

AMBER 2.0: A Dataset for Naturalistic Settings with HMD-Based RSVP Tasks [★]

Muhammad Ahsan Awais¹[0000–0001–8722–5787], Tomas
Ward¹[0000–0002–6173–6607], and Graham Healy¹[0000–0001–6429–6339]

Insight Research Ireland Centre for Data Analytics, School of Computing, Dublin
City University, Dublin, Ireland.

`muhammad.awais2@mail.dcu.ie`

Abstract. This paper introduces an EEG (Electroencephalography) dataset collected using the Rapid Serial Visual Presentation (RSVP) paradigm combined with an HMD (head-mounted display), to simulate real-world conditions for Brain-Computer Interfaces (BCIs). EEG signals were recorded under both clean and intentionally contaminated conditions, where participants engaged in activities to intentionally introduce artifacts deleterious to system performance such as head movements, body movements, and talking, simulating real-world sources of noise.

Unlike the conventional RSVP-BCI method that displays stimuli on a monitor, we investigate the use of a head-mounted display, and compare both approaches under clean and noisy conditions to assess their performance differences.

Target versus non-target classification using multiple machine learning models revealed a decline in ROC-AUC scores under noisy conditions compared to clean settings. Additionally, our comparative analysis between HMD-based and monitor-based BCIs, supported by paired t-tests, confirmed that the differences in classification performance were not statistically significant, reinforcing the feasibility of using HMDs for real-world BCI applications.

By incorporating a head-mounted display setup, this dataset provides a valuable resource for developing artifact-resistant Brain-Computer Interfaces applicable to gaming, metaverse environments, and assistive technologies. Researchers can leverage this dataset to explore signal-denoising techniques and enhance the robustness of EEG-based BCIs in naturalistic conditions that use HMDs. The datasets presented in this study can be found at: <https://zenodo.org/records/14750342>

Keywords: EEG · BCI · Artefacts · P300 ERPs · RSVP · Noise · EEG Dataset .

[★] This work was supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289 P2 and CHIST-ERA under grant number (CHISTERA IV 2020 - PCI2021-122058-2A

1 Introduction

Advances in brain-computer interfaces (BCIs) have created opportunities in fields such as assistive technologies, neurorehabilitation, and cognitive state monitoring [1] [2]. However, to extend the utility of BCIs beyond controlled laboratory environments, the need for approaches robust to common sources of noise in EEG signals and datasets to support this has grown significantly. Such datasets are critical for developing models that can generalize to real-world scenarios, where environmental variables and user interactions may differ significantly from traditional laboratory settings.

Considering these requirements, the AMBER (Advancing Multimodal Brain-computer interfaces for Enhanced Robustness — A dataset for naturalistic settings) [3] dataset was developed. This dataset utilizes the Rapid Serial Visual Presentation (RSVP) paradigm as a means of eliciting P300 responses. RSVP is a widely used experimental method in neuroscience and psychology, where a sequence of stimuli (i.e., images) is presented at high speed [4]. Participants are tasked with identifying specific target stimuli embedded within the sequence. RSVP has proven to be an effective tool for eliciting P300 components in EEG, a key feature in BCI systems, due to its ability to reliably evoke distinct neural responses to target and non-target stimuli.

Despite the extensive use of RSVP paradigms for P300 elicitation, most studies rely on traditional computer monitor displays [5] [6] [7]. This experimental setup offers a consistent and familiar visual environment, reducing distractions and delivering accurate data for P300 detection. However, it may not adequately represent real-world conditions, particularly in scenarios requiring mobility or immersion. Real-world applications of BCIs frequently go beyond the boundaries of a standard laboratory setting, necessitating systems that work in dynamic and engaging environments. For example, BCIs designed for immersive gaming or other settings that mimic real-world situations [8] [9] [10].

