Artificial speech synthesizer control by brain-computer interface

Jonathan S. Brumberg [1,2], *Philip R. Kennedy*[2], *Frank H. Guenther*[1,3]

[1] Department of Cognitive and Neural Systems, Boston University, USA
[2] Neural Signals Inc., Duluth, GA, USA
[3] Division of Health Sciences and Technology, Harvard University-Massachusetts Institute of Technology, USA

brumberg@cns.bu.edu, phlkennedy@neuralsignals.com, guenther@cns.bu.edu

Abstract

We developed and tested a brain-computer interface for control of an artificial speech synthesizer by an individual with near complete paralysis. This neural prosthesis for speech restoration is currently capable of predicting vowel formant frequencies based on neural activity recorded from an intracortical microelectrode implanted in the left hemisphere speech motor cortex. Using instantaneous auditory feedback (< 50 ms) of predicted formant frequencies, the study participant has been able to correctly perform a vowel production task at a maximum rate of 80-90% correct.

Index Terms: speech synthesis, brain computer interface

1. Introduction

Artificial speech synthesizers, for the most part, have been used primarily as a theoretical research tool for investigating normal speech production in humans. Notable exceptions include devices to replace glottal pulse activity for laryngectomy patients and text-to-speech synthesizers incorporated into augmentative and alternative communication (AAC) devices. Recent progress in the field of brain computer interfacing (BCI) has broadened the user base for such AAC devices to include those with extreme, near complete, paralysis. In particular, methods involving electromyographic (EMG) [1] and electroencephalographic (EEG) [2-6] potentials are bridging the gap between individuals with near-total paralysis and the external, communicating world. Both types of communication restoration systems have been used to decode and predict intended discrete letter and word choices via a brain computer interface. However, these systems are often bulky (e.g. EEG) and slow, often requiring minutes to produce a single word. These production rates make such devices completely unsuitable for real-time fluent speech communication between a BCI user and another party.

In response to this unfilled need for fast artificial speech production, our research group has developed a real-time artificial vocal tract prosthesis for paralyzed individuals lacking speech production capabilities [7-9]. The research described within this report introduces the first ever neural prosthesis for speech restoration. It utilizes a wireless intracortical microelectrode [9,10] to obtain neural activity related to speech production, maps it into the speech domain through a "neural decoder" and synthesizes instantaneous auditory feedback in less than 50 ms. Not only does our speech prosthesis provide real-time acoustic output for fluent speech production, but the continuous feedback allows the user to perform online error correction for mispronounced speech sounds. This type of control is far more natural than keyboard selection and "backspace" deletion of erroneous predictions characteristic of alternative neural interfacing AAC devices.

The details of the decoder, subject and implant will be discussed in Section 2. Preliminary results from our first speech prosthesis prototype will be presented in Section 3. Section 4 will include a brief discussion of the results along with future research directions made apparent from our initial investigation. We finish with a concluding summary in Section 5.

2. Subjects and Methods

Speech restoration by wireless intracortical microelectrode BCI requires implantation of an extracellular recording electrode and electrical hardware for amplification and transmission brain activity. For our methods, the choice of implantation site, electrode type and decoding algorithms are crucially important for the success of the neural prosthesis. The complete neural prosthesis is illustrated in Figure 1 including implant location, wireless telemetry subsystems, signal processing, neural decoding and artificial speech synthesis modules.

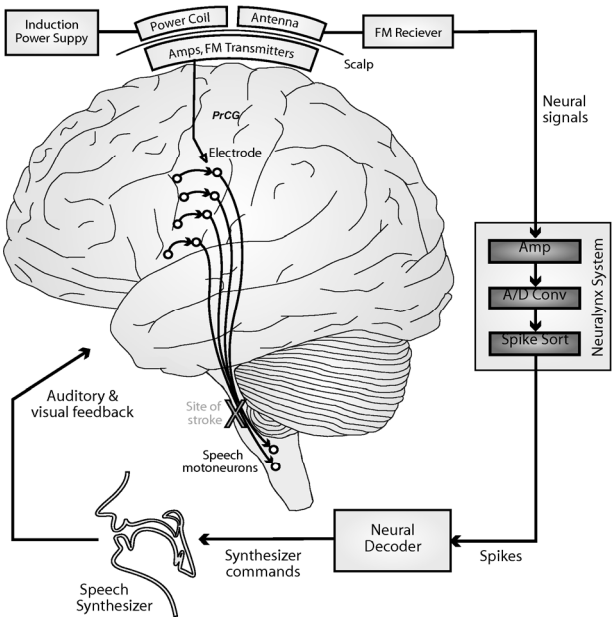

Figure 1. *A schematic of the neural prosthesis indicating the stroke location (labeled with an "X"), electrode implantation site, wireless telemetry, signal*

processing, neural decoding and artificial speech synthesis.

