A Noninvasive Brain-Computer Interface for Real-Time Speech Synthesis: The Importance of Multimodal Feedback

Jonathan S. Brumberg, Kevin M. Pitt, and Jeremy D. Burnison

***Abstract*— We conducted a study of a motor imagery brain-computer interface (BCI) using electroencephalography to continuously control a formant frequency speech synthesizer with instantaneous auditory and visual feedback. Over a three-session training period, sixteen participants learned to control the BCI for production of three vowel sounds (/i/ [heed], /ɑ/ [hot], and /u/ [who'd]) and were split into three groups: those receiving unimodal auditory feedback of synthesized speech, those receiving unimodal visual feedback of formant frequencies, and those receiving multimodal, audio-visual (AV) feedback. Audio feedback was provided by a formant frequency artificial speech synthesizer, and visual feedback was given as a 2-D cursor on a graphical representation of the plane defined by the first two formant frequencies. We found that combined AV feedback led to the greatest performance in terms of percent accuracy, distance to target, and movement time to target compared with either unimodal feedback of auditory or visual information. These results indicate that performance is enhanced when multimodal feedback is meaningful for the BCI task goals, rather than as a generic biofeedback signal of BCI progress.**

***Index Terms*— Brain-computer interfaces, electroencephalography, neural engineering, neural prosthesis, speech synthesis.**

I. Introduction

BRAIN-COMPUTER interfaces (BCI) for communication have a number of possible applications including silent speech [1] and restoration of communicative output for individuals with profound speech and motor impairments [2], [3]. Most often, BCIs are thought of as a method of last resort for individuals who can not speak due to neurological disease or injury, and can not use existing forms of augmentative and alternative communication [4], [5]. Individuals commonly considered for BCI intervention include those with quadriplegia with anarthria due to locked-in syndrome [6] or neurodegenerative diseases such as amyotrophic lateral sclerosis.

Most non-invasive BCI applications for communication employ discrete typing interfaces using event-related potentials such as the P300 [7]–[9], steady brain oscillations (steady state visually evoked potential; [10]), or neurological rhythms related to cognitive-motor processing [11]–[13]. These approaches have been largely successful and have provided a new avenue for written communication for at-home users [8], [9], with spoken communication possible through text-to-speech synthesis. Another approach to BCI-based communication is to directly decode intended speech from neurological activity using intracranial recordings [14]. Only a few studies of non-invasive BCIs have attempted to decode speech using using electroencephalography (EEG), and have focused on discrete prediction and selection of internally spoken words [15] and vowels [16] from a small dictionary of possible items.

In contrast to typing and single word/vowel prediction, which rely on discrete selections, the motor and acoustic bases of speech production and perception are the result of continuous processes of speech-motor articulation and generation/reception of a time-varying acoustic signal [17]. The continuous auditory output of speech production is then transformed into discrete phonetic, syllabic, and word-based representations for speech perception and language comprehension [18]. These same processes hold true for perception of self-produced speech; continuous movements of the speech articulators are used to generate an acoustic signal that is perceived by oneself to maintain error-free productions, and to adjust to errors when they occur [19], [20]. During speech, movements of the speech articulators alter the vocal tract resonance, which is measured acoustically as formant frequencies, or formants [17]. Most monophthong vowels can be represented by just the first two formants, making them an ideal low-dimensional representation of speech for BCI applications. Additionally, they can be readily synthesized into artificial speech sounds [21] with minimal computational overhead [22].

We propose an alternative BCI system for communication in which EEG activity related to motor imagery is used to directly and continuously control a two-dimensional (2D) formant vector that can be presented graphically as a 2D cursor, or synthesized in real-time for instantaneous auditory feedback. Other studies have investigated formant synthesis for auditory feedback via intracortical neural prosthesis for speech [22] and surface electromyography (sEMG) [23]. In the

Manuscript received May 23, 2017; revised January 5, 2018; accepted February 14, 2018. Date of publication February 21, 2018; date of current version April 6, 2018. This work was supported by the National Institute on Deafness and Other Communication Disorders under Grant R03-DC011304. *(Corresponding author: Jonathan S. Brumberg.)*

J. S. Brumberg is with the Graduate Neuroscience Program, Department of Speech-Language-Hearing: Sciences and Disorders, The University of Kansas, Lawrence, KS 66045 USA (e-mail: brumberg@ku.edu).

K. M. Pitt is in the Department of Speech-Language-Hearing: Sciences and Disorders, The University of Kansas, Lawrence, KS 66045 USA.

J. D. Burnison was with the Graduate Neuroscience Program, The University of Kansas, Lawrence, KS 66045 USA.

Digital Object Identifier 10.1109/TNSRE.2018.2808425

current study, we extend these efforts and investigate EEG as a possible signal modality, and limb-motor imagery as the mental control strategy, which draws from prior work on continuous cursor control via BCI [24]–[26]. Further, we explicitly examine the role of feedback modality on BCI performance by examining three separate participant groups, those receiving unimodal feedback of visual or auditory information, and those receiving multimodal audio-visual information.

Prior studies investigating BCI feedback modality have found that unimodal feedback of visual information typically leads to enhanced BCI performance, and that unimodal auditory feedback suffers in terms of percent accuracy and learning time [25], [26]. Past studies involving auditory feedback for motor BCIs used generic audio stimuli as an indicator of BCI performance and success. Combined multimodal audio-visual feedback has also underperformed relative to unimodal visual feedback, possibly due to an overreliance on visual information [23], [26]. Prior work also suggests that visual feedback using cursor position information will likely lead to greater performance compared to auditory feedback alone [25] though it is possible that after a short training period, performance will be equal for both unimodal feedback conditions. In our experiment, we hypothesize that the addition of speech-related auditory information (instead of generic audio signals) to conventional 2D cursor visual feedback will enhance BCI performance in a vowel production task.

