CORTICAL NEURAL PROSTHETICS

Andrew B. Schwartz
Departments of Neurobiology and Bioengineering, University of Pittsburgh,
Pittsburgh, Pennsylvania 15203; email: abs21@pitt.edu

Key Words  neuronal populations, chronic microelectrodes, prostheses, arm control, extraction algorithms

Abstract  Control of prostheses using cortical signals is based on three elements: chronic microelectrode arrays, extraction algorithms, and prosthetic effectors. Arrays of microelectrodes are permanently implanted in cerebral cortex. These arrays must record populations of single- and multiunit activity indefinitely. Information containing position and velocity correlates of animate movement needs to be extracted continuously in real time from the recorded activity. Prosthetic arms, the current effectors used in this work, need to have the agility and configuration of natural arms. Demonstrations using closed-loop control show that subjects change their neural activity to improve performance with these devices. Adaptive-learning algorithms that capitalize on these improvements show that this technology has the capability of restoring much of the arm movement lost with immobilizing deficits.

INTRODUCTION

Microelectrodes embedded chronically in the cerebral cortex hold promise for using neural activity to control devices with enough speed and agility to replace natural, animate movements in paralyzed individuals. Known as cortical neural prostheses (CNPs), devices based on this technology are a subset of neural prosthetics, a larger category that includes stimulating, as well as recording, electrodes. For many years, patients have been implanted with stimulation-based devices designed to activate neurons in different parts of the CNS. This class of neural prostheses is now used extensively in applications to restore hearing and alleviate the symptoms of Parkinson’s disease. However, the mirror technology used to record signals from neurons has been applied rarely to human patients and to this point is found only in research settings. These devices, called brain-computer interfaces (BCIs), link the brain to the external world by computer processing the recorded neural signal to extract the subject’s command to control an external device. For those who are movement impaired, recording-based neural prostheses may enable communication or movement. Likely beneficiaries of this evolving technology include people paralyzed by head or spinal-cord trauma or those with deficits caused by stroke, amyotrophic lateral sclerosis (ALS), cerebral palsy, and multiple sclerosis.
paralyses may range from complete—with no respiratory or eye movements—to quadri- and paraplegia. Whereas some BCI devices are designed for communication only (i.e., electroencephalography (EEG)-based, word-spelling programs), CNPs, by using single-cell activity, aim to restore movement as well. The scope of this review is limited to CNPs, specifically those designed to record signals in a form that can be used by the subject to control arm movement.

BACKGROUND

The basic theorem in this field states that within the discharge pattern of cortical neurons, there exists a rather direct representation of the desired movement. This movement image has been documented only in the last 20 years. Before this, investigators assumed that the primary motor cortex, the cortical region most closely related to movement, drove muscle activation, directly. Even though this region seemed to be anatomically and topographically organized into a body map on the cortical surface, recorded signals from a purely muscle-based coordinate system would be computationally difficult to transform into natural movements of the limb. Certainly the transformation of muscle activation to muscle force alone is a difficult nonlinear problem. Given a force and a muscle length, the way in which a limb is displaced following contraction of a specific muscle is dependent on limb geometry, the orientation of the limb relative to external forces (loads and gravity), and the history (inertia) of the moving segments at the time of contraction. Factor in the complexity of a redundant muscle system with many effectors operating simultaneously, and the problem of calculating the hand’s trajectory from a sample of muscle activations becomes a very difficult engineering problem. Yet these are precisely the problems confronting designers of functional electrical stimulation (FES) systems. Electrical stimulation of paralyzed muscles must produce coordinated shortening of the muscles around the limb’s joints that move the attached segments to achieve the proper end point displacement. Current CNPs provide a control signal in end point coordinates (see below). This signal, combined with the problems just enumerated and the general difficulty inherent in long-term electrical activation of limb muscles, is why initial implementations have been based on the control of artificial rather than real arms.

End point (the end of the arm or the hand) movement is represented simply in the activity of motor cortical cells. Georgopoulos and his colleagues showed that motor cortical discharge rate was directionally tuned as monkeys made reaching movements (Georgopoulos et al. 1982, 1986). Movement direction was measured at the hand when all seven degrees of freedom in the shoulder, arm, and wrist were free to move. The tuning function (cosine-shaped) relating discharge rate to direction is broad, covering all movement directions, and shows that each cell changes its discharge rate for all directions, or conversely, that all cells actively code each direction (simultaneous activity). By itself, the tuning function of a single cell is not very useful for decoding direction because a single direction will correspond...
to more than one discharge rate, and the broadness of the function means that small fluctuations of discharge rate will correspond to large changes in direction. However, specific directions were well predicted if weighted responses from many cells were added together vectorially (Georgopoulos et al. 1984) using a linear method termed the population vector algorithm (PVA). This direction was instantaneously and continuously represented in the cortical activity throughout movement (Schwartz 1992). The magnitude as well as the direction of this neural vector representation was highly correlated with movement velocity (Georgopoulos et al. 1988, Moran & Schwartz 1999). With these properties, the movement trajectory (time pattern of hand positions) could be extracted from the population activity for reaching movements and for a variety of drawing tasks (Schwartz 1993, Schwartz 1994, Schwartz & Moran 1999).