Building upon the foundation of AMBER, the AMBER 2.0 dataset is introduced in this paper to simulate more immersive real-world scenarios. The primary innovation in AMBER 2.0 is the use of a head-mounted display (HMD) for RSVP image presentation, replacing the traditional monitor setup. This change was motivated by the need to explore BCI performance in virtual and augmented reality environments, which are increasingly relevant in applications such as assistive technology, gaming, and training simulations.

AMBER 2.0 reflects the growing demand for datasets that replicate real-world BCI scenarios. By incorporating HMD-based testing, it facilitates research into BCIs for virtual and augmented reality settings. Moreover, it allows comparative studies on the impact of display technology on BCI performance and enables the development of robust algorithms tailored to immersive environments.

2 Material and Methods

Ten healthy participants (aged 20–35 years) from Dublin City University participated in the study, with ethical approval from the DCU Research Ethics Committee (DCUREC/2021/175). EEG data were recorded using an ANT-Neuro eego sports mobile EEG system with a 32-channel cap based on the 10-20 system. CPz served as the reference channel, and electrode impedance was maintained below 15 k Ω . Data were sampled at 1000 Hz with a 500 Hz lowpass filter and saved in EDF format. The RSVP experiment was displayed to the participants using the "VisionHMD bigeyes H3" head-mounted display device.

AMBER 2.0 includes data collected from four sessions for each participant, where each session contains seven blocks, resulting in a total of 28 blocks across all four sessions. Each session starts with three baseline blocks: eyes open (B1), eye movement (B0), and eyes closed (B2), followed by the RSVP experiments conducted in random order. A description of each task in a single session is given in Table 1.

Table 1. Description of tasks performed in a single session

No.	Task ID	Task Description	Duration
1.	B1	Baseline Eyes Open	60s
2.	B0	Baseline Eyes Movement	10s
3.	B2	Baseline Eyes Close	60s
4.	X2	RSVP Task	90s
5.	X4	RSVP + Body Movement	90s
6.	X6	RSVP + Talking	90s
7.	X8	RSVP + Head Movement	90s

As seen in Table 1, the RSVP experimental sessions are of two different types:

1. The standard RSVP image search paradigm, where subjects perform the target search task while wearing a head-mounted display (X2);
2. The RSVP paradigm with participant-induced noise, where subjects follow a protocol to generate three distinct types of EEG artifacts—head movement, body movement, and talking—while performing the RSVP task (X4, X6, and X8).

Similar to AMBER, a total of 360 images per RSVP block were presented to participants in AMBER 2.0. However, these images were displayed using a head-mounted display instead of a monitor screen. A total of 144/1296 target/standard trials were captured for each intentionally as well as non-intentionally contaminated RSVP task per participant. Additional information on target and non-target markers is detailed in a corresponding CSV file containing 360 rows, with labels "1" indicating the presence of a target image and "2" indicating a non-target image.

An example of a participant wearing the HMD device and performing the experiment is shown in Figure 1. It also illustrates the RSVP task, where images of cars serve as the target class, while other images, such as dogs, cats, and flowers, are used as non-target or standard classes.

