34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Audio-visual feedback for electromyographic control of vowel synthesis

Eric Larson, Howard P. Terry, and Cara E. Stepp

Abstract—**We describe the design and testing of a human machine interface to use surface electromyography (sEMG) collected from a covert location in response audio-visual feedback. Using sEMG collected from the Auricularis Posterior muscle, N=5 healthy participants participated in 6 sessions over multiple days to learn to transition from visual and vowel synthesis feedback to vowel synthesis feedback alone. Results indicate that individuals are able to learn sEMG control of vowel synthesis using auditory feedback alone with an average of 67% accuracy and that this skill can also generalize to new vowel targets. Control of vowel synthesis using covertly-recorded sEMG is a promising step toward more reliable mobile human machine interfaces for communication.**

I. Introduction

Current investigations into human-machine interfaces (HMI) for communication in noisy or hostile environments have concentrated on brain signals as measured by electroencephalography (EEG). Although EEG is the commonly used portable brain imaging technology currently available, it suffers from poor signal-to-noise ratios, usually requires multiple electrodes and can be heavily degraded by body movement and environmental factors.

The surface electromyography (sEMG) signal is several orders of magnitude larger in amplitude, and thus well-suited for obtaining reliable data in mobile applications. However, sEMG is often avoided for body-machine interfaces since its use often requires 1) healthy innervation of musculature that is often unavailable in cases of spinal cord injury or stroke and 2) cooption of otherwise useable control function. For instance, individuals with severely limited motor function may have some residual control in limited body areas, but would prefer to directly use that control for motor output (e.g., a button press or switch) rather than to use sEMG control. However, recent work has shown promise for the use of the Auricularis Posterior (AP) muscle for HMI control [1]. The AP is vestigial and thus can provide a control signal without interfering with other bodily functions. In addition, it is located behind the ear, which allows discreet electrode placement. Furthermore, the AP muscle is spared in many individuals even with very severe paralysis, suggesting that techniques taking advantage of AP-control for HMI could be a promising direction for assistive technology for locked-in patients.

Much HMI work has focused on control using visual feedback, typically in a 2D plane [2-4]. Constant attention to visual feedback comprises a substantial cognitive load and can detract from using typical vision to perform simultaneous tasks. Use of auditory feedback for HMI control has the benefit of potentially allowing simultaneous performance of visually-dependent tasks, and has been attempted by other groups with some success [e.g., 5, 6].

This work describes the design of a HMI to test the ability of healthy individuals to learn two-dimensional control using sEMG recorded from their AP muscles in response to auditory feedback provided in the form of synthesized vowels. We hypothesized that participants would be able to quickly learn to control the HMI using audio-visual feedback and that they would be able to maintain that control using auditory control alone after visual feedback was removed.

II. Methods

We acquired sEMG signals bilaterally from the AP muscle in five able-bodied participants (age range 20 – 31 years, 3 female). All participants provided written consent for participation in the study and were compensated $10/hr for their time. The study protocol was approved by the Institutional Review Board of Boston University. Participants were trained to modulate their muscle activity (and resulting measured sEMG power) to move a cursor in two dimensions representing the f1-f2 formant space, producing different target vowel sounds. Continuous auditory feedback was provided that was changed based on the cursor movement in the f1-f2 plane. Participants were

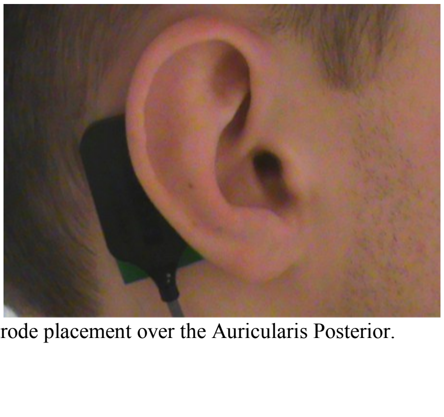

Fig. 1. Electrode placement over the Auricularis Posterior.

Manuscript received March 15, 2012. This work was supported by Boston University.