II. Methods

A. Participants

We recruited eighteen adults from the University of Kansas to participate in a three-session, BCI-based vowel production study (age range: 21 – 36, mean 27.5 years; 14 female; 4 left-handed). All participants were native speakers of American English, self-reported normal speech, language and hearing with no known neurological or neuromotor impairments, and received monetary compensation for their time. All study procedures were approved by the Institutional Review Board of the University of Kansas, and all participants provided their informed consent prior to engaging in study activities.

B. Procedure

1) Study Protocol: The participants were grouped into three conditions, those receiving: unimodal auditory feedback of synthesized vowel sounds (AO), unimodal visual feedback of the 2D vowel formants (VO), and multimodal, audio-visual feedback of vowel sounds (AV). Figure 1 shows a visual depiction of the relationship between formants and vowels used in American English. In this study, the cursor represents the instantaneous formant estimate and the text locations represent the target vowel centers on the formant plane (only UW [who'd], AA [hod], and IY [heed] were used for training participants in the present study). Examples of the synthesized vowels are provided in the Supplementary Material.

During the initial session, participants first completed a screening questionnaire to ensure they met the inclusion criteria of native fluency in American English, and did not have any metallic cranial implants. The remaining study procedures were identical for all three sessions. Training data was collected at the start of each session in which participants were presented with thirty repetitions of the vowel stimuli (synthesized, visualized formants, or both) in random order for three seconds each and were instructed to make one of three movements using kinesthetic motor imagery in response to the vowel stimuli: left-hand for the vowel UW, right-hand for AA, and bilateral feet for IY. Model weights were then estimated for a Kalman filter BCI decoder (section II-C) based on the training data and used for online BCI control.

Online test trials involved a short 1.5 s presentation of the target vowel, followed by a one-second blank interval, and a six-second response period in which participants were instructed to make the appropriate imagined movement to direct the BCI toward the target. Real-time visual (cursor movements) and/or audio (synthesized vowel sounds) feedback was provided to the participants in the response period, which began with formant values in the center of the vowel space (e.g., a neutral vowel sound). Participants were instructed to initiate motor imagery as soon as the BCI decoder began producing an output in order to move the decoded formants away from initial neutral values such that they match the target stimulus (UW, AA, or IY) while avoiding other vowels (e.g., AE: [had]), and to hold the BCI output within the target region as long as possible. In the visual domain, holding the target in place requires participants to match cursor positions with circular target regions on the screen, while in the auditory domain, synthesized BCI outputs must be matched to the auditory memory of the target vowel. This task is analogous to other visual 2D center-out tasks (e.g., [24]) and 1D cursor control with multimodal audio-visual feedback (e.g., [25]) using sensorimotor-based BCIs, though based on speech-related audio and visual feedback. Decoding ended after six seconds regardless of the output accuracy, and participants were given a random 3–5 s break between each stimulus. Most participants completed four runs[1] of thirty trials (ten trials per stimulus) in the online test period.

2) Stimulus Feedback: Visual feedback of the 2D formant plane was provided on a computer screen. For clarity, vowel stimuli were represented by a two-character label (UW, AA, IY) in a circular target region according to a set of average formants in Hz (UW: (300, 870), AA: (730, 1090), IY: (270, 2290)). Formants were log-normalized to equalize representations of low and high frequencies in the BCI [17]. Figure 1 shows eight monophthong vowels in American English, including the vowels IY, AA, and UW (and AE) used in this study. The vowel circle was highlighted during training to indicate the target stimulus. During test runs, the vowel circle was highlighted during the stimulus phase only, and the decoded formants were displayed in real-time as a moving cursor on the formant plane in the response period.

Auditory feedback was provided through speakers positioned to the front-left and front-right of participants seated in a sound-treated booth. All auditory feedback (stimulus & response phases) was synthesized using a formant speech synthesizer (Snack Sound Toolkit, KTH Royal Institute of

[1] Some participants only completed two runs in the first session due to completion of the screening questionnaire and training protocol.

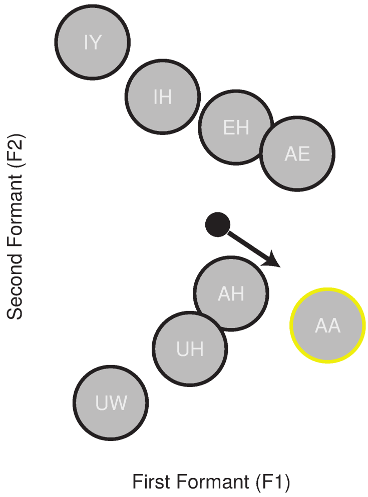


Fig. 1. The 2D formant plane with example vowels and a cursor to indicate the current decoded output. In this case, participants imagine a right-hand movement to move the predicted formants to the AA target to produce the vowel in "hot". The cursor position is instantaneously synthesized into auditory output along with visual feedback in the AV condition, and without visual feedback in the AO condition.

Technology) in real-time using the same values for visual feedback. Synthesized formants were provided as continuous auditory feedback for the duration of the online test run response period, and used a constant 2500 Hz for the third formant frequency and 120 Hz for the fundamental frequency.

3) EEG Acquisition: All EEG recordings were obtained using a 64-channel acquisition system (g.HIAmp, g.tec) with 64 active electrodes for recording the EEG (two earlobe electrodes were designated as references), and one stimulus trigger channel for alignment of recorded signals. The ground electrode was placed at location FPz according to the 10-10 standard [27], all electrodes were referenced to the left earlobe, and the electrooculogram was monitored together with the EEG from electrodes AFz, AF7, AF8, FP1, FP2, F9, and F10. All signals were recorded at a 256 Hz sampling rate with a notch filter from 58–62 Hz to eliminate the powerline artifact, stored for offline analysis, and transmitted to a networked computer for real-time BCI analysis.

