Surface electromyographic control of speech synthesis

Meredith J. Cler, Alfonso Nieto-Castanon, Frank H. Guenther, Cara E. Stepp, *Member, IEEE*

Abstract— **Individuals with very high spinal cord injuries (e.g. C1-C3) may be ventilator-dependent and therefore unable to support speech breathing. However, their facial musculature is intact, given that these muscles are innervated by cranial nerves. We developed a system using surface electromyography (sEMG) recorded from facial muscles to control a phonemic interface and voice synthesizer and tested the system in healthy individuals. Users were able to use five facial gestures to control an onscreen cursor and the phonemic interface. Users had mean information transfer rates (ITRs) of 59.5 bits/min when calculating ITRs using the number of phonemes selected. To compare with orthographic systems, ITRs were also calculated using the equivalent number of letters required to spell the selected word. With this calculation, users had a mean ITR of 70.1. Results are promising for further development and testing in individuals with high spinal cord injuries.**

I. Introduction

Of the estimated 273,000 individuals in the U.S. who have spinal cord injury (SCI) [1], 54.1% have cervical injuries (e.g. C1-C7). These individuals are therefore unable to use their arms and hands to use a mouse and keyboard [2]; furthermore, those who have very high cervical injuries (e.g. C1-C3) may be ventilator dependent and unable to produce natural speech. The only communication modalities available clinically for these individuals with very high SCI are eye-tracking devices and mouth sticks. Brain controlled devices using electroencephalography (EEG) are also accessible, but are very slow and unreliable for practical use [3].

However, individuals with even very high SCI have unimpaired facial musculature, because these muscles are innervated by cranial nerves. Here we describe a system that leverages this spared muscle function. We tested the ability of healthy individuals to use facial surface electromyography (sEMG) to control a cursor in order to select phonemes that were then synthesized with a concatenative speech synthesizer. sEMG requires intact muscle control, but has a much higher signal to noise ratio than EEG [4].

Research supported by CELEST, an NSF Science of Learning Center (SBE-0354378) and NIDCD grants DC002852 and DC012651.

M.J. Cler is with the Graduate Program for Neuroscience-Computational Neuroscience, Boston University, Boston, MA 02215 USA (e-mail: mcler@bu.edu).

A. Nieto-Castanon is with the Department of Speech, Language, and Hearing Sciences, Boston University, Boston, MA 02215 USA (e-mail: alfnie@gmail.com).

F.H. Guenther is with the Departments of Speech, Language, and Hearing Sciences and Biomedical Engineering, Boston University, Boston, MA 02215 USA (e-mail: guenther@bu.edu).

C.E. Stepp is with the Departments of Speech, Language, and Hearing Sciences and Biomedical Engineering, Boston University, Boston, MA 02215 USA (phone: 617-353-7487; fax: 617-353-5074; e-mail: cstepp@bu.edu)

II. Methods

A. Participants

Six healthy adults who reported no history of speech, language, or hearing disorders and were native speakers of American English participated. Participants had not engaged in previous sEMG research, nor were they familiar with phonemic spellers. The participants (3 male) had a mean age of 20.7 years (SD = 0.9). All participants completed written consent in compliance with the Boston University Institutional Review Board.

B. Experimental Design

All participants had one 90-minute training session. After the participant's skin was cleaned with alcohol and then exfoliated with adhesive tape, five single differential sEMG sensors were placed on the surface of the skin. Electrodes were placed parallel to underlying muscle fibers (see Fig. 1 and Table I). Electrodes were placed to record activity of muscles that were activated during particular facial gestures. The sEMG signal was then mapped to a cursor movement (see Table I).

Participants were all able isolate these facial gestures. Gestures were chosen to correspond with the movement of the cursor: when the users contracted muscles at the top of their face, the cursor moved up, and when they contracted muscles in the left cheek and mouth, the cursor moved left.

Figure 1. Electrode placement

Users were given a three-minute introduction and free use of the phonemic interface using a standard mouse, followed by a sequence of calibrations for the sEMG mouse, and 45 trials of interaction with the interface and the sEMG mouse. Each trial began with the presentation of one of 45

common American English words. Stimuli were presented both aurally and visually; auditory stimuli were generated with the interface offline and the visual stimuli was a phonemic spelling of the word (e.g. if the stimulus was "voice", a participant heard the word synthesized by the interface and saw the word spelled phonemically: "v-oi-s"). After the stimulus was presented, users selected the given phonemes with the phonemic speller, using their sEMG signals to control the cursor. When participants finished selecting the phonemes, the selected phonemes were synthesized, giving auditory feedback to the user.

