EEG and Eye-Tracking Error-Related Responses During Predictive Text Interactions: A BCI Case Study

Sophia K. Mehdizadeh[1], Edward Cutrell[2], R. Michael Winters[2], Nemanja Djuric[3], Yang Cheng[4], Ivan J. Tashev[2], and Yu Te Wang[2*]

Abstract— **Brain-computer interfaces (BCIs) employ various paradigms which afford intuitive, augmented control for users to navigate digital technologies. In this study we explore the application of these BCI concepts to predictive text systems: commonplace interactive and assistive tools with variable usage contexts and user behaviors. We conducted an experiment to analyze user neurophysiological responses under these different usage scenarios and evaluate the feasibility of a closed-loop, adaptive BCI for use with such technologies. We recorded electroencephalogram (EEG) and eye tracking (ET) data from participants while they completed a self-paced typing task in a simulated predictive text environment. Participants completed the task with different degrees of reliance on the predictive text system (completely dependent, completely independent, or their choice) and encountered both correct and incorrect text generations. Data suggest that erroneous text generations may evoke neurophysiological responses that can be measured with both EEG and pupillometry. Moreover, these responses appear to change according to users' reliance on the predictive text system. Results show promise for use in a passive, hybrid, BCI with a closed-loop, adaptive framework, and support a neurophysiological approach to the challenge of real-time human feedback on system performance.**

I. INTRODUCTION

Brain-computer interfaces (BCIs) utilize various physiological modalities, signal analysis paradigms, and system frameworks [1] allowing users to interact with computer-based systems through their neurophysiological processes. In this study, we explore the application of these concepts to commonplace interactive systems within digital technologies. Given that interest in BCI and increasingly interactive, intelligent human-computer interfaces continues to grow [2], we conduct a case study analyzing user neurophysiological responses while interacting with such systems. For this case study, we focus on predictive text: a widely implemented assistive tool within many of today's digital technologies. Yet many choose not to use predictive text, which may reflect the variable usage contexts and individual user differences of these systems [3], [4]. We design a study to explore these scenarios through individuals' behavioral and neurophysiological responses and discuss the application of our findings within the context of a passive BCI and closed-loop, adaptive framework.

Passive BCIs operate under a cognitive monitoring-style paradigm in which qualities of a user's mental state are inferred through real-time analysis of their physiological signals [5]. The type of analysis we will focus on is detection of error-related potentials (ErrPs). The ErrP is a signal pattern studied in electroencephalogram (EEG) data and a popular paradigm of passive BCIs [1]; studies have found this potential is evoked soon after a person makes an error [6], observes another party making an error [7], or interacts with a system which misinterprets their intentions [8], [9]. Our aim is first to evaluate if an ErrP is evoked in participants' EEG when an intelligent text entry system incorrectly predicts their intended message. Furthermore, we explore if a related response can be found in participants' eye tracking (ET) data and discuss the feasibility of a passive, EEG-ET hybrid BCI [1]. Findings that support real-time error detection and classification of individuals' system usage could be implemented in a closed-loop framework, passively providing real-time human feedback to improve individuals' interactions with these systems over time [7], [10].

II. METHOD

A. Stimulus and Task Preparation

This study uses a simulated predictive text generation interface developed using PsychoPy [11]. The sentences participants type into the interface are experimentally controlled, and the text "predictions" participants encounter as they are typing are generated pseudo-randomly under certain criteria chosen to mimic interactions of real predictive text systems. 450 common English sentences, ranging 4-8 words long, are adapted from an in-lab dataset. We use the spaCy Python library to identify 1) nouns, verbs, adjectives, and adverbs in each sentence that are 2) more than three letters long, and that are 3) not the first word of the sentence. These words are the set of possible correct ("match") generations made by the simulated predictive text environment. For some trials, the environment may instead generate incorrect ("mismatch") words, which are selected from a set of different words of the same part of speech.

B. Data Acquisition

Ten healthy adults, fluent in English and with normal or corrected-to-normal vision, are compensated for participating in this experiment. All participants sign an informed consent form approved by the Microsoft Research Ethics Review Program Team. We record 32-channel EEG using the Brain Products LiveAmp system with active gel electrodes at a

[1] S. K. Mehdizadeh is with the ATLAS Institute, University of Colorado Boulder, Boulder, CO 80309 USA. This work was done as a research intern at Microsoft Research, Redmond, WA 98052, USA

[2] Y. T. Wang (*corresponding author, yutewang65@gmail.com), E. Cutrell, R. M. Winters, and I. J. Tashev are with Microsoft Research, Redmond, WA 98052 USA

