33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Brain-Machine Interfaces for Real-time Speech Synthesis

Frank H. Guenther and Jonathan S. Brumberg

Abstract—**This paper reports on studies involving brain-machine interfaces (BMIs) that provide near-instantaneous audio feedback from a speech synthesizer to the BMI user. In one study, neural signals recorded by an intracranial electrode implanted in a speech-related region of the left precentral gyrus of a human volunteer suffering from locked-in syndrome were transmitted wirelessly across the scalp and used to drive a formant synthesizer, allowing the user to produce vowels. In a second, pilot study, a neurologically normal user was able to drive the formant synthesizer with imagined movements detected using electroencephalography. Our results support the feasibility of neural prostheses that have the potential to provide near-conversational synthetic speech for individuals with severely impaired speech output.**

I. Introduction

RECENT years have seen the development of brain-machine interfaces (BMIs) that allow the user to manipulate computer cursors, virtual keyboards, and even external robotic arms. The neural inputs to these systems most commonly come from either intracortical electrodes or electroencephalography (EEG). In the current paper we focus on a specific subset of brain-machine interfaces: those aimed at restoring speech communication via real-time speech synthesis. Specifically, we describe two studies that utilize brain-machine interfaces to control the sound output of a formant synthesizer. The first utilizes a specialized intracranial electrode developed by Philip Kennedy and colleagues at Neural Signals, Inc. to collect electrical signals that are decoded into the first two formant frequencies of a speech signal. The second utilizes a 48-channel EEG system to drive the formant synthesizer. Future directions in BMI for real-time speech synthesis are then outlined.

II. Study 1: An Intracortical BMI for Real-Time Vowel Synthesis

In December 2004, a locked-in brain stem stroke volunteer, ER, was implanted with a two-channel Neurotrophic Electrode [1] in speech motor cortex with the primary goal of decoding the neural activity related to speech production and providing an alternative means for communication. 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.

The implant was located in an area of motor cortex which was related to the movements of the speech articulators; we hypothesized that this region would allow the user to naturally control a real-time speech synthesizer. Localization of the implant involved first conducting a pre-operative fMRI study in which ER participated in imagined picture naming and word repetition tasks in an fMRI protocol. The task revealed increased BOLD response in much of the normal speech production network, and the implantation site was chosen as the area of peak activity on the ventral precentral gyrus (location of the speech motor cortex). Details of the implantation procedure can be found elsewhere [1].

Neural spike clusters were identified from the multi-unit extracellular potential recorded from the Neurotrophic Electrode. Briefly, the extracellular potentials were first bandpass filtered (300-6000 Hz), then a voltage threshold was applied (±10μV) using the Cheetah data acquisition system (Neuralynx, Inc., Bozeman, MT). Threshold crossings were taken as putative action potentials, and an approximately 1 ms (or 32-point; hardware-dependent) data segment sampled at 30 kHz around each crossing was saved for classification analysis. All spike waveforms were classified on-line using a convex-hull technique (SpikeSort3D, Neuralynx, Inc.) according to manually defined regions obtained from previous offline analysis. These cluster regions, once stabilized, were reused for each recording session. In the current study, 56 spike clusters were identified across two recording channels (N1=29, N2=27), although this is likely an overestimate of the number of unique neural sources as some clusters may represent the same parent neural source and others may represent externally generated noise.

The BMI for real-time control of the formant frequency-based speech synthesizer used cluster firing rates smoothed via a continuous filter approach (see [2] for details). The neural decoder translated the average firing rates from the 56 clusters into estimated values of the first and second formant frequencies of the intended utterance utilizing a Kalman filter [3] based decoding algorithm. Specifically, the formant space position (i.e. the first two formant frequencies, or simply formants) and velocities (i.e. 1st derivative of formants) were decoded from normalized unit firing rates. A similar continuous filter decoder was developed by Kim and colleagues [4] to decode hand movement kinematics from human subjects. The formant

Manuscript received March 26, 2011. This work was supported in part by the National Institutes of Health (grants DC002852 and DC007683) and by CELEST, an NSF Science of Learning Center (SMA-0835976).

F. H. Guenther is with the Departments of Speech, Language, & Hearing Sciences and Cognitive & Neural Systems, Boston University, Boston, MA 02215 USA (phone: 617-353-5765; fax: 617-353-7755; email: guenther@bu.edu).