TABLE I. ELECTRODE PLACEMENT

Electrode Number	Electrode Placement	Muscle Group	Cursor Action
1	Left of mouth	Risorius and orbicularis oris	Move left
2	Right of mouth	Risorius and orbicularis oris	Move right
3	Above eyebrow	Frontalis	Move up
4	Chin	Mentalis	Move down
5	Side and slightly below eye	Orbicularis oculi	Click

C. Phonemic Speller and Speech Synthesizer

The speech synthesizer used in this study was originally developed for use on a touch screen computer by sliding a finger between phoneme "keys" that are laid out in a circular keyboard layout based roughly on articulatory features (constriction locations and degrees). The synthesizer was configured to start a new word each time the finger is placed on the screen at the location of a phoneme key. This phoneme acts the first phoneme in the word, and the remaining phonemes of the word are indicated by sliding the finger (without lifting it) to the subsequent phoneme keys. After all phonemes in the word are indicated in this manner, the finger is lifted off the tablet, at which point the word is synthesized via a proprietary concatenative synthesis process and played over the computer's speaker. The interface consisted of fourteen vowels and diphthongs, surrounded by twenty-four consonants.

In the current study, touch screen control was replaced with an interface specifically designed for sEMG control, in which the user clicked on each phoneme individually and then clicked another button when ready for the program to synthesize the series of phonemes selected.

D. Data Acquisition

sEMG signals were pre-amplified and filtered with Bagnoli-2 EMG systems (Delsys, Boston, MA), which were set to a gain of 1000 and included a band-pass filter with roll-off frequencies of 20 and 450 Hz. Simultaneous sEMG signals were digitally recorded with National Instruments hardware and custom Python software at 5000 Hz. The data was collected with a window size of 100 ms, over which the root mean square (RMS) was calculated.

E. Calibration

After users became familiar with the phonemic interface, the sEMG mouse system was calibrated. Users were asked to make each facial gesture twice (i.e., left left; right right; up up; down down; blink blink), until they were able to produce a clean calibration run with minimal muscle coactivation. The maximum RMS from each electrode during the calibration run was then used to calculate thresholds using a set of electrode-specific multipliers, determined in pilot testing. The multipliers were 0.3 for the left, right, and up electrodes, 0.5 for the down electrode, and 0.7 for the blink electrode, and represented how much activation was required for the system to recognize muscle activation as indicating a deliberate gesture. For example, the participants were required to produce an activation that was at least 70% of the maximum blink RMS from their individual calibration in order for the system to register a blink.

F. sEMG Mouse

The custom sEMG mouse was written in Python and allowed the user to move in any 360º direction by using facial gestures, in isolation or in combination.

The RMS of the sEMG signals was calculated from each electrode every 100 ms and checked against thresholds. If the RMS from the blink electrode was higher than the blink threshold, the cursor was clicked to select a phoneme. Otherwise, the movement of the cursor was calculated with (1) and (2). The RMS values from the left, right, up, and down electrodes (RMS_R, RMS_L, RMS_U, RMS_D) were divided by thresholds from calibrations. This resultant ratio was then squared and used as the magnitude of the cursor movement in those four directions [5]. To convert these magnitudes to a change in x and y cursor position, the left movement was subtracted from the right and the up was subtracted from down, and these x and y direction values were multiplied by a scalar that was identical for all users (speed in (1) and (2)).

$$\Delta x = [(RMS_R/threshold_R)^2 - (RMS_L/threshold_L)^2] \times speed \quad (1)$$

$$\Delta y = [(RMS_D/threshold_D)^2 - (RMS_U/threshold_U)^2] \times speed \quad (2)$$

G. Performance Measure

In this experiment, performance was measured using information transfer rate (ITR). This value, calculated using Wolpaw's method, represents both speed and accuracy [6]. ITRs were calculated for each trial using (3); in this equation, N is 38, the number of phonemic targets on the screen, and A is accuracy from 0 to 1. The output of this equation is bits per selection, which must then be converted to bits per minute by multiplying (3) by the selection rate in selections per minute.

$$bits / selection = \log_2(N) + A \times \log_2(A) + (1 - A) \times \log_2((1 - A) / (N - 1)) \quad (3)$$

In this experiment, selection rate, in selections per minute was calculated in two ways. First, we calculated the

number of selections, ranging from three to six phonemes, divided by the time the user took to make those selections. In order to benchmark the system to equivalent orthographic methods, we then recalculated selections per minute as the number of letters in the orthographical spelling of the word, divided by the time it took the user to spell the word phonemically.