[3] N. Djuric is with Microsoft, Belgrade, 11000, Serbia

[4] Y. Cheng is with Microsoft, Redmond, WA 98052 USA

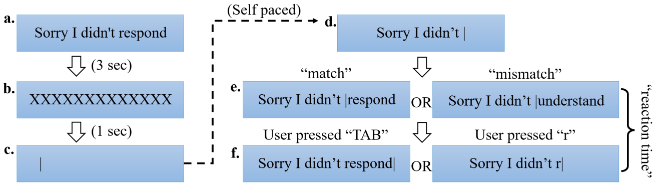

Fig. 1. Experiment setup. All data, keystrokes, and stimulus timing information are recorded simultaneously using Lab Streaming Layer.

sampling rate of 500 Hz, and ET data using a screen-mounted Tobii Pro Nano at a sampling rate of 60 Hz. During the experiment, participants sit at a desk in an artificially-lit room with a computer monitor approximately 24 inches away. They use a standard QWERTY keyboard and their head is supported by a chin rest. Participants are never told the system is not a real predictive text algorithm.

At the start of each trial one of the 450 stimulus sentences is displayed inside an un-editable textbox for three seconds (Fig. 1a), then visually masked [12] for one second (Fig. 1b). Once the mask disappears (Fig. 1c), participants type the target sentence verbatim (Fig. 1d). For each trial, the experiment script randomly selects one suitable position in the sentence (defined above) to generate a "prediction" as the participant is typing. Approximately 50% of the time the generated word will match what the participant is attempting to type, while the other 50% of the time the word will be a mismatch (Fig. 1e). Participant behavior following text generation is governed by one of three possible "user scenarios" given to participants at the start of each run:

- Dependent (reliant): Participants evaluate if the prediction is correct, and if so, always press TAB to accept it (Fig. 1f, left). If incorrect, then they reject it by continuing to type the correct word (Fig. 1f, right).
- Independent (not reliant): Participants still observe generated predictions, but reject them all and type the word themselves, even if the prediction is correct.
- Free choice: Participants decide for themselves how they would like to use (or not use) the predictive text.

Each participant completes three runs (50 sentences per run) under each user scenario (3 runs x 3 scenarios = 9 runs total). Participants are instructed to keep their typing speed and pattern consistent regardless of the user scenario.

C. Data Analysis

Bad trials, including system lag > 100 msec in displaying the visual stimulus [13] or participants not engaging with the text generation as instructed, are removed from analysis. From the logged keystrokes, we calculate participant reaction time (RT) for each trial as the time between text generation appearing on the screen to the next key press (Fig. 1f).

EEG data are notch filtered at 60, 120, 180, and 240 Hz to remove powerline noise, and bandpass filtered from 1-10 Hz [6], [7], [8]. EEG channels are re-referenced to an average reference. One epoch is extracted from each trial, starting 100 msec prior to text generation and ending 1000 msec after text generation. Currently, we examine the epochs from channel "Fz" only. Prior work has consistently characterized ErrP by fronto-central midline channels [6], [7], [8]. From the recorded ET features, we focus on pupil diameter. Pupil data are preprocessed by a moving average filter with 50-msec window size [14]. One epoch is extracted from each trial, starting 100 msec prior to text generation and ending 2000 msec after text generation. Bad trials identified from the behavioral analysis, and epochs where more than 500 msec of ET data are missing are discarded from both physiological analyses (479 match, 268 mismatch trials). Three participants' data are excluded due to excessive EEG artifacts. Any missing datapoints in the remaining valid ET epochs are linearly interpolated. From the cleaned ET data, we calculate participants' percent change in pupil diameter (PCPD) at each timepoint, t, using their average pupil diameter across all runs (μ) as their baseline [15]. For example, the PCPD for participant p is illustrated in (1).

$$PCPD_p(t) = \frac{Pupil\ diam_p(t) - \mu_p}{\mu_p} * 100\% \qquad (1)$$

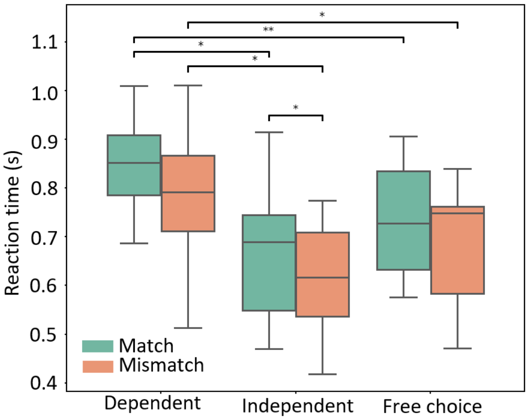

Fig. 2. Differences in RT between three user scenarios (x-axis) and match (green) vs. mismatch (orange) text conditions. Statistically significant results from pairwise tests (with Holm-Bonferroni corrections) are indicated [(*) p<0.05; (**) p<0.01].

