CHAPTER EIGHT

Restoring upper extremity function with brain-machine interfaces

Samuel R. Nason, Matthew J. Mender, Joseph G. Letner, Cynthia A. Chestek, and Parag G. Patil

*Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, United States
Robotics Graduate Program, University of Michigan, Ann Arbor, MI, United States
Neuroscience Graduate Program, University of Michigan, Ann Arbor, MI, United States
Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, United States
Department of Neurosurgery, University of Michigan Medical School, Ann Arbor, MI, United States
Department of Neurology, University of Michigan Medical School, Ann Arbor, MI, United States
*Corresponding author: e-mail address: pgpatil@med.umich.edu

Contents

1. Introduction 154
2. Neural signal recordings 156
  2.1 Noninvasive recording 158
  2.2 Brain surface recording 159
  2.3 Penetrating extracellular recording 160
3. Signal processing and prosthetic control 162
  3.1 Filtering, amplifying, and digitizing 163
  3.2 Processing and decoding 166
  3.3 Controlling prosthetic actuators 171
4. Emerging technologies 174
5. Conclusions 176
References 177

Abstract

One of the most exciting advances to emerge in neural interface technologies has been the development of real-time brain-machine interface (BMI) neuroprosthetic devices to restore upper extremity function. BMI neuroprostheses, made possible by synergistic advances in neural recording technologies, high-speed computation and signal processing, and neuroscience, have permitted the restoration of volitional movement to patients suffering the loss of upper-extremity function. In this chapter, we review the scientific and technological advances underlying these remarkable devices. After presenting an introduction to the current state of the field, we provide an accessible technical discussion of the two fundamental requirements of a successful neuroprosthesis:
signal extraction from the brain and signal decoding that results in robust prosthetic control. We close with a presentation of emerging technologies that are likely to substantially advance the field.

1. Introduction

The advent of digital signal processing and high-speed computation has opened the possibility of restoring lost upper-extremity function through brain-machine interface (BMI) neuroprosthetic devices. By capturing brain activity, predicting underlying intentions from that activity, and controlling a prosthetic device accordingly, BMI neuroprostheses emerge as a promising therapy for the 27 million existing patients worldwide with spinal cord injury and the nearly 1 million new patients injured annually (James et al., 2019). Presently, BMI devices can guide robotic arms and, in some cases, activate paralyzed extremities with volitional control in a limited way in a laboratory environment. Table 1 details some of the major accomplishments made by groups investigating brain-machine interfaces across the past 40 years, from studies in animals to present human clinical trials. Going forward, the promise of the field is encouraging.

BMI technologies have arisen from both publicly and privately sponsored research efforts. The Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative of the National Institutes of Health (NIH) has funded hundreds of academic, neuroscientific, and neuroprosthetic research programs, yielding major advances in neural interface technologies and the understanding of brain signaling. In addition, research performed in the private sector has spawned numerous companies (Emotiv Inc., Neuralink Corp., Kernel, Paradromics Inc., InBrain Neuroelectronics) and corporate ventures (Google Brain, Facebook Reality Labs) seeking to enable volitional interaction with prostheses, computers, vehicles, and video games through thought.

In this chapter, we will first explore various portable methods of extracting information from the central nervous system that are capable of predicting motor intention. Then, we will focus on the potential of high-bandwidth intracortical recording to provide the information needed to restore accurate and precise movement. Neuroprostheses have progressed drastically in the past 20 years, from slow predictions (Carmena et al., 2003; Serruya et al., 2002; Shenoy et al., 2003; Taylor et al., 2002) to control systems using all limbs of the body (Willett et al., 2020). As such, we will conclude our chapter with some discussion of the newest neuroprosthetic technologies and some projections for the coming years.
<table>
<thead>
<tr>
<th>Year</th>
<th>Subject</th>
<th>Milestone</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>Monkey</td>
<td>Primary motor cortex neurons change activity levels depending on the directions of arm movements</td>
<td>Georgopoulos, Kalaska, Caminiti, and Massey (1982)</td>
</tr>
<tr>
<td>1996</td>
<td>Cat</td>
<td>The Utah microelectrode array is presented for single unit recording with 100 electrodes</td>
<td>Nordhausen, Maynard, and Normann (1996)</td>
</tr>
<tr>
<td>1999</td>
<td>Rat</td>
<td>Real-time control of a robotic arm using intracortical brain activity</td>
<td>Chapin, Moxon, Markowitz, and Nicolelis (1999)</td>
</tr>
<tr>
<td>2006</td>
<td>Human</td>
<td>Human implanted with a Utah array uses a BMI</td>
<td>Hochberg et al. (2006)</td>
</tr>
<tr>
<td>2008</td>
<td>Monkey</td>
<td>Monkey uses a single-neuron BMI to control functional electrical stimulation</td>
<td>Moritz, Perlmutter, and Fetz (2008)</td>
</tr>
<tr>
<td>2011</td>
<td>Human</td>
<td>Continued Utah array functionality for 1000 days after implantation in a human</td>
<td>Simeral, Kim, Black, Donoghue, and Hochberg (2011)</td>
</tr>
<tr>
<td>2012</td>
<td>Monkey</td>
<td>Monkeys intuitively control functional electrical stimulation of their paralyzed arms using a BMI</td>
<td>Ethier, Oby, Bauman, and Miller (2012)</td>
</tr>
<tr>
<td>2012</td>
<td>Human</td>
<td>Human uses a BMI to control a robotic arm for self-drinking</td>
<td>Hochberg et al. (2012)</td>
</tr>
</tbody>
</table>

Continued
2. Neural signal recordings

Diagnosing neurological disorders by recording the activity of the central nervous system has become common practice in medicine. Across the past several decades, several recording modalities have become critical components of many diagnostic procedures, some of which are illustrated in Fig. 1. From non-invasive to invasive deep-brain methods, which are compared in Fig. 2, each has prevalence in the realm of neuroprosthetic devices, which we discuss below.
Fig. 1 Relative invasiveness of various portable neural recording modalities. Scale of the drawing is not realistic. ECoG, electrocorticography; EEG, electroencephalography.

