Advances in Cognitive Neural Prosthesis: Recognition of Neural Data with an Information-Theoretic Objective

Zoran Nenadic
Department of Biomedical Engineering
Department of Electical Engineering and Computer Science
University of California
Irvine, CA 92697

Daniel S. Rizzuto and Richard A. Andersen
Division of Biology
California Institute of Technology
Pasadena, CA 91125

Joel W. Burdick
Division of Engineering and Applied Science
California Institute of Technology
Pasadena, CA 91125

Abstract

We give an overview of recent advances in cognitive-based neural prostheses, and point out the major differences with respect to commonly used motor-based brain-machine interfaces. While encouraging results in neuroprosthetic research have demonstrated the proof of concept, the development of practical neural prostheses is still in the phase of infancy. To address complex issues arising in the development of practical neural prostheses we review several related studies ranging from the identification of new cognitive variables to the development of novel signal processing tools.

In the second part of this chapter, we discuss an information-theoretic approach to the extraction of low-dimensional features from high-dimensional neural data. We argue that this approach may be better suited for certain neuroprosthetic applications than the

traditionally used features. An extensive analysis of electrical recordings from the human brain demonstrates that processing data in this manner yields more informative features than off-the-shelf techniques such as linear discriminant analysis. Finally, we show that the feature extraction is not only a useful dimensionality reduction technique, but also that the recognition of neural data may improve in the feature domain.

11.2 Introduction

The prospect of assisting disabled individuals by using neural activity from the brain to control prosthetic devices has been a field of intense research activity in recent years. The nature of neuroprosthetic research is highly interdisciplinary, with the brain-machine interfaces (BMIs) playing the central role. Although the development of BMIs can be viewed largely as a technological solution for a specific practical application, it also represents a valuable resource for studying brain mechanisms and testing new hypotheses about brain function.

Up to date, the majority of neuroprosthetic research studies have focused on deriving hand trajectories by recording their neural correlates, primarily, but not exclusively, from the motor cortex (Wessberg et al. (2000); Serruya et al. (2002); Taylor et al. (2002); Carmena et al. (2003); Mussa-Ivaldi and Miller (2003)). The trajectory information contained in the action potentials of individual neurons is decoded and the information is used to drive a robotic manipulator or a cursor on a computer screen. We refer to this neuroprosthetic approach as “motor-based.” Additionally, progress has been made in interfacing electroencephalographic (EEG) signals and assistive devices for communication and control (Wolpaw et al. (2002)). These noninvasive techniques are commonly termed brain-computer interfaces (BCIs) (Wolpaw and McFarland (2004); Pfurtscheller et al. (2003c)).

While remarkable success in the development of BMIs has been achieved over the past decade, practical neural prostheses are not yet feasible. Building a fully operational neuroprosthetic system presents many challenges ranging from long-term stability of recording implants to development of efficient neural signal processing algorithms. Since the full scope of prosthetic applications is still unknown and it is unlikely that a single BMI will be optimal for all plausible scenarios, it is important to introduce new ideas about the types of signals that can be used. It is also important to address the many technological challenges that are currently impeding the progress toward operational neural prostheses. To this end, the neuroprosthetic research effort of our group spans several related directions including cognitive-based BMIs, decoding from local field potentials (LFPs), identification of alternative cognitive control signals, electrophysiologic recording advances, and development of new decoding algorithms.

In section 11.3, we give a brief overview of these research efforts. More details can be found in the relevant literature cited. In section 11.4, we discuss novel information-theoretic tools for extraction of useful features from high-dimensional neural data. Experimental results with electrically recorded signals from the human brain are presented in section 11.5, and the advantages of our technique over traditional ones are discussed. Concluding remarks are given in section 11.6.

11.3 Advances in Cognitive Neural Prosthesis

The motor-based approach, although predominantly used, is certainly not the only way of using brain data for neuroprosthetic applications. Shenoy et al. (2003) argue that neural activity present before or even without natural arm movement provides an important source of control signals. In nonhuman primates, these types of neural signals can be found, among other areas, in parietal reach region (PRR) of the posterior parietal cortex (PPC). PPC is an area located at an early stage in the sensory-motor pathway (Andersen et al. (1997)), and is involved in transforming sensory inputs into plans for actions, so-called "sensory-motor integration." In particular, PRR was shown to exhibit directional selectivity with respect to planned reaching movements (Snyder et al. (1997)). Moreover, these plans are encoded in visual coordinates (also called retinal or eye-centered coordinates) relative to the current direction of gaze (Batista et al. (1999)), thus providing extrinsic spatial information and underscoring the cognitive nature of these signals. We refer to this approach to neural prostheses as "cognitive-based." The human homologue of PRR has recently been identified in functional-magnetic-resonance imaging experiments (Connolly et al. (2003)).

11.3.1 Cognitive-Based Brain-Machine Interfaces

The cognitive-based approach to neural prostheses does not require the execution of arm movements; its true potential lies in assisting paralyzed individuals who are unable to reach but who are capable of making reaching plans. It has been shown through a series of experiments (Musallam et al. (2004)) that monkeys easily learn to control the location of a computer cursor by merely thinking about movements. Briefly, the monkeys were shown a transient visual cue (target) at different screen locations over multiple trials. After the target disappeared, the monkeys were required to plan a reach movement to the target location without making any arm or eye movements. This stage of the experiment is referred to as the "delay" or "memory period." The action potentials (spike trains) of individual neurons from PRR were collected during the memory period and were decoded in real time to predict the target location. If the correct location was decoded, a feedback was provided to the animals by illuminating the target location and the animals were rewarded. The trials were aborted if the animals made eye or arm movements during the memory period. This ensured that only cognitive and not motor-related signals were used for decoding, thus underscoring the potential of the cognitive-based approach for severely paralyzed patients.