Why is it so desirable to extract a trajectory signal from the brain? The trajectory of the end point contains natural characteristics of animate motion. Examples of these invariant features are the bell-shaped velocity profile (Morasso 1981) of reaching movement and the two-thirds power law pertaining to drawing and handwriting (Viviani & Terzuolo 1982). Although prosthetic devices can be effective without operating like natural limbs, the embodiment of these characteristics is desirable in terms of biomechanical compatibility with other body parts, ease of control, and aesthetics.

APPLICATION

These experimental results show that accurate predictions of arm movement can be generated by recoding activity from a population of cortical neurons. For this to be a real-time control signal, parallel recordings must be made from multiple electrodes. In addition, this signal should be dependable—units (preferably the same ones) should be able to be discriminated for years from each electrode. Recording single-unit activity this way requires chronic intracortical implantation of microelectrodes. Alternatively, noninvasive EEG scalp electrodes can record electric fields useful for prosthetic control. However, aside from a brief comparison to EEG prostheses, this review concentrates on intracortical, single-unit activity used for CNPs.

Most of the work in this field presumes that movement-related information in the recorded unit activity is encoded as firing rate. For single units, this rate is the inverse of the time interval between action potentials of a single neuron, and for multunit clusters the analogous interval is measured between amplitude crossings of the group’s summed electrical signal. Alternatively, power within a frequency band of multunit activity may be taken as an activity measure. Although information may be contained in the synchronous activity of neurons (Butler et al. 1992, Hatsopoulos et al. 1998, Mazurek & Shadlen 2002), so far the amount of information in such a code seems relatively small. Extraction algorithms convert firing rate to movement displacement. Typically, firing rates from many different units
recorded simultaneously are used as input to the algorithm, and hand position is the output. Real-time decoding is needed for prosthetic control, so instantaneous firing rate calculated in a bin (e.g., 10–100 ms wide) is fed continuously to the algorithm, producing a continuous stream of hand positions. CNPs operate by recording multiple channels of single-unit activity simultaneously, conditioning these signals, usually by discriminating spike activity from the recorded signal, processing the spike trains with an extraction algorithm to generate a movement trajectory, and finally, feeding the extracted movement trajectory to a computer graphics display or a robot arm controller. Note, the control scheme does not end with movement generation. Rather, this is an example of closed-loop control where subjects observe the generated movement and modify their underlying neural activity to change the movement in a continuous fashion.

Every CNP is composed of three building blocks. The first step is to record the type of neural activity from which a consistent control signal can be extracted. Extraction is based on the concept of a neural code—recorded signals need to be deciphered and related to a desired movement. Finally, the control made possible by the extracted signal is implemented either in a computer display, an active prosthetic arm or other mechanical device, or by using electrical activation of the subject’s muscles. In summary, the three CNP components are

1. microelectrodes and recording electronics. Chronic electrodes provide many individual recording sites implanted permanently in the cerebral cortex. The recording electronics condition and discriminate the recorded signal. An excellent review on this technology is available (Schmidt 1999).
2. extraction algorithms. These are computer programs running in real time that take the conditioned data (e.g., action potential events or spike times) and convert them to end point positions.
3. actuators. These can be animated computer displays, movement of a robot arm, or activation of muscles in a subject’s own arm.

This review addresses the first two topics. The control of virtual-reality, computer displays and tele-robotic actuators are beyond the scope of this review.

ELECTRODES

Microwires

The first chronic recording electrodes were microwires. Developed over the last 40 years, these electrodes consist of fine wires 20 to 50 microns in diameter. They are generally composed of stainless steel or tungsten and insulated with teflon or polyimide. The tips may be etched or ground, but more often they are simply cut with a pair of scissors, leaving a planar recording surface. The wires can be arranged as arrays (for instance, in two rows of eight wires) by soldering them to a small connector (Williams et al. 1999, Nicolelis et al. 1999). Spacing between
wires (100–300 microns) is maintained either with polyethylene glycol or methyl methacrylate. Wire arrays are surgically implanted in the anesthetized animal during a procedure that takes 8 to 10 h (rhesus monkey, four arrays, 64 electrodes). The cortex is exposed through a hole in the skull. After investigators remove the dura, the array is advanced slowly with a micromanipulator (100 microns/min) to minimize dimpling of the cortical surface. Although some anesthetics (i.e., ketamine) permit enough spontaneous or sensory-invoked activity in motor cortex to serve as an indicator of how far to advance the arrays, with gas inhalants such as fluorothane, spontaneous neural activity is not detectable with microwires in the motor cortex, making the optimal insertion depth uncertain. This uncertainty may leave the tips in a layer of cortex where it is difficult to record unitary activity and may be a key reason for subsequent recording failures. For exposed macaque cortex, the best depths are ≤2 mm below the surface. The shafts of the arrays are glued to the bone, so the depth of the electrode tip relative to the skull is permanently fixed. However, the cortical surface may move after the surgery, perhaps rebounding, if it had been dimpled, or shrinking, if swelling had taken place. The causes and dynamics of this phenomenon are not understood, but shrinking or expanding the brain, however slight, is a key issue because it will change the relative position of the electrode tip, perhaps moving it to different cortical layers or into the underlying white matter.

**Silicon Micromachined Microprobes**

A number of silicon substrate microprobes have been developed (Hetke & Anderson 2002, Jones et al. 1992). Two types are reviewed here. The first are planar devices from the University of Michigan, and the second is the array developed at the University of Utah.

