Towards Restoration of Articulatory Movements: Functional Electrical Stimulation of Orofacial Muscles

Tanja Schultz[1‡], Miguel Angrick[1‡], Lorenz Diener[1], Dennis Küster[1], Moritz Meier[1], Dean J. Krusienski[2], Christian Herff[3] and Jonathan S. Brumberg[4]

***Abstract*— Millions of individuals suffer from impairments that significantly disrupt or completely eliminate their ability to speak. An ideal intervention would restore one's natural ability to physically produce speech. Recent progress has been made in decoding speech-related brain activity to generate synthesized speech. Our vision is to extend these recent advances toward the goal of restoring physical speech production using decoded speech-related brain activity to modulate the electrical stimulation of the orofacial musculature involved in speech. In this pilot study we take a step toward this vision by investigating the feasibility of stimulating orofacial muscles during vocalization in order to alter acoustic production. The results of our study provide necessary foundation for eventual orofacial stimulation controlled directly from decoded speech-related brain activity.**

I. INTRODUCTION

Silent speech interfaces (SSI, [1]) are designed to enable users to perform spoken communication in the absence of an intelligible airborne acoustic speech signal. Common examples of SSI include speech decoding, i.e. the transformation into text, and direct speech synthesis from speech-related biosignals [2]. SSI are particularly suitable for individuals without physical impairments using techniques that record orofacial muscle activity via surface electromyography [3], [4], and tongue kinematics from ultra-sound measurements of the oral cavity [5] and permanent magnet articulography [6]. However, individuals with severe neuromotor impairments due to stroke, head trauma, infection or inflammation of the facial nerve, observations at the periphery may not be sufficient, leaving only the measurement of brain activity to supply requisite signals for a speech prosthesis. Recent studies show that the decoding of articulatory and segmental features from intracranial recordings of brain activity is possible (see [7], [8] for a review) as well as speech synthesis from invasive [9], [10] and non-invasive [11] recordings of neural activity.

In the limb-motor domain, neural prostheses have evolved from controlling robotic arms toward brain-controlled functional electric stimulation (FES) enabling reaching and grasping movements in a patient with tetraplegia using their own arm [12]. Targeted spinal cord stimulation has even been used to restore walking ability [13]. We envision speech production, as a form of complex motor control, can draw from these new areas of FES research for restoring movements of the orofacial musculature. In particular, targeted stimulation of the orofacial musculature will cause the vocal articulator to change configuration and result in appropriate changes in speech acoustic output (e.g., stimulation of the zygomaticus major muscle will retract the lips as in the vowel [i]). In this pilot study, we investigate the feasibility of stimulating a subset of orofacial muscles involved in speech production as a first step toward developing neural speech prosthesis that may be used to eventually control the upper vocal tract articulators using decoded speech-related brain activity.

II. MATERIAL AND METHODS

Two experiments were conducted to (1) identify orofacial muscle movements corresponding to speech articulation using surface electromyography (EMG) following our prior work [3], and then (2) to investigate the acoustic and kinematic effects of targeted stimulation of those muscles. The pilot stimulation experiment used a speech production paradigm with three conditions: (1) self-controlled stimulation (SCS), (2) external-controlled stimulation (ESC) and a no-stimulation, reference condition (REF). Four male and one female healthy subjects between 27 and 55 years participated voluntarily in this feasibility study and gave written consent to the recording of EMG signals and functional electrical stimulation of their orofacial muscles. The experimental procedures are compliant with the principles outlined in the Declaration of Helsinki. The analyses and results presented in this paper are based on the only subject (male, 31), who had completed the full set of trials for both EMG and FES recordings.

A. EMG-based Analysis of Orofacial Muscle Activity

We first examined the EMG signals of individual orofacial muscles during voluntary production of two vowels (V) and two vowel-consonant-vowel (VCV) combinations. Each production was recorded as a single trial spoken in isolation with preceeding and succeeding silence. For this study, we limited the analysis to the production of two vowels [a], [e] and two vowel-consonant-vowel productions [ava] and [eve]. The motivation for the selection of these particular pronunciations in given in section III.