4) Offline and Online Preprocessing: Offline analysis of training data and real-time analysis of online test trials involved similar data preprocessing procedures. For training data analysis, a 1 Hz high-pass filter was applied to the raw data followed by electrooculographic artifact removal via independent components analysis (ICA). Visual inspection of the power spectral density of four sensorimotor electrodes C3, C4, CP3, and CP4 was used to identify participant-specific ranges for the μ and β bands used to control the BCI. Participants without an observable μ or β rhythm were identified for post-analysis. The processed signal was rereferenced to the common average, then split into three filtered streams using a fourth-order butterworth filter: 1) low-pass filtered at 30 Hz to identify any remaining suprathreshold (± 150 μV) artifacts, and band-pass filtered according to the observed 2) μ and 3) β bands. The μ and β band power was then estimated using the Hilbert transform for training the Kalman filter decoder. A new decoder was trained at the beginning of every session to account for any possible changes in electrode position or relative amplitude of the μ and β rhythms.

During test trials, stored values from the offline analysis were used to implement online filtering and band-power estimation. ICA was not used since current methods for ICA rejection require as much as three seconds of data [28], which is too slow for real-time audio-visual feedback of speech sounds.[2] Instead, any recorded signal greater than ± 150 μV magnitude was marked as an artifact and signaled the decoder to stop processing incoming data until the artifact ended.

C. BCI Decoder

A Kalman filter decoding algorithm [22] was used to convert recorded sensorimotor rhythms into a 2D formant velocity control vector, which was integrated for audio and/or visual feedback. In this implementation, the output state model represents velocities in the 2D formant plane and is given by:

$$x[n] = Ax[n-1] + w[n]. \tag{1}$$

The 2 × 2 state matrix A describes the *a priori* probabilities of future formant velocities $x[n]$ based on past estimates $x[n-1]$, and is represented by a first-order autoregressive model. The likelihood model represents the relationship between formant velocities and anticipated modulations in the sensorimotor rhythm over all features (e.g., combined μ and β band power for all electrodes) and is given by:

$$y[n] = Hx[n] + q[n]. \tag{2}$$

The N × 2 likelihood matrix H is a linear model of the relationship between the observed sensorimotor rhythm $y[n]$ and a given formant velocity vector $x[n]$. The error terms in Equations 1 and 2 are Gaussian random variables $N(0, W)$ and $N(0, Q)$, respectively, and the 2 × 2 matrix W is the state model residual covariance while the N × N matrix Q is the likelihood model residual covariance.

The choice of decoding features was made based on the presence or absence of either the μ or β rhythm per participant. Participants with no observable μ rhythm were removed from this analysis. Likelihood models and covariance matrices were fit to both μ and β band features for participants with observable μ and β rhythms, and only to the μ band features for participants with no observable β rhythm. The Kalman filter decoder *a posteriori* estimates of formant velocity were integrated in real-time, downsampled to 32 Hz, and provided to the participants as audio and/or visual feedback depending on participant group assignment (AO, VO, or AV).

D. Performance Analysis

Decoded formants began in the center of the vowel space defined by the vowel triangle of UW, AA, and IY, and participants were required to change the output formants to match the target vowel using motor imagery. Visually, a circle and text label on the screen identified the ideal formant locations for each vowel, and the auditory target stimulus provided a

[2] Auditory delays of 100-200 ms can induce speech disruptions [29].

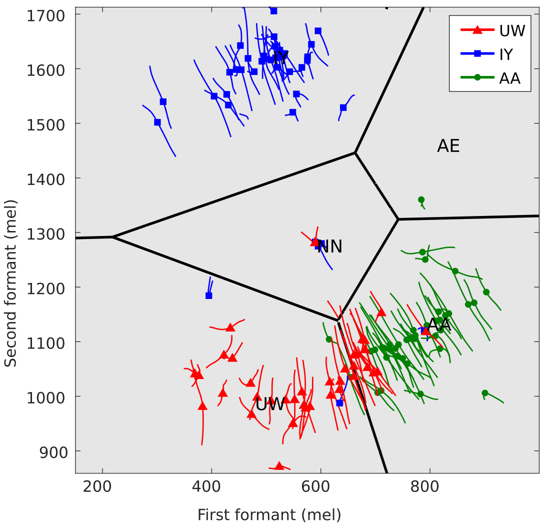


Fig. 2. An example of trial trajectories for an AV group participant. The LDA classification regions for vowel formants are labeled for the four possible outcomes (IY, UW, AA, and AE) and neutral (NN). Markers represent the closest point to the vowel target and code for each ground truth label (triangle: UW, circle: AA, square: IY), with tails to represent the 500 ms around the endpoint position (±250 ms).

synthesized version of the ideal target vowel sound. Speech perception is variable and adaptive, and though the auditory targets were chosen to represent ideal vowel exemplars, they are an arbitrary choice since humans often accept non-ideal auditory stimuli as representative of specific phonemes [30]. Therefore, we used a publicly available data set of two productions of ten vowel sounds from both men (N = 33) and women (N = 28) [31] to train a linear discriminant analysis classifier for estimating the accuracy of decoded formant frequencies. We further simulated participants' decision-making process for determining when synthesized outputs were acceptable by grouping vowels into five major categories: high-front (e.g., IY), high-back (e.g., UW), low-back (e.g., AA), low-front (e.g., AE), and neutral, and determined the phonetic label associated with all time points in the response period after leaving the neutral vowel region (see Figure 2).