E. Larson is with the Speech and Hearing Sciences Department, University of Washington, Seattle, WA 98107 USA.

H. P. Terry is with the Speech, Language, and Hearing Sciences Department at Boston University, Boston, MA 02215 USA.

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

trained to move the cursor to hit vowel targets first using audio-visual feedback, and then using only auditory feedback as outlined below.

A. Electromyography

Participants sat upright in a quiet room in front of a laptop screen while sEMG signals were recorded. The sEMG recordings were pre-amplified and filtered using a Delsys™ Bagnoli system set to a gain of 1000, with a band-pass filter with roll-off frequencies of 20 Hz and 450 Hz. sEMG signals were then digitized at 16-bit resolution using a Fast Track Pro USB (M-Audio, Inc., USA) sampling at 44,100 Hz. Using the RTAudio suite [7], signals were processed in 2048-sample segments. Each segment was windowed using a Hanning window, a DFT was performed using FFTW, and the power of each input channel (from DC – 1000 Hz) was calculated by summing the squared magnitude of the corresponding frequency components.

The skin behind the ear of each participant was prepared for electrode placement by cleaning the surface with an alcohol pad and "peeling" (exfoliation) with tape to reduce electrode-skin impedance, noise, DC voltages, and motion artifacts. A Delsys ™ 2.1 differential surface electrode was placed over each AP, parallel to underlying muscle fibers (see Fig. 1). Each electrode each consisted of two 10-mm silver bars with an inter-bar distance of 10-mm.

For each participant, the maximum voluntary contraction (MVC) was used to transform the power of each sEMG electrode into a corresponding position in the f1-f2 plane for the vowel training sessions. Specifically, this was done by first linearly mapping activations from 10% – 70% MVC of the first and second sEMG electrodes onto the f1 (300-1200 Hz) and f2 (600-3400 Hz) axes, respectively. The resulting (*x,y*) position was then hard-limited to be at or within the axis bounds. To de-noise the muscle activation measurements, these positions were smoothed over time using a decaying exponential filter (one-pole IIR) with a 1 sec time constant.

B. Participant feedback and vowel targets

Different forms of auditory and/or visual feedback were provided to the participants based on the training session. During the participant-controlled portion of each trial (i.e., after the cue period), real-time auditory feedback was always provided, consisting of a continuous vowel sound synthesized using the Synthesis ToolKit [7] implementation of a Klatt synthesizer [8]. The f1 and f2 formants in the Klatt synthesizer were modulated based on the participant's cursor position in the f1-f2 plane. Low-level Gaussian white noise was also added to the auditory feedback to mask residual background noise.

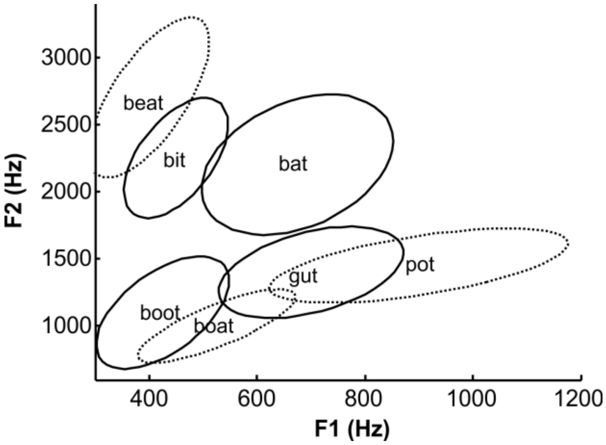

Fig. 2. Target ellipse locations in the f1-f2 plane. Solid lines indicate training targets and dotted lines indicate generalization targets.