With vision being the main sensory modality of the posterior parietal cortex (Blatt et al. (1990); Johnson et al. (1996)), PRR is likely to continue receiving appropriate error signals after paralysis. In the absence of proprioceptive and somatosensory feedback (typically lost due to paralysis), visual error signals become essential in motor learning. Musallam et al. (2004) have shown that the performance of a PRR-operated prosthesis improved over the course of several weeks. Presumably, the visual feedback allowed the monkeys to learn how to compensate for decoding errors.

After reaching goals are decoded, trajectories can be computed from low-level trajectory instructions managed by smart output devices, such as robots, computers, or vehicles, using supervisory control systems (Sheridan (1992)). For example, given the Cartesian coordinates of an intended object for grasping, a robotic motion planner can determine the detailed joint trajectories that will transport a prosthetic hand to the desired location (Andersen et al. (2004a)). Sensors embedded in the mechanical arm can ensure that the commanded trajectories are followed and obstacles are avoided, thereby replacing, at least to some degree, the role of proprioceptive and somatosensory feedback.

11.3.2 Local Field Potentials

LFPs represent the composite extracellular potential from perhaps hundreds or thousands of neurons around the electrode tip. In general, LFPs are less sensitive to relative movement of recording electrodes and tissues; therefore, LFP recordings can be maintained for longer periods of time than single cell recordings (Andersen et al. (2004b)). However, LFPs have not been widely used in BMIs, perhaps because of the assumption that they do not correlate with movements or movement intentions as well as single cell activity. Recent experiments in monkey PPC, in particular the lateral intraparietal (LIP) area and PRR, have demonstrated that valuable information related to the animal's intentions can be uncovered from LFPs. For example, it has been shown that the direction of planned saccades in macaques can be decoded based on LFPs recorded from area LIP (Pesaran et al. (2002)). Moreover, the performances of decoders based on spike trains and LFPs were found to be comparable. Interestingly, the decoding of behavioral state (planning vs. execution of saccades) was more accurate with LFPs than with spike trains. Similar studies have been conducted in PRR. It was found that the decoding of the direction of planned reaches was only slightly inferior with LFPs than with spike trains (Scherberger et al. (2005)). As with LIP studies, it has also been shown that LFPs in this area provide better behavioral state (planning vs. execution of reaching) decoding than do spike trains.

While the decoding of a target position or a hand trajectory provides information on ***where*** to reach, the decoding of a behavioral state provides the information on ***when*** to reach. In current experiments, the time of reach is controlled with experimental protocol by supplying a "go signal." Practical neural prostheses cannot rely on external cues to initiate the movement; instead this information should be decoded from the brain, and future BMIs are likely to incorporate the behavioral state information. Therefore, it is expected that LFPs will play a more prominent role in the design of future neuroprosthetic devices.

11.3.3 Alternative Cognitive Control Signals

The potential benefits of a cognitive-based approach to neural prosthesis were demonstrated first through offline analysis (Shenoy et al. (2003)) and subsequently through closed loop (online) experiments (Musallam et al. (2004)). Motivated by previous findings of reward prediction based on neural activity in various brain areas (Platt and Glimcher (1999); Schultz (2004)), Musallam et al. (2004) have demonstrated that similar cognitive variables can be inferred from the activity in the macaques' PRR. In particular, they have found significant differences in cell activity depending on whether a preferred or nonpreferred reward was expected at the end of a trial. The experiments included various preferred versus nonpreferred reward paradigms such as citrus juice versus water, large amount versus small amount of reward, and high probability versus low probability of reward. On each day, the animal learned to associate one cue with the expectation of preferred reward and another cue with nonpreferred reward. The cues were randomly interleaved on a trial-by-trial basis. This study demonstrated that the performance of brain-operated cursor control increases under preferred reward conditions, and that both the reach goals and the reward type can be simultaneously decoded in real time.

The ability to decode expected values from brain data is potentially useful for future BMIs. The information regarding subjects' preferences, motivation level, and mood could be easily communicated to others in a manner similar to expressing these variables using body language. It is also conceivable that other types of cognitive variables, such as the patient's emotional state, could be inferred by recording activity from appropriate brain areas.

11.3.4 Neurophysiologic Recording Advances

One of the major challenges in the development of practical BMIs is to acquire meaningful data from many recording channels over a long period of time. This task is especially challenging if the spike trains of single neurons are used, since typically only a fraction of the electrodes in an implanted electrode array will record signals from well-isolated individual cells (Andersen et al. (2004b)). It is also hard to maintain the activity of isolated units in the face of inherent tissue and/or array drifts. Reactive gliosis (Turner et al. (1999)) and inadequate biocompatibility of the electrode's surface material (Edell et al. (1992)) may also contribute to the loss of an implant's function over time.