EMG signals were recorded with sampling rate of 2048 Hz using seven pairs of single Ag/Ag-Cl electrodes in a classical bipolar setting with a minimum of 1.5 cm center-to-center inter-electrode spacing (Quattrocento, OT Bioelettronica). The acquisition system uses a differential amplifier to suppress noise (fixed gain 150V/V, $33mV_{pp}$ input range at 16 bits

[1] Cognitive Systems Lab, University of Bremen, Bremen, Germany tanja.schultz@uni-bremen.de, ‡These authors contributed equally

[2] ASPEN Lab, Biomedical Engineering Dept., Virginia Commonwealth University, Richmond, VA, USA djkrusienski@vcu.edu

[3] School for Mental Health and Neuroscience, Maastricht University, The Netherlands c.herff@maastrichtuniversity.nl

[4] Speech and Applied Neurosciene Lab, Speech-Language-Hearing Dept., University of Kansas, Lawrence, KS, USA brumberg@ku.edu

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

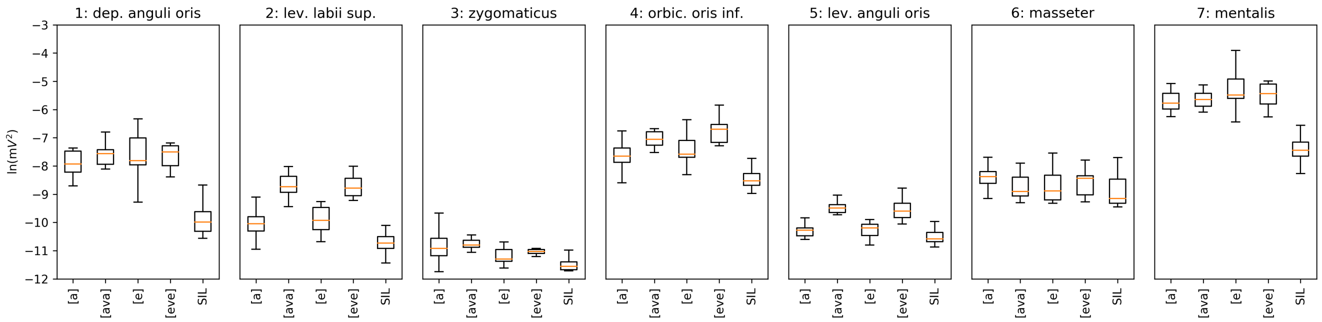

Fig. 1. EMG activities of facial muscles during V and VCV productions. During silence trials (SIL), participants were asked to relax all facial muscles.

resolution). A low-pass filter with a 500 Hz cutoff frequency was applied to avoid aliasing and a 10 Hz high-pass filter was used to reduce movement artifacts. A self-adhesive button ground electrode placed on the left wrist provided a common reference. As shown in Figure 2, the electrodes were positioned to obtain the EMG signals of seven facial muscles involved in speech articulation:

(1) *depressor anguli oris*
(2) *levator labii superioris*
(3) *zygomaticus major*
(4) *orbicularis oris inferior*
(5) *levator anguli oris*
(6) *masseter*
(7) *mentalis*

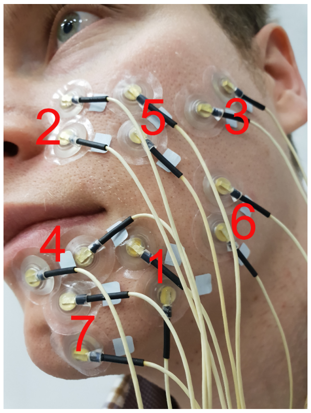

Fig. 2. EMG electrode positioning for Facial Muscle Activity Analysis

Following acquisition, we computed the signal power $p(t) = |s(t)|^2$ in 32 ms windows with 10 ms overlap. Figure 1 compares the signal power of the EMG activity for the seven facial muscles during V and VCV productions as well as baseline silence trials (SIL) in which participants were asked to relax their facial muscles. For each muscle, a boxplot shows the median, interquartile range, minimum and maximum of the signal power (y-axis). Signal power varies significantly across the different muscles, corresponding to their size and activity during production. While producing the sound [a], neither the *levator labii superior*, which elevates the upper lip, nor the *levator anguli oris*, which elevates the oral angle, are engaged and thus show power values similar to the relaxed position (SIL). Similary, low power is observed in both muscles and the *zygomaticus major* during production of [e]. In contrast, the *masseter* which enables a forced closure of the mouth, shows large variation even in SIL trials, implying a less relaxed jaw position.