Fig. 2 Relative specificity to single neurons of various portable electromagnetic neural recording modalities. Neurons as displayed in the inset are realistically dense at 40,000 neurons/mm³. Neuron distributions and densities in the main image are not realistic.
2.1 Noninvasive recording

Noninvasive recording modalities require minimal to no alteration of the person to record brain activity. Without crossing the skin barrier, noninvasive recordings avoid many of the infection, scarring, and surgical risks that are inherent to more invasive recording methods. This makes noninvasive recording methods the most readily accessible to any person, as the risks of complications are low. Technologies commonly used in clinical spaces, such as functional magnetic resonance imaging, functional near-infrared spectroscopy, and magnetoencephalography, have substantial literature demonstrating their capabilities to measure brain activity noninvasively. However, the seconds to minutes of latency between cognitive effort and representation in the measured signals, coupled with massive equipment needs, render these methods infeasible for real-time and portable neuroprostheses. Consequently, we will begin our discussion with electroencephalography (EEG).

Electroencephalography is a non-invasive recording technique that measures the electromagnetic byproducts of brain activity instantaneously. As neurons in the brain fire and signals traverse axons and dendrites, small-amplitude electric currents are generated along with coupled magnetic fields, which are also weak in magnitude. When thousands to millions of neurons in close proximity become active in synchrony, which is typical during cognitive activation of a brain region, the magnetic fields from individual neurons superimpose to create larger-magnitude magnetic fields. These larger-magnitude magnetic fields generate small electrical currents on the scalp, which is measured by EEG. EEG signals are recorded using metallic electrodes several millimeters to centimeters in diameter that make electrically conductive connections with the scalp. Different frequency bands of the recorded activity, such as the delta (\(\sim<4\) Hz), theta (\(\sim4–8\) Hz), alpha (\(\sim8–15\) Hz), beta (\(\sim15–30\) Hz), and gamma (\(\sim30–40\) Hz) bands, are assumed to represent various types of cognitive effort, enabling a neuroprosthesis to predict intention of the user from strength, or power, of these bands.

Due to the simplicity of the EEG cap, EEG systems are lightweight, safe to equip, and portable enough for use with a neuroprosthesis. EEG-based BMIs were originally proposed in the 1980s, with the P300 speller as the first brain-controlled computer (Farwell & Donchin, 1988). With relative ease of implementation and use, they have been thoroughly explored and usually are limited to communication applications, which can accommodate high error rates. Classification accuracies of EEG-based BMIs alone can be fairly high in the range of 90%, but their information throughputs
(a common measure of performance for all types of BMIs) are just 0.71 bits per second (Speier, Arnold, Lu, Deshpande, & Pouratian, 2014). This level of information is useful for rehabilitation after stroke (Hodkin et al., 2018) and some commercial uses, such as controlling basic computer functions and video games. Unfortunately, the relatively low throughput, reliance on conductive and gelled electrodes (which are difficult to place with paralysis), and requirement to be motionless (Kline, Huang, Snyder, & Ferris, 2015) make EEG usage for functional restoration difficult.

2.2 Brain surface recording

Noninvasive recordings are substantially displaced from the source neurons, making it difficult to determine the specificity of a particular recording to some behavior. Invasive recordings attempt to address the challenges of spatial specificity by placing electrodes closer to neurons via surgical implantation. Up to 60% of people with paralysis, from spinal cord injury, ALS, and other neurological disorders, have reported they would be willing to undergo surgery to enable use of a neuroprosthesis (Blabe et al., 2015). All invasive recording methods developed to date capture electrophysiological signals from brain activity. The least invasive of such procedures involves placing an array of electrodes, similar in organization to an EEG array but with smaller electrode sizes, on the surface of the dura mater. Implanting the electrodes beneath dura and in direct contact with the surface of the brain is more invasive but brings the electrodes closer to the neurons being recorded.

These types of invasive recordings are called epidural and subdural electrocorticography (ECoG). ECoG electrode arrays are organized in similar fashion to EEG electrode arrays, but with smaller contact sizes, flatter materials to fit between the skull and the brain, and generally greater electrode quantities. Similar to EEG recordings, ECoG recordings represent the compound activity of neurons, but on the order of hundreds to thousands of cells as the electrode sizes are generally smaller and the electrodes are closer to the neural populations. The same general frequency bands of population activity are found in ECoG and EEG.

For BMI applications, ECoG is adept at detecting movement effort with nearly 100% accuracy due to the greater neuronal specificity. A popular application of ECoG-based BMIs is the prediction of hand and finger movements. While many continuous control experiments predict one degree of freedom (Chao, Nagasaka, & Fujii, 2010; Flint, Rosenow, Tate, & Slutzky,
most of them rely on projecting a relatively binary movement/no movement signal onto a continuous movement dimension, instead of directly predicting position along that dimension. Others have attempted to classify attempted postures of the hand instead of the positions of the fingers, but required seconds of latency to make accurate predictions (Chestek et al., 2013; Pistohl, Schulze-Bonhage, Aertsen, Mehring, & Ball, 2012). Alternatively, using ECoG recordings to predict speech has been quite successful, enabling prediction of speech in human clinical trials (Anumanchipalli, Chartier, & Chang, 2019; Makin, Moses, & Chang, 2020). Since speech is spread across a wide area of cortex, the wide expanse of ECoG arrays makes speech prediction a particularly productive application for ECoG.