Fixed-geometry implants, routinely used for chronic recordings in BMIs, are not well suited for addressing the above issues. Motivated by these shortcomings, part of our research effort has been directed toward the development of autonomously movable electrodes that are capable of finding and maintaining optimal recording positions. Based on recorded signals and a suitably defined signal quality metric, an algorithm has been developed that decides when and where to move the recording electrode (Nenadic and Burdick (2006)). It should be emphasized that the developed control algorithm and associated signal processing steps (Nenadic and Burdick (2005)) are fully unsupervised, that is, free of any human involvement, and as such are suitable for future BMIs. Successful applications of the autonomously movable electrode algorithm using a meso-scale electrode testbed have recently been reported in Cham et al. (2005) and Branchaud et al. (2005).

The successful implementation of .autonomously movable electrodes in BMIs will be beneficial for several reasons. For example, electrodes can be moved to target specific neural populations that are likely to be missed during implantation surgery. Optimal recording quality could be maintained and the effects of cell migration can be compensated for by moving the electrodes. Finally, movable electrodes could break through encapsulation and seek out new neurons, which is likely to improve the longevity of recording.

Clearly, the integration of movable electrodes with BMIs hinges upon the development of appropriate micro-electro-mechanical systems (MEMS) technology. Research efforts to develop MEMS devices for movable electrodes are under way (Pang et al. (2005a,b)).

11.3.5 Novel Decoding Algorithms

In mathematical terms, the goal of decoding algorithms is to build a map between neural patterns and corresponding motor behavior or cognitive processes. Because of the randomness inherent in the neuro-motor systems, the appropriate model of this map is probabilistic. In practical terms, decoding for cognitive-based BMIs entails the selection of the intended reach target from a discrete set of possible targets. Consequently, the decoder is designed as a classifier, where observed neural data is used for classifier training.

Recent advances in electrophysiologic recordings have enabled scientists to gather increasingly large volumes of data over relatively short time spans. While neural data ultimately is important for decoding, not all data samples carry useful information for the task at hand. Ideally, relevant data samples should be combined into meaningful features, while irrelevant data should be discarded as noise. For example, representing a finely sampled time segment of neural data with a (low-dimensional) vector of firing rates, can be viewed as an heuristic way of extracting features from the data. Another example is the use of the spectral power of EEG signals in various frequency bands, for example, *p*-band or β-band (McFarland et al. (1997a); Pfurtscheller et al. (1997)), for neuroprosthetic applications such as BCIs.

In the next section, we cast the extraction of neural features within an information-theoretic framework and we show that this approach may be better suited for certain applications than the traditionally used heuristic features.

11.4 Feature Extraction

Feature extraction is a common tool in the analysis of multivariate statistical data. Typically, a low-dimensional representation of data is sought so that features have some desired properties. An obvious benefit of this dimensionality reduction is that data becomes computationally more manageable. More importantly, since the number of experimental trials is typically much smaller than the dimension of data (so-called small-sample-size problem (Fukunaga (1990))), the statistical parameters of data can be estimated more accurately using the low-dimensional representation.

Two major applications of feature extraction are representation and classification. Feature extraction for representation aims at finding a low-dimensional approximation of data, subject to certain criteria. These criteria assume that data are sampled from a common probability distribution, and so these methods are often referred to as blind or unsupervised. Principal component analysis (PCA) (Jolliffe (1986)) and independent component analysis (ICA) (Jutten and Herault (1991)) are the best-known representatives of these techniques. In feature extraction for classification, on the other hand, each data point's class memberchip is knnwn, and thus the method is considered supervised. Low-dimensional features are found that maximally preserve class differences measured by suitably defined criteria. Linear discriminant analysis (LDA) (Duda et al. (2001)) is the best known representative of these techniques. Once the features are extracted, a classifier of choice can be designed in the feature domain.[1]

A common heuristic approach to feature extraction is to rank individual (scalar) features according to some class separability criterion. For example, informative neural features are those that exhibit stimulus-related tuning, that is, they take significantly different values when conditioned upon different stimuli. The feature vector is then constructed by concatenating the several most informative features. While seemingly reasonable, this strategy is completely ignorant of the joint statistical properties of the features and may produce highly suboptimal feature vectors. More elaborate algorithms exist for the selection of scalar features (Kittler (1978)), but they are combinatorially complex (Cover and Campenhout (1977)) and their practical applicability is limited.

Another popular strategy for analyzing spatiotemporal neural signals is to separate the processing in the spatial and temporal domain. Data are first processed spatially, typically by applying off-the-shelf tools such as the Laplacian filter (McFarland et al. (1997a); Wolpaw and McFarland (2004)), followed by temporal processing, such as autoregressive frequency analysis (Wolpaw and McFarland (2004); Pfurtscheller et al. (1997)). However, the assumption of space-time separability is not justified and may be responsible for suboptimal performance. In addition, while spectral power features have clear physical interpretation, there is no reason to assume that they are optimal features for decoding. Rizzuto et al. (2005) have recently demonstrated that decoding accuracy with spectral power features could be up to 20 percent lower than a straightforward time domain decoding.

In the next two subsections, we introduce a novel information-theoreticcriterion for feature extraction conveniently called "information-theoretic discriminant analysis" (ITDA). We show that informative features can be extracted from data in a linear fashion, that is, through a matrix manipulation.[2] For spatiotemporal signals, the feature extraction matrix plays the role of a spatiotemporal filter and does not require an assumption about the separability of time and space. Moreover, the features are extracted using their joint statistical properties, thereby avoiding heuristic feature selection strategies and computationally expensive search algorithms.