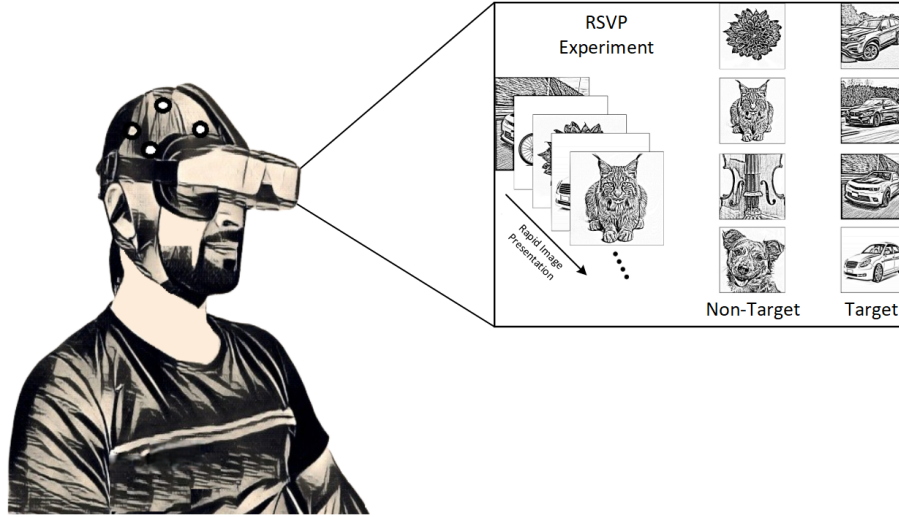


Fig. 1. A participant wearing electrode cap and the HMD device while performing the AMBER 2.0 experiment

2.1 Data Records

The raw EEG data from 10 participants is stored in EDF format, containing the complete EEG signals and task events. Each participant has 7 EDF files (3 baselines and 4 RSVP) per session, totaling 28 files across four sessions. Additionally, RSVP tasks include target and standard image triggers provided in 16 CSV marker files per participant. Together, these EDF and CSV files enable comprehensive analysis of EEG signals and events during traditional and artefactual RSVP tasks.

3 Technical Validation

Ensuring the reliability and consistency of the collected RSVP-P300 dataset is crucial for its usability in further research. This section presents signal quality assessments and classification performance evaluations to validate the dataset under different conditions (e.g. clean versus noise).

We present grand averages across all subjects at channel Pz in Figure 2 to visualize the structure of EEG signals in the P300 RSVP paradigm. This figure compares the traditional RSVP condition with the three intentionally contaminated conditions—body movement, head movement, and talking artifacts—displaying all four conditions within a single plot. In this representation, solid lines represent the grand average of target epochs, while dotted lines represent the grand average of standard epochs.

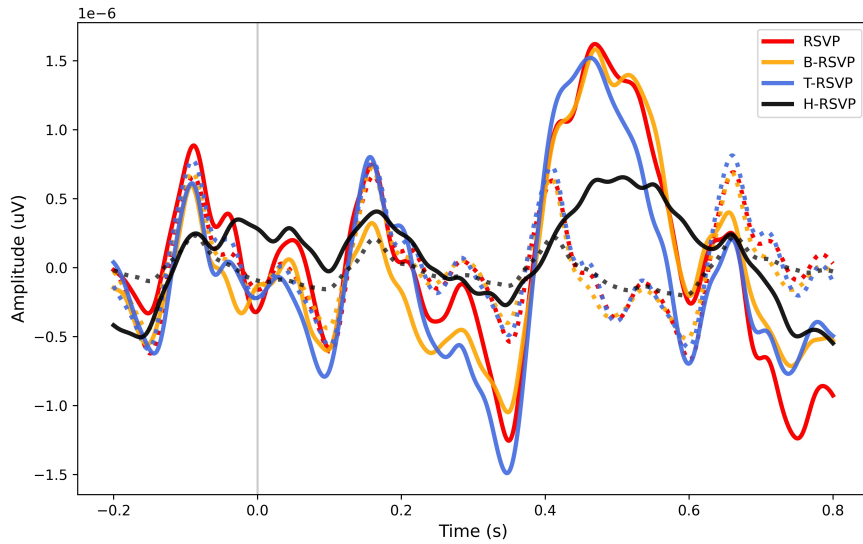


Fig. 2. Grand average comparisons of target vs. standard epochs at channel Pz under different conditions. Solid lines represent target epochs, while dotted lines indicate standard epochs. T—talking, B—body movement, H—head movement.

The figure highlights the challenge of assessing the impact of artifacts on P300 detection solely through visual inspection of averaged epochs. While the differences between conditions are not always apparent at a glance, these plots confirm the presence of distinguishable target and non-target ERP (Event-related Potential) responses, reinforcing their suitability for evaluating classification models. By analyzing how noise affects recognition performance, this work provides a foundation for future research on noise mitigation techniques. Ultimately, achieving comparable classification performance in both clean and artifact-prone environments will be crucial for developing EEG-based systems that are robust enough for real-world applications.