ECoG recording electrodes have also experienced some technological advancements in recent years. Not only are electrode sizes decreasing, enabling greater quantities of channels and greater specificity to neurons (Viventi et al., 2011), some sizes have become so small that single-neuron action potentials have been recorded from the surface (Khodagholy et al., 2015). It has been hypothesized using field modeling that this capability requires very small electrodes in very close contact to the brain, where the potentials are likely originating from superficial layer I neurons (Hill, Rios, Sudhakar, Lempka, & Chestek, 2018).

### 2.3 Penetrating extracellular recording

The most invasive neural recording options place electrodes among the cells in the brain. This enables simultaneous capture of population activity and the activity of individual neurons within the electrophysiological recordings. In contrast to the disc-shaped electrodes discussed thus far, electrodes that penetrate brain tissue are needle-shaped probes with one or more recording sites. As individual neurons are activated, they produce an action potential that creates a voltage gradient in the nearby extracellular space, which is detectable within approximately 100–150 μm from the neuron (Henze et al., 2000; Moffitt & McIntyre, 2005). If a shank with a recording site is within that range, its voltage recording may represent each action potential’s voltage gradient with what are called “spikes” in the signal, or visible perturbations that are higher frequency compared to the population activity bands discussed for EEG and ECoG. By recording the spiking rates of individual neurons, these most invasive modalities of recording are also the most specific to behavior.
The most invasive penetrating electrodes are flexible cylinders tens of centimeters in length with multiple contacts. These electrodes provide deep-brain stimulation to structures such as the ventral intermediate nucleus of thalamus, subthalamic nucleus, and zona incerta to reduce symptoms of motor neurodegenerative disorders, such as Essential Tremor and Parkinson’s Disease (Benabid, Chabardes, Mitrofanis, & Pollak, 2009). However, these electrodes can also provide electrophysiological recordings of these deep brain structures related to motor intentions (Swann et al., 2018). Due to the contact size of typical deep-brain stimulation macroelectrodes, the ability to distinguish the activity of individual neurons is difficult with so many active neurons nearby the large electrode area. Instead, often the local field potentials representing the aggregate activity of the nearby hundreds of neurons drives the voltage recordings. On the other hand, deep-brain microelectrodes have electrode areas that are small enough to record the spiking activity of individual neurons. Some studies have used these electrodes to obtain BMI control signals (Patil, Carmena, Nicolelis, & Turner, 2004; Tan et al., 2016). Thalamus and neighboring structures are promising targets for deep-brain neuroprostheses. Descending cortical and cerebellar motor commands can be captured entirely within a few cubic millimeters, suggesting that just one or two deep-brain microelectrode arrays could capture all signals for a whole-body neuroprosthesis. However, limited site quantities and surgical risks associated with heavily invasive deep-brain implants are challenges that must be overcome before widespread clinical translation (Bullard et al., 2020).

More superficial implants lower the surgical risk of extracellular recording. Michigan-style probes (Wise, Anderson, Hetke, Kipke, & Najafi, 2004) allow many simultaneous recordings with hundreds to thousands of planar electrodes and have had tremendous success across the neuroscientific community. One shank with nearly one thousand electrodes, such as the new Neuropixels probe of similar design to the Michigan-style probe (Jun et al., 2017), can provide highly spatially specific signals along columns of neural processes. Recordings along columns of neurons are extremely helpful for neuroscientific investigations because neural signals traversing axons and dendrites from one neuron to another can be tracked precisely in space and time. However, for a high-dimensional neuroprosthesis, sampling many independent neurons across multiple columns that represent those behavioral dimensions is necessary.

Single-recording-site penetrating electrodes enable simultaneous recording from multiple places in the brain. Historically, microwires were implanted individually to obtain simultaneous recordings from separate cortical areas,
and either left-in-place chronically or explanted following each experiment (Georgopoulos, Kalaska, Crutcher, Caminiti, & Massey, 1984; Taylor et al., 2002). Then, microfabricated devices like the Utah microelectrode array (or the Neuroport array approved for use in humans) were developed for simpler chronic implantation of electrodes by providing an organized bed of single-site shanks on a rectangular platform to cover a wide cortical area (Nordhausen et al., 1996). The shanks, which are generally all the same 1.1 or 1.5 mm length with a 0.4 mm pitch for cortical implants, target layer V neurons in motor cortex, which will tend to be pyramidal cells driving corticospinal neurons. Since the technology is already approved for acute use in humans to localize sources of epilepsy and it allows spatially specific recording of the spiking activity of many independent neurons, Utah arrays have become the standard implant for chronic intracortical BMI studies in nonhuman primates and humans, enabling self-feeding (Ajiboye et al., 2017; Hochberg et al., 2012; Velliste et al., 2008), high-speed typing (Gilja et al., 2012, 2015; Kao, Nuyujukian, Ryu, & Shenoy, 2017; Pandarinath et al., 2017), and control of their own limbs despite paralysis (Ajiboye et al., 2017; Bouton et al., 2016; Capogrosso et al., 2016; Ethier et al., 2012).

3. Signal processing and prosthetic control

Brain–machine interfaces have not yet translated to full-time use because of several inherent issues with experimental systems. Current clinical and experimental BMIs require a lab cart with all of the computers, require that the user remain tethered to the recording system with all of the high-bandwidth data being recorded, and consume hundreds of watts (several car batteries worth) of electrical power to run the necessary computations. This is an astounding amount of power in comparison to typical implantable medical devices like pacemakers and deep-brain stimulators, which require just tens of microwatts. Relaxation of all of these requirements is critical for therapeutic BMIs, and recent work has suggested promising solutions to many of these problems. Below, we discuss the minimum required components of a BMI in terms of their history, current investigations, and opportunities for improvement in the future.