11.4.1 Linear Supervised Feature Extraction

In general, linear feature extraction is a two-step procedure: (1) an objective function is defined and (2) a full-rank feature extraction matrix is found that maximizes such an objective. More formally, let $\mathbf{R} \in \mathbb{R}^n$ be a random data vector with the class-conditional probability density function (PDF) $f_{\mathbf{R}|\Omega}(\boldsymbol{r} \,|\, \omega_i)$, where the class random variable (RV) $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$ is drawn from a discrete distribution with the probability $P(\omega_i) \triangleq P(\Omega = \omega_i)$, $\forall i = 1, 2, \ldots$, c. For example, $\mathbf{R}$ could be a matrix of EEG data from an array of electrodes sampled in time and written in a vector form. The class variable could be the location of a visual target, or some cognitive task such as imagination of left and right hand movements (Pfurtscheller et al. (1997)). The features F $\in \mathbb{R}^m$ are extracted as

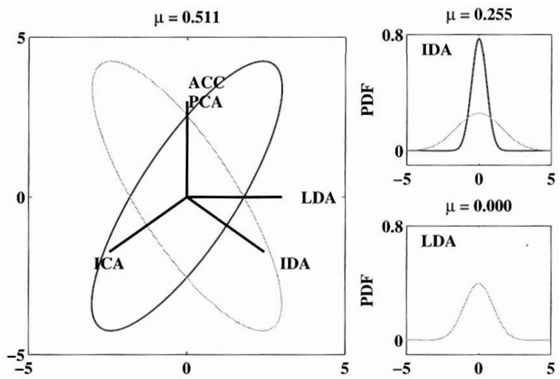

Figure 11.1 (Left) Two Gaussian class-conditional PDFs with $P(\omega_1) = P(\omega_2)$, represented by 3-Mahalanobis distance contours. The straight lines indicate optimal 1D subspace according to different feature extraction methods: PCA, ICA, LDA, ITDA and approximate Chernoff criterion (Loog and Duin (2004)) ACC. (Right) The PDFs of optimal 1D features extracted with ITDA and LDA.

$F = \boldsymbol{T}R$, where $\boldsymbol{T} \in W^{m \times n}$ is a full-rank feature extraction matrix found by maximizing a suitably chosen class separability objective function $J(\boldsymbol{T})$.

Many objective functions have been used for supervised feature extraction purposes. In its most common form, LDA, also known as the Fisher criterion (Fisher (1936)) or canonical variate analysis, maximizes the generalized Rayleigh quotient (Duda et al. (2001)). Under fairly restrictive assumptions, it can be shown that LDA is an optimal[3] feature extraction method. In practice, however, these assumptions are known to be violated, and so the method suffers from suboptimal performance. A simple example where LDA fails completely is illustrated in figure 11.1. Another deficiency of LDA is that the dimension of the extracted subspace is at most $c - 1$, where c is the number of classes. This constraint may severely limit the practical applicability of LDA features, especially when the number of classes is relatively small.

Kumar and Andreou (1998) have developed a maximum-likelihood feature extraction method and showed that these features are better suited for speech recognition than the classical LDA features. Saon and Padmanabhan (2000) used both Kullback-Leibler (KL) and Bhattacharyya distance as an objective function. However, both of these metrics are defined pairwise, and their extension to multicategory cases is often heuristic. Loog and Duin (2004) have developed an approximation of the Chernoff distance, although their method seems to fail in some cases (see figure 11.1).

Mutual information is a natural measure of class separability. For a continuous RV R and a discrete RV Ω, the mutual information, denoted by $\mu I(\mathsf{R}; \mathrm{R})$, is defined as

$$\mu I(\mathsf{R};\Omega) \triangleq H(\mathsf{R}) - H(\mathsf{R}\,|\,\Omega) = H(\mathsf{R}) - \sum_{i=1}^{c} H(\mathsf{R}\,|\,\omega_i)\, P(\omega_i) \qquad (11.1)$$

where $H(\mathsf{R}) \triangleq -\int f_{\mathsf{R}}(\boldsymbol{r}) \log(f_{\mathsf{R}}(\boldsymbol{r}))\, d\boldsymbol{r}$ is Shannon's entropy. Generally, higher mutual information implies better class separability and smaller probability of rnisclassification. In particular, it was shown in Hellman and Raviv (1970) that $\varepsilon_{\mathsf{R}} \leq 1/2\ [H(\Omega) - \mu I(\mathsf{R};\Omega)]$, where $H(\Omega)$ is the entropy of Ω and ε_{R} is the Bayes error. On the other hand, the practical applicability of the mutual information is limited by its computational complexity, also known as the curse of dimensionality, which for multivariate data requires numerical integrations in high-dimensional spaces. Principe et al. (2000) explored the alternative definitions of entropy (Renyi (1961)), which, when coupled with Parzen window density estimation, led to a computationally feasible mutual information alternative that was applicable to multivariate data. Motivated by these findings, Torkkola developed an information-theoretic feature extraction algorithm (Torkkola (2003)), although his method is computationally demanding and seems to be limited by the curse of dimensionality. Next, we introduce a feature extraction objective function that is based on the mutual information, yet is easily computable.