J. S. Brumberg is with the Department of Cognitive & Neural Systems, Boston University, Boston, MA 02215 USA (email: brumberg@bu.edu).

frequencies were then used to drive an artificial speech synthesizer which played the synthesized vowel waveform from the computer speakers with a total system delay of less than 50 ms from neural firing to sound output. Low system delays are necessary for fluent speech, as auditory feedback delays more than 200 ms are known to disrupt normal speech production [5].

Control of the real-time speech BMI is analogous to two-dimensional cursor control as previously demonstrated using intracranial BCIs in monkeys and humans. The main difference lies in the nature of the "task space"; speech is represented in the auditory domain whereas cursor movement is carried out in the visuo-spatial domain. Speech production and perception are naturally acoustic tasks; as such auditory feedback is much more informative regarding our ongoing speech movements than visual feedback (which is lacking during self-generated speech). Formant frequencies are a natural choice for the auditory representation in our system for several reasons. First, the Neurotrophic Electrode was implanted near the border between left premotor and primary motor cortices in a location believed to be involved in planning upcoming utterances according to an established a neurocomputational model of speech production, the Directions into Velocities of Articulators (DIVA) model [6]. The DIVA model posits that speech motor trajectories are planned as formant frequency trajectories by the premotor cortex. Second, the formant frequencies of speech are highly correlated with movements of the vocal tract articulators; e.g., changes in the first formant frequency are strongly related to upward/downward movements of the tongue and/or jaw. Finally, formant frequencies provide a convenient, low-dimensional representation that can be used to synthesize many speech sounds, including all of the vowels.

We first verified that formant frequency information was encoded in the neural signals from the implanted electrode. To do this, the subject was presented with artificially synthesized vowel sequences, played over computer speakers, consisting of repetitions of three different vowels (AA [hot], IY [heat] and UW [hoot]) interleaved with a neutral vowel sound (AH [hut]). The vowels and vowel-transitions were synthesized using a formant synthesizer according to predetermined formant trajectories. The subject was asked to attempt to speak along with the vowel sequence stimulus that was being presented. The data obtained in this paradigm was used for offline calibration of the real-time Kalman filter neural decoder. Parameters for the Kalman filter decoder were estimated by performing a least squares regression of unit firing rates and the vowel sequence formant trajectories. An offline analysis of the training data was performed to determine the correlations between two-fold cross-validated optimal linear combinations of ensemble unit firing rates and formant frequencies. This analysis found statistically significant correlations in both F1 ($r=0.49$, $p<0.001$) and F2 ($r=0.57$, $p<0.001$), verifying the DIVA model prediction of a formant frequency representation in the speech motor cortex.

We then allowed the subject to control the formant synthesizer directly using the BMI. He was first presented with a synthetic vowel-to-vowel sequence such as AH-AA and instructed to listen only during stimulus presentation. These stimuli were limited to two vowels (V1 V2) where V1 was always AH and V2 was randomly selected between the three vowels AA, IY and UW. A production period followed in which the subject was instructed to attempt to produce the vowel sequence. During the production period, the real-time neural decoder was activated and new formant frequencies were predicted from brain activity related to the production attempt every 15 milliseconds. These formant frequencies were input to the formant synthesizer, which produced sound output at a delay of 50 ms from the neural signals.

Over 25 experimental sessions, the subject attained 70% correct production on average after approximately 15-20 practice attempts per session. Fig. 1 illustrates the within-session learning effect. Production trials were grouped into blocks (roughly four blocks of six trials per session) and analyzed for endpoint production accuracy and error. Early trials (Block 1) show relatively poor performance which statistically significantly increases by Block 4 ($p < 0.05$; t-test of zero slope as a function of block). Vowel sequence endpoint error, defined as the Euclidean distance from the endpoint formant pair to the target vowel, significantly decreased from the session start to termination ($p < 0.05$; t-test of zero slope as a function of block). A detailed description of the methods and results of this study can be found elsewhere [2], [7].

These results show it is possible for a human subject to use a real-time BMI utilizing continuous formant frequency speech synthesis. Although the speech sounds used produced here are rudimentary, they indicate the promise of the direct speech synthesis BMI approach, particularly when one considers that 100-channel intracortical electrode arrays are now available for human implantation (as opposed to the 2-channel electrode used here).