2.1. Subject

The participant is a 25 year old male suffering from locked-in syndrome as a result of a brain stem stroke incurred at age 16. Locked-in syndrome is characterized by near-complete motor paralysis with only small amounts of voluntary control over the ocular muscles, though the remainder of the brain responsible for cognition and sensation is left largely intact. Indeed, our subject retains voluntary motor behavior only through slow vertical movements of the eyes, allowing him to answer yes/no questions. His audition and somatic sensation were not noticeably impaired though his visual perception has suffered due to an inability to control the eyes in a conjugate fashion.

The implant location was determined according to the areas of highest blood oxygen level dependent (BOLD) response during a preoperative functional magnetic resonance imaging (fMRI) investigation of attempted speech production. Large-scale activation of the speech production network was observed with a peak activity on the left ventral precentral gyrus. This area has been identified as a region on the border of the premotor and primary motor cortex for the speech articulators (i.e. lips, tongue and jaw movements) and will hereafter be referred to as the speech motor cortex.

The subject was implanted over 4 years ago at age 21, approximately 5 years after becoming locked-in due to brain stem stroke. The implantation procedure was approved by the Food and Drug Administration (IDE G960032), Neural Signals, Inc. Institutional Review Board, and Gwinnett Medical Center Institutional Review Board. Informed consent was obtained from the participant and his legal guardian prior to implantation. Further details concerning the implantation procedure can be found elsewhere [9].

2.2. Implant

The implanted electronics of the neural prosthesis consist of a recording electrode and amplification and wireless telemetry systems. The recording electrode used in the device is a chronically implantable two-channel Neurotrophic Electrode [9,10]. This electrode is unique among multiunit extracellular electrodes.

Standard multiunit extracellular electrodes come in an array (multielectrode array, MEA) configuration with many tens of penetrating tips which are capable of recording one to two neural action potentials, or spikes, per tip [11-14]. These electrode arrays are implanted on the surface of the cerebral cortex, where they float while attached to amplification and transmission (most utilize wired connections while our implant is the only wireless system) hardware. One major design issue with this type of microelectrode configuration is that the array tends to move with respect to the brain over time. Such movement leads to variable spike amplitudes due to changing distances between the electrode tip and originating neuron. In addition, the movement often leads to scar tissue formation around the electrode tip, known as gliosis, which further degrades the recorded neural signal.

The Neurotrophic Electrode is a microwire style electrode with recording wires placed within a 1 mm long glass cone filled with a neurotrophic growth factor [10]. The growth factor promotes neurite growth from nearby neurons (as far away as 600 μm) allowing simultaneous recording of multiple neural sources [10,15,16]. Additionally, the neurite ingrowth creates a bridge of neural tissue, effectively anchoring the recording electrodes within the cortical surface and eliminating both gliosis and signal recording artifacts resulting from electrode micromovements. Finally, the Neurotrophic Electrode is the only chronic microelectrode to be implanted for over four years in human subjects (approximately 4.5 years as of this writing).

2.3. Decoding Algorithm

We chose to decode neural signals into speech representations according to a discrete-time filtering method as opposed to discrete classification of intended phoneme or word productions (e.g. EMG and EEG methodologies). According to previous neuroimaging and computational modeling studies of speech production [17,18], the speech premotor cortex was found to represent acoustic features of intended speech sounds. Therefore, we utilized formant frequencies, the resonant frequencies of the vocal tract, as a continuously variable parameter used for control of a formant-based artificial speech synthesizer [19].

Specifically, a Kalman filter [20] was trained according to a "center-out and back" vowel-to-vowel calibration sequence in a two-dimensional formant frequency space acoustically presented to the subject. Our center-out task is taken from the seminal work by Georgopoulos and colleagues which investigated the tuning properties of motor cortical neurons to arm and hand kinematics (e.g. hand movement velocities) [21]. During the calibration sequence, the subject was instructed to attempt to speak along with the stimulus as it was produced by a computer. A visual representation of the 2-D formant frequency plane is shown in Figure 2 (right). The UH (hut) sound was chosen as the "center" point with the remaining vowels IY (heat), A (hot) and OO (hoot) as the periphery with periphery target regions approximately 150 Hz in F1 and 300 Hz in F2. The sequence consisted of eight center-out repetitions of each peripheral vowel with steady vowel periods lasting 1 s. and transition periods lasting 300 ms. for a total sequence length of 63.4 seconds. The calibration signal was artificially generated utilizing formant frequency trajectories between the center sound and each periphery target. We used an offline training algorithm to determine the model parameters based on the relationship between the calibration signal formant frequencies and their velocities (1^{st} time derivative) and the observed neural firing rates.