11.4.2 Information-Theoretic Objective Function

Throughout the rest of the article we assume, that the class-conditional densities are Gaussian, that is, $\mathrm{R}\,|\,\omega_i \sim \mathcal{N}(\boldsymbol{m}_i, \boldsymbol{\Sigma}_i)$, with positive definite covariance matrices. The entropy of a Gaussian random variable is easily computed as

$$H(\mathsf{R}\,|\,\omega_i) = \frac{1}{2}\log((2\pi e)^n |\boldsymbol{\Sigma}_i|)$$

where $|\Sigma|$ denotes for the determinant of the matrix Σ. To complete the calculations required by (11.1), we need to evaluate the entropy of the mixture PDF $f_{\mathsf{R}}(\boldsymbol{r}) \triangleq \sum_i f_{\mathsf{R}|\Omega}(\boldsymbol{r}\,|\,\omega_i)P(\omega_i)$. It is easy to establish that $\mathrm{R} \sim (\mathrm{m}, \boldsymbol{\Sigma})$, where

$$\mathrm{m} = \sum_{i=1}^{c} \boldsymbol{m}_i P(\omega_i) \quad \text{and} \quad \boldsymbol{\Sigma} = \sum_{\mathrm{i=1}}^{c} \left[\boldsymbol{\Sigma}_i + (\boldsymbol{m}_i - \mathrm{m})(\boldsymbol{m}_i - \boldsymbol{m})^{\mathrm{T}}\right] P(\omega_i). \qquad (11.2)$$

Note that unless the class-conditional PDFs are completely overlapped, the RV R is non-Gaussian. However, we propose a metric similar to (11.1) by replacing $H(\mathsf{R})$ with the entropy of a Gaussian RV with the same covariance matrix $\boldsymbol{\Sigma}$:

$$\mu(\mathsf{R};\Omega) \triangleq H_g(\mathsf{R}) - \sum_{i=1}^{c} H(\mathsf{R}\,|\,\omega_i)P(\omega_i) = \frac{1}{2}\left[\log(|\boldsymbol{\Sigma}|) - \sum_{i=1}^{c}\log(|\boldsymbol{\Sigma}_i|)P(\omega_i)\right] \qquad (11.3)$$

where $H_g(\mathsf{R})$ is the Gaussian entropy. Throughout the rest of the article, we refer to this metric as a p-metric.

We will explain briefly why the p-metric is a valid class separability objective. For a thorough mathematical exposition, the reader is referred to Nenadic (in press). If the class-conditional PDFs are fully overlapped, that is, $\boldsymbol{m}_1 = . \cdot . = \boldsymbol{m}_c$ and $\Sigma_1 = \ldots = \Sigma_c$, it follows from (11.2) and (11.3) that $\mu(\mathsf{R};\mathrm{R}) = 0$. Also note that in this case $\mathrm{R} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$, thus $\mu(\mathsf{R};\mathrm{R}) = \mu I(\mathsf{R};\mathbf{O})$.On the other hand, if the class-conditional PDFs are different, R deviates from the Gaussian RV, so the p-metric $\mu(\mathsf{R};\Omega)$ can be viewed as a biased version of $\mu I(\mathsf{R};\Omega)$, where $\mu(\mathsf{R};\mathrm{R}) \geq \mu I(\mathsf{R};0) \geq 0$ because for a fixed covariance matrix, Gaussian distribution maximizes the entropy $[H_g(\mathsf{R}) \geq H(\mathsf{R})]$. As the classes are more separated, the deviation of R from a Gaussian RV increases, and the p-metric gets bigger. It turns out that this bias is precisely the negentropy defined as $\bar{H}(\mathsf{R}) \triangleq H_g(\mathsf{R}) - H(\mathsf{R})$, which has been used as an objective function for ICA applications (see Hyvärinen (1999) for survey). Therefore, ITDA can be viewed as a supervised version of ICA. Figure 11.1 confirms that ICA produces essentially the same result as our method (note the symmetry of the example), although the two methods are fundamentally different (unsupervised vs. supervised). Figure 11.1 also shows the p-metric in the original space and subspaces extracted by ITDA and LDA.

The p-metric has some interesting properties, many of which are reminiscent of the Bayes error ε_{R} and the mutual information (11.1). We give a brief overview of these properties next. For a detailed discussion, refer to Nenadic (in press). First, if the class-conditional covariances are equal, the p-metric takes the form of the generalized Rayleigh quotient; therefore, under these so-called homoscedastic conditions, ITDA reduces to the classical LDA method. Second, for a two-class case with overlapping class-conditional means and equal class probabilities (e.g., figure 11.1), the p-metric reduces to the well known Bhattacharyya distance. Like many other discriminant metrics, the μ-metric is independent of the choice of a coordinate system for data representation. Moreover, the search for the full-rank feature extraction matrix T can be restricted to the subspace of orthonormal projection matrices without compromising the objective function. Finally, the p-metric of any subspace of the original data space is bounded above by the p-metric of the original space. These properties guarantee that the following optimization problem is well posed. Given the response samples $\mathrm{R} \in \mathbb{R}^n$ and the dimension of the feature space m, we find an orthonormal matrix $\mathrm{T} \in \mathbb{R}^{m\times n}$ such that the p–metric $\mu(\mathsf{F};0)$ is maximized

$$T^* = \arg\max_{T\in\mathbb{R}^{m\times n}} \{\mu(\mathsf{F};\Omega) : \mathrm{F} = \mathrm{T}\mathrm{R}\} \quad \text{subject to} \quad \boldsymbol{T}\boldsymbol{T}^{\mathrm{T}} = \boldsymbol{I}. \tag{11.4}$$

Based on our discussion in section 11.4.2, it follows that such a transformation would find an m-dimensional subspace, where the class separability is maximal. Interestingly, both the gradient $\partial\mu(\mathsf{F};\Omega)/\partial\boldsymbol{T}$ and the Hessian $\partial^2\mu(\mathsf{F};\Omega)/\partial\boldsymbol{T}^2$ can be found analytically (Nenadic (in press)), so the problem (11.4) is amenable to Newton's optimization method.