III. Results

Participants had mean phonemic ITRs between 46.8 and 74.1 bits/min with a mean of 59.5 (SD = 9.8) (see Fig. 2 and Table II). In order to compare this system with orthographic systems, ITRs were also calculated with the equivalent number of selections required to spell the word orthographically. With this calculation, participants had mean ITRs ranging from 56.8 to 88.3 bits/min with a mean of 70.1 bits/min (SD = 12.3).

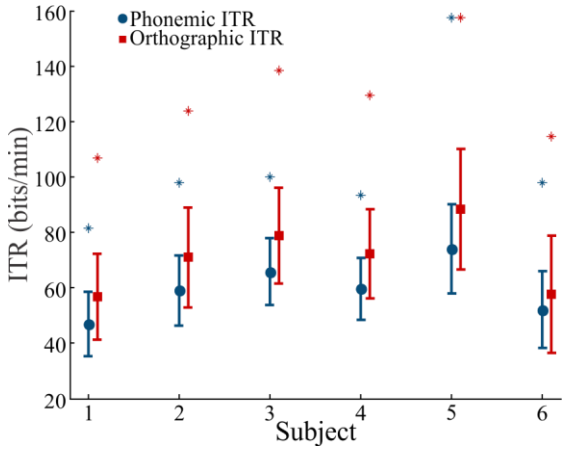

Figure 2. Mean information transfer rate measured in bits per minute. Error bars are 95% confidence intervals of the mean. Blue data show ITR calculated with direct phonemic selections per minute. Red data show ITR calculated with equivalent orthographic selections per minute. Asterisks show maximum trial ITR.

TABLE II. Comparisons to Other Systems

System	ITR Range (bits/min)	Example References
Eye-tracking (includes predictive methods)	60-222	[7-10]
Mechanical switch (e.g. mouth stick; includes predictive methods)	96-198	[8]
Other sEMG systems (continuous muscle control)	5.4-51	[5, 11-13]
Invasive BCIs	5.4-69	[14, 15]
Non-invasive BCIs	1.8-24	[15-18]

IV. Discussion

As anticipated, participants had mean ITRs that were much higher than non-invasive BCIs and other sEMG systems (see Table II). ITRs were also higher than or equivalent to invasive BCIs. ITRs in this study were comparable to many eye-tracking and mouth stick methods, including those that use prediction. sEMG may offer advantages over eye-tracking systems. For example, performance using eye-tracking systems is degraded with any head movement or changes in ambient lighting, and their use may cause fatigue [8].

A. Benefits and Limitations of Phonemic Interface

Most similar assistive devices use orthographic spellers rather than the phonemic interface presented here. Using phonemes rather than letters has the potential to allow greater personality and expression as well as different accents. Use of phonemic output could always be combined with pre-packaged speech to text software (e.g., Dragon) to compose text messages. However, using an orthographic speller directly requires user literacy, which can be problematic in individuals with severe motor impairment [19]. Conversely, by allowing users to combine sequences of phonemes, novel conversational items can be generated without that user having knowledge of orthographic spelling.

Further, many words have fewer phonemes than letters, which reduces the number of selections that must be made and therefore the time it takes to communicate the same amount of information (e.g. VOICE becomes V-OI-S). The stimuli in this study were all common English five-letter words. When converted to the phoneme system used in this study, the number of required selections ranged from 3 to 6 with a mean of 4.3 selections per word (SD = 0.8). Fifty-seven percent of the words took fewer than five selections to spell phonetically, and 98% took fewer than six selections.