The Kalman filter model parameters were saved for real-time usage similar to previously reported studies involving online prediction of arm and hand kinematics [22]. The Kalman filter decoder was then applied in real-time to convert neural firing rates, sampled every 15 ms, into formant frequencies for instantaneous speech synthesis. The total signal acquisition to synthesizer output processing time was less than 50 ms. This system delay is comparable to the natural delay from speech production to self perception in healthy subjects. Slower feedback delay times (> 200 ms) have been shown to disrupt fluent speech in normal subjects [23] and we would expect similar disfluencies with any speech prosthesis user subject to extremely delayed speech feedback.

3. Results

We tested the decoder performance in an online neural decoding paradigm. Again, vowel sequences were artificially generated according to formant frequency trajectories between

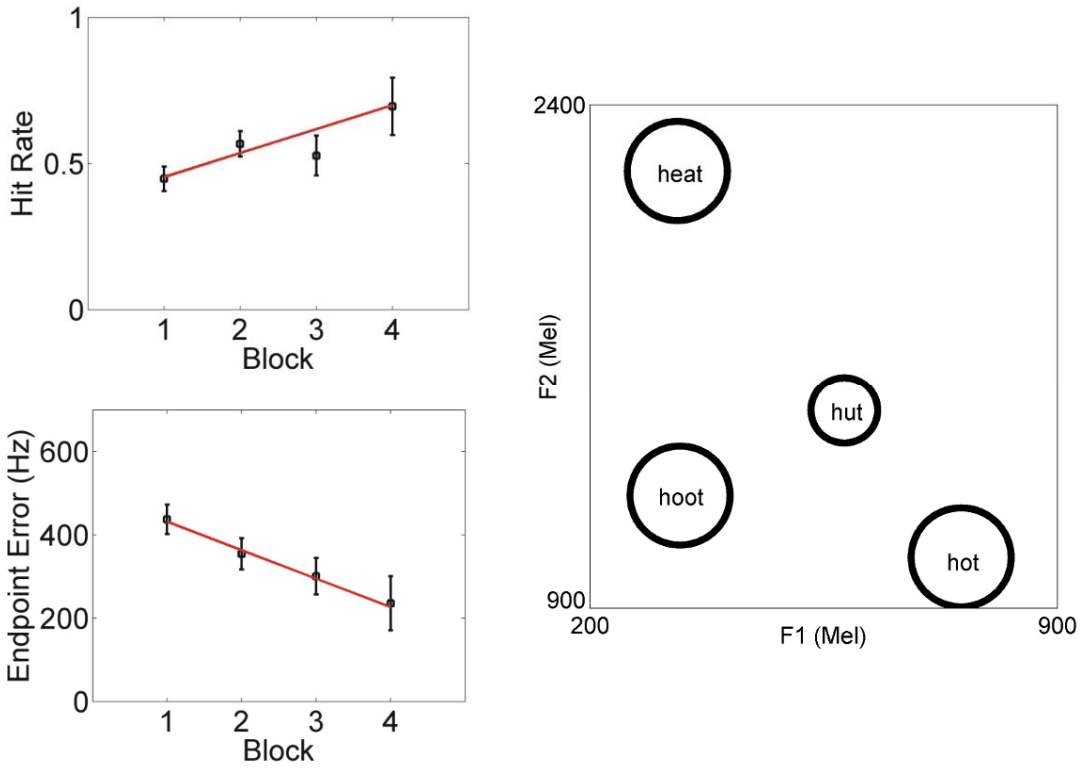

Figure 2. *(Left) Endpoint vowel accuracy, (top) hit rate, and (bottom) Euclidean distance in the F1-F2 plane from target in Hz. (Right) 2-D formant plane with example vowel targets and associated classification region.*

a "center" sound (UH) and one of three periphery vowels (A, IY or OO). The subject was instructed to listen to a randomly selected two vowel sequence (e.g. UH (hut) – IY (heat)) and to repeat the sound after a GO instruction. The trial time was limited to six seconds and ended in success when the subject produced the endpoint vowel sound within the specified formant frequency target region. A trial was marked as a failure if the subject could not produce the endpoint vowel sound and hold it within the target region for 500 ms. The subject participated in 25 sessions over a 5 month period in which he performed between 5-34 trials per session. Each online testing session was divided into 1-4 trial blocks with short rest times between blocks.