The dwell times within each vowel region in Figure 2 were computed, and trials were marked correct if (1) the dwell time exceeded 500 ms and (2) the vowel region matched the target vowel. Trials were marked incorrect if dwell times exceeded 500 ms, but decoded formants were in other, non-target vowel regions, or no dwell times exceeded 500 ms. Accuracy was also measured using the Euclidean distance on the formant plane from trial endpoints to the target vowel centers. We also recorded movement time to target and target production duration as measures of performance. Participants were able to produce any combination of formant values, including those not specifically trained. Therefore, participants were able to produce sounds in the AE category in addition to those in the trained vowel categories (IY, AA, and UW). As a result, the task was truly a four alternative task, and participants were required to avoid AE productions for accurate performance.

We evaluated the statistical significance of each performance measure (accuracy, endpoint distance, movement time, and duration) using a three-factor mixed models analysis of variance to determine the effects of feedback (between-subjects, levels: AO, VO, AV), sessions (within-subjects, levels: 1, 2, 3), and runs within a session (within-subjects, levels: 1 – 4). Participants were included in all models as a random factor. Statistical significance was assessed with a criteria of $\alpha < 0.05$, and any significant effects were further analyzed using pairwise multiple comparisons with the Tukey correction.

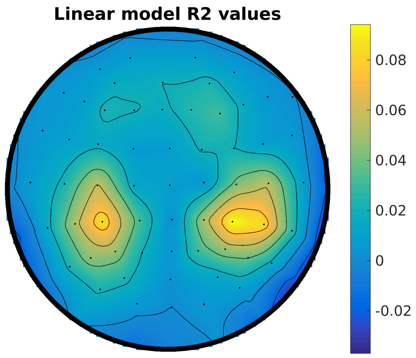


Fig. 3. The contributions of each electrode to the Kalman filter likelihood model, r^2, are plotted on a topographic representation of the scalp. The anterior scalp is toward the top of the image and the posterior toward the bottom. The left and right scalp are on the left and right, respectively.

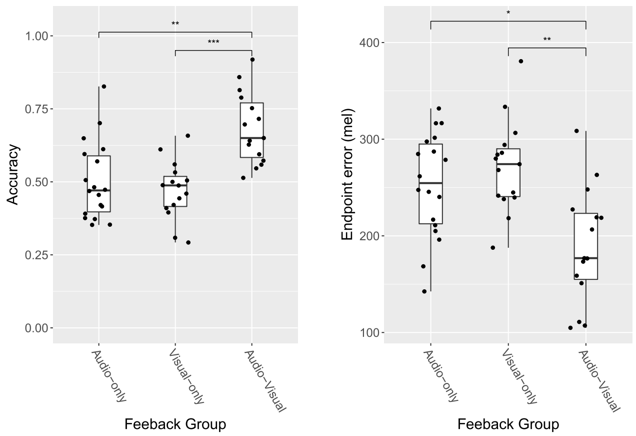
III. Results

A. SMR Bands and Model Training

The μ band was not observed for two participants (P02, P17), and the power spectral density was similar in shape to a $1/F$ distribution (F is frequency). These participants were excluded from any further analysis. Additionally, the β rhythm was not observed in ten participants in at least one experimental session. The Kalman filter decoder was trained only on μ band features if the β band was not observed in order to estimate the most parsimonious decoding model, and on their combination otherwise. The average range of the μ rhythm was 8.13 – 13.31 Hz and 17.60 – 25.50 Hz for the β rhythm. Figure 3 shows the r^2 values for each μ band channel in the Kalman filter likelihood model fit for one participant. In this example, electrodes that contributed most to decoding were over the left and right sensorimotor regions and midline frontal regions on the scalp.

B. BCI Performance

1) Accuracy and Endpoint Distance: A generalized linear mixed model, with a logit link function, was used to assess the effects of feedback type, runs, and sessions on BCI production accuracy. We first determined that an interaction model (feedback x session and feedback x run) did not provide any additional explanatory power over a main effects model alone ($\chi^2(10) = 7.94$, $p = 0.63$), therefore, we examined the mixed model without interactions. Examination of the fixed factors revealed a main effect of feedback type ($\chi^2(2) = 17.70$, $p < 0.001$) but not for sessions or runs. A *post-hoc* Tukey's multiple comparisons test indicated that the AV group performed with greater accuracy than both the



(a) (b)

Fig. 4. (a) Classification accuracy of produced vowels, and (b) Endpoint error from the end of a vowel production to the center of the target vowel. Boxplots show the interquartile range with median values per feedback group and data points represent the per session performance of each participant. ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$.

VO ($z = 3.85, p < 0.001$) and AO ($z = 3.48, p < 0.01$) groups (average accuracy, VO: 47.2%, AO: 50.1%, AV: 68.3%). We also found an average improvement of 3.0% accuracy by including only features that were observed (i.e., based on μ-band features only if the β rhythm was not present). A linear mixed model of endpoint distance from target revealed similar results, with a statistically significant effect of feedback type ($F(2, 13.17) = 5.77$, $p < 0.05$) only. Multiple comparisons testing revealed smaller endpoint distances for the AV group compared to the VO ($z = -3.23$, $p < 0.01$) and AO ($z = -2.59$, $p < 0.05$) groups, indicating AV productions more closely approached the center of the vowel target region than the other two groups. These results are shown in Figures 4a and 4b. An example of BCI productions are shown in Figure 2 with individual trial productions plotted over vowel regions identified by the black line borders: top-left (IY), bottom-left (UW), bottom-right (AA), and top-right (AE). In this example most productions were in the direction of the target vowel, though some mismatches are noted. A video of example participant trials with audio-visual feedback is included in the Supplementary Material.