III. Stimulation of Orofacial Muscles

The results of the EMG analysis describes patterns of muscle activity that are used for production of the selected V, VCV tokens. While several muscles were involved, here we focus the proof-of-concept experiment on *zygomaticus major* stimulation for the following reasons: (i) this muscle elevates and draws the angle of the mouth upward and laterally to raise the upper lip in labiodental and bilabial fricatives, and thus is relevant to the production of V[v]V combinations in this study. For this reason we focused our experiment on production of V[v]V combinations in the context of two vowel conditions, i.e. an open vowel [a] and a close-mid vowel [e]. Furthermore, (ii) the muscle is reasonably isolated from other orofacial muscles, reducing the risks of cross-stimulation, (iii) it is easy to locate as it originates from the zygomatic bone and inserts at the angle of the mouth, and (iv) it is distant from critical areas like the eye and the carotid artery, thus avoiding risks and unpleasant side-effects of stimulation.

A. FES Device, Electrodes, Stimulation Parameters

Functional electrical stimulation of the *zygomaticus major* was performed bilaterally using a *MOTIONSTIM 8* device[1].

For each muscle, two electrodes with a 32 mm diameter (Kraut+Timmermann) were positioned between the *zygomaticus major* origin and insertion according to common bipolar facial electromyography guidelines [14]. We adjusted the size and shape of the electrodes to avoid overlaps when placed on the face. The positioning of the electrodes is comparable to channel 3 as shown in Figure 2. All stimulation events had 1.5 sec duration, with a 0.5 sec ramp (i.e. 0.5 sec ramp-up, 0.5 sec full stimulation, 0.5 sec ramp-down), with an impulse width of 120 μs, frequency of 40Hz, and strength of 15 mA.

B. Stimulation Study Design and Setup

During the experiment, participant vocalizations with and without FES were recorded with a Rode NT-1 condensor microphone (distance about 0.4 m). In each session, participants wearing noise cancelling headphones were asked to produce a neutral, audible vowel by voicing (i.e., exciting the vocal folds) while relaxing the muscles of the upper vocal

[1] The authors are very grateful to Medel Medizinische Elektronik Handelsgesellschaft mbH for providing the device so quickly and easily

tract. Video and audio were recorded using the Lab Streaming Layer (LSL) middleware, which provides acquisition, network transmission and software-based time-synchronization of the data streams. Each recording session contained three parts in which the *zygomaticus major* was stimulated during vocalization using (1) self-controlled stimulation (SCS) and (2) external-controlled stimulation (ESC) using a manual push-button trigger device. The third condition served as a control and did not involve any stimulation (REF). In the first condition, participants control the initiation of stimulation while in the second, stimulation was controlled externally by an experiment observer.

Participants completed 5 consecutive trials separated by a small break (minimum, 3 s) to relax the muscles between stimulations. Each trial started with the production of either the vowel [a] or [e] without stimulation, followed by stimulation (SCS, ECS) or voluntary movements (REF) for producing the consonant [v] and concluded in vowel production again. For the analysis, we considered 3 s windows with a 1 s pre-stimulus period, 1.5 s of stimulation and 0.5 s post-stimulus. During the recording of reference trials, the length for the speech production of the consonant were approximated by the subject.

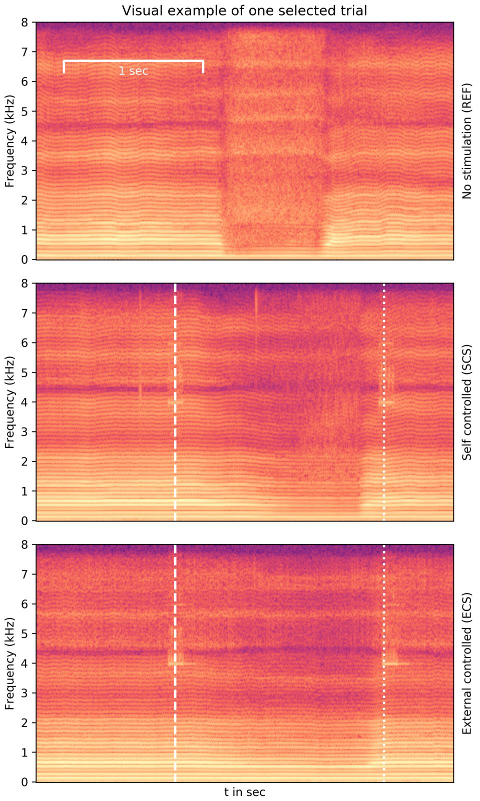

Fig. 3. Spectrograms of voluntary articulation (REF, top), self-controlled stimulation (SCS, middle), and external-controlled stimulation (ECS, bottom); dashed lines indicate the start and dotted lines the end of stimulation.