The Michigan probe is somewhat unique, in that boron diffusion is used as an initial processing step of the silicon wafer to delineate the shape of the probe. A number of steps are used to deposit silicon dioxide and silicon nitride for insulation, and this is followed by photolithography to pattern the interconnects and recording sites. Iridium is deposited on the exposed recording sites as the electrode surface. This fabrication allows for a wide variety of probe shapes and configurations. A standard probe for chronic implants consists of four parallel, dagger-like shanks connected to a microsilicon ribbon cable. The shanks are 15 microns thick and 50 to 100 microns wide, with shank tips spaced 150 microns apart. The probe designed for monkey recording is 3.8 mm long and has 4 recording shafts placed along the shaft. The ribbon cable is flexible and has a connector at the end. Probes are implanted with a pair of forceps through the open dura. The connector is glued to the skull, but in contrast to the microwires, the semiflexible ribbon cable allows the probes to “float” in the brain (e.g., move up and down with the cortex as it pulses). Because the multiple recording sites are placed along the shaft, at least some of the sites will be situated at cortical depths desirable for good extracellular recordings.
Investigators used a completely different approach for the probe designed at the University of Utah, which is now commercially available through Cyberkinetics, Inc. The fabrication of this device begins with a solid block of silicon. Checkerboard slices with a microsaw are cut most of the way through the block. Etching of the block then results in a three-dimensional, $10 \times 10$ array of needles on a $4 \times 4$-mm square. Additional processing applies metal and insulation layers. The final array has a recording site at the tip of each shank with its interconnect running down the shank and through the back of the block, where gold pads are located for wire bonding to leads projecting to a skull connector. The 35- to 75-micron-long recording tips are platinum, with an impedance of 100 to 500 kOhms. The shank lengths range from 1.0 to 1.5 mm. Implantation of the array is achieved by injecting the array through the reflected dura with a special high-speed device that overcomes the inertia of the cortex. Leads are flexible enough to allow the array to “float” on the cortical surface. This design has the advantage of placing a relatively large number of recording sites in a compact volume of cortex. Furthermore, conventional wisdom suggests that a recording site at the tip is ideal in terms of sampling the potential field of an action potential and is the place least likely to experience tissue damage from electrode insertion (see below). However, with a single recording site at a fixed cortical depth, the Utah array suffers from the same placement problem as microwires. The length of the shanks is limited to 1.5 mm because of the way the device is fabricated from a single block of silicon.

Tissue Reactions

One common problem of all chronically implanted electrodes is that of the tissue-electrode interface. Any object inserted into the brain damages the parenchyma. During insertion, blood vessels are disrupted and microhemorrhage is common. Neurons are either ripped or sliced as the electrode is inserted. Microglia derived from monocytes are activated, and astrocytes begin to proliferate, which forms a loose encapsulation around the electrode for a considerable distance (100–200 microns). A poorly understood cascade of signaling events stemming from disruption of vessels and the blood-brain barrier takes place, which leads to an infiltration of nonlocal cellular elements, immune components, and epithelial cell proliferation. Local changes in the extracellular concentration of potassium and calcium may silence the activity of nearby neurons through a local mechanism similar to spreading depression (Somjen 2001). Most histological studies report that neuron density is near normal at distances within 100 microns of the electrode after several weeks. However, mechanical considerations would suggest that the kill zone around the electrode could be larger if the electrode is not inserted exactly at 90° perpendicular to the cortical surface (Edell et al. 1992). Initially, few neurons could be found within 50 to 100 microns of the implant site, but the cells looked normal outside this zone. After four to six weeks, this sparse zone decreased, with healthy-appearing neurons apparent closer to the electrode. Early on, astrocytes, identified
by glial fibrillary acetic protein (GFAP) could be seen in the margins of the sheath, with processes infiltrating and surrounding the implant site. As neurons filled around the implant, the perielectrode sheath became more compact with heavy GFAP staining, but cell bodies and processes could not be readily distinguished in the highly compressed tissue.

Astrocytes, the most prevalent type of glial cell, play a role in supporting brain tissue and nourishing neurons (for instance, they supply neurons with lactate derived from glucose). They also play a role in scar formation. Another type of glia, the microglia, are CNS analogs of macrophages. These are mobile cells that engulf fragments of damaged cells. Microglia are the other major component of the electrode encapsulation (Szarowski et al. 2003). Another component of reactive astrocyte response, identified by the marker vimentin, was present in the encapsulation (Szarowski et al. 2003). Vimentin is thought to be specific to immature astrocytes, and in this study, cells staining positive for this marker had long thin processes extending more than 100 microns into the sheath. These astrocytes were found much less frequently than those positive for GFAP, and they formed a thinner layer than their counterparts. Organization of the vimentin layer followed a slightly different time course than the GFAP astrocytes and microglia, both of which had similar morphology within the sheath over time.

This multicomponent process following electrode insertion is still not well understood. The shape and size of the electrode, and the way it is inserted, are probably critical factors in the type of damage imparted (Edell et al. 1992). A cylindrical shape may be ideal for pushing blood vessels and cells away without damage. However, because of the filamentous nature of neural tissue, forces from this displacement may propagate through the tissue, resulting in tearing, stretching, and compression. Alternatively, a conical electrode shape with a sharp tip will cut a hole through the tissue, leaving adjacent tissue intact as the electrode slips by. But a pressure band will eventually build in front of the electrode as some of the tissue is compressed, and this pressure will be transmitted to nearby tissue. A flat, sword-like tip may be preferable because its slicing action would cause minimal tissue compression. Recently, investigators compared different electrode shapes and insertion techniques (Szarowski et al. 2003). Although the size and shape of the electrode made a difference in the reaction around the electrode in the first one to two weeks after implantation, histology taken at longer intervals showed little difference except for the volume of tissue affected. Insertion technique appeared to make little difference.