2) Movement Time and Production Duration: Movement time for accurate productions was examined using a linear mixed model to determine the effects of feedback type, number of runs, and sessions. We first log-transformed movement time to improve data conditioning for normality assumptions of the statistical test. We then determined that an interaction model did not provide any additional explanatory model over a main effects model alone ($\chi^2(10) = 13.14$, $p = 0.22$). The results of the main effects analysis revealed a statistically significant effect of run ($F(3, 108.66) = 13.04$, $p < 0.001$), and no other statistically significant effects of feedback type and session number. A pairwise comparison of the number of runs using Tukey's correction for multiple comparisons indicated that participants reduced movement time as the number of runs within a session increased (Run 2 < Run 1, $p < 0.001$; Run 3 < Run 1, $p < 0.001$; Run 4 < Run 1, $p < 0.001$). A linear mixed model analysis of production duration indicated no differences between any of the main effects.

IV. Discussion

We examined the performance of a formant frequency speech synthesizer BCI with instantaneous auditory and visual feedback. All of the participants in this study were naïve to BCI methods and none had prior BCI experience. The results are discussed below with respect to our initial hypotheses regarding the effect of feedback modality on BCI performance, and in relation to prior reports of 1D and 2D cursor control BCIs with visual and/or audio feedback [24]–[26].

A. Effect of Feedback Modality

Overall, we found that receiving either visual or audio feedback alone was equally useful for speech synthesizer BCI control, but resulted in lower performance compared to combined audio-visual feedback. Both participant groups receiving unimodal feedback were able to operate the BCI with similar accuracy (VO: 47.2%, AO: 50.1%) in our four-choice task (producing either an IY, AA, or UW, and avoiding AE yields a 25% chance rate) and production endpoint distance (VO: 272 mel, AO: 253 mel).[3] Participants who received multimodal audio-visual feedback outperformed both the unimodal feedback groups in terms of percent accuracy (68.3%) and production endpoint distance (190 mel). From these results we can conclude that multimodal feedback was beneficial for successful continuous control of the speech synthesizer BCI, and moreso than either unimodal visual or auditory feedback.

The results of the present study are consistent with prior results from two studies on the effects of multimodal, audio-visual feedback on formant speech synthesizer BCI control [22], [23], [32]. In the first study, a participant with locked-in syndrome controlled a formant speech synthesizer BCI using an intracortical microelectrode implant [22], [32]. In the second experiment, participants without neurological impairments used bilateral orofacial sEMG to control a formant speech synthesizer with either audio feedback alone, or with combined audio-visual feedback [23]. In both studies, visual feedback was similar to the present study (e.g., a moving cursor on the 2D formant plane), and audio feedback was provided by synthesized vowel sounds based on instantaneous formant estimates. The implant participant in the first study learned to control the speech synthesizer with nearly 70% accuracy, and reduced endpoint error to 233 Hz over 25 sessions [22], and in the second study, the participants who received audio-visual feedback achieved 88% average accuracy during training, while the audio-only group reached just 49%. The average BCI performance of the AV group in the current study was similar to those found in the implant study for both accuracy and endpoint distance to target (190 mel = 128 Hz), and the AO group in the current study reached an average accuracy similar to that reported by the sEMG participants.

[3] The Mel-scale is a logarithmic transformation of frequency that is matched to the just noticeable difference in pitch perception in humans [17].

Both previous studies of formant synthesizer control examined the effect of unimodal versus multimodal feedback on performance. Guenther *et al.* [22]reported that in some sessions only auditory feedback was provided for controlling the synthesizer, and that performance was not statistically significantly different from sessions with audio-visual feedback. Hands *et al.* [23] examined this effect more rigorously by training some participants with unimodal audio feedback and others with multimodal, audio-visual feedback, and found enhanced performance during multimodal feedback training. In later sessions, visual feedback was incrementally removed from participants who received audio-visual feedback, which resulted in an equalization of accuracy between groups due to a reduction in performance from participants trained with audio-visual feedback [23]. Hands *et al.* [23] speculate that the decrease in performance by the audio-visual feedback group was due to an over-reliance on visual information for controlling the sEMG synthesizer, but note the utility of audio feedback In the present study, our finding of equal performance between the unimodal feedback groups suggests no effect of over-emphasis of one modality over another.

Two other studies have examined continuous BCI control with auditory feedback of performance using EEG via sensorimotor rhythm [25] and slow cortical potentials [26]. Nijboer *et al.* [25] investigated auditory feedback using a two-choice continuous motor BCI with unimodal feedback of either visual (cursor position) or auditory (harp or bongo sounds corresponding to changes in the sensorimotor rhythm) information. Participants in this study who received visual feedback reached greater performance levels than those receiving auditory feedback, with an average percent accuracy of 74.1% and 56.0% respectively [25]. However, Nijboer *et al.* [25] note that this asymmetry was largely due to differences in the first six of nine recording blocks, after which performance was not statistically different between the participant groups (last three blocks accuracy – visual: 71.0%, auditory: 64.3%). Furthermore, they note that by the final block, four participants in each of the feedback conditions surpassed the 70% accuracy benchmark for a binary selection task [25]. Hinterberger *et al.* [26] participants received either audio, visual, or combined audio-visual feedback of the slow cortical potential amplitude. The visual feedback was a 1D cursor controlled in the vertical dimension to match one of two vertically arranged targets (similar to [25]). The audio feedback consisted of high and low pitch piano tones that were rapidly played in response to changes in the polarity of the slow cortical potential. In this study, though some participants in each condition achieved successful operation of the BCI (out of 16 participants per group, six obtained 70% accuracy with visual feedback, four with audio feedback, and two with combined feedback), the evidence suggests that visual feedback was superior [26]. Further, the authors explain that the audio feedback of piano tones may have engaged neural processing for recognizing harmonies and melodies in the resultant feedback signal, leading to reduced efficacy of audio or audio-visual feedback. The use of speech-related multimodal feedback in the current study appears to have led to enhanced learning, rather than reduced, or equal performance found in other audio-visual feedback BCI paradigms [22], [26].