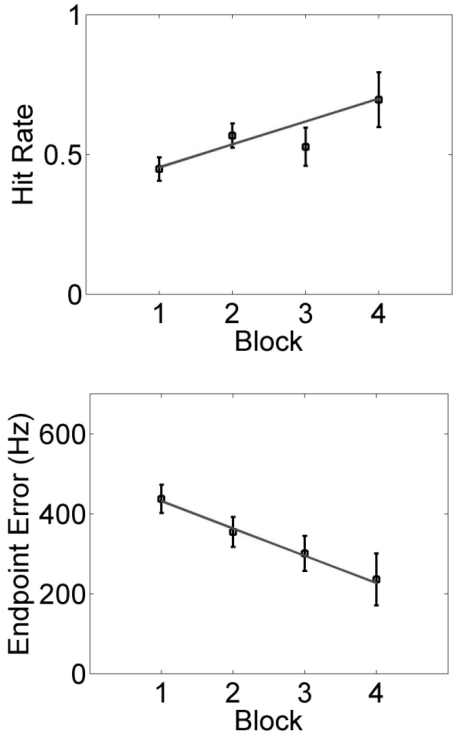

Fig. 1. Improvements over time in success rate (top) and average endpoint error (bottom) as a function of block when using the intracortical BMI in the vowel production task.

III. Study 2: A Non-Invasive BMI for Real-Time Vowel Synthesis

Although intracortical BMIs offer the promise of gaining maximal information from individual neurons, the need for neurosurgery reduces the potential application of these systems to individuals with locked-in syndrome at present. Noninvasive BMIs using EEG have also proven useful for providing communication, typically via a typing process on a computer (see [8] for a review). Here we examine the potential of EEG for real-time speech synthesis.

In this pilot study, a single subject participated in a single 1-hour session in which he performed one off-line acquisition run for training the BMI decoder and three on-line BMI control runs. In the acquisition protocol, the subject was visually and acoustically presented with 20 repetitions of three vowel sounds (60 total trials) in random order, each 3 seconds in duration: AA, IY, or UW. For each vowel, the subject was asked to perform a different imagined motor action to elicit a sensorimotor rhythm (SMR) response: left hand movement for UW, right hand movement for AA and foot pressing for IY. Limb imagery was used for reliability and known EEG response; future work will involve speech imagery. The vowel was represented visually on a computer screen display of the formant plane (see Fig. 2) and acoustically by formant synthesizer output of the 2D formant pair using the Snack Sound Toolkit (KTH Royal Institute of Technology) through pneumatic insert earphones (Model ER-1, Etymotic Research, Inc.). After completing the acquisition protocol, the EEG activity of all electrodes were band-pass filtered between 5 and 25 Hz to capture the mu (8-12 Hz) and beta (12-25 Hz) frequency ranges which are known to be modulated by motor imagery. The filtered EEG was then common-average referenced and a running root-mean-squared (RMS) magnitude was taken in overlapping windows of 100 ms. The RMS magnitude and target formant values were then used to train the coefficients of a Kalman filter decoder in which the desired outputs were the first (F1) and second (F2) formants of the target vowel sound. Leave-one-out cross-validation was used to estimate the decoder weights; the data were split into 60 trials, and the decoder was repeatedly trained on 59 trials and tested on the remaining trial. Off-line decoder training resulted in a set of weights which map SMR band power into a two dimensional formant frequency. The locations of highest contribution to the formant mapping occurred over the left and right sensorimotor regions as expected given the limb motor imagery control paradigm.

The trained Kalman filter decoder was used in a real-time feedback protocol in which the subject was first presented with a randomly selected 1.5 s target vowel sound visually and acoustically and instructed to perform the relevant imagined motor action during an 8 s response period following a random waiting interval (1-2 s in duration). During the response period, visual and/or auditory feedback corresponding to the decoder predicted formants was provided to the subject. Visual feedback consisted of a yellow cursor moving on the graphical formant plane. Auditory feedback was generated by the formant synthesizer. A trial was labeled correct when the predicted formants were within a circular target region 1.5 barks (a logarithmic frequency scale used in speech) in diameter and otherwise labeled incorrect. Two runs of 10 repetitions per vowel were conducted in which the subject received both auditory and visual feedback of the vowel sound and one run of 10 repetitions per vowel with only auditory feedback, for a total of 90 trials. The formant predictions at the end of each trial are shown in Fig. 2. AA trials are represented by squares, IY by circles and UW by triangles. Endpoints within the appropriate circular target are correct; those outside the target are incorrect. The mean accuracy over all trials was 0.71 and was not significantly different between audio-visual (0.75) and audio-only (0.63) trials, indicating the pilot subject did not heavily rely on visual information.