11.5 Experimental Results

In this section, we compare the performances of LDA and ITDA on a dataset adopted from Rizzuto et al. (2005). The data represents intracranial encephalographic (iEEG) recordings from the human brain during a standard memory reach task (see figure 11.2). It should be noted that iEEG signals are essentially local field potentials (see section 11.3.2). At the start of each trial, a fixation stimulus is presented in the middle of a touchscreen and the participant initiates the trial by placing his right hand on the stimulus. After a short fixation period, a target is flashed on the screen, followed by a memory period. After the memory period, the fixation stimulus is extinguished, which signals the participant to reach to the

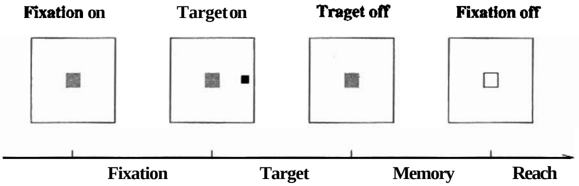

Figure 11.2 The timeline of experimental protocol.

memorized location (formerly indicated by the target). The duration of fixation, target, and memory periods varied uniformly between 1 and 1.3 s. The subject had 8 electrodes implanted into each of the following target brain areas: orbital frontal cortex (OF), amygdala (A), hippocampus (H), anterior cingulate cortex (AC), supplementary motor cortex (SM), and parietal cortex (P). The total number of electrodes in both hemispheres was 96. The targets were presented at 6 different locations: 0^o, 60^o, 120^o, 180^o, 240^o, 300^o; these locations respectively correspond to right, top right, top left, left, bottom left, and bottom right position with respect to the fixation stimulus. The number of trials per stimulus varied between 69 and 82, yielding a total of 438 trials. The electrode signals were amplified, sampled at 200 Hz and bandpass filtered. Only a few electrodes over a few brain areas showed stimulus-related tuning according to the location of the target. The goal of our analysis is to decode the target location and the behavioral state based on the brain data. Such a method could be used to decode a person's motor intentions in real time, supporting neuroprosthetic applications. All decoding results are based on a linear, quadratic, and support vector machine (SVM) classifier (Collobert and Bengio (2001)) with a Gaussian kernel.

11.5.1 Decoding the Target Position

To decode the target position, we focused on a subset of data involving only two target positions: left and right. While it is possible to decode all six target positions, the results are rather poor, partly because certain directions were consistently confused. The decoding was performed during the target, memory and reach periods (see figure 11.2). All decoding results are based on selected subsegments of data within 1 s of the stimulus that marks the beginning of the period. figure. 11.3 shows that only a couple of electrodes in both left and right parietal cortex exhibit directional tuning, mostly around 200 ms after the onset of the target stimulus. In addition, there is some tuning in the SM and OF regions. Similar plots (not shown) are used for the decoding during memory and reach periods.

For smoothing purposes and to further reduce the dimensionality of the problem, the electrode signals were binned using a 30 to 70 ms window. The performance (% error) of the classifier in the feature domain was evaluated through a leave-one-out cross-validation; the results are summarized in table 11.1. Note that the chance error is 50 percent for this particular task. For a given classifier, the performance of the better feature extraction method is shown in boldface, and the asterisk denotes the best performance per classification task. Except for a few cases (mostly with the quadratic classifier), the performance of the ITDA method is superior to that of LDA, regardless of the choice of classifier. More

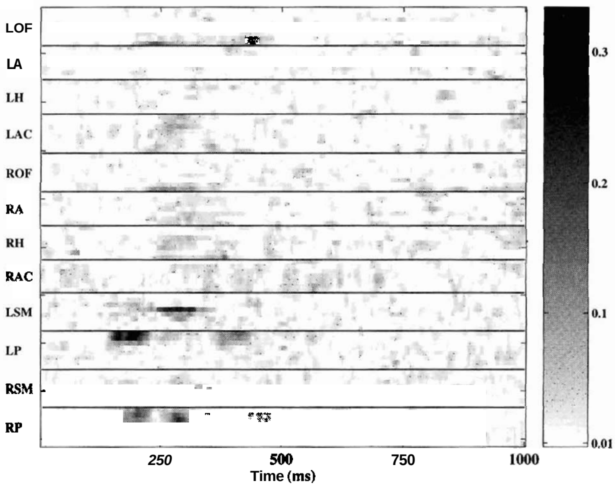

Figure 11.3 The distribution of the μ-metric over individual electrodes during the target period. The results are for two-class recognition task, and are based on 162 trials (82 left and 80 right). Different brain areas are: orbital frontal (OF), amygdala (A), hippocampus (H), anterior cingulate (AC), supplementary motor (SM), and parietal (P), with the prefixes L and R denoting the left and right hemisphere.

importantly, ITDA provides the lowest error rates in all but one case (target, SM), where the two methods are tied for the best performance. We note that all the error rates are significantly smaller ($p < 0.001$) than the chance error, including those during the memory period, which was not demonstrated previously (Rizzuto et al. (2005)). Also note that, in general, the SVM classifier is better combined with both ITDA and LDA features than are the linear and quadratic classifiers.