The second component of the implant reaction is a chronic process, taking place on a slower timescale after implantation, and results in the formation of a tight cellular sheath around the electrode (Turner et al. 1999). Whereas the sheath was completely formed around the insertion site, composed of loosely packed cells and 2 layers thick in the first 2 weeks postinsertion, by 6 to 12 weeks, the sheath had become thinner and tighter with 4 to 6 layers of compact, dense cells with small nuclei, effectively isolating the probe from the brain tissue (Szarowski et al. 2003). A study using impedance spectroscopy (Williams 2001) showed that electrode
impedance was well correlated with sheath density. This increase in tip impedance following implantation is one of the reasons that microwire recordings have been successful. The diameter of the wire’s exposed surface is 50 microns, which is generally too large for good isolation of single units. As the sheath forms, the amount of exposed surface is reduced, which raises the electrode impedance. These higher impedance recordings are more in range with conventional extracellular recordings and make it possible to isolate single action-potential waveforms. This principle may apply to the Utah arrays, which have relatively large recording sites at the shank tips. In contrast, this type of encapsulation is an apparent problem for the Michigan probes, which, in the past, tended to record good action potentials for the first one to three weeks after implantation, after which time the signal degraded.

Some of these issues are being addressed with new technology. Ideally, one would like to regulate the extraelectrode environment so that sprouting neurites are attracted to the vicinity of the recording surface before encapsulation takes place (Kennedy et al. 2000). Presently, investigators are studying modifications to the electrode surface using hydrogels, silk-like polymers, and nanotubes. These structures can be bound to bioactive compounds, such as neurotrophins, that attract growing neurites (Cui et al. 2001). Conversely, the inflammatory response can be reduced. Lipid microtubes (Meilander et al. 2001, Zhong et al. 2001) can deliver molecules to block transcription factors and downregulate genes controlling release of proinflammatory cytokines (Manna & Aggarwal 1998). Systemic and local dexamethasone administration was effective at reducing the density of the peri-electrode sheath, whereas cyclosporin A was found ineffective (Shain et al. 2003). Investigators made no attempt to correlate this with improved recording conditions. Dexamethasone inhibits astrocyte hyperplasia and can also be incorporated into coatings applied to the electrode. The diffusion of these molecules into the tissue can be regulated so that a therapeutic dose is maintained over time. This technology may achieve the desired electrode-neuron-glial sheath sandwich that could lead to permanent long-term recording of neural activity.

Experience with these probes suggests that encapsulation is a major factor in the deterioration of recording conditions. In the past, this deterioration may have been due to the configuration of the probe, the planar recording sites on the side of the shaft, and/or the size of the recording site (Hetke & Anderson 2002, Schmidt et al. 1997). That the recording deterioration was due to encapsulation is supported by the observation that the recording sites could be reactivated by passing current through the electrode, which reduces the impedance by removing tissue or opening tunnels through the encapsulation (Schmidt et al. 1993a,b). Better results have been reported, recently. Investigators have maintained good recordings in the guinea pig for more than one year (Kipke et al. 2003). In our laboratory we have been recording good units for more than ten months from implanted probes in monkey cortex. One consistent finding in comparing the success of chronic recordings in rhesus monkeys, as compared to either guinea pigs or rats, is that it is easier to obtain consistent high-yield recordings in the rodent. This finding may be related
to differences in the reaction to implantation or to the differences in cortical folding between the rodent (lissencephalic) and primate (gyrencephalic).

EXTRACTION ALGORITHMS

To generate a prosthetic movement signal, information contained in the parallel recordings is transformed from the domain of spikes/sec to extrinsic (i.e., Cartesian) coordinates. Approaches for extracting information divide into two broad categories: inferential methods and classifiers. Inferential methods are model based—they depend on some understanding of underlying mechanisms. To illustrate this, we use a somewhat unrelated example from the vestibular system. The response generated by receptors in the semicircular canals to angular acceleration was predicted by modeling the endolymph, cupula, and canal as a torsional pendulum. On the basis of this mechanical model, the canal input, angular acceleration, is integrated to become angular velocity at the output. This model was confirmed by recordings of axonal activity in the eighth nerve, which were correlated with angular velocity. In contrast, classifier methods need not consider any mechanism. Rather they rely on a consistent representation of the parameter to be extracted. Imagine a set of three neurons recorded as a monkey moves to targets in eight directions. Considering each neuron as a binary element (on or off), these three neurons could encode each direction unambiguously. Each combination of the neuronal pattern would be assigned to one target direction. Then, observation of the neurons as a pattern obtained by simultaneously recording could be used to predict movement direction. In this ideal example, no assumptions of mechanism, or even of continuity between patterns, are necessary. However, if the task was changed, for instance by adding or removing targets, there is no guarantee that the eight-target code would still apply. In reality, neurons are not bistable but change their firing rates continuously. Their firing rate–movement tuning functions are unimodal, broad, and noisy, with no clear transition between on and off states. Furthermore, in practice, to find a perfect set of neurons that encode a single parameter in an orthogonal manner (i.e., 3 neurons for direction in 3D space) is difficult. These realistic conditions have led to more sophisticated classifiers than the ideal example given here.