The phonemes used in this interface had an organization based roughly on articulatory features, but finding the phonemes was not nearly as quick as one would expect finding letters in an alphabetic layout or the common QWERTY keyboard layout would be. Just as individuals must learn where all the letters are on a QWERTY keyboard when they learn to type, so would users of these systems. In addition, participants in this study were not required to translate their desired word into a phonemic spelling, but were rather prompted with a sequence of phonemes to select. Extra training time would also be expected in order for users to fluently map their intended words to the phonemes used in this interface.

B. Future Improvements

In the future, ITRs from this system could be improved by a variety of methods. First, as noted above, end-users of this system will require training in order to maximize their ITRs. Other studies show that training can increase ITRs by almost fifty percent [9]. Training on this system would include motor learning to use the sEMG mouse, becoming more familiar with the placement of the phonemes within the interface, and learning to translate their intended words into the set of 38 phonemes available in this interface (not required for participants of this study). Further study is needed to determine optimal training protocols and the potential effects of longer such interaction on user fatigue.

Additionally, the described sEMG mouse has a relatively simple algorithm and relies on the users to learn to control

their facial gestures. For users with motor disorders, a more sophisticated algorithm may be required. Ideally, users would be asked to make any facial gesture that they felt intuitively corresponded to left, right, up, down, and click actions. Then a machine-learning algorithm would be trained to recognize these gestures and translate them to the appropriate cursor actions.

Finally, many similar systems employ some kind of predictive models. The most basic way to include prediction with this system would be to use a phonemic dictionary and predict which phonemes are most likely to be chosen, given last selected phoneme. Then the predicted phoneme or set of phonemes could be highlighted or displayed on the screen separately, much like cellular phone text prediction. A machine-learning algorithm could also be implemented here to use the end-user's history rather than a dictionary. Adding even a basic predictive model could improve the final ITRs by as much as 100% [20, 21].

V. Conclusions

Individuals with high spinal cord injuries may require human-machine interfaces to communicate. Individuals with SCI have intact facial musculature and control, as the required muscles are innervated by cranial nerves. In this paper, we presented a system using surface electromyography recorded from facial muscles to control a phonemic interface and voice synthesizer.

Healthy participants were able to interact with the phonemic interface by using five facial gestures to control an onscreen cursor. To calculate information transfer rate, we used Wolpaw's method and found that users had mean ITRs of 59.5 bits/min when calculating each selected phoneme as one selection. ITRs were also calculated using the equivalent number of letters needed to spell the word orthographically as the number of selections made. With this method, users had a mean ITR of 70.1 bits/min. ITRs calculated both ways were higher than non-invasive BCIs, other sEMG systems, and invasive BCIs. ITRs were comparable to eye-tracking systems that use prediction. Future development to add predictive methods to this system are anticipated to increase ITRs significantly.

Acknowledgment

The authors would like to thank Lenny Varghese for his assistance with the sEMG mouse.

References

[1] "Spinal Cord Injury Facts and Figures at a Glance," *The National SCI Statistical Center,* February 2013.

[2] A. B. Jackson, M. Dijkers, M. J. Devivo, and R. B. Poczatek, "A demographic profile of new traumatic spinal cord injuries: change and stability over 30 years," *Arch Phys Med Rehabil,* vol. 85, pp. 1740-8, Nov 2004.

[3] O. Tonet, M. Marinelli, L. Citi, P. M. Rossini, L. Rossini, G. Megali, and P. Dario, "Defining brain–machine interface applications by matching interface performance with device requirements," *J Neurosci Methods,* vol. 167, pp. 91-104, 2008.

[4] C. E. Stepp, "Surface electromyography for speech and swallowing systems: measurement, analysis, and interpretation," *J Speech Lang Hear Res,* vol. 55, pp. 1232-46, Aug 2012.

[5] M. R. Williams and R. F. Kirsch, "Evaluation of head orientation and neck muscle EMG signals as command inputs to a human-computer interface for individuals with high tetraplegia," *IEEE Trans Neural Syst Rehabil Eng,* vol. 16, pp. 485-96, Oct 2008.

[6] J. R. Wolpaw, N. Birbaumer, W. J. Heetderks, D. J. McFarland, P. H. Peckham, G. Schalk, E. Donchin, L. A. Quatrano, C. J. Robinson, and T. M. Vaughan, "Brain-computer interface technology: a review of the first international meeting," *IEEE Trans Rehabil Eng,* vol. 8, pp. 164-73, Jun 2000.