11.5.2 Decoding the Behavioral State

As discussed in section 11.3.2, for fully autonomous neuroprosthetic applications it is not only important to know ***where*** to reach, but also ***when*** to reach. Therefore, the goal is to decode what experimental state (fixation, target, memory, reach) the subject is experiencing, based on the brain data. To this end, we pooled the data for all six directions, with 438 trials per state, for a total of 1,752 trials. As with the target decoding, all the decoding results are based on selected subsegments of data within 1 s of the stimulus that marks the beginning of the period. Figure 11.4 shows that only a subset of electrodes exhibits state tuning (mostly the electrodes in the SM area during the second part of the trial state period). In addition, there is some tuning in the AC, H, and P areas. The data were further smoothed by applying a 40 to 50 ms window. The performance (% error) of the classifier in the feature space was evaluated through a stratified twenty-fold cross-validation (Kohavi (1995)), and the results are summarized in table 11.2.

Table 11.1 The average decoding errors and their standard deviations during the target, memory and reach periods. The columns represent the brain area, the number of electrodes N_e, the period (ms) used for decoding, the bin size (ms), the size of the data space (n), the type of the classifier (L-linear, Q-quadratic, S-SVM). The size of the optimal subspace (m) is given in the parentheses. Note that LDA is constrained to m $= 1$.

Period	Area	N_e	Time	Bin	n	Class.	LDA		(m)	ITDA		(m)
target	OF	4	160–510	70	20	L	6.17	± 0.24	(1)	**4.94***	± 0.22	(1)
						Q	**6.17**	± 0.24	(1)	8.02	± 0.27	(1)
						S	6.17	± 0.25	(1)	**4.94***	± 0.22	(1)
	P	2	150–450	50	12	L	7.41	± 0.26	(1)	**6.79***	± 0.25	(1)
						Q	8.02	± 0.27	(1)	**7.41**	± 0.26	(1)
						S	7.41	± 0.26	(1)	**6.79***	± 0.25	(2)
	SM	2	100–450	70	10	L	14.20	± 0.35	(1)	**13.58***	± 0.34	(3)
						Q	14.20	± 0.35	(1)	**13.58***	± 0.34	(2)
						S	**13.58***	± 0.34	(1)	**13.58***	± 0.34	(3)
	SM,P	2	120–520	40	20	L	5.56	± 0.23	(1)	**4.32***	± 0.20	(1)
						Q	**5.56**	± 0.23	(1)	**5.56**	± 0.23	(1)
						S	4.94	± 0.22	(1)	**4.32***	± 0.20	(1)
memory	OF	3	240–330	30	6	L	29.63	± 0.46	(1)	**28.40***	± 0.45	(1)
						Q	30.25	± 0.46	(1)	**28.40***	± 0.45	(2)
						S	31.48	± 0.47	(1)	**29.01**	± 0.46	(1)
	P	4	610–730	30	16	L	33.95	± 0.48	(1)	**32.72**	± 0.47	(1)
						Q	**33.33**	± 0.47	(1)	35.80	± 0.48	(1)
						S	31.48	± 0.47	(1)	**29.63***	± 0.46	(4)
	SM	2	250–370	30	8	L	29.63	± 0.45	(1)	**29.01**	± 0.46	(6)
						Q	29.63	± 0.46	(1)	**25.93**	± 0.44	(3)
						S	29.63	± 0.46	(1)	**24.69***	± 0.43	(4)
	SM, P,A	3	620–680	30	6	L	28.40	± 0.45	(1)	**26.54***	± 0.44	(1)
						Q	**27.16**	± 0.45	(1)	28.40	± 0.45	(1)
						S	27.16	± 0.45	(1)	**26.54***	± 0.44	(1)
reach	OF	2	270–420	50	6	L	10.49	± 0.31	(1)	**9.26**	± 0.29	(1)
						Q	10.49	± 0.31	(1)	**9.88**	± 0.30	(1)
						S	9.88	± 0.30	(1)	**8.64***	± 0.28	(1)
	OF	4	250–550	50	24	L	6.79	± 0.25	(1)	**6.17**	± 0.24	(1)
						Q	**6.79**	± 0.25	(1)	**6.79**	± 0.25	(1)
						S	6.17	± 0.24	(1)	**4.94***	± 0.22	(22)

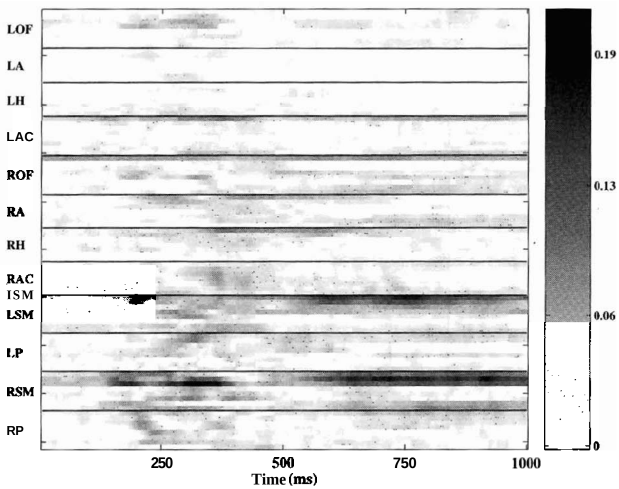

Figure 11.4 The distribution of the μ-metric over individual electrodes. The results are for four-class recognition task based on 1,752 trials (438 trials per state).