Two error measures were computed for speech production accuracy. First, we determined the average target acquisition rate for each trial block as shown in Figure 2 (top left). Second, we computed the average endpoint error in terms of Euclidean distance from the last predicted formant frequency to the target center as shown in Figure 2 (bottom left). Both measures showed statistically significant improvement ($p<0.05$; t-test with null hypothesis of zero slope as a function of block number) within each testing session. This increase in performance indicated that the subject successfully learned to control the 2-D neural decoder for artificial speech synthesis.

4. Discussion

The neural prosthesis described within this report is the first ever brain computer interface for continuous control over an artificial speech synthesizer. Our results establish the feasibility for continuous brain control for speech restoration in real-time by profoundly paralyzed and mute individuals. Our subject was capable of producing vowel sounds at average rates approaching 80% within each testing session (approximately two hours), with a maximum of 89% by the end of the final block of the final session.

The results presented here also demonstrate that neural signals recorded from the speech motor cortex can be used in a formant frequency prediction paradigm. Such a conclusion implies that activity in this region of cortex is modulated by the production of sounds represented by formant frequencies. This finding is very important for the advancement of our understanding of speech production in humans as it directly identifies both a neural correlate of speech production as opposed to indirect methods such as fMRI investigations.

Despite the success of our initial testing of the neural speech prosthesis, a number of outstanding issues remain for further study. First, the system, as currently implemented, is capable only of vowel prediction and vowel synthesis as a direct result of formant frequency decoding and synthesis. Therefore, our next step is to extend our neural prosthesis to account for other speech sound types (e.g. consonants). We expect this modification to be straightforward as formant frequencies are strongly related to speech articulator configurations [24].

Additionally, the speech prosthesis does not yet contain any mechanism for speech onset and termination control (i.e. phonation). An investigation into the possibility of phonation control by the neural decoder is already underway and is planned to be incorporated into the next major device version. We expect both the phonation and articulatory control paradigms will greatly increase the user's ability to quickly and accurately control an artificial speech synthesizer.

5. Conclusions

We reported on the development and performance of the first neural prosthesis for artificial speech production controlled by a paralyzed individual suffering from locked-in syndrome. The neural decoder was described as an intracortical microelectrode device with wireless transmission of recorded signals which performed a mapping between neural firing rates in the speech motor cortex and intended speech utterances. Predicted formant frequencies were synthesized in real-time with a total system delay less than 50 ms. for instantaneous auditory feedback and closed-loop BCI control. The subject in our study learned to control the speech synthesizer using the brain-computer interface by the end of each testing session.

6. Acknowledgements

Supported by the National Institute on Deafness and other Communication Disorders (R01 DC07683; R44 DC007050-02) and the National Science Foundation (SBE-0354378). We would like to thank the participant and his family for their participation in our investigations.

7. References

[1] E.J. Wright, S.A. Siebert, P.R. Kennedy, and J.L. Bartels, "Novel method for obtaining electric bio-signals for computer interfacing," *Neuroscience Meeting Planner 2008*, Wash: 2008.

[2] D.J. Krusienski, E.W. Sellers, F. Cabestaing, S. Bayoudh, D.J. McFarland, T.M. Vaughan, and J.R. Wolpaw, "A comparison of classification techniques for the P300 Speller," *Journal of Neural Engineering*, vol. 3, 2006, pp. 299-305.

[3] D.J. Krusienski, E.W. Sellers, D.J. McFarland, T.M. Vaughan, and J.R. Wolpaw, "Toward enhanced P300 speller performance," *Journal of Neuroscience Methods*, vol. 167, Jan. 2008, pp. 15-21.

[4] E.W. Sellers, D.J. Krusienski, D.J. McFarland, T.M. Vaughan, and J.R. Wolpaw, "A P300 event-related potential brain-computer interface (BCI): The effects of matrix size and inter stimulus interval on performance," *Biological Psychology*, vol. 73, Oct. 2006, pp. 242-252.