Inferential Methods

A population of neural activity measured as spike occurrences can be represented as a vector. In this example, each dimension of the vector corresponds to a particular neuron, and the value or magnitude of each dimension is proportional to that cell’s firing rate. This vector then would have a direction and magnitude in a neural space. The purpose of the extraction algorithm is to transform this vector into a corresponding vector in movement space. The direction of each dimension of the neural vector has no physical meaning other than acting as a neuron label. Models
(for instance, based on directional tuning functions) are used to transform this neural vector into a movement vector.

**POPULATION VECTOR ALGORITHM** As introduced earlier, the relation between movement direction and firing rate in the motor cortex can be described with the cosine function

\[ D - b_o = A \cdot \cos \theta \]  

This function is equivalent to the expression

\[ D - b_o = b_x m_x + b_y m_y + b_z m_z \]  

or to the dot product of the two vectors \( B \) and \( M \), where \( D \) is the discharge rate of the studied unit, \( A \) is the amplitude of the tuning function, \( \theta \) is the angle between the cell’s preferred direction, the movement direction \( B \) is a vector in the unit’s preferred direction with a magnitude equal to the amplitude of the tuning function, and \( M \) is a unit vector in the movement direction. The cell’s maximum firing rate is the tuning function amplitude plus \( b_o \). The coefficients of the \( B \) vector (the unit’s preferred direction) are typically found by regressing movement direction to discharge rate. Most commonly, this relation comes from data gathered during a task where the subject reaches from a center-start position to a set of targets arranged circumferentially around the start position so that the reaches occur uniformly in different directions.

The population vector is formed by combining weighted contributions from each unit along its preferred direction. For a particular movement, the \( B \) vector of each unit is normalized to the firing rate of the cell during that movement. This contributory vector, \( C_i \), of the \( i \)th unit, is added with those of the other \( N \) units in the population to form the population vector.

The population vector can be generated in small bins (i.e., 20 ms) and correlates very well with the hand’s velocity throughout an arm movement. These vectors can be added together, tip to tail, to form a neural trajectory that predicts and matches the arm’s trajectory (Georgopoulos et al. 1988). In the past, these experiments were executed by recording units one at a time, summing their responses together to form the population vector. For prosthetic control, the vector output must be generated in real time by recording from electrode arrays using electronics capable of processing many units simultaneously.

The population vector algorithm can be an efficient decoder if two ideal conditions are met: The preferred directions of the recorded units should be distributed uniformly in space with radially symmetric tuning functions (Georgopoulos et al. 1988). Considering the small number of units recorded simultaneously with chronic electrodes, these conditions usually are not met. Often units are not well tuned, and those that do not form a symmetric distribution of preferred directions. These problems are accentuated in prosthetic control, where trials cannot be averaged and movement commands are calculated in small bins. These issues were addressed in the modified population vector algorithm used by
Taylor et al. (2002). Instead of weighting each unit’s contribution to the population vector only by its discharge rate, a number of additional factors were used. As subjects used their neural activity to direct a cursor to center-out targets, \( X \), \( Y \), and \( Z \) weighting coefficients were calculated iteratively. This procedure was considered coadaptive because the weighting coefficients were adjusted as the subject was learning to make brain-controlled movements. By accounting for changes in the tuning characteristics of the units as the subject learned, this algorithm was very effective with relatively few units. Unlike the original population vector algorithm, this algorithm reduced the contribution of poorly tuned units by scaling the coefficients according to the unit’s cosine fit. One way that units commonly deviated from the cosine function was that firing rates in the preferred direction were not always 180° from the antipreferred direction. Using separate coefficients when rates were above or below the mean compensated for this noncosine behavior. Nonuniformity of the directional distribution was addressed directly by normalizing the \( X \), \( Y \), and \( Z \) contributions and was addressed indirectly by emphasizing coefficients that produced good movements in all parts of the workspace.

Other extraction techniques are based on more formalized optimization algorithms. The optimized linear estimator (Salinas & Abbott 1994) is similar to the population vector algorithm in that individual cell responses are added vectorially to give a single population vector. The difference is that each unit’s preferred direction is calculated using a multiple regression across all the units simultaneously to give optimal fits to the sampled movement directions. A correlation matrix between the firing rates of all cells is inverted to give the global least-squares fit of the population. Compared to the population vector, this method has the advantage of correcting for nonuniform direction distributions. Other investigators used a similar linear approach with time-shifted firing rates regressed to hand position (Wessberg et al. 2000).

**KALMAN FILTER**  
The methods described to this point are considered static models because they consider the velocity at each time step to be independent. Dynamic state models have been developed to account for correlations between nearby velocities (or movement increments).