To quantitatively assess the separability of target and non-target trials, we trained and evaluated two classifiers: Bayesian Ridge and Support Vector Machine (SVM). Classification performance was measured using the area under the receiver operating characteristic curve (ROC-AUC) for each subject.

Before classification, the EEG data was preprocessed to enhance signal quality. A bandpass filter (0.1–30 Hz) was applied to remove unwanted frequencies, followed by downsampling to 100 Hz to reduce computational complexity. The signals were then re-referenced to the common average reference (CAR) to minimize channel-specific noise. Finally, the data was epoched from -0.2 to 0.8 seconds relative to stimulus onset.

Python, with the MNE library (v1.6.0), was used to implement the preprocessing pipeline, while Scikit-learn (v1.3.2) [11] was employed for training and evaluating Bayesian Ridge and SVM models, ensuring robust performance assessment using ROC-AUC metrics.

The classification results presented in Table 2 reveal notable differences in performance across conditions and classifiers. In the traditional or Non-intentionally Contaminated (NIC) condition, where no deliberate artifacts were introduced, both Bayesian Ridge and SVM achieved their highest performance, with average ROC-AUC scores of 0.81. This indicates that, in an ideal setting, both classifiers effectively distinguish target from non-target trials.

Table 2. ROC-AUC scores for each participant under clean and noisy conditions using Bayesian Ridge Regression (BRR) and Support Vector Machine (SVM).

Subject	Traditional RSVP		Artefactual RSVP					
			Body Movement		Talking		Head Movement	
	BRR	SVM	BRR	SVM	BRR	SVM	BRR	SVM
1	0.85	0.84	0.85	0.86	0.89	0.87	0.47	0.66
2	0.90	0.89	0.83	0.82	0.85	0.82	0.50	0.62
3	0.74	0.74	0.71	0.72	0.88	0.85	0.55	0.59
4	0.83	0.84	0.81	0.80	0.79	0.88	0.67	0.73
5	0.81	0.79	0.51	0.67	0.74	0.76	0.52	0.57
6	0.60	0.66	0.65	0.65	0.60	0.70	0.49	0.55
7	0.81	0.81	0.68	0.70	0.76	0.76	0.58	0.62
8	0.80	0.79	0.78	0.75	0.83	0.82	0.63	0.65
9	0.85	0.85	0.85	0.86	0.79	0.82	0.70	0.73
10	0.92	0.91	0.87	0.81	0.84	0.82	0.68	0.69
Average	0.81	0.81	0.75	0.76	0.80	0.81	0.58	0.64

In the body movement condition, the classification performance showed a moderate decline compared to the NIC condition. Bayesian Ridge dropped to an average AUC of 0.75, while SVM slightly outperformed it with 0.76. This suggests that while body movements introduce artifacts, they do not completely disrupt target vs. non-target discrimination. Some subjects maintained high clas-

sification performance despite the artifacts, as seen in Bayesian Ridge’s highest AUC of 0.87 and SVM’s highest of 0.86. However, for certain subjects, the impact was more pronounced, with Bayesian Ridge dropping as low as 0.51 and SVM to 0.65, suggesting individual variability in tolerance to body movement artifacts.

A similar trend was observed in the talking condition, where classification performance remained relatively high but still lower than in the NIC condition. Bayesian Ridge yielded an average AUC of 0.80, while SVM maintained a slight advantage at 0.810. This suggests that while speech-related muscle artifacts affect EEG signals, they do not substantially degrade P300 classification.

The most significant performance drop occurred in the head movement condition, where Bayesian Ridge’s classification ability was heavily compromised, with an average AUC of 0.579. SVM, although still affected, performed better with an average AUC of 0.641. This drastic reduction in classification performance suggests that head movements introduce severe noise, leading to difficulty in detecting the P300 signal. Compared to the NIC condition, the AUC drop of 0.232 for Bayesian Ridge and 0.171 for SVM highlights the significant impact of head movements on EEG signal integrity and classification reliability.