Intracortical brain–machine interfaces are generally composed of several key components, which are portrayed in Fig. 3. First, the tissue interface directly acquires electrophysiological signals from the brain, as discussed above. These voltage signals must be filtered and then amplified so they can be converted into digital signals. This allows general computing
hardware to perform computations on those measurements. The first in the chain of computations is feature extraction, which extracts the components from neural signals, known as neural features, that are easiest to use to make predictions. The second is the prediction algorithm that converts the neural features into a behavioral prediction, usually referred to as a “decoder.” Finally, those behavioral predictions are transmitted to a prosthesis to perform the intended command of the user.

3.1 Filtering, amplifying, and digitizing

Typically referred to as the analog front-end (AFE), the hardware required to filter, amplify, and digitize intracortical signals has been a large topic of investigation for decades. Although modern experimental BMIs employ several computers, recent simulations show that the power consumption of the minimum required BMI components (as discussed previously) is dominated by the AFE (Nason et al., 2020). Consequently, it is critical to reduce the power consumption of the AFE to translate BMIs into widespread clinical use.
At minimum, experimental AFEs first filter and amplify the incoming neural signals (Chandrakumar & Markovic, 2017a, 2017b; Chen, Blauuw, & Sylvester, 2014; Harrison & Charles, 2003; Jang et al., 2018; Lopez et al., 2014; Mahajan et al., 2015; Muller et al., 2014; Ng & Xu, 2016; Shen, Lu, & Sun, 2018). The first filter in the chain is a band-pass filter that eliminates frequencies outside of those relevant to brain activity. Typical BMI AFEs inherently filter out frequencies below approximately 0.1 Hz and above approximately 10 kHz to permit passing of population activity that exists in the single to hundreds of Hz and single neuron activity that exists in the single kHz range. Any electromagnetic interference from radio, Wi-Fi, or cellular signals and the baseline voltage of the body’s tissue are therefore attenuated by this first stage filter. Then, groups typically apply an additional set of filters to extract one or both of the population activity (several filters with all cutoffs below approximately 500 Hz) and the single neuron activity (high-pass filter with a cutoff near 200–300 Hz).

In addition to reducing noise, front-end filters also amplify their neural signal inputs. Amplification is necessary to bring the neural signals, which are typically on the order of tens to hundreds of microvolts in amplitude, to voltages at which the downstream electronic circuits generally operate, which are on the order of volts. Conveniently, this amplification factor can be incorporated into the design of one or both of the filter stages. We can further compress the circuitry to save electrical power consumption with another optimization. Instead of two stages of filters, which is helpful for on-the-fly configuration for experimentation, one appropriately configured filter would suffice in a clinical BMI. This eliminates some unnecessary hardware and can substantially cut the electrical power consumption. One final improvement to the efficiency of the amplifiers and filters is related to the inherent advancements in transistor technology. Improved fabrication technologies, reduced transistor sizes, and more electrically efficient materials have enabled drastic improvements in the power efficiencies of neural amplifiers across the past several decades.

The last stage of the AFE is the analog-to-digital converter (ADC). The ADC receives the filtered and amplified analog neural signal and periodically converts it to a digital one for use with computers. The rate at which conversions occur can be controlled, but it should be at least double the highest recorded frequency to enable valid and artifact-free reconstruction in the digital domain (Shannon, 1949). The resolution of the conversions can also be controlled with hardware customizations and is generally set to 16 bits. Such a high resolution allows the detection of sub-microvolt changes in the
recorded voltages while maintaining a dynamic range large enough to com-
pensate for millivolt scale recording artifacts, which can result from some
movements or electrical stimulation (see Section 3.3 for a discussion of this).
Although ADCs are critical to the operation of a BMI, the fabrication and
design changes that have resulted in improved power efficiency are largely
independent of the field of neuroprostheses. As such, we discuss only the
selection characteristics of ADCs that can provide power-savings.

In many BMI usage cases, there are rarely electrical artifacts in the signals.
In usage cases that inherently generate artifacts, some of which will be dis-
cussed in Section 3.3, there are methods to record neural signals while
avoiding those artifacts. The 16-bits of ADC resolution typically used in
BMI recording systems scales into the millivolt range to accommodate for
these artifacts, which is an order of magnitude larger than the largest of
signals recorded from intracortical electrodes. Consequently, 16-bits is a
luxurious resolution that is generally overdesigned for most BMI applica-
tions. In one analysis of the ADC requirements for BMIs, it was found that
8 effective bits of resolution would suffice to accurately capture neural spikes
(Even-Chen et al., 2020). A 50% reduction in the quantity of bits, from 16 to
8, would result in an estimated 99.997% reduction in the power consump-
tion of each ADC (Kester, 2004; Murmann, 2020). Additionally, whether
by hardware or software, reduction in the quantity of bits reduces the com-
putational complexity down-stream of the ADC, which will be discussed
later.