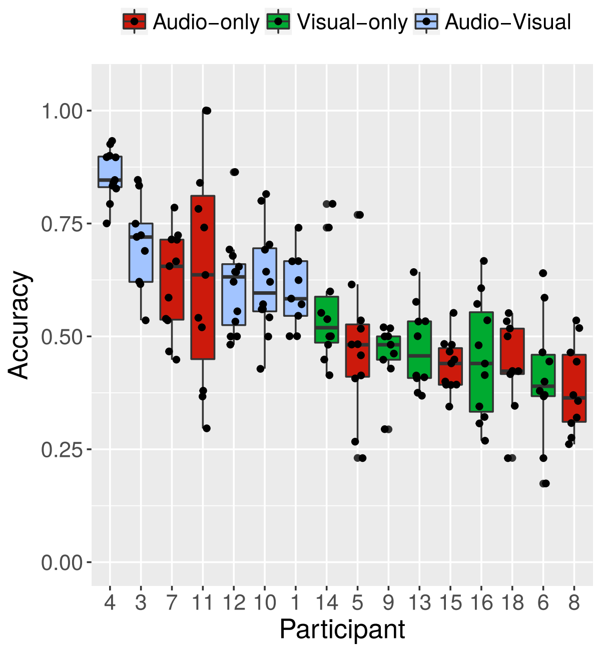


Fig. 5. Percent accuracy for each participant; boxplots are color-coded for the audio-visual (AV), audio alone (AO), or visual alone (VO) feedback groups.

The present study differs in a few key areas and extends these past approaches to decoding in two dimensions rather than one (similar to [24]), and replaces generic auditory feedback (harps & bongos [25]; piano tones [26]) with speech-related feedback that is meaningful for brain networks involved in perception of one's own speech [23]. Finally, the use of speech-related audio feedback is hypothesized to engage neural processes in support of the vowel production task rather than in conflict to enhance performance.

B. Effect of Practice/Fatigue

We found no effects of session for any of our performance measures indicating that participants, independent from feedback modality, did not show evidence of learning over the three recording sessions. This result is in contrast to general findings of improvements to BCI performance across sessions [22], [23], [26], though notably we did not see a decrease in performance either. We did find a reduction in movement time within sessions (similar to [22]) suggesting a short-term improvement. Since we did not find any effects of sessions, a first-last analysis as in [25] was not warranted; however, when examining the range of performance we found nine participants with percent accuracy over the two-class 70% benchmark (all five AV participants, three AO, and one VO) in their best run, eight participants with mean accuracy greater than 50%, and all participants performed with at least 50% accuracy in at least one run (see Table I and Figure 5).

Hinterberger *et al.* [26] note that it is common for BCI performance to remain unchanged within the first three training sessions in the slow cortical potential paradigm,

TABLE I

SUMMARY OF PERCENT ACCURACY FOR EACH PARTICIPANT OVER ALL SESSIONS AND RUNS

Subj	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	18
Group	AV	AV	AV	AO	VO	AO	AO	VO	AV	AO	AV	VO	VO	AO	VO	AO
Mean	60.0	70.4	85.8	47.1	40.6	62.1	38.6	46.1	62.0	64.6	61.4	47.6	55.7	43.6	45.1	43.8
Min	50.0	53.6	75.0	23.1	17.4	44.8	26.1	29.4	42.9	29.6	48.1	36.8	41.4	34.5	26.9	23.1
Max	**74.1**	**84.6**	**93.3**	**76.9**	64.0	**78.6**	53.6	52.0	**81.5**	**100.0**	**86.4**	64.3	**79.3**	55.2	66.7	55.2

with learning occurring over longer durations. Similarly, the accuracy reported for the first three sessions in [22, Fig. 7D] do not appear to change, with greater reductions in endpoint error attributed to long-term learning over 25 sessions. Further, the participants in [24] achieved high accuracy in a 2D cursor control task with unimodal visual feedback after considerable training (between 22 – 68 sessions totaling 8.8 – 27.2 hours), and all participants had previous BCI experience (between 19 – 91 hours). It is possible that additional training is needed [33] in our paradigm to see the effects of long-term learning on intersession performance.

C. Alternative to Existing BCIs for Communication

Nearly all non-invasive BCIs for communication employ typing or spelling for communicative output (e.g., [7], [13]; see [3] for a review), in which vocal output is possible via text-to-speech synthesis. This type of approach draws from existing practices in augmentative and alternative communication for facilitating message creation through selection of letters, words, or phrases, or via symbols for those who lack sufficient literacy skills [4], [5]. On the other hand, intracortical BCIs for communication are increasingly focused on restoring speech directly from neural signals related to speech production [22], [34]–[38]. A small number of non-invasive BCI studies have also focused on direct classification of speech [15], [16], though these attempts have not used real-time instantaneous feedback for either BCI output or as a continuous maintenance signal of performance for improving decoding (cf. real-time cursor position). Receiving timely, task-relevant auditory feedback of speech is critical for learning to control the speech production system, both from a developmental [39] and from a computational modeling perspective [19], [20]. In both examples, auditory feedback serves to alert the speaker to errors in production in order to issue corrective motor commands and learn new speech sounds. There is additional evidence that multimodal audio-visual feedback can further improve speech learning outcomes [40], [41].