The results further suggest that while P300 detection remains feasible in the presence of minor body and talking artifacts, head movements significantly impair classification performance, making artifact mitigation crucial for real-world applications. The variability in individual performance further emphasizes the need for adaptive models that can compensate for movement-induced artifacts.

To further contextualize the results we compared the ROC-AUC scores from the AMBER 2.0 dataset with those from the previous AMBER dataset [3]. Both datasets followed the same protocol, but the previous version used a monitor display instead of an HMD for the RSVP task. The classification results for Bayesian Ridge Regression (BRR) and SVM in AMBER were comparable to those in AMBER 2.0, with only minor variations across conditions, as shown in Table 3.

Notably, the ROC-AUC scores in AMBER dataset [3] remained consistent across clean and noisy conditions, showing trends similar to AMBER 2.0. To statistically confirm this consistency, we conducted paired t-tests comparing the ROC-AUC scores from AMBER (monitor display) and AMBER 2.0 (HMD display) across four conditions: traditional RSVP, RSVP with body movement, RSVP with talking, and RSVP with head movement. Table 4 presents the statistical results, where a significance threshold of $p > 0.05$ was used for all comparisons. The p-values for all conditions are above this threshold, indicating that the differences in classification performance between AMBER and AMBER 2.0 are not statistically significant.

This indicates that transitioning to an HMD display did not introduce significant variability or degrade classification performance. These findings suggest that HMD-based RSVP tasks are as effective as traditional monitor-based setups, reinforcing their potential for real-world BCI applications.

Table 3. ROC-AUC scores from the previous AMBER dataset [3], where the RSVP task was presented on a monitor display.

Subject	Traditional RSVP		Artefactual RSVP					
			Body Movement		Talking		Head Movement	
	BRR	SVM	BRR	SVM	BRR	SVM	BRR	SVM
1	0.67	0.71	0.49	0.57	0.77	0.79	0.43	0.51
2	0.96	0.92	0.93	0.88	0.90	0.87	0.87	0.81
3	0.81	0.81	0.79	0.80	0.84	0.84	0.52	0.55
4	0.87	0.86	0.88	0.87	0.89	0.90	0.57	0.64
5	0.91	0.90	0.89	0.87	0.89	0.89	0.52	0.77
6	0.91	0.88	0.79	0.76	0.83	0.85	0.49	0.72
7	0.84	0.83	0.54	0.72	0.82	0.80	0.48	0.65
8	0.84	0.80	0.56	0.73	0.81	0.80	0.52	0.68
9	0.92	0.89	0.87	0.89	0.90	0.89	0.84	0.83
10	0.91	0.91	0.86	0.86	0.50	0.80	0.50	0.66
Average	0.86	0.85	0.76	0.80	0.82	0.84	0.57	0.68

The observed impact of noise underscores the importance of developing adaptive artifact-handling techniques, whether through preprocessing, model fine-tuning, or hardware improvements. The dataset provides a valuable benchmark for researchers aiming to enhance EEG classification methods by testing their models under different levels of real-world noise. By offering a structured evaluation of classifier performance across multiple conditions, this work paves the way for more practical, movement-tolerant EEG-based applications, bridging the gap between controlled laboratory studies and real-world deployment.

Table 4. Paired t-test results comparing the ROC-AUC scores between AMBER (monitor display) and AMBER 2.0 (HMD display) across four conditions.