The power consumption of both amplifiers and ADCs is directly tied to
the bandwidth of the signals they are recording. Intracortical local field
potentials require low recording and amplification bandwidths, similar to
ECoG and EEG neural activity bands, which implies reduction in the power
consumption for those amplifiers and ADCs. Unfortunately, since they rep-
resent the net activity of hundreds of cells, their relationship to behavior
is more abstract than single unit spiking rates. This may explain why local
field potential-based BMIs have not achieved the performance levels of
spike-based BMIs (Flint, Wright, Scheid, & Slutzky, 2013; Milekovic
et al., 2018; Perge et al., 2014; So, Dangi, Orsborn, Gastpar, & Carmena,
2014; Stavisky, Kao, Nuyujukian, Ryu, & Shenoy, 2015; Wang et al.,
2014). Alternatively, some groups have investigated reducing the bandwidth
of spike recordings. One study found that the standard 0.25–7kHz band-
width typically used to detect threshold crossings resulting from spikes is
unnecessarily large, and a 1–2kHz bandwidth more than suffices to maintain
high decoding performance (Even-Chen et al., 2020). We also previously
found that a more reduced 0.3–1 kHz bandwidth, denoted the spiking band power, predicts behavior comparably well to higher-bandwidth decoders (Irwin et al., 2016), with the additional benefits of greater signal-to-noise ratio and greater specificity to single neurons (Nason et al., 2020). By reducing amplifier and ADC bandwidth to the 0.3–1 kHz spiking band, we found that the AFE could undergo a 90% reduction in power consumption with the custom circuits described here (Irwin et al., 2016; Nason et al., 2020).

Although custom analog circuits have great promise toward promoting next generation BMIs that can be used anywhere, challenges remain that must be addressed before they can be deemed safe for general use with humans. Customized amplifiers like those cited here are difficult to fabricate reliably without being thoroughly tested in a variety of conditions and licensed by a commercial vendor. Only a few academic neural amplifiers have been used regularly in preclinical studies with animals. Cortera Neurotechnologies, Inc. (Berkeley, CA, USA), have developed one such device capable of recording and stimulation (Johnson et al., 2017). The RHD series of devices produced by Intan Technologies (Los Angeles, CA, USA) is another example of an integrated electrophysiological sensor that has translated to experimental use (Harrison & Charles, 2003; Harrison et al., 2007). The amplifiers have maintained their customizability while keeping consumption low, requiring 1.07 mW/channel for typical 10 kHz bandwidth or 0.14 mW/channel for 0.3–1 kHz spiking band power bandwidths. Although not yet used for many translational studies to date, they have been incorporated into embedded BMIs for animal studies (Bullard et al., 2019) and tested with humans (Wang et al., 2019).

Existing and future analog circuits provide clear improvements to current clinical BMIs, but they require sufficient iteration and functional testing and validation prior to their commercialization and full-time use with BMIs.

### 3.2 Processing and decoding

Once neural signals are digitized, BMIs employ computational algorithms that convert the recorded brain activity into a person’s intentions. These algorithms are called decoders as they decode the encoded neural information into a more usable control signal. The encoded information for intracortical recordings is often some representation of each electrode’s spiking or firing rate, broadly called features in the machine learning community, which are extracted via some processing discussed in the subsequent paragraph. The different effectors that can be controlled by these neural features
will be discussed in detail below, but some brief examples include computer
cursors and robotic or natural arms. BMIs often predict positions and veloc-
ities from the neural features to control these effectors, but in cases where
limb movements are being predicted, BMIs can also predict the activations
of the limb’s muscles through a biomechanical model (Naufel, Glaser,
Kording, Perreault, & Miller, 2019). These predictions are made at a phys-
iologically relevant frequency (>10 predictions per second) so that move-
ment of the effector appears continuous and the visual feedback is readily
provided. As was discussed previously, neural data from most recording
modalities have been used to predict these control signals, but intracortical
recordings have been most successful in predicting accurate and precise
real-time control signals due to the better time and spatial resolution.

The first processing stage in an intracortical BMI is the extraction of rel-
levant neural features from the digitized neural data. The bandwidth of the
digitized neural data often governs what types of neural features can be
extracted. If the data is low bandwidth, or filtered below approximately
500Hz, then local field potentials, such as power in the delta, theta, alpha,
beta, and gamma bands, are the primary feature. Filtering for these features
can be present in the AFE or otherwise be implemented in processing.
Within each band, some estimate of the signal power is calculated by inte-
grating the squared signal over some time period for each channel. Then, a
subset of these features can be used to predict movement in a BMI (Flint,
Lindberg, Jordan, Miller, & Slutzky, 2012). The period over which these
calculations are computed generally corresponds to the frequency at which
an updated prediction is required, often on the order of tens of milliseconds.
Since the relationship between local field potentials and complex behavior is
less directly related, as previously discussed, intracortical BMIs more often
use spiking rates of the recorded neurons as the neural feature.

Spiking rates can be extracted from digitized signals that have been
filtered between approximately 300Hz and 10kHz. The spiking rate of
an individual unit, or one or more neurons firing in synchrony with mor-
phologically similar spikes, can be extracted through a process called spike
sorting, discussed in detail by others (Lewicki, 1998; Rey, Pedreira, &
Quiroga, 2015). After sorting, the spikes of each unit are counted in similar
periods, yielding one feature per unit. Unfortunately, spike sorting requires
substantial processing, either manual or algorithmic, that makes implemen-
tation on translational BMIs difficult due to their limited space for compu-
tations. As such, many groups have transitioned to using just the threshold
crossing rates of each electrode without sorting the spikes, as it is presumed
that the neurons nearby one electrode encode similar information (Ventura, 2008). It has also been demonstrated that simply using the multiunit threshold crossing rates even for neuroscientific analyses is valid with negligible information losses (Trautmann et al., 2019). Unfortunately, extraction of threshold crossing rates requires the high power AFEs discussed in the previous section, making them difficult to fit in the budget for portable and low-power BMIs. Hence, lower bandwidth proxies for spiking rates, like the 0.3–1kHz spiking band power discussed in the previous section, may be more realistic features for a portable BMI.