The results of this pilot study indicate that, with current EEG-based BMI methods, it is possible to control a formant synthesizer to produce vowel targets. The speech "movements" produced by the current BMI are not yet fast and accurate enough to mimic normal human speech, but our initial results indicate that further research into EEG-based BMIs for real-time speech synthesis are warranted.

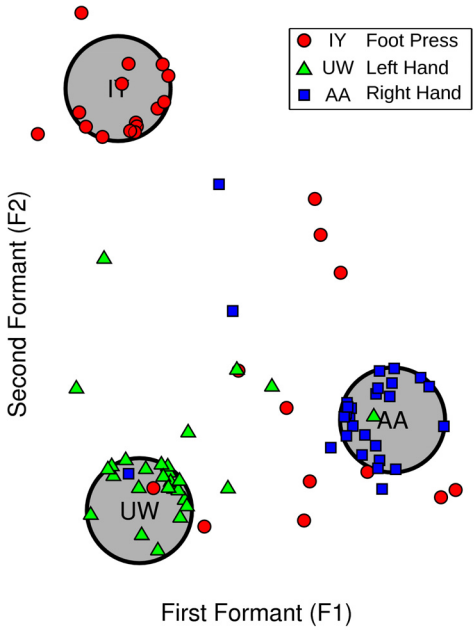

Fig. 2. Summary of pilot EEG-based formant synthesis BMI results.

IV. Future Directions: Beyond Vowel Synthesis

The use of a formant synthesizer in the current studies limits the speech output to continuously voiced speech segments, namely vowels, diphthongs, semi-vowels, and glides. Most consonants cannot be produced without very precise manipulations of many parameters in a formant synthesizer, a process that is not suited to current BMI technologies, which are currently limited to low-dimensional control applications. In computer simulations of the DIVA model of speech production [6], we have demonstrated that it is possible to produce intelligible speech involving both vowels and consonants using a biologically based controller that plans movements of an articulatory synthesizer (e.g. [9]) using a 3-dimensional auditory planning space. We are currently developing real-time, low-dimensional speech synthesizers based on this concept.

Further improvements in intracortical BMIs for speech communication will also occur as the electrode channel capacity of BMI systems continues to increase, and as our understanding of the neural representations underlying speech improves. Based on our initial results with BMIs that control of real-time speech synthesizers, as well as the impressive demonstrations of accurate computer cursor control with high channel capacity BMIs, we believe that BMIs which allow locked-in patients to produce synthetic speech at near-conversational rates are possible in the very near future.

Acknowledgment

We thank Philip Kennedy, Dinal Andreason, Princewill Ehirim, Hui Mao, Joe Wright, Jason Tourville, Mikhail Panko, Robert Law, and Alfonso Nieto-Castanon for their contributions to the research described herein.

References

[1] 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, pp. 168-176, 2008.

[2] 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, e8218, December 2009.

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

[4] 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, CNE'07. 3rd International IEEE/EMBS Conference*, pp. 486—489, 2007.

[5] D. G. MacKay, "Metamorphosis of a critical interval: Age-linked changes in the delay in auditory feedback that produces maximal disruption of speech," *Journal of the Acoustical Society of America*, vol. 43, pp. 811-821, 1968.

[6] 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, pp. 280-301, 2006.

[7] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther, "Brain-computer interfaces for speech communication," Speech Communication, vol. 52, pp. 367-379, 2010.

[8] J. S. Brumberg and F. H. Guenther, "Development of speech prostheses: current status and recent advances," *Expert Review of Medical Devices*, vol. 7, pp. 667-679, 2010.

[9] S. Maeda, "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," in *Speech Production and Speech Modelling*, W. J. M. Hardcastle and A. Marchal, Eds. Dordrecht: Kluwer, 1990, pp. 131-149.