III. Results

A. Reaction Time

We average RTs within match and mismatch conditions of the three user scenarios for each participant (Fig. 2). A two-way repeated measures ANOVA reveals significant overall effect on RT of the user scenario [$F(2,16) = 11.89$, $p<0.001$] and whether the generated word is a match or mismatch to the target sentence [$F(1,8) = 13.73$, $p = 0.006$], with no significant interaction effects [$F(2,16) = 0.49$, $p = 0.621$]. When participants are heavily relying (dependent) on the predictive text, their response patterns can be clearly distinguished from when they are not relying [match: $t(8) = 3.28$, $p = 0.045$; mismatch: $t(8) = 4.48$, $p = 0.010$], or only partially relying [match: $t(8) = 4.72$, $p = 0.009$; mismatch: $t(8) = 3.14$, $p = 0.045$], on the system. Furthermore, trends suggest that participants have faster RTs to incorrect compared to correct text generations across all user scenarios [dependent: $t(8) =$

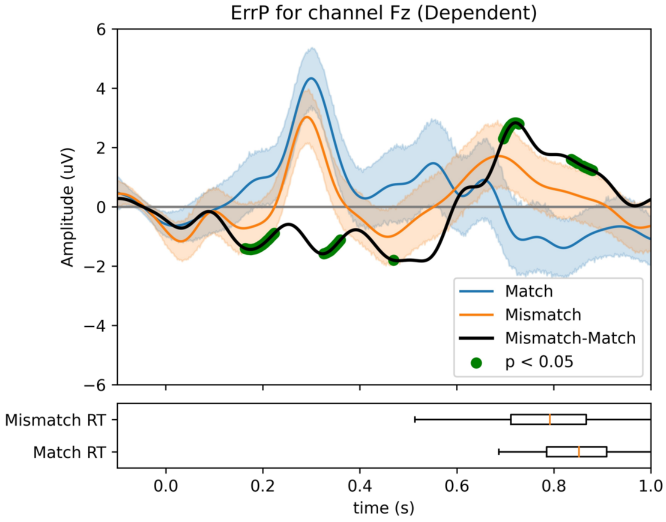

Fig. 3. Dependent user scenario only. Top: Grand average ERPs for match (blue) and mismatch (orange) conditions, with 95% confidence intervals, and mismatch-minus-match difference ERP (black). Timepoints where match and mismatch ERPs differ significantly are highlighted in green (WSR test, corrected). Bottom: RTs shown in reference to ERPs.

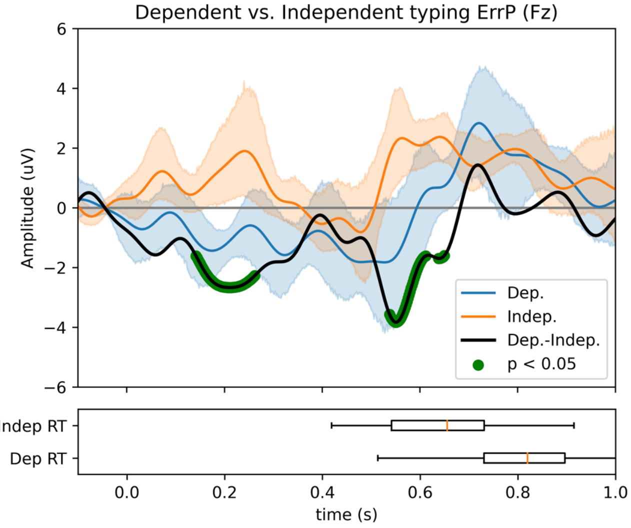

Fig. 4. Top: Grand average difference ERPs (mismatch-minus-match) for dependent (blue) and independent (orange) scenarios with 95% confidence intervals. Dependent-minus-independent difference ERP in black. Timepoints where dependent and independent ERPs differ significantly are in green (WSR test, corrected). Bottom: RTs shown in reference to ERPs.

2.56, p = 0.067; independent: $t(8)$ = 3.64, p = 0.020; free choice: $t(8)$ = 2.26, p = 0.067].