After feature extraction, there are a number of different decoding algorithms to translate the observed neural data into a command signal. Before any particular algorithm can be used in a BMI, it first must undergo its own specific training procedure that optimizes the decoder’s parameters for best prediction of those control signals. These parameters, sometimes called weights, represent how any particular measurement of neural activity can be translated to a control signal from some number of operations. Early on, it was observed that neurons modulate their firing rate with the direction of movement (Georgopoulos et al., 1982), which laid the groundwork that control signals could be predicted from weighted sums of the observed neural signals. The oldest decoding algorithm is perhaps the population vector algorithm based on this work (Georgopoulos, Schwartz, & Kettner, 1985). Early nonhuman primate BMI experiments used population vectors to control a cursor in real-time (Taylor et al., 2002), but the algorithm was mostly replaced by other linear/weighted-sum decoders shortly after. Linear regression decoders, which no longer make the same population vector assumption that each recorded neuron contributes equally to a behavior (Chase, Schwartz, & Kass, 2009), allowed nonhuman primates to control computer cursors (Serruya et al., 2002) and a robotic arm (Carmena et al., 2003). Ridge regression decoders, which employ a regularization to the linear regression to train more representative weights and avoid some mathematical issues during training, have enabled human participants to control multiple degrees of freedom of a robotic arm (Collinger et al., 2013). Additionally, Kalman filters (Kalman, 1960), which optimally combine predictions based on prior behaviors and predictions based on current neural data, have demonstrated great success in predicting nonhuman primate (Gilja et al., 2012; Irwin et al., 2017; Mulliken, Musallam, & Andersen, 2008; Wu, Shaikhouni, Donoghue, & Black, 2004) and human behavior (Ajiboye et al., 2017; Gilja et al., 2015; Malik, Truccolo, Brown, & Hochberg, 2011; Pandarinath et al., 2017; Simeral et al., 2011).
Linear decoding models have enabled impressive accomplishments with BMIs, but they maintain an inherent assumption that the relationship between brain activity and behavior is linear. They cannot accommodate the possibility where the behavior is not well described by a weighted sum of the neural activity measurements. It may ultimately be possible to achieve better performance with a nonlinear model that better represents what is likely to be a nonlinear relationship outside of a limited context. Some groups have already explored this by implementing computational neural networks for the prediction of behavior. Training and using neural network decoders are complex procedures, though the recent advancements in high-speed and parallel computing have made that process feasible for heavy computational machines. An early recurrent neural network demonstrated great cursor control in nonhuman primates (Sussillo et al., 2012), and more recent networks have improved BMI control in humans (Schwemmer et al., 2018). Though these studies have given neural networks promise for BMI control, their computational complexity must drop drastically before they can become relevant to portable or even implantable BMIs.

Whether an algorithm is linear or nonlinear, there are several additives that can potentially reduce the computational complexity and improve the decoding performance. Closing the feedback loop for real-time BMIs creates a very different control system than if the existence of feedback is ignored, providing an opportunity for further optimization of a BMI decoder’s parameters. This has driven researchers to employ a technique called intention training that incorporates some assumptions about the user’s behavioral intentions when performing some action. One example is recalibrated feedback intention training (ReFIT), originally proposed for usage with the Kalman filter but contains principles that can be applied to any BMI decoder (Gilja et al., 2012). After using the Kalman filter in a BMI task, the decoder’s predictions, corrected based on the assumption that the user intended to acquire targets as quickly as possible, are then used to train an updated Kalman filter, theoretically with a better linear representation of how neural activity maps to intended behavior. Several variations of this technique have been tested (Orsborn, Dangi, Moorman, & Carmena, 2012; Willett et al., 2018), all showing improvements in performance compared to decoder models trained without consideration for visual feedback in monkeys and humans. These studies have motivated a piecewise linear model to better train BMI decoders on the first try, without needing an intention training step (Willett et al., 2019). As these techniques all occur
during training, which can take advantage of computationally intense and power-hungry machines, they remain very relevant for low-power portable BMIs.

Another computational addition to BMI decoders is dimensionality reduction, where the hundreds of neural features being recorded can be compressed to just the features that best represent the neural population’s activity (Churchland et al., 2012; Cunningham & Yu, 2014; Shenoy, Sahani, & Churchland, 2013). Principal components analysis (PCA) is an example of a dimensionality reduction technique, where the PCA dimensions would first be trained, then the incoming neural features would first be decomposed into the PCA dimensions in real-time prior to being decoded into behavior. This can have the benefit of denoising the neural data and may reduce power consumption in the decoder by reducing the number of features to process. PCA and similar techniques have also been used to address waveform amplitude and firing rate changes of neurons in motor cortex (Chestek et al., 2007, 2011), which had traditionally made all decoder models lose their accuracy over time despite signals still being available across years (Simeral et al., 2011). By realigning neural features across days using a manifold/subset of neural population dimensions (Gallego, Perich, Chowdhury, Solla, & Miller, 2020), latent factor analysis via dynamical systems (LFADS) (Pandarinath et al., 2018), or other techniques (Degenhart et al., 2020), high-performance decoding using the originally trained weights remains possible. However, it remains unclear if the cuts in computational complexity resulting from compression of neural features are great enough to warrant the additional computations inherent to the dimensionality reduction algorithm.

To summarize, the advent of more advanced processing computers and decoding algorithms has sparked a remarkable improvement in decoding performance across the past two decades. Today, linear decoders have been established as the tried-and-true method for predicting behaviors from brain activity. The computational simplicity of linear decoders makes them promising candidates for implementation on a portable and low-power BMI, as the algorithms are the appropriate complexity to fit on embedded processors, similar to what is in a deep-brain stimulator. Newer decoders, such as neural networks, have shown promise for the future capabilities of BMI decoders, but they require substantial reductions in computational complexity before they can be implemented full-time in BMIs. Although including dimensionality reducing techniques like PCA may add some substantial computations to a BMI, perhaps reducing the number of input
features to a decoder may result in an overall lower complexity and provide room for more complex algorithms to be implemented on portable and low-power BMIs.