These methods have great potential because they use features of the movement or behavior that are predictable in their own right, in addition to predictions based on neural activity. Because many features of movement are regular (speed changes smoothly, direction tends to be constant at high speed, etc.), these state conditions can lend great power to the overall prediction. The prior hand velocities combined with a history of firing rates are incorporated into the Kalman filter (Paninski et al. 2004) to predict future hand velocity. The state model for velocity is often very simple, i.e., random walk. Used this way, velocities change smoothly—the velocity of the hand does not change much from one instant to the next and is characteristic of trajectories during reaching movements. Formally, \( x_k \) is the velocity at time step \( k \), and \( x_{k+1} \) is the velocity at the next time step. The discharge rate at \( k \) is \( z_k \). The state and observation equations are
Velocity can be expressed by a three-dimensional vector \((X, Y, Z)\); \(z_k\) is a vector of the current firing rates (at time step \(k\)) for the \(N\) simultaneously recorded units. \(A\) is a matrix, for example, and is an identity matrix if the state model is a random walk. \(H\) is a matrix of coefficients to convert the velocity into discharge rates and consists of an \(X\), \(Y\), and \(Z\) coefficient for each recorded unit (i.e., preferred direction). The error in velocity between the estimated (Equation 3) and actual value is \(w_k\) with a distribution specified with a covariance matrix \(\Theta\). For the discharge rates, the error is \(v_k\) with a covariance of \(\Lambda\).

Of course, the state model can be more complex, for instance by taking into account the bell-shaped velocity profile and the tendency for straight arm movements during reaching or by including terms for position and acceleration in addition to velocity.

The Kalman filter works iteratively in steps. It begins by guessing the velocity from an initial distribution specified by a mean \(\hat{x}_0\) and covariance \(\Sigma_0\). This choice is then used to get \(\hat{x}_{k-1}\), the state estimate (Equation 5).

\[
\hat{x}_{k-1} = A\hat{x}_{k-1} + w_k
\]

and

\[
z_k = Hx_k + v_k.
\]

\(\hat{x}_{k-1}\) is the estimate of velocity in the previous time step \(k - 1\) and has a covariance of \(\Sigma_{k-1}\).

The current velocity estimate at step \(k\) uses \(\hat{x}_{k-1}\) and the velocity model, \(A\):

\[
\hat{x}_k = A\hat{x}_{k-1}.
\]

The covariance of estimate errors in the state estimate of velocity is

\[
P^-_k = A\Sigma_{k-1}A^T + \Theta.
\]

The prediction of velocity using \(\hat{x}_k^\sim\) and present discharge rates is

\[
\hat{x}_k = \hat{x}_k^\sim + K_k(z_k - H\hat{x}_k^\sim),
\]

where

\[
K_k = P^-_kH^T(HP^-_kH^T + \Lambda)^{-1}.
\]
The state estimate is then updated with the spike activity to give the new estimate of velocity, $\hat{x}_k$, using Equation 7. This measurement update uses the difference between the actual, $z_k$, and estimated, $H\hat{x}_{k-1}$, discharge rates multiplied by the gain factor, $K$, defined in Equation 8. The $^{-1}$ in this equation denotes pseudoinverse, and $^T$ signifies transpose. The current velocity estimate, $\hat{x}_k$, is then used as the time update and becomes $\hat{x}_{k-1}$ in the next iteration.

Investigators have elaborated this general procedure using statistical methods. The estimates of the Kalman filter can be considered the peak value of a probability density function. For example, the density functions for each step of the recursion steps can be defined in the following way: $\hat{x}_{k-1}$ is the mean of $p(x_{k-1}|z_k-1, z_{k-2}, \ldots z_1)$, $\hat{x}_k$ is the mean of $p(x_k|z_{k-1}, z_{k-2}, \ldots z_1)$, and $\hat{x}_k$ is the mean of $p(x_k|z_k, z_{k-1}, \ldots z_1)$.

The generation of these density functions for nonlinear, non-Gaussian dynamic state models is a current area of interest in statistical research (Brockwell et al. 2004).

Classifiers
The methods described above generate predictions based on prior knowledge of how neural activity is related to movement and, in some cases, the general structure of arm movements. Algorithms based on pattern recognition do not require such prior knowledge. The self-organizing feature map (SOFM) is an example of a classifier. These artificial networks depend on a consistent relation between neural firing rates and movement. SOFMs can be visualized as a single layer of elements or nodes (artificial neurons), each of which is connected to an input vector with a set of connection weights. In one example of this scheme (Lin et al. 1997), each of the $n$ weights corresponded to a recorded unit’s discharge rate ($n$ recorded units in the sampled population). Initially, each of the network element’s $n$-dimensional weight vectors was set randomly. In an initial step, the input vector consisting of the recorded neurons’ firing rates was compared to each of the artificial element’s weight vectors. The element with the weight vector closest to the input vector was declared as the winner and modified to resemble the input vector. Its neighbors’ weights were also moved closer to the input vector. This process was repeated for successive time points until clusters of similar elements were created on the surface. These clusters had similar weight vectors and were distinct from other clusters. In an identification step, each cluster was assigned to a direction. At this point the network is considered trained. New input vectors were fed in, and the predicted movement direction was chosen as the labeled direction of the cluster closest to the input vector.

Other networks using back-propagation (Wessberg et al. 2000) and nonlinear maximum-likelihood estimation (Pouget et al. 1998) have been used successfully to convert populations of neural activity to predicted movement.

Another type of decoder was developed using snippets of arm trajectory and time windows of recorded discharge rates (Isaacs et al. 2000). Each 200-ms window contained 10 sequential discharge rates from a single neuron and corresponded to a
trajectory snippet. Adjacent windows from the same neuron and windows between different neurons recorded simultaneously were compared in a covariance matrix and categorized with principal components. Used as a decoder, an eigenvector calculated from the training data was multiplied by a novel window of discharge rates to give a principal component. This component was compared with a dictionary of principal components from the training set, and the closest match was declared the winner. The corresponding trajectory snippet was then taken as the current trajectory prediction.