B. EEG ErrP

We first analyze data from the dependent user scenario only (when participants are most reliant on the text generations) to determine if an ErrP is evoked by mismatch text predictions. Fig. 3 shows the grand average Event-Related Potentials (ERPs) for match and mismatch conditions within this scenario, and the mismatch-match difference. Although the majority of the response does not reach statistical significance, some key features can be identified. In the difference ERP we see a broad negative deflection around 480 msec and a broad positive peak around 720 msec after text generation, both of which are ErrP features that have been identified in prior studies on interaction-specific ErrPs [8], [9]. Preceding this are two positive peaks at about 250 and 390 msec (do not reach statistical significance according to the Wilcoxon signed rank test) and two negative peaks at about 170 and 320 msec (significant, $p<0.05$), again resembling the interaction ErrP finding from [9]. Next, we explore how these features change when participants are in the independent user scenario (Fig. 4). Since these trials differ in participants' reliance on the text generations to complete the task, we expect a difference in how they perceive system errors. A Wilcoxon signed rank (WSR) test identifies two windows (140-260 msec and 540-650 msec) during which the dependent and independent ErrP differ significantly. Although statistical significance is limited, results suggest that features of the EEG response may vary according to participants' reliance on the system.

C. Pupil Dilation Response

Fig. 5 shows participants' PCPD averages from the dependent and independent user scenarios. In the dependent scenario, before any participant keyboard response, we see evidence of greater pupil dilation for generated mismatches, and constriction for matches (this difference does not reach statistical significance according to the WSR test). In contrast, there is almost no difference between these conditions at this time in the independent scenario, suggesting a pupillometry response to erroneous text during greater system reliance. After the keyboard response, we must consider an important detail for the dependent scenario: a generated match word will remain on the screen after the user responds with TAB, but a mismatch word will disappear when the user types the first letter of the correct word (Fig. 1f). It is possible that this visual difference could amount to a difference in luminance, making it unclear how much of the pupil response observed after participants' keypress in the dependent scenario is due to their perception of the generated text versus this potential confounder. However, it is worth noting that there are no such differences in visual presentation for the independent user scenario (regardless of if the generated word is a match or mismatch, participants always type the following letter, and the generated text always disappears from the display). Interestingly, our data still shows a slight difference in pupil response between conditions in the time after participants' keypress.

IV. CONCLUSIONS AND DISCUSSIONS

In this study, we analyzed behavioral and neurophysiological responses while participants completed a self-paced typing task in a simulated predictive text environment with different degrees of reliance on the system (completely dependent, completely independent, or their choice). An analysis of RT supports that participants were in fact significantly altering their behavior according to the reliance prompts they were given, validating the experimental conditions.

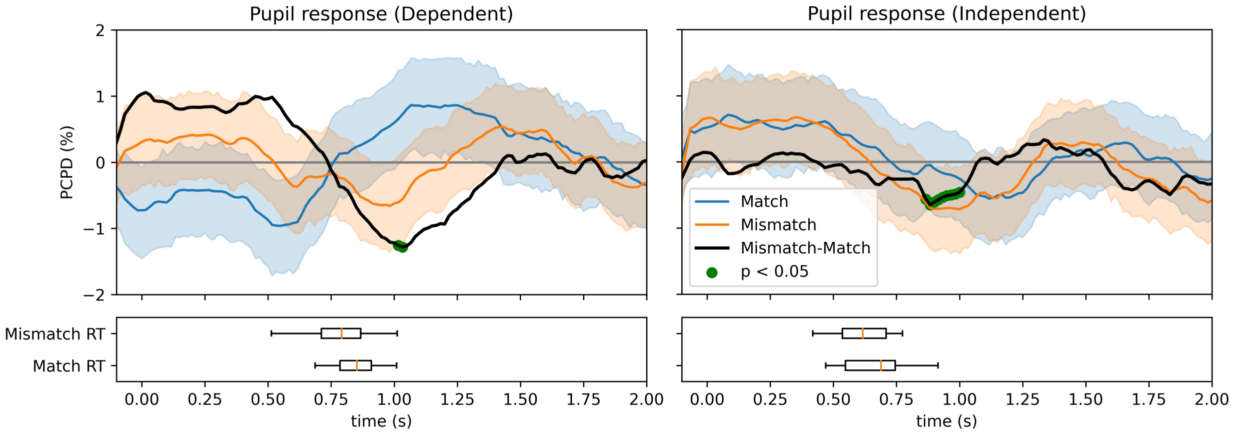

Fig. 5. Top: Grand average PCPD for match (blue) and mismatch (orange) conditions with 95% confidence intervals, and mismatch-minus-match difference (black) for dependent (left) versus independent (right) user scenarios. Timepoints where match and mismatch responses differ significantly are highlighted in green (WSR test, corrected). Bottom: RTs shown in reference to pupil response for dependent and independent scenarios.