### 3.3 Controlling prosthetic actuators

The final stage of a brain–machine interface is the prosthesis. The prosthesis can take many forms: a surrogate arm and hand that is the colloquial usage of the term prosthesis, or a tool, such as a computer or a speaker, that provides some special functionality to the user. Ultimately, the goal of the prosthesis is to execute the intentions of the user. The following paragraphs discuss the different prostheses currently being investigated (along with images in Fig. 4) as well as their projections for usage outside of laboratory environments.

**Fig. 4** A variety of prosthetic actuators used in literature. Virtual prostheses include computer cursors (A, Gilja et al., 2015), keyboards (B, Jarosiewicz et al., 2015), and virtual limbs (C, Collinger et al., 2013). Robotic prostheses (D, Collinger et al., 2013, E, Hochberg et al., 2012) include various types of robotic arms. Stimulation prostheses activate muscular tissue via electrical stimulation on the skin surface (F, Bouton et al., 2016), via intramuscular electrodes (G, Ajiboye et al., 2017), or via spinal electrodes (H, Capogrosso et al., 2016).
Virtual environments provide a cheap, customizable, and informative medium through which to test and use BMIs. By simulating realistic environments, humans and nonhuman primates have been able to control virtual arms and hands (Ajiboye et al., 2017; Irwin et al., 2017) with nothing but their brain activity, creating an expectation of capability when testing with real limbs. Additionally, as personal computers and smartphones have become prevalent in modern life, it becomes increasingly important for a person using a neuroprosthesis to be able to efficiently interact with those devices. Instead of controlling a prosthesis to then manipulate electronic devices, several groups have investigated skipping the middleman and decoding intended body movements directly into control signals for computers. Some of the first mainstream neuroprosthetic investigations provided nonhuman primates with the ability to control a computer cursor in multiple dimensions (Carmena et al., 2003; Serruya et al., 2002; Taylor et al., 2002). More recently, monkeys and humans have been able to type messages on keyboards at high character rates (Gilja et al., 2015; Nuyujukian, Kao, Ryu, & Shenoy, 2016; Pandarinath et al., 2017), generally by using imagined arm movements to control a cursor by way of the BMI. One study in particular enabled a person to control a tablet computer through several applications with an intracortical BMI, demonstrating the usefulness of brain-computer interfaces for people with paralysis (Nuyujukian et al., 2018). Restoring computer usage returns substantial independence to people with paralysis in terms of interacting with digital media. However, the inability to directly manipulate the real world may leave something to be desired for BMI-controlled computer users.

Despite the unnaturalness of the colloquial prosthesis, it does enable interaction with the real world, and robotic limbs are one step more natural. Often affiliated with people with amputations, modern robotic limbs can perform most of the functions of the natural hand in terms of motor control and sensation. Foundational studies demonstrated the ability to control various robotic arms just with the decoder outputs (Carmena et al., 2003; Chapin et al., 1999; Hochberg et al., 2012, 2006). Many recent studies have employed the Modular Prosthetic Limb (Johns Hopkins University, Applied Physics Laboratory, Baltimore, MD, USA), which supports both motor output (Collinger et al., 2013; Wodlinger et al., 2015) and sensory feedback (Flesher et al., 2016). There are a handful of other robotic hands that could fit brain-machine interface applications (LUKE Arm, Mobius Bionics, LLC, Manchester, New Hampshire, USA; i-Limb, Össur hf., Reykjavík, Iceland; TASKA Hand, TASKA Prosthetics, Riccarton, Christchurch, New
Zealand; BeBionic Hand, Ottobock Healthcare, Duderstadt, Germany; Ability Hand, Psyonic, Inc., Champaign, IL, USA), but require the attachment of arms for full mobility. A central motivator for the use of robotic upper extremities is that they do not require surgical intervention for repairs, making them lower risk prostheses to return some function to people with paralysis. However, the robotic limb, which is not necessarily lightweight, would require hefty mounting hardware for attachment to a wheelchair or even to the user if future rehabilitation technology enables independent mobility. The addition of a robotic limb while maintaining the natural limbs of the person is bulky, reducing maneuverability in smaller spaces and potentially making usage cumbersome. Further, embodiment of a robotic limb is a difficult balance for engineers to maintain. Accurate and precise motor control coupled with naturalistic sensory feedback that may result from continued research may be sufficient for most to embody a robotic limb, but the current state of robotic arms and hands do not match the capabilities of the natural upper extremity.

Consequently, several groups have investigated methods through which the natural limb can be reanimated for use, despite paralysis via the nervous system. Functional electrical stimulation (FES) is a technology that can reanimate paralyzed muscles by delivering electrical stimulation to a muscle or the nerve innervating a muscle. Coordinating contractions of several muscles via electrical stimulation can return function to any degree of freedom of the body. To date, several studies have demonstrated voluntary FES of the upper extremity using pattern recognition from residual functional muscle (Kilgore et al., 2008; Kilgore, Peckham, Thrope, Keith, & Gallaher-Stone, 1989; Memberg et al., 2014; Peckham, Mortimer, & Marsolais, 1980). More recently, groups have enabled nonhuman primates (Ethier et al., 2012; Moritz et al., 2008) and people with paralysis (Ajiboye et al., 2017; Bouton et al., 2016) to control their upper extremities using BMI-controlled FES. Recent work investigating simultaneous individuation of digits within the hand would work well with a BMI-controlled FES system, as the predictions of the finger movements could be used as a direct control signal for electrical stimulation (Aggarwal et al., 2008; Nason, Vaskov, Patil, & Chestek, 2019). However, as mentioned previously, there are a number of challenges with recording brain activity during neuromuscular stimulation regarding the large amplitude artifacts each stimulation pulse generates. Fortunately, several means for mitigating the effects of artifacts have been presented. Intramuscular electrodes could be used in place of surface electrodes as they generate smaller amplitude stimulation
artifacts in intracortical recordings (Young et al., 2018). Additionally, presence of the stimulation artifacts can often be predicted based on the status of the BMI, allowing artifact-free recording in between periods of stimulation. Otherwise, common-average referencing (Ludwig et al., 2009) or a linear regression reference (Young et al., 2018) have been shown to nearly completely eliminate the artifact in the intracortically recorded signals for a relatively small increase in computational complexity.