Table 11.2 The average behavioral state decoding errors and their standard deviations with pooled data (6 directions, 4 trial states). Note that LDA is constrained to $m \leq 3$.

Area	N_e	Time	Bin	n	Class.	LDA		(m)	ITDA		(m)
SM	4	500–1000	50	40	L	24.70	±0.04	(3)	**24.17**	±0.04	(4)
					Q	24.82	±0.04	(3)	**24.58**	±0.04	(5)
					S	24.76	±0.04	(3)	**23.99'**	±0.04	(4)
SM	3	120–400	40	21	L	35.36	±0.06	(3)	**35.06**	±0.05	(9)
					Q	36.25	±0.05	(3)	**31.31***	±0.05	(12)
					S	35.42	±0.06	(3)	**31.43**	±0.06	(14)
SM, AC,H	4	250–500	50	20	L	29.23	±0.06	(3)	**28.75**	±0.06	(3)
					Q	28.99	±0.06	(3)	**27.74***	±0.06	(5)
					S	28.93	±0.06	(3)	**27.74***	±0.06	(5)
P	4	200-350	50	12	L	48.69	±0.06	(3)	**47.86**	±0.05	(10)
					Q	**48.99**	±0.07	(3)	50.89	±0.05	(10)
					S	49.70	±0.05	(3)	**47.68***	±0.04	(10)

Note that the chance error is 75 percent for this particular task. Except for one case, the classification accuracy with ITDA features is superior to LDA features, regardless of the classifier choice. Additionally, the best single performance always is achieved with the ITDA method. Note that the best decoding results are obtained from the SM area in the interval [500–1000] ms. Interestingly, we were able to decode the trial states from the parietal area, although the accuracy was considerably lower (just above 50 percent).

11.5.3 Discussion

Based on the analyzed data, we conclude that the classification with ITDA features is more accurate than the classification with LDA features, with an improvement as high as 5 percent. In rare cases where LDA provides better performance, the quadratic classifier was used. This could mean that LDA features fit the quadratic classifier assumptions (Gaussian classes, different covariance matrices) better than do ITDA features. Nevertheless, ITDA features are in general better coupled to the quadratic classifier than are LDA features. The advantages are even more apparent when ITDA is used in conjunction with the linear and SVM classifier. Similar behavior was observed when ITDA was tested on a variety of data sets from the UCI machine learning repository (Hettich et al. (1998)). Details can be found in Nenadic (in press).

In all cases, the best performance is achieved in a subspace of considerably lower dimension than the dimension of the original data space, n. Therefore, not only is the classification easier to implement in the feature space, but the overall classification accuracy is improved. While theoretical analysis shows that dimensionality reduction cannot improve classification accuracy (Duda et al. (2001)), the exact opposite effect is often seen in dealing with finitely sampled data.

Like many other second-order techniques, for example, LDA or ACC, ITDA assumes that the class-conditional data distribution is Gaussian. Although this assumption is likely to be violated in practice, it seems that the ITDA method performs reasonably well. For example, the performance in the original space with the SVM classifier is Gaussian-assumption free, yet it is inferior to the SVM classifier performance in the ITDA feature space. Likewise, it was found in Nenadic (in press) that unless data is coarsely discretized and the Gaussian assumption is severely violated, the performance of ITDA does not critically depend on the Gaussian assumption.

11.6 Summary

We have reviewed recent advances in cognitive-based neural prosthesis. The major differences between the cognitive-based and the more common motor-based approach to BMIs have been discussed. To maximize information encoded by neurons, better understanding of multiple brain areas and the types of signals the brain uses are needed. Part of our research effort is to identify sources of information potentially useful for neuroprosthetic applications. Other research efforts are focused on technological issues such as the stability of recording, the development of unsupervised signal analysis tools, or the design of complex decoding algorithms.

The decoding of neural signals in cognitive-based BMIs reduces to the problem of classification. High-dimensional neural data typically contains relatively low-dimensional useful signals (features) embedded in noise. To meet computational constraints associated with BMIs, it may be beneficial to implement the classifier in the feature domain. We have applied a novel information-theoretic method to uncover useful low-dimensional features in neural data. We have demonstrated that this problem can be posed within an optimization framework, thereby avoiding unjustified assumptions and heuristic feature selection strategies. Experimental results using iEEG signals from the human brain show that our method may be better suited for certain applications than are the traditional feature extraction tools. The study also demonstrates that iEEG signals may be a valuable alternative to spike trains commonly used in neuroprosthetic research.

Acknowledgments

This work is partially supported by the National Science Foundation (NSF) under grant 9402726 and by the Defense Advanced Research Projects Agency (DARPA) under grant MDA972-00-1-0029. Z. Nenadic also acknowledges the support from the University of California Irvine (UCI) set-up funds. The authors would like to thank the anonymous reviewers for their constructive criticism. The authors would also like to acknowledge the editors for their timely processing of this manuscript.

Notes

E-mail for correspondence: znenadic@uci.edu

(1) Consistent with engineering literature (Fukunaga (1990)), we consider the feature extraction as a preprocessing step for classification. Some authors, especially those using artificial neural networks, consider feature extraction an integral part of classification.

(2) Recently, a couple of nonlinear feature extraction methods have been proposed (Roweis and Saul (2000); Tenenbaum et al. (2000)) where features reside on a low-dimensional manifold embedded in the original data space. However, linear feature extraction methods continue to play an important role in many applications, primarily due to their computational effectiveness.

(3) Optimality is in the sense of Bayes.