Practical Considerations

Extraction algorithm development can be summarized with a few statements. First, linear methods are very effective in extracting movement information from the recorded population. Second, methods that account for asymmetrical samples can effectively compensate for small sample size. Third, methods relying on training sets must produce robust outputs. The training set should be generated from a range of movements that represent the same range of movements for which the prosthetic device will be used. This problem can be addressed with algorithms that are adaptive, for example by changing weighting coefficients to maximize success rates. Finally, algorithms that capture the features of natural movement such as smoothness, segmentation, and curvature-speed tradeoffs are desirable because these more natural movements may be inherently more controllable.

The issue of controllability is important. Prosthetic devices are closed-loop devices. Subjects generate an output (i.e., neural signal) and then watch the device move. Observation of the device’s performance closes the control loop, and subjects have the opportunity to change their output to advance the device to the goal. Learning, in the form of modified neural output, can dramatically improve prosthetic performance (Carmena et al. 2003, Serruya et al. 2002, Taylor et al. 2002) compared to open-loop decoding (in which the subject does not observe movement of the device). The output of an extraction algorithm needs to be understood by the subject, who should be able to alter neural activity in a consistent way to achieve a predictable change in the way the prosthesis moves. This consideration is likely to be more important than the details and performance of a given extraction algorithm developed and evaluated with open-loop data.

OTHER BCIs

Recording individual action potentials requires an invasive surgical procedure to place the electrodes. EEG, which uses scalp electrodes to record signs of electrical brain activity noninvasively, is one type of BCI technology being implemented presently in human subjects. Although many electrodes are placed around the head, often only the electrodes over the sensorimotor cortex record signals useful for brain control. EEG activity is the complex sum of many neuronal currents with complex geometries filtered through the brain, skull, and scalp. Several methods
are used to extract movement-related intentions from these signals, including frequency decomposition, recognition of movement-related evoked potentials, and slow cortical potentials. These methods allow subjects to control computer cursors in one dimension in word-spelling and sentence-construction tasks and may be useful in developing a communication interface. Although it is still an open question whether this technology can generate signals for more complex movements, a recent report showed that a well-trained subject could use the EEG approach to perform a two-dimensional, center-out task (Wolpaw & McFarland 2003). A detailed review of this subject has been written recently (Wolpaw et al. 2002).

More localized electrical activity can be recorded when electrodes are placed between the dura and pia, as practiced in cortical mapping prior to epilepsy surgery. Information derived from these signals holds the promise of controlling more complex prosthetic movements (Leuthardt et al. 2003, Rohde et al. 2002).

This work implies that the more closely the recorded signal represents activity of single units, the more useful it is for movement control.

PAST AND CURRENT CNP STATUS

In a lecture at Oxford in 1963 W. Gray Walter (1963), a pioneer in the use of EEG and mobile automata, reported having used brain signals recorded from human motor cortex to operate a slide projector. This was the first example of CNP feasibility. Much later, in 1996, investigators demonstrated continuous movement control with initial, open-loop monkey experiments (Perepelkin & Schwartz 1996, Schwartz et al. 1996), which showed that populations of units recorded simultaneously with microwires in both right and left hemispheres could be used to generate population vectors for prosthetic arm control. In the ensuing four years, investigators continued most of this work in open-loop conditions, calculating population vectors posthoc or using real-time population vectors to drive a robot arm without feedback to the monkey (Isaacs et al. 2000, Wessberg et al. 2000).

More recently, a closed-loop CNP was first used in a rat to move a simple lever for reward (Chapin et al. 1999). Since then, CNPs using small populations of cell activity were shown effective for controlling closed-loop, two- (Serruya et al. 2002) and three-dimensional (Taylor et al. 2002) movements. As described in the previous section, subjects saw the result of the extraction process in real time and learned to modify their neural activity to improve their performance on the task (moving a computer cursor to specified targets). This, compared to open-loop control, dramatically improved their overall ability to reach the target quickly and accurately. In the Taylor et al. (2002) study, monkeys moved a cursor in three-dimensional, virtual-reality targets, performing the task either by moving their hands (hand control) or with their arms restrained (brain control). When switching between these tasks, the preferred directions of the chronically recorded units changed. Although there was no global pattern to these shifts, they were consistent from day to day, and the size of the shifts increased over days as the animals’ performance improved. The better linear fit of the activity patterns to the
cosine-tuning function was also coincident with this improvement. A coadaptive algorithm designed to track these learning-induced changes in neural activity was very effective. One animal had a success rate of more than 80% in daily training sessions (including periods of inattention) with a population of 64 units. This animal, using brain control, performed consistently for many minutes, reached novel targets on the first attempt, and moved with a speed and accuracy approaching normal arm movements.

This closed-loop task was modified to include an electro-mechanical robot arm in the control loop (Taylor et al. 2003). Instead of signaling cursor movement, the extracted cortical velocity predictions were streamed to a robot controller. Position of the robot arm was tracked three-dimensionally, and these movements were fed back to the graphics routine so that the animal saw the movement of the robot arm as cursor movement in the virtual reality display. Because the robot arm did not move exactly as commanded, the cursor movement was perturbed. However, the monkey learned to correct these perturbations using visual feedback during the movement to achieve a high level of performance. Similar results were found in a later study of two-dimensional movements combined with isometric grip regulation (Camena et al. 2003).