Condition	Bayesian Ridge Regression (BRR)		SVM	
	t value	p value	t value	p value
Traditional RSVP	1.405	0.193	1.394	0.197
Body Movement	0.092	0.929	0.77	0.461
Talking	0.354	0.731	1.472	0.175
Head Movement	-0.1	0.922	1.066	0.314

4 Conclusion

The AMBER 2.0 datasets provide comprehensive resources for advancing BCI research in naturalistic and real-world environments. By addressing the limitations of lab-constrained datasets, this dataset aim to facilitate the development

of robust and reliable BCIs capable of handling the complexities of everyday use. AMBER 2.0 introduces immersive scenarios beyond the boundaries of a standard laboratory setting, through HMD-based RSVP tasks, marking significant strides toward practical BCI applications. The dataset serves as a benchmark for evaluating BCI algorithms and exploring the interplay of environmental factors, task demands, and user performance, paving the way for innovative applications in neurotechnology.

HMDs provide a more immersive and portable alternative to monitor-based setups, opening new possibilities for real-world applications of BCI systems.

This study introduces and examines the newly developed HMD-based RSVP dataset, emphasizing its potential for advancing subject-independent classification and enabling BCIs to function effectively outside the lab.

Acknowledgements We would like to express our gratitude to all the individuals who participated in the EEG experiments for their valuable contribution to this research. We are thankful for their time and effort in making this research possible.

Ethics Statement The studies involving humans were approved by the Dublin City University (DCU) Research Ethics Committee (DCUREC/2021/175). The study was conducted under local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Data Availability Statement The datasets presented in this study can be found at: <https://zenodo.org/records/14750342>

References

1. Robinson, N., Mane, R., Chouhan, T., and Guan, C.: Emerging trends in BCI-robotics for motor control and rehabilitation. *Current Opinion in Biomedical Engineering*, 20, p100354 (2021)
2. Mane, R., Chouhan, T., and Guan, C.: BCI for stroke rehabilitation: motor and beyond. *Journal of neural engineering*, 17(4), p041001 (2020)
3. Awais, M.A., Redmond, P., Ward, T.E. and Healy, G.: AMBER: advancing multi-modal brain-computer interfaces for enhanced robustness—A dataset for naturalistic settings. *Frontiers in Neuroergonomics*, 4, p.1216440 (2023)
4. Zhou, Q., Zhang, Q., Wang, B., Yang, Y., Yuan, Z., Li, S., Zhao, Y., Zhu, Y., Gao, Z., Zhou, J. and Wang, C.: RSVP-based BCI for inconspicuous targets: detection, localization, and modulation of attention. *Journal of Neural Engineering* 21(4), p.046046 (2024)
5. Lian, J., Qiao, X., Zhao, Y., Li, S., Wang, C. and Zhou, J.: EEG-Based Target Detection Using an RSVP Paradigm under Five Levels of Weak Hidden Conditions. *Brain Sciences* 13(11), p.1583 (2023)
6. Li, B., Lin, Y., Gao, X. and Liu, Z.: Enhancing the EEG classification in RSVP task by combining interval model of ERPs with spatial and temporal regions of interest. *Journal of Neural Engineering* 18(1), p.016008 (2021)

7. Awais, M.A., Ward, T., Redmond, P. and Healy, G.: From lab to life: assessing the impact of real-world interactions on the operation of rapid serial visual presentation-based brain-computer interfaces. *Journal of Neural Engineering* 21(4), p.046011 (2024)
8. Hughes, A. and Jorda, S.: Applications of biological and physiological signals in commercial video gaming and game research: a review. *Frontiers in Computer Science*, 3, p.557608 (2021)
9. Rao, S., Raju, D. C., Sai, P. N., Prabhakar, M., Tiwari, A., Dhatteval, J. S., and Shukla, U. K.: Real-Time Collaborative Gaming using Multiagent Brain-Computer Interfaces. In: *International Conference on Communication. Security and Artificial Intelligence (ICCSAI)*, pp. 705–709. IEEE, Greater Noida (2023)
10. Dillen, A., Omid, M., Díaz, M. A., Ghaffari, F., Roelands, B., Vanderborcht, B., and De Pauw, K.: Evaluating the real-world usability of BCI control systems with augmented reality: a user study protocol. *Frontiers in Human Neuroscience*, 18, p.1448584 (2024)
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., and Duchesnay, É.: Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, pp. 2825–2830 (2011)