TABLE I. VOWEL TARGET ELLIPSES IN THE F1-F2 (X-Y) PLANE

IPA	Ellipse Definition					
	Token[a]	x_c	y_c	Δx	Δy	(°)
/ɪ/	Bit	450	2250	80	450	-7
/ɛ/	Bat	675	2200	165	525	-7.5
/u/	Boot	425	1100	100	425	-10
/ʌ/	Gut	700	1400	150	350	-15
/i/*	Beat	400	2700	80	600	-7.5
/o/*	Boat	525	1000	80	300	-25
/a/*	Pot	900	1450	175	350	-45

a. Words were shown to participants instead of the IPA symbols for simplicity. Ttarget vowel ellipses were defined by their center in the f1-f2 (x_c, y_c) space and their extent along the f1 and f2 axes (Δx, Δy) prior to rotation (in degrees). *signifies targets used during the generalization session (session 6)

The visual feedback provided varied across sessions and consisted of three types. First, the cursor position was indicated by a gray dot. Second, an ellipse indicated the acceptable range of positions for the current target vowel in the f1-f2 plane, where the ellipse darkened whenever the cursor was inside the target area (ellipse). Ellipse locations for all targets are shown in Fig. 2. Third, a token word representing the vowel sound (e.g., "beat" for /i/) was shown, either at the center of the corresponding vowel's ellipse if the ellipse was visible, or at the center of the f1-f2 plane if the ellipse was not visible. Seven different target vowels were used, with corresponding target ellipses in the f1-f2 plane as summarized in Table I. Target ellipse characteristics were chosen based on speech production data from men, women, and child speakers of American English [9].

C. Training sessions

Participants were trained to control the vowel output using sEMG activation over six sessions occurring in 5 consecutive days, each lasting between 35 and 60 minutes depending on participant performance. Sessions 1 – 4 each occurred on separate days, whereas sessions 5 and 6 occurred on the same day. These sessions were designed to gradually to teach participants to coordinate muscle activity based on audio-visual feedback, eventually only using auditory feedback to hit trained and untrained vowel targets:

1) Basic sEMG control training. In this preliminary training session, participants learned basic muscle contraction control separately for each sEMG electrode. First, the maximum voluntary contraction (MVC) for each sEMG electrode was recorded, and the total power was calculated. Participants then performed a game that required them to move a cursor on the left side of the screen in one dimension (vertically) to reach targets located at 33%, 66%, and 100% MVC. This approximately 2 min exercise was repeated for both electrodes for six iterations.

2-3) Vowel training: audio and full visual feedback. In the second and third sessions (which were identical), Participants learned to use their muscles to move a cursor to different target vowel ellipses in the f1-f2 plane. The participants were shown a 2-D plot of the f1-f2 plane, and the cursor began frozen in the lower-left corner. At the start of each trial, a target ellipse appeared on the screen, with the token word centered on the ellipse (e.g., "bit"), and the corresponding vowel sound was played to the participant for two seconds. This was followed by two seconds of silence, after which point the cursor location became controlled by the sEMG signals, with the visual cursor position and audio vowel feedback changing based on the sEMG activation. Whenever the participant's cursor moved inside the target ellipse, the ellipse darkened to indicate correct positioning. Participants were then given up to 15 sec to move the cursor into the target vowel ellipse. A 1-15 sec inter-trial interval was added, and four different targets were trained ("bit", "bat", "boot", and "gut") for 30 trials, yielding 120 trials. Participants were aware that upcoming sessions would be performed using no visual feedback and many tried to prepare during session 3 by performing trials with their eyes closed.

4) Vowel training: audio and limited visual feedback. The fourth session was identical to the second and third sessions, except that the real-time vowel position cursor was not shown. Thus, in each trial the participants had to move the cursor to the target vowel ellipse, relying on the audio feedback to move the cursor inside the ellipse.

5) Vowel training: audio feedback only. The fifth session was the same as the fourth session, except that there were no visual cues at all. For each trial, participants were cued with the vowel sound and the printed word token (e.g., "bit"), but the target ellipse was not shown and the word token was always centered on the f1-f2 plane rather than at the location of the target ellipse.

6) Generalization: audio feedback only. The sixth session was the same as the fifth, in that they could only use the auditory vowel sound feedback to determine if the cursor position was in the correct location (ellipse). However, here the previously used four targets were replaced by three untrained targets ("beat", "boat", and "pot"), each tested on 40 trials instead of 30. This tested the ability of participants to generalize the auditory training to novel vowel sounds.

At the beginning of the third through sixth sessions, participants were given 1-3 minutes to arbitrarily move the visible cursor around the f1-f2 plane (with no designated targets) to re-familiarize themselves with basic cursor control.