There are a number of prosthetic devices that can be controlled by BMIs, ranging from computers to implantable hardware reanimating natural limbs. While present research aims to bring these technologies to full-time usage, perhaps the next stage of BMI-controlled prostheses will enable simultaneous usage of multiple functional devices.

4. Emerging technologies

Existing technologies, if they were available in a wireless implantable form factor, can already begin to restore at least a few degrees of freedom of movement for at least a few years (Aflalo et al., 2015; Ajiboye et al., 2017; Bouton et al., 2016; Collinger et al., 2013; Hochberg et al., 2012, 2006; Pandarinath et al., 2017). This can be viewed as a promising start for the field when the first real-time control with visual feedback using multi-electrode arrays was only two decades ago. There is again a wide variety of early preclinical technologies that may lead to large performance increases within the next two decades, some of which are showcased in Fig. 5. There are opportunities for both integrated circuits and novel electrodes to enable order of magnitude improvements in brain machine interface performance. With the traditional neural signal circuit architecture described above, i.e., a low noise amplifier followed by a 16-bit ADC, followed by a method of spike extraction and often spike sorting, there are not a lot of remaining tools to improve the power consumption. Existing approaches are already fairly well optimized (Harrison et al., 2007). However, if you move outside of that architecture, for example by identifying low frequency correlates of spikes rather than spikes directly (Irwin et al., 2016; Lim et al., 2020; Nason et al., 2020) or by heavily optimizing sampling rates and communication bit rates (Even-Chen et al., 2020), order of magnitude reductions in power consumption are possible. Also, it is possible to use front ends with a very small number of transistors that may somewhat compromise signal quality in return for size, i.e., the “neural dust” approach (Seo et al., 2016), that could one day enable implanting many more channels. Specifically, prototype
Fig. 5  Emerging neural recording technologies. Novel small-scale wireless implantable electronic devices (left). (A) Ultrasonic neural dust (Seo et al., 2016). (B) Radio-frequency (RF) coupled Neurograins (Lee et al., 2019). (C) Magnetoelectric (ME) stimulators (Singer et al., 2020). (D) Infrared recording MOTEs (Lee et al., 2018). (E) Infrared recording probes (Lim et al., 2020). Novel electrode technologies (right). (F) High-density opto-electrophysiological silicon probes (Mendrela et al., 2018). (G) High-density oversampled silicon probes (Scholvin et al., 2016). (H) Neuropixels probe (Jun et al., 2017). (I) Carbon fiber microelectrode array (Patel et al., 2016).
devices exist using ultrasonic power (Seo et al., 2016), RF with optimized small antennas (Lee et al., 2019), magneto-electric power (Singer et al., 2020), and optical power (Lee et al., 2018; Lim et al., 2020).

In terms of electrodes, the Utah array was developed in the late 90s (Nordhausen et al., 1996). This is now in widespread use in humans, with 48 implants prior to 2018 (Bullard et al., 2020). The robustness of this design has made it difficult to improve upon, since whatever displaces it must have a lifetime measured in years (Chestek et al., 2011; Downey, Schwed, Chase, Schwartz, & Collinger, 2018; Simeral et al., 2011) and compete with many person years of existing safety data. However, electrodes in pre-clinical testing have promising characteristics that could make their way into clinical use after a period of hardening over the next decade. Channel count is now far higher in rodent experiments than NHP or humans due to devices such as Neuropixels with 384 channels (Jun et al., 2017), and other high density silicon probes (Mendrela et al., 2018; Scholvin et al., 2016; Yang, Lee, Villagracia, & Masmanidis, 2020). Similarly, there are a large number of novel electrode designs that are smaller than neurons, after early work noted that this can enable a drastic reduction in bleeding and scarring (Nguyen et al., 2014; Seymour & Kipke, 2007). Specifically, “ultrasmall” electrodes have been created using carbon fibers (Guitchounts, Markowitz, Liberti, & Gardner, 2013; Kozai et al., 2012; Patel et al., 2016), silicon carbide (Knaack et al., 2016), PDMS (Cohen-Karni, Timko, Weiss, & Lieber, 2009), nanoelectronic thread (Luan et al., 2017), neuron-like electronics (Yang et al., 2019), parylene (Rodger et al., 2008), thermal drawing (Canales et al., 2015), and high density wire bundles (Obaid et al., 2020). This could one day enable multielectrode arrays that achieve very high channel count despite less overall damage than the 100 channel devices in use today. One can imagine a future where neural interfaces become a first line treatment for some illnesses rather than a last resort.

5. Conclusions

Neuroprostheses have progressed from spastic and noisy cursor control in nonhuman primates to humans with paralysis naturally controlling their own paralyzed limbs and communicating with distant loved ones through a computer using brain-machine interfaces. This exciting field has exploded in the past two decades, funded by the Brain Initiative and commercial and industrial interest. Translation of neuroprostheses to full-time use with humans has many challenges ahead, but the openness
of the Food and Drug Administration (FDA) has enabled incredible preliminary accomplishments to validate neuroprostheses as a safe and beneficial therapy to people with paralysis. Continued investigation via FDA-approved human clinical trials may eventually transition neuroprostheses from the last option to a first treatment.

References