Interestingly, our results also show a trend of faster RT to erroneous compared to correct text generations across all conditions. This was particularly unexpected for the independent scenario, where correctness of the generated text should have the least influence over participants' response.

Results from the EEG analysis suggest that incorrect text generations evoke a measurable response resembling interaction-style ErrPs reported by prior studies [8], [9]. Future work will continue to explore how users' reliance on the interactive system may modulate this response. Results from our exploration of the pupil dilation response appear to parallel those from EEG (although less pronounced), suggesting that system misinterpretations may also evoke a measurable pupillometry response that varies according to users' reliance on the system. However, it is not clear from the current experimental paradigm which mechanism(s) underlie this response, and more work is also needed to rule out any possible confound of screen luminance.

Overall, data suggest that the complex usage scenarios of intelligent interactive interfaces may evoke distinct neurophysiological responses that could be used within a passive, EEG-ET hybrid BCI. In future work, we intend to test this by developing an online EEG-ET classifier for system error detection as well as users' system reliance. Prior research in closed-loop, adaptive BCI has demonstrated how such classifiers provide real-time human feedback on system performance, which can then be used to further train and improve the interactive system with continued use [7], [10]. As the interactive complexity of modern technologies continues to increase, so does the amount of training data they need to support those features. Through our case study on predictive text systems, we propose and show support for a BCI approach to this challenge.

References

[1] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of EEG-based brain–computer interface paradigms," J. Neural Eng., vol. 16, no. 1, p. 011001, 2019.

[2] G. A. M. Vasiljevic and L. C. de Miranda, "Brain–Computer Interface Games Based on Consumer-Grade EEG Devices: A Systematic Literature Review," International Journal of Human–Computer Interaction, vol. 36, no. 2, pp. 105–142, 2020.

[3] K. Palin, A. M. Feit, S. Kim, P. O. Kristensson, and A. Oulasvirta, "How do People Type on Mobile Devices?: Observations from a Study with 37,000 Volunteers," in Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services, Taipei Taiwan, 2019, pp. 1–12.

[4] P. O. Kristensson and T. Müllners, "Design and Analysis of Intelligent Text Entry Systems with Function Structure Models and Envelope Analysis," in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama Japan, 2021, pp. 1–12.

[5] T. O. Zander and C. Kothe, "Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general," J. Neural Eng., vol. 8, no. 2, p. 025005, 2011.

[6] C. B. Holroyd and M. G. H. Coles, "The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity," Psychol Rev, vol. 109, no. 4, pp. 679–709, 2002.

[7] R. Chavarriaga and J. del R. Millan, "Learning From EEG Error-Related Potentials in Noninvasive Brain-Computer Interfaces," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2010, vol. 18, pp. 381–388.

[8] P. W. Ferrez and J. del R. Millan, "Error-Related EEG Potentials Generated During Simulated Brain–Computer Interaction," in IEEE Transactions on Biomedical Engineering, 2008, vol. 55, pp. 923–929.

[9] A. F. Salazar-Gomez, J. DelPreto, S. Gil, F. H. Guenther, and D. Rus, "Correcting robot mistakes in real time using EEG signals," in 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017, pp. 6570–6577.

[10] K.-J. Chiang et al., "A Closed-loop Adaptive Brain-computer Interface Framework: Improving the Classifier with the Use of Error-related Potentials," presented at the International IEEE/EMBS Conference on Neural Engineering, 2021.

[11] J. Peirce et al., "PsychoPy2: Experiments in behavior made easy," Behav Res, vol. 51, pp. 195–203, 2019.

[12] S. L. Macknik and M. S. Livingstone, "Neuronal correlates of visibility and invisibility in the primate visual system," Nat Neurosci, vol. 1, no. 2, pp. 144–149, 1998.

[13] S. K. Card, T. P. Moran, and A. Newell, The psychology of human-computer interaction. Hillsdale, New Jersey, USA: Erlbaum, 1983.

[14] J. Klingner, R. Kumar, and P. Hanrahan, "Measuring the task-evoked pupillary response with a remote eye tracker," in Proceedings of the 2008 symposium on Eye tracking research & applications, New York, NY, USA, 2008, pp. 69–72.

[15] J.-L. Kruger, E. Hefer, and G. Matthew, "Measuring the impact of subtitles on cognitive load: eye tracking and dynamic audiovisual texts," in Proceedings of the 2013 Conference on Eye Tracking South Africa, New York, NY, USA, 2013, pp. 62–66.