomedical Engineering Systems and Technologies* (A. Fred, J. Filipe, and H. Gamboa, eds.), vol. 52 of *Communications in Computer and Information Science*, pp. 200–210, Springer Berlin Heidelberg, 2010.

[115] J. F. Delgado Saa and M. Çetin, "Hidden conditional random fields for classification of imaginary motor tasks from eeg data," in *Proceedings of 19th European Signal Processing Conference*, EUSIPCO, 2011.

[116] J. F. Delgado Saa and M. Çetin., "A latent discriminative model-based approach for classification of imaginary motor tasks from eeg data," *Journal of Neural Engineering*, vol. 9, no. 2, p. 026020, 2012.

[117] D. Kelly, J. McDonald, and C. Markham, "Evaluation of threshold model hmms and conditional random fields for recognition of spatiotemporal gestures in sign language," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pp. 490 –497, 27 2009-oct. 4 2009.

[118] H.-S. Yoon, J. Soh, Y. J. Bae, and H. S. Yang, "Hand gesture recognition using combined features of location, angle and velocity," *Pattern Recognition*, vol. 34, no. 7, pp. 1491 – 1501, 2001.

[119] K. P. Murphy and M. A. Paskin, "Linear time inference in hierarchical hmms," in *IN PROCEEDINGS OF NEURAL INFORMATION PROCESSING SYSTEMS*, 2001.

[120] S. Fine, Y. Singer, and N. Tishby, "The hierarchical hidden markov model: Analysis and applications," *Machine Learning*, vol. 32, pp. 41–62, 1998. 10.1023/A:1007469218079.

[121] J.-M. Cano-Izquierdo, J. Ibarrola, and M. Almonacid, "Improving Motor Imagery Classification With a New BCI Design Using Neuro-Fuzzy S-dFasArt," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 20, pp. 2 –7, jan. 2012.

[122] L.-P. Morency, A. Quattoni, and T. Darrell, "Latent-Dynamic Discriminative Models for Continuous Gesture Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition. CVPR '07.*, pp. 1 – 8, june 2007.

[123] R. Mohammadi, A. Mahloojifar, and D. Coyle, "A combination of pre- and post-processing techniques to enhance self-paced bcis," *Advances in Human-Computer Interaction*, vol. 2012, 2012.

[124] G. Townsend, B. Graimann, and G. Pfurtscheller, "Continuous eeg classification during motor imagery-simulation of an asynchronous bci," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 12, no. 2, pp. 258–265, 2004.

[125] K. J. Miller, C. J. Honey, D. Hermes, R. P. Rao, M. denNijs, and J. G. Ojemann, "Broadband changes in the cortical surface potential track activation of functionally diverse neuronal populations," *NeuroImage*, vol. 85, Part 2, no. 0, pp. 711 – 720, 2014. ¡ce:title¿New Horizons for Neural Oscillations¡/ce:title¿.

[126] R. Flamary and A. Rakotomamonjy, "Decoding finger movements from ECoG signals using switching linear models," *Frontiers in Neuroscience*, vol. 6, no. 29, 2012.

[127] W. Wang, A. D. Degenhart, J. L. Collinger, R. Vinjamuri, G. P. Sudre, P. Adelson, D. L. Holder, E. Leuthardt, D. Moran, M. L. Boninger, A. Schwartz, D. Crammond, E. C. Tyler-Kabara, and D. Weber, "Human motor cortical activity recorded with Micro-ECoG electrodes, during individual finger movements," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pp. 586–589, 2009.

[128] Z. Wang, Q. Ji, J. W. Miller, and G. Schalk, "Prior knowledge improves decoding of finger flexion from electrocorticographic signals.," *Frontiers in Neuroscience*, vol. 5, p. 127, 11/2011 2011.

[129] D. J. Krusienski, D. J. McFarland, and J. R. Wolpaw, "Value of amplitude, phase, and coherence features for a sensorimotor rhythm-based braincomputer interface," *Brain Research Bulletin*, vol. 87, no. 1, pp. 130 – 134, 2012.