In the current study we developed a BCI that addresses each of the issues: (1) direct speech output using a non-invasive, EEG-based sensorimotor paradigm, (2) instantaneous auditory feedback to engage the neural processes involved in speech production, and (3) low-dimensional output that can be used generatively to produce a range of speech sounds. The BCI system examined in this study is similar to the intracortical and sEMG systems described in [22] and [23], but uses the EEG sensorimotor rhythm to control the two formants. Using formants in the BCI speech synthesizer allows for (1) instantaneous synthesis of artificial speech sounds for real-time feedback (e.g., [21]), and (2) an intuitive representation of a range of speech sounds (nine American English vowels can be produced with just F1 and F2, diphthongs are created by movements from one (F1,F2) location to another [17]).

The speech synthesizer used in the current study is able to continuously play speech sounds with an update overhead less than 30 ms, which is important to avoid any possible disruptions to speech sound production due to delayed auditory feedback [29]. In addition, decoding in a low-dimensional articulatory-acoustic domain is generalizable to all speech sounds, which can be sequenced together to form more complex syllables and words. Discrete speech classification, on the other hand, requires new dictionary items for each new potential selection. The dictionary can be minimized by considering articulatory-acoustic features (e.g., [34], [42]), but still minimally depend on 10–30 different features (e.g., place, manner, voicing, tongue advancement, tongue height), which may be difficult to represent using the information present in EEG signals. In this way, continuous decoding based acoustic or articulatory features may be advantageous for EEG decoding due to the relaxed requirement for only a low number of degrees of freedom, e.g., two or three formants, or 3–5 articulatory features of tongue, jaw and lip movements. The tradeoff using a continuous decoding framework is that generalization will likely require a high degree of skill and extensive training. We have shown that it is possible to control a BCI device using motor-imagery for production of synthesized vowels in real time, and additional research is needed to fully explore participants' ability to generalize control of trained sounds to those untrained (cf. [23]).

V. CONCLUSIONS

We reported on a sensorimotor BCI with real-time auditory and visual feedback of speech sounds that has the potential to serve as a method of communication for individuals with severe neuromotor impairments. The results of our study support past research on continuous-output motor BCIs and specifically support other attempts for instantaneous speech synthesis using biophysiological signals. Interestingly, we found that combined audio-visual feedback was an advantage for successful BCI control, which is in contrast to past studies that found the two modalities may compete with each other and reduce performance. Key to our implementation is a congruence between audio and visual feedback, which supports the overall task goal of speech production. The relatively low number of participants per feedback group was a limitation in our study, though all participants were assigned to groups in a random fashion and a mixed effects analysis was used to account for individual variations. The resulting analyses had sufficient power to detect statistically significant differences in the current cohort despite the limited sample size. Future work with larger numbers of participants over greater BCI training durations will be helpful for identifying any possible changes in BCI performance as a result of feedback-specific BCI learning. In addition, we focused on controlling directional changes in the 2D formant plane with emphasis on three trained vowels, and participants were told to avoid production of a fourth vowel. Future studies should further examine generalization to other, non-trained vowels, without the addition of any other motor-imagery control strategies (i.e., using combinations left-right hand and bilateral foot movements to produce non-trained formants). Additional work is also needed to examine the possibility of controlling a synthesizer capable of production consonants as well as vowels [43], and on using speech-motor imagery (of the jaw, lips, and tongue), rather than limb-motor imagery, to control the speech synthesizer.

References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, “Silent speech interfaces,” *Speech Commun.*, vol. 52, no. 4, pp. 270–287, 2010.

[2] 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, no. 6, pp. 767–791, 2002.

[3] J. S. Brumberg, K. M. Pitt, A. Mantie-Kozlowski, and J. D. Burnison, “Brain–computer interfaces for augmentative and alternative communication: A tutoria,” *Amer. J. Speech-Lang. Pathol.*, vol. 27, no. 1, pp. 1–12, 2018.

[4] S. Fager, D. Beukelman, M. Fried-Oken, T. Jakobs, and J. Baker, “Access interface strategies,” *Assist. Technol.*, vol. 24, no. 1, pp. 25–33, 2012.

[5] D. Beukelman and P. Mirenda, *Augmentative and Alternative Communication: Supporting Children and Adults With Complex Communication Needs*, 4th ed. Baltimore, MD, USA: Paul H. Brookes, 2013.

[6] F. Plum and J. B. Posner, *The Diagnosis of Stupor and Coma*. Philadelphia, PA, USA: F.A. Davis.

[7] E. Donchin, K. M. Spencer, and R. Wijesinghe, “The mental prosthesis: Assessing the speed of a P300-based brain-computer interface,” *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 2, pp. 174–179, Jun. 2000.

[8] E. W. Sellers and E. Donchin, “A P300-based brain–computer interface: Initial tests by ALS patients,” *Clin. Neurophysiol.*, vol. 117, no. 3, pp. 538–548, 2006.

[9] E. M. Holz, L. Botrel, T. Kaufmann, and A. Kübler, “Long-term independent brain-computer interface home use improves quality of life of a patient in the locked-in state: A case study,” *Arch. Phys. Med. Rehabil.*, vol. 96, no. 3, pp. S16–S26, 2015.

[10] O. Friman, T. Luth, I. Volosyak, and A. Graser, “Spelling with steady-state visual evoked potentials,” in *Proc. 3rd Int. Conf. Neural Eng. (EMBS)*, Kohala Coast, HI, USA, May 2007, pp. 354–357.

[11] A. Kübler *et al.*, “The thought translation device: A neurophysiological approach to communication in total motor paralysis,” *Experim. Brain Res.*, vol. 124, no. 2, pp. 223–232, 1999.

[12] G. Purtscheller and C. Neuper, “Motor imagery and direct brain-computer communication,” *Proc. IEEE*, vol. 89, no. 7, pp. 1123–1134, Jul. 2001.

[13] B. Blankertz *et al.*, “The berlin brain-computer interface: EEG-based communication without subject training,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 147–152, Jun. 2006.