Currently, work is underway to demonstrate brain control of reaching and grasping using direct vision of an anthropomorphic robot arm (Helms Tillery et al. 2003, Schwartz 2003). A child-sized motorized prosthetic arm with a 3-degrees-of-freedom (DOF) shoulder and a 1-DOF elbow is mounted near the monkey’s shoulder. The monkey will reach out to a piece of food at different locations in the three-dimensional workspace and grasp it with a simple gripper before bringing the food back to its mouth. For training purposes, portions of the task can be automated. For instance, the arm can be computer-guided to reach for and grasp the food, followed by a brain-controlled retrieval by the monkey.

Recording technology continues to improve. Currently, the Michigan silicon probes continue to record good unit activity following their implantation more than six months ago in macaque cortex. The Utah probe has been modified with new insulating materials and connector technology (Cyberkinetics Inc.), leading to an increase in the number of units recorded per array. Finally, the microwire approach has been extended with new fabrication techniques (Nicolelis et al. 2003). This field is now at the point where more reliable implants can be combined with the control of elaborate prosthetic effectors capable of producing near-natural arm movement.

FUTURE WORK AND PROSPECTS

On the basis of a number of studies, cortical neural prostheses will be feasible in generating natural movements either with artificial effectors or intrinsic muscle activation. The information contained in the recorded signal, consisting of individual action potentials recorded simultaneously from the cerebral cortex, can be
extracted in real time and be used to make purposeful movements. Several groups are now preparing studies to apply this technology to humans.

The largest remaining obstacle to the successful implementation of cortical neural prostheses is the chronic recording electrode. The electrode should be able to record single-unit activity reliably for many years, be relatively easy to implant, and be capable of recording many action potentials from different cells in a small volume of cortex. Designs such as those produced by the University of Michigan, where multiple recording sites are placed on each shank, have the advantage of such dense sampling. On the other hand, the Utah probe can achieve high-density sampling by spacing many shanks close together. These probes have an ideal placement of the recording surface at the shank tip but do not have multiple sites along each shaft. A potential compromise would be to arrange planar silicon probes in a high-density array, a project that has produced several prototypes to date (Hetke & Anderson 2002).

As described above, technology is being developed to regulate the tissue/electrode interface associated with arrays implanted in the brain. Investigators have demonstrated the potential of this technology in human patients with the cone electrode (Kennedy et al. 2000). This electrode is a capillary tube filled with growth factor or peripheral nerve extract. Also in the tube are the exposed ends of two microwires, which act as differential electrodes. Neurites that sprout in response to the electrode penetration are attracted to the interior of the tube, through which they grow and form synaptic connections to other neurons. The axon is permanently trapped next to the recording electrode. Although only a few channels of multiunit data were recorded, this activity was used by locked-in ALS patients for communicating with a spelling/letter-board program. One patient used this method for more than a year.

An ethical issue arises as we work toward implementing these devices in disabled patients. Paralyzed patients are motivated to volunteer as experimental subjects. Surgeons and researchers are eager to implant chronic electrodes. The question persists, at what point should this imperfect technology be applied? This common issue in bioengineering has been addressed with other neural prostheses (for example, cochlear implants, deep brain stimulators, and visual prostheses). Presently, an informal survey (A. Schwartz, personal communication) of the laboratories using CNP suggests that, on average, a chronic electrode implanted in monkey cortex has only a 40% to 60% chance of recording unit activity. Although each lab has an example of an all-star animal with good recordings for multiple years, electrode recordings usually deteriorate after several months. How much improvement in this technology is needed before human experimentation is warranted? Considering the rapid improvements in electrode technology, a better understanding of the biology associated with electrode interaction within the brain, the financial race of investment-driven development, and the desires of patients and researchers to try this technology, humans likely will be implanted with these chronic devices in the next few years. The technology used in these devices, the care taken to develop the correct surgical procedures, and the peripheral
technologies associated with the electrodes—connectors, telemetry, spike conditioning, and real-time computing—should be well considered, not only by regulatory bodies but also by the scientists, engineers, and patients who will be using them. The marketing aspect of this technology is rarely considered. There is a trade-off between the severity of a patient’s deficit and the relative efficacy of a CNP. Locked-in patients will benefit from any device that allows them to communicate better, whereas patients with C5 spinal cord lesions may still have arm mobility and would only benefit from a device that restored natural arm movements combined with a degree of grasping. How many patients within these categories would be willing to undergo surgical implantation before this technology is commercially viable? This answer depends, to a large degree, on the viability of the electrodes, the information content of the recorded signals, and the engineering needed to make natural, agile effectors. Furthermore, it will be important to develop effective training procedures for patients to acquire the skills to use these devices. Certainly, companies have already been formed with the expectation that these conditions will be met. In the mean time, those of us working in the laboratory have the exciting prospect of direct access to the previously inaccessible neuronal substrate of human skill, with the potential generation of scientific discovery of fundamental aspects of learning and cognition.

ACKNOWLEDGMENTS

The author thanks Dawn Taylor, Stephen Helms Tillery, and Rob Isaacs for their contributions to the experimental results cited here. Furthermore, the Harrington Department of Bioengineering at Arizona State University, the Neural Prosthesis Program at NIH, and the Whitaker Foundation supported the author’s research discussed in this review.

The Annual Review of Neuroscience is online at http://neuro.annualreviews.org

LITERATURE CITED


Georgopoulos AP, Kettner RE, Schwartz AB. 1988. Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. J. Neurosci. 8:2928–37