[7] L. A. Frey, K. P. White, and T. E. Hutchinson, "Eye-Gaze Word Processing," *IEEE Trans Systems Man Cybernetics,* vol. 20, pp. 944-950, 1990.

[8] D. J. Higginbotham, H. Shane, S. Russell, and K. Caves, "Access to AAC: present, past, and future," *Augment Altern Commun,* vol. 23, pp. 243-57, Sep 2007.

[9] S. S. Liu, A. Rawicz, S. Rezaei, T. Ma, C. Zhang, K. Lin, and E. Wu, "An Eye-Gaze Tracking and Human Computer Interface System for People with ALS and Other Locked-in Diseases," *J Med Biol Eng,* vol. 32, pp. 111-116, 2012.

[10] P. Majaranta, I. S. MacKenzie, A. Aula, and K. Raiha, "Effects of feedback and dwell time on eye typing speed and accuracy," *Univ Access Inf Soc,* vol. 5, pp. 199-208, 2006.

[11] E. Larson, H. P. Terry, M. M. Canevari, and C. E. Stepp, "Categorical vowel perception enhances the effectiveness and generalization of auditory feedback in human-machine-interfaces," *PLoS One,* vol. 8, p. e59860, 2013.

[12] S. Vernon and S. S. Joshi, "Brain-muscle-computer interface: mobile-phone prototype development and testing," *IEEE Trans Inf Technol Biomed,* vol. 15, pp. 531-8, Jul 2011.

[13] C. Choi, B. C. Rim, and J. Kim, "Development and Evaluation of a Assistive Computer Interface by sEMG for Individuals with Spinal Cord Injuries," presented at the IEEE International Conference on Rehabilitation Robotics, Zurich, 2011.

[14] P. Brunner, A. L. Ritaccio, J. F. Emrich, H. Bischof, and G. Schalk, "Rapid Communication with a "P300" Matrix Speller Using Electrocorticographic Signals (ECoG)," *Front Neurosci,* vol. 5, p. 5, 2011.

[15] N. J. Hill, T. N. Lal, M. Schroder, T. Hinterberger, B. Wilhelm, F. Nijboer, U. Mochty, G. Widman, C. Elger, B. Scholkopf, A. Kubler, and N. Birbaumer, "Classifying EEG and ECoG signals without subject training for fast BCI implementation: comparison of nonparalyzed and completely paralyzed subjects," *IEEE Trans Neural Syst Rehabil Eng,* vol. 14, pp. 183-6, Jun 2006.

[16] M. S. Treder, N. M. Schmidt, and B. Blankertz, "Gaze-independent brain-computer interfaces based on covert attention and feature attention," *J Neural Eng,* vol. 8, p. 066003, Dec 2011.

[17] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin Neurophysiol,* vol. 113, pp. 767-91, Jun 2002.

[18] F. Nijboer, E. W. Sellers, J. Mellinger, M. A. Jordan, T. Matuz, A. Furdea, S. Halder, U. Mochty, D. J. Krusienski, T. M. Vaughan, J. R. Wolpaw, N. Birbaumer, and A. Kubler, "A P300-based brain-computer interface for people with amyotrophic lateral sclerosis," *Clin Neurophysiol,* vol. 119, pp. 1909-16, Aug 2008.

[19] D. A. Koppenhaver and D. E. Yoder, "Literacy issues in persons with severe speech and physical impairments," in *Issues and research in special education*. vol. 2, R. Ross-Gaylord, Ed., ed New York, NY: Teachers College Press, Columbia University, 1992, pp. 156-201.

[20] H. Trinh, A. Waller, K. Vertanen, P. O. Kristensson, and V. L. Hanson, "iSCAN: A Phoneme-based Predictive Communication Aid for Nonspeaking Individuals," presented at the ASSETS'12: Proceedings of the ACM SIGACCESS Conference on Computers and Accessibility, Boulder, CO, 2012.

[21] K. Vertanen, H. Trinh, A. Waller, V. L. Hanson, and P. O. Kristensson, "Applying Prediction Techniques to Phoneme-based AAC Systems," presented at the NAACL-HLT 2012 Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), Montreal, Canada, 2012.