III. Results

The ability of participants to correctly achieve target locations within the allowed 15 sec while in the various feedback characteristics was characterized as the percentage of success and is shown in Fig. 3. Starting with the full audio-visual feedback condition in session 2, participants exhibited a high level of performance (mean = 76.3%) and maintained this level through session 3 (mean = 77.3%), despite the fact that many participants were engaging in preparatory visual deprivation. Removal of online visual feedback in session 4 showed a detrimental effect, reducing performance to an average of 64.2%. Further removal of visual target locations did not show a further deleterious effect (mean = 65.7%). In the final session, use of a test set resulted in consistently poorer performance (mean = 46.3%). Average reaction times ranged from 3.3 s (session 3) to 4.6 s (session 6), all well within the 15 s opportunity given to participants.

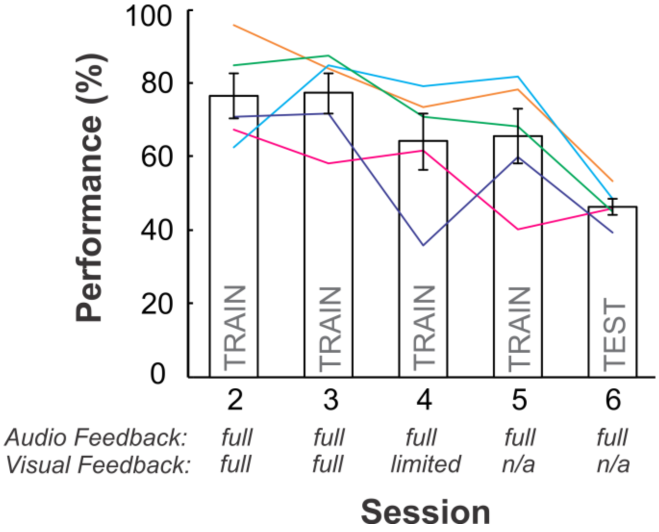

Fig. 3. Bars indicate mean performance across sessions (+/- 1 standard error). Individual performance levels of the 5 participants are shown by the lines.

Individual performance of participants varied somewhat, but displayed an overall trend of slightly reduced performance in the absence of visual feedback and a noticeable reduction with the use of a test set in session 6.

IV. Discussion

The previous work by Brumberg and colleagues reported the first ever cortically-controlled HMI for continuous control over an artificial speech synthesizer, showing usability in a paralyzed individual suffering from locked-in syndrome [6]. Using intra-cortically placed electrodes for control, their participant was able to learn to reach vowel targets with roughly 45 – 70% average accuracy over the multiple sessions of training. In their study, the first 10 sessions were performed using audio feedback alone and the final 5 included both visual and audio feedback. Using our training set, our results showed comparatively good average accuracies of 64 – 77%. This performance is promising given the difference in control signal: central versus peripheral. We have extended the exciting invasive techniques to include a methodology that is relatively inexpensive, non-invasive, and appropriate for mobile settings.

Our experimental paradigm also included a generalization task (session 6). The participants in this study were able to generalize their auditory-visuo-motor mappings to achieve new target locations, although performance did drop from a mean of 65.7% to 46.3%. This performance drop could be entirely due to the limits to generalization or due to the wider excursions necessary to reach the test targets relative to the training targets (see Fig. 2). Nevertheless, we anticipate that this flexible learning of categorical boundaries could provide promising feedback for a range of HMI applications. In our future work we plan to study the role of categorical perception on performance by contrasting the methods of this work with target locations that are not located at perceptually salient locations.

Along these same lines, the current target ellipse locations were set based on speaker production data [9]. It is possible that setting ellipse locations based on individual speaker perception ellipses will improve performance. Future work will examine this by eliciting and utilizing perceptual vowel categories for each participant.

The training protocol in this study utilized a multi-step approach over 5 days. Participants started with target and online visual and auditory feedback with the goal of using only auditory target and online feedback by the end of the experiment. We chose this paradigm with the goal of maximizing the new auditory-motor mapping necessary for the participants to perform the task. However, participants did not show obvious learning between sessions 2 and 3, nor was there a large reduction in performance between sessions 4 and 5. Given this lack of an obvious training effect, it is possible that similar performance could be attained with a shortened training protocol. We will explore this possibility in our future work.