IV. Experimental Results

In this section we compare experimental results in the three conditions (REF, SCS, and ECS) based on (1) the visual inspection of the spectrograms for each production, (2) a correlation analysis, and (3) a similarity calculation based on Euclidean distance scores accumulated over time-alignments given by the Dynamic Time Warp (DTW) algorithm. For all analyses, the acoustic signals were digitized using 16 kHz sampling and 16-bit resolution. After windowing using a Blackman window with a segment length of 32 ms and an shift of 0.5 ms, spectral features were calculated based on the short-time Fast Fourier Transform (FFT) resulting in 257 bins covering 31 Hz each.

A. Visual Comparison of Spectral Features

We computed the spectrogram for a 3 s audio segment of each trial, in which the first second corresponds to the production before stimulation, followed by 1.5 s of stimulation and 0.5 s after stimulation. All trials were labeled according to the stimulation condition: voluntary articulation (REF), self-controlled stimulation (SCS), and external-controlled stimulation (ECS). Figure 3 displays the spectrograms of the recorded audio signals for one selected trial of producing [ava]. It is clearly visible that both SCS and ECS impact the articulation during the period of stimulation of the *zygomaticus major*, which caused the raising of the lips to form a sound that is perceived as a voiced labiodental fricative [v] or a voiced bilabial fricative/approximant [β].

B. Correlation Analysis

Pearson correlations were calculated between the spectral features of REF and each of the stimulation conditions (SCS & ECS) for the tokens [ava] and [eve]. Individual trials from SCS and ECS were compared with all REF spectrograms from the same VCV by computing correlation coefficients across all spectral bins and using the mean as a representative score. In order to analyze the quality of the consonant [v] in SCS/ECS VCV trials, we compared the spectral coefficients prior to stimulation with the first 1 s segment of the corresponding REF spectrogram. Similarly, we compared the entire 3 s spectrogram of VCV trials.

Figure 4 summarizes both the distributions of correlation coefficients across pre-stimulus spectral features as well as from the complete trial. Here, the pre-stimulus correlation coefficients are used as a baseline. In most cases we achieve reasonable high correlations across the pre-stimulus spectral coefficients with regards to their reference voluntary articulations as expected. The correlation coefficients are lower for the complete spectrogram indicating the auditory output during FES did not completely mimic the [v] production from REF trials, but somewhat overlap the pre-stimulus distributions. Further investigation is needed to interpret results for [eve] production in the ECS condition.

C. Time Alignments and Distance Scores

To determine the similarities between the spectrograms of REF trials to stimulation trials, we time-aligned spectrograms in the SCS & ECS conditions to REF using Dynamic Time Warping (DTW) using the best matching REF trial as reference. Frame-wise Euclidean distances were then

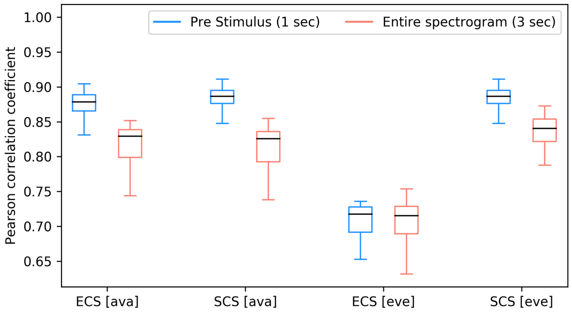

Fig. 4. Boxplot for Pearson Correlation Coefficients; Pre-stimulus accumulates over all frames prior to stimulus onset; Entire spectrogram accumulates over all frames of the complete 3 sec trial

accumulated along the corresponding DTW-path. We used the distance of the n REF trials to each other as a baseline, where for each trial i the best DTW score is calculated over all trials $i \neq j, \quad i, j = 1 \dots n$. Manual cross-checks showed that DTW-paths closely follow the diagonal, indicating that sound patterns are reliably reproduced with respect to timings.