[5] F. Nijboer, E. Sellers, J. Mellinger, M. Jordan, T. Matuz, A. Furdea, S. Halder, U. Mochty, D. Krusienski, T. Vaughan, J. Wolpaw, N. Birbaumer, and A. Kübler, "A P300-based brain-computer interface for people with amyotrophic lateral sclerosis," *Clinical Neurophysiology*, vol. 119, Aug. 2008, pp. 1909-1916.

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

[7] J.S. Brumberg, A. Nieto-Castanon, F.H. Guenther, J.L. Bartels, E.J. Wright, S.A. Siebert, D.S. Andreasen, and P.R. Kennedy, "Methods for construction of a long-term human brain machine interface with the Neurotrophic Electrode," *Neuroscience Meeting Planner 2007*, Wachington, DC: 2008.

[8] F.H. Guenther, J.S. Brumberg, A. Nieto-Castanon, J.L. Bartels, S.A. Siebert, E.J. Wright, J.A. Tourville, D.S. Andreasen, and P.R. Kennedy, "A brain-computer interface for real-time speech synthesis by a locked-in individual implanted with a Neurotrophic Electrode," *Neuroscience Meeting Planner 2008*, Washington, DC: 2008.

[9] J.L. Bartels, D. Andreasen, P. Ehirim, H. Mao, S. Seibert, E.J. Wright, and P.R. Kennedy, "Neurotrophic electrode: Method of assembly and implantation into human motor speech cortex," *Journal of Neuroscience Methods*, vol. 174, Sep. 2008, pp. 168-176.

[10] P.R. Kennedy, "The cone electrode: a long-term electrode that records from neurites grown onto its recording surface," *Journal of Neuroscience Methods*, vol. 29, 1989, p. 181—193.

[11] A. Hoogerwerf and K. Wise, "A three-dimensional microelectrode array for chronic neural recording," *Biomedical Engineering, IEEE Transactions on*, vol. 41, 1994, pp. 1136-1146.

[12] K. Jones, P. Campbell, and R. Normann, "A glass/silicon composite intracortical electrode array," *Annals of Biomedical Engineering*, vol. 20, Jul. 1992, pp. 423-437.

[13] E.M. Maynard, C.T. Nordhausen, and R.A. Normann, "The Utah Intracortical Electrode Array: A recording structure for potential brain-computer interfaces," *Electroencephalography and Clinical Neurophysiology*, vol. 102, Mar. 1997, pp. 228-239.

[14] P.J. Rousche and R.A. Normann, "Chronic recording capability of the Utah Intracortical Electrode Array in cat sensory cortex," *Journal of Neuroscience Methods*, vol. 82, Jul. 1998, pp. 1-15.

[15] P.R. Kennedy, R.A.E. Bakay, and S.M. Sharpe, "Behavioral correlates of action potentials recorded chronically inside the Cone Electrode," *NeuroReport*, vol. 3, 1992, p. 605—608.

[16] P.R. Kennedy, S.S. Mirra, and R.A.E. Bakay, "The cone electrode: ultrastructural studies following long-term recording in rat and monkey cortex," *Neuroscience Letters*, vol. 142, 1992, p. 89—94.

[17] F.H. Guenther, M. Hampson, and D. Johnson, "A theoretical investigation of reference frames for the planning of speech movements," *Psychological Review*, vol. 105, 1998, p. 611—633.

[18] F.H. Guenther, S.S. Ghosh, and J.A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain and Language*, vol. 96, 2006, p. 280—301.

[19] D.H. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, 1980, p. 971—995.

[20] R.E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, 1960, p. 35—45.

[21] A.P. Georgopoulos, J.F. Kalaska, R. Caminiti, and J.T. Massey, "On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex," *Journal of Neuroscience*, vol. 2, 1982, p. 1527—1537.

[22] S. Kim, J.D. Simeral, L.R. Hochberg, J.P. Donoghue, G.M. Friehs, and M.J. Black, "Multi-state decoding of point-and-click control signals from motor cortical activity in a human with tetraplegia," *Neural Engineering, 2007. CNE'07. 3rd International IEEE/EMBS Conference*, 2007, p. 486—489.

[23] D.G. MacKay, "Metamorphosis of a Critical Interval: Age-Linked Changes in the Delay in Auditory Feedback that Produces Maximal Disruption of Speech," *The Journal of the Acoustical Society of America*, vol. 43, Apr. 1968, pp. 811-821.

[24] K.N. Stevens, *Acoustic Phonetics*, MIT Press, 2000.