Regardless of the feedback paradigm, it is interesting that participants were able to carefully modulate the activity of the AP, given its vestigial nature. This finding corresponds well with other recent work utilizing AP sEMG for 2D cursor control [1]. Surprisingly, although participants were randomly recruited via flyers as healthy speakers of American English, out of the five participants, 4 were able to produce some voluntary movement of their ears prior to the training provided in session 1. However, there are some obvious weaknesses in the use of the AP for HMI control. For instance, /bit/ was by far the most difficult target for most participants. This is likely because it required the most independent activation (large activations from the left site and minimal activations from the right site). The AP was chosen for this study due to the pragmatic benefits for potential applications provided by its vestigial nature and covert location. However, it is precisely its vestigial nature that makes it difficult to learn to control compared to other muscles. Preliminary work in our lab indicates that a variety of other recording sites may offer a substantial increase in performance and our future work will compare performance using a variety of sites.

V. Conclusion

We have developed an HMI using sEMG from the AP that can be controlled with accuracies up to 77% using audio-visual feedback. Use of the AP provides a covert recording location that doesn't interfere with other motor function. Use of auditory feedback shows promising results and has the benefit of potentially allowing simultaneous performance of visually-dependent tasks in a variety of users. Use of sEMG from the AP for multidimensional control of vowel synthesis could provide reliable mobile human machine interfaces for human communication.

Acknowledgment

The authors thank Alan Pacheco, Margaux Canevari, and Elias Thorp for technical assistance and beta testing of the interface used in this project and Dr. Jonathan Brumberg for helpful conversations. This study was funded by Boston University and NIH NIDCD training grant T32DC000018 (EDL).

References

[1] C. Perez-Maldonado, A. S. Wexler, and S. S. Joshi, "Two-dimensional cursor-to-target control from single muscle site sEMG signals," *IEEE Trans Neural Syst Rehabil Eng,* vol. 18, pp. 203-9, Apr 2010.

[2] J. Fruitet, D. J. McFarland, and J. R. Wolpaw, "A comparison of regression techniques for a two-dimensional sensorimotor rhythm-based brain-computer interface," *J Neural Eng,* vol. 7, p. 16003, Feb 2010.

[3] G. Schalk, K. J. Miller, N. R. Anderson, J. A. Wilson, M. D. Smyth, J. G. Ojemann, D. W. Moran, J. R. Wolpaw, and E. C. Leuthardt, "Two-dimensional movement control using electrocorticographic signals in humans," *J Neural Eng,* vol. 5, pp. 75-84, Mar 2008.

[4] Y. Li, J. Long, T. Yu, Z. Yu, C. Wang, H. Zhang, and C. Guan, "An EEG-based BCI system for 2-D cursor control by combining Mu/Beta rhythm and P300 potential," *IEEE Trans Biomed Eng,* vol. 57, pp. 2495-505, Oct 2010.

[5] F. Nijboer, A. Furdea, I. Gunst, J. Mellinger, D. J. McFarland, N. Birbaumer, and A. Kubler, "An auditory brain-computer interface (BCI)," *J Neurosci Methods,* vol. 167, pp. 43-50, Jan 15 2008.

[6] F. H. Guenther, J. S. Brumberg, E. J. Wright, A. Nieto-Castanon, J. A. Tourville, M. Panko, R. Law, S. A. Siebert, J. L. Bartels, D. S. Andreasen, P. Ehirim, H. Mao, and P. R. Kennedy, "A Wireless Brain-Machine Interface for Real-Time Speech Synthesis," *PLoS ONE,* vol. 4, p. e8218, December 2009.

[7] G. Scavone and P. Cook, "RtMIDI, RtAudio, and a Synthesis (STK) Update," in *Proceedings of the International Computer Music Conference*, Barcelona, Spain, 2005.

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

[9] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *J Acoust Soc Am,* vol. 97, pp. 3099-111, May 1995.