

Investigating self-introspection in an auditory ERP-based BCI: A pilot experiment

Man Yang¹, Jordy Thielen¹, Michael Tangermann¹

¹Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands

E-mail: man.yang@donders.ru.nl

ABSTRACT: The decoding performance of a brain-computer interface (BCI) depends on both, the quality of the classification model and the signal-to-noise ratio of the brain signals. In this pilot study, we introduce *self-introspection* as a subjective measure of a user's awareness of their own BCI performance, i.e., their ability to anticipate how well the BCI will work. We investigated this concept in an auditory event-related potential (ERP)-based BCI paradigm across six experimental sessions of a single participant. After each trial, the participant rated the expected BCI feedback using a continuous mechanical slider, before receiving the true BCI feedback. This enabled a direct comparison between subjective self-introspection scores and BCI system scores. Results indicate a positive relationship between these scores across sessions, suggesting that the user had an internal sense of his BCI performance. As the present study included a single participant, the findings should be interpreted as a feasibility demonstration. Nevertheless, they show that subjective system performance can be estimated within an auditory ERP BCI paradigm, providing a potential method for studying user learning and adaptation in BCI use.

INTRODUCTION

Electroencephalography (EEG) is widely used in brain-computer interface (BCI) research due to its non-invasiveness, high temporal resolution, and relatively low cost. EEG-based BCIs translate users' neural activity into signals that can be decoded by computational models and used to provide feedback.

Among different BCI modalities, auditory BCIs provide an important alternative when visual attention is limited or unavailable, especially in assistive and rehabilitation contexts. For example, auditory BCIs have been explored as communication tools for individuals with severe motor impairments such as amyotrophic lateral sclerosis (ALS), where visual BCIs may be difficult to use due to impaired eye movement or visual function [1]. In addition, auditory paradigms have been investigated for language rehabilitation in patients with aphasia [2]. These applications highlight the potential of auditory BCI paradigms as assistive and rehabilitative technologies.

Many auditory BCIs rely on variants of the oddball paradigm to elicit event-related potentials (ERPs) like

the P300 response. Previous studies demonstrated that spatially distributed sound sources can reliably evoke such responses. A prominent example is the AMUSE paradigm, in which stimuli are presented from multiple loudspeakers at different spatial locations around the participant [3]. This AMUSE paradigm has been used for practical applications, including auditory spelling systems [4] and BCI-supported language rehabilitation for patients with stroke-induced aphasia [2].

Despite these advances, BCI performance often shows variability within and across sessions. Factors influencing performance are often categorized into system-related and user-related factors [5, 6]. While system-related improvements such as signal processing and machine learning have been widely studied, user-related factors remain comparatively underexplored. In particular, it remains unclear how users internally evaluate their performance during a BCI task. One possible aspect of this internal evaluation is the user's ability to estimate the outcome of a BCI task. In this study, we refer to this ability as *self-introspection*. Understanding this ability may provide insight into user-system interaction and may offer a complementary readout of how well users monitor and adapt their mental strategy during BCI use.

Therefore, the present study investigates self-introspection in an auditory ERP-BCI paradigm adapted from Kojima et al [7]. A single participant performed a language-related auditory BCI task while the BCI decoded the recorded EEG. After each trial, the participant first rated their perceived performance before receiving the BCI feedback. By examining the relationship between subjective self-introspection and system performance, this study aims to investigate whether users can reliably estimate their BCI performance in the absence of explicit feedback and whether this measure is feasible as a future tool for studying learning and adaptation in BCI use.

MATERIALS AND METHODS

Participant: One participant (male, 23 years old) was included in this pilot study. The participant was familiar with general EEG and BCI knowledge, but had no prior knowledge of the specific experimental design used in this study. Written informed consent was obtained before participation. Ethical approval for the experimental

protocol was granted by the local ethics committee of the Faculty of Social Sciences, Radboud University, and all procedures adhered to the approved guidelines. The presented data from a single participant represents an initial pilot stage of a larger study on self-introspection in the context of an auditory BCI.

Materials: EEG data were collected with 32 active electrodes positioned according to the international 10-10 system and sampled at 500 Hz. Data were amplified using a BrainAmp amplifier (Brain Products GmbH). Data acquisition was handled by the BrainAmpSeries Connector application (version 