[14] S. Chakrabarti, H. M. Sandberg, J. S. Brumberg, and D. J. Krusienski, “Progress in speech decoding from the electrocorticogram,” *Biomed. Eng. Lett.*, vol. 5, no. 1, pp. 10–21, 2015.

[15] P. Suppes, Z.-L. Lu, and B. Han, “Brain wave recognition of words,” *Proc. Nat. Acad. Sci. USA*, vol. 94, no. 26, pp. 14965–14969, 1997.

[16] C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike, “Single-trial classification of vowel speech imagery using common spatial patterns,” *Neural Netw.*, vol. 22, no. 9, pp. 1334–1339, 2009.

[17] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA, USA: MIT Press, 2000.

[18] P. Indefrey and W. J. M. Levelt, “The spatial and temporal signatures of word production components,” *Cognition*, vol. 92, nos. 1–2, pp. 44–101, 2004.

[19] E. Golfinopoulos, J. A. Tourville, and F. H. Guenther, “The integration of large-scale neural network modeling and functional brain imaging in speech motor control,” *NeuroImage*, vol. 52, no. 3, pp. 862–874, 2010.

[20] G. Hickok, J. F. Houde, and F. Rong, “Sensorimotor integration in speech processing: Computational basis and neural organization,” *Neuron*, vol. 69, no. 3, pp. 407–422, 2011.

[21] D. H. Klatt, “Software for a cascade/parallel formant synthesizer,” *J. Acoust. Soc. Amer.*, vol. 67, no. 3, p. 971, 1980.

[22] F. H. Guenther *et al.*, “A wireless brain-machine interface for real-time speech synthesis,” *PLoS ONE*, vol. 4, no. 12, p. e8218, 2009.

[23] G. L. Hands, E. Larson, and C. E. Stepp, “Effects of augmentative visual training on audio-motor mapping,” *Hum. Movement Sci.*, vol. 35, pp. 145–155, Jun. 2014.

[24] J. R. Wolpaw and D. J. McFarland, “Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans,” *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 51, p. 17849, 2004.

[25] F. Nijboer *et al.*, “An auditory brain–computer interface (BCI),” *J. Neurosci. Methods*, vol. 167, no. 1, pp. 43–50, 2008.

[26] T. Hinterberger *et al.*, “A multimodal brain-based feedback and communication system,” *Experim. Brain Res.*, vol. 154, no. 4, pp. 521–526, 2004.

[27] R. Oostenveld and P. Praamstra, “The five percent electrode system for high-resolution EEG and ERP measurements,” *Clin. Neurophysiol.*, vol. 112, no. 4, pp. 713–719, 2001.

[28] S. Halder *et al.*, “Online artifact removal for brain-computer interfaces using support vector machines and blind source separation,” *Comput. Intell. Neurosci.*, vol. 2007, 2007, Art. no. 82069. doi: 10.1155/2007/82069.

[29] G. Fairbanks and N. Guttman, “Effects of delayed auditory feedback upon articulation,” *J. Speech Hear Res.*, vol. 1, no. 1, pp. 12–22, 1958.

[30] P. Iverson and P. K. Kuhl, “Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling,” *J. Acoust. Soc. Amer.*, vol. 97, no. 1, pp. 553–562, 1995.

[31] G. E. Peterson and H. L. Barney, “Control methods used in a study of the vowels,” *J. Acoust. Soc. Amer.*, vol. 24, no. 2, pp. 175–184, 1952.

[32] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther, “Brain-computer interfaces for speech communication,” *Speech Commun.*, vol. 52, no. 4, pp. 367–379, 2010.

[33] J. N. Mak and J. R. Wolpaw, “Clinical applications of brain-computer interfaces: Current state and future prospects,” *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 187–199, 2009.

[34] J. S. Brumberg, E. J. Wright, D. S. Andreasen, F. H. Guenther, and P. R. Kennedy, “Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex,” *Frontiers Neurosci.*, vol. 5, p. 65, May 2011.

[35] S. Kellis, K. Miller, K. Thomson, R. Brown, P. House, and B. Greger, “Decoding spoken words using local field potentials recorded from the cortical surface,” *J. Neural Eng.*, vol. 7, no. 5, p. 056007, 2010.

[36] E. M. Mugler *et al.*, “Direct classification of all American English phonemes using signals from functional speech motor cortex,” *J. Neural Eng.*, vol. 11, no. 3, p. 035015, 2014.

[37] S. Martin *et al.*, “Decoding spectrotemporal features of overt and covert speech from the human cortex,” *Frontiers Neuroeng.*, vol. 7, p. 14, May 2014.

[38] C. Herff *et al.*, “Brain-to-text: Decoding spoken phrases from phone representations in the brain,” *Frontiers Neurosci.*, vol. 9, pp. 1–11, Jun. 2015.

[39] D. K. Oller and R. E. Eilers, “The role of audition in infant babbling,” *Child Develop.*, vol. 59, no. 2, pp. 441–449, 1988.

[40] W. F. Katz and S. Mehta, “Visual feedback of tongue movement for novel speech sound learning,” *Frontiers Hum. Neurosci.*, vol. 9, p. 612, Nov. 2015.

[41] A. Suemitsu, J. Dang, T. Ito, and M. Tiede, “A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning,” *J. Acoust. Soc. Amer.*, vol. 138, no. 4, pp. EL382–EL387, 2015.

[42] F. Lotte *et al.*, “Electrocorticographic representations of segmental features in continuous speech,” *Frontiers Hum. Neurosci.*, vol. 9, p. 97, Feb. 2015.

[43] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, “Real-time control of an articulatory-based speech synthesizer for brain computer interfaces,” *PLoS Comput. Biol.*, vol. 12, no. 11, p. e1005119, 2016.