The Euclidean distances REF-SCS and REF-ECS were averaged over all corresponding trials. Levene's tests showed no significant deviations from a normal distribution of error variances. We thus performed standard statistical analyses (t-tests) to compare the distances obtained in all three conditions. Our results showed that ECS resulted in a significantly (p = .036) larger Euclidean distance than REF (ECS = 5175.9; REF = 4467.5), suggesting substantial differences in the auditory signal compared to unmodified voluntary articulation. We further observed a statistical trend toward a significantly larger distance for ECS compared to SCS (p = .102; SCS = 4624.4). However, we found no significant differences between SCS and REF (p = .411). Our results thus suggest that external-controlled stimulation of the zygomaticus major muscles may not yet achieve an auditory signal that is indistinguishable from REF. We further expect that a more statistically powerful future manipulation might reveal significant differences also emerge between ECS and SCS. In contrast, it appears that SCS is relatively stable and difficult to distinguish from REF activity. We speculate that this might be explained by subjects automatically compensating in response to this type of stimulation by means of the auditory feedback loop.

The results confirm our expectations, namely that voluntary articulations are relatively stable across trials (lowest distance for REF), while self-controlled stimulation leads to articulation patterns that are closer to REF than articulation patterns from external-controlled stimulation. Furthermore, the ratios between the scores reveal that SCS patterns are within 3.5% of the baseline scores, and ECS within a 15.9%. This reasonable closeness is confirmed by listening to some of the stimulated speech output, in which the stimulated production is correctly perceived. Importantly, this preliminary analysis is based on a small sample and the results must be interpreted with caution. Nevertheless, we find these results to be very encouraging with respect to the possibility of using ECS to modulate orofacial muscle activity and the resulting speech output.

V. Conclusion

We presented a pilot feasibility study for understanding the acoustic consequences of orofacial FES during speech production toward the goal of restoring the ability to physically produce speech for individuals with severe speech-motor impairments. Development of an orofacial FES system that can accurately control the speech articulators may eventually be controlled directly by decoded speech-related brain activity. Mounting evidence suggests that decoding speech articulations from neural recordings is possible; therefore, our present study provides some of the first evidence that FES stimulation may be used to effectively control the speech articulators. Experimental results of this pilot shed light on the challenges and potential of such a brain-driven active voice prosthesis.

References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. S. Brumberg, "Silent Speech Interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270 – 287, 2010.

[2] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.

[3] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session Independent Non-Audible Speech Recognition Using Surface Electromyography," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico, 2005, pp. 331–336.

[4] L. Diener, C. Herff, M. Janke, and T. Schultz, "An initial investigation into the real-time conversion of facial surface emg signals to audible speech," in *IEEE EMBC*, 2016, pp. 888–891.

[5] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[6] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "Direct speech generation for a silent speech interface based on permanent magnet articulography," in *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*, 2016, pp. 96–105.

[7] C. Herff and T. Schultz, "Automatic speech recognition from neural signals: a focused review," *Frontiers in neuroscience*, vol. 10, p. 429, 2016.

[8] S. Chakrabarti, H. M. Sandberg, J. S. Brumberg, and D. J. Krusienski, "Progress in speech decoding from the electrocorticogram," *Biomedical Engineering Letters*, vol. 5, no. 1, pp. 10–21, 2015.

[9] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, "Speech synthesis from ecog using densely connected 3d convolutional neural networks," *bioRxiv*, p. 478644, 2018.

[10] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Intelligible speech synthesis from neural decoding of spoken sentences," *bioRxiv*, p. 481267, 2018.

[11] J. S. Brumberg, K. M. Pitt, and J. D. Burnison, "A non-invasive brain–computer interface for real-time speech synthesis: The importance of multimodal feedback," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 4, pp. 874–881, 2018.

[12] A. B. Ajiboye, F. R. Willett, D. R. Young, W. D. Memberg, B. A. Murphy, J. P. Miller, B. L. Walter, J. A. Sweet, *et al.*, "Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration," *The Lancet*, vol. 389, no. 10081, pp. 1821–1830, 2017.

[13] F. B. Wagner, J.-B. Mignardot, C. G. Le Goff-Mignardot, R. Demesmaeker, S. Komi, M. Capogrosso, A. Rowald, I. Seáñez, M. Caban, *et al.*, "Targeted neurotechnology restores walking in humans with spinal cord injury," *Nature*, vol. 563, no. 7729, p. 65, 2018.

[14] A. J. Fridlund and J. T. Cacioppo, "Guidelines for human electromyographic research," *Psychophysiology*, vol. 23, no. 5, pp. 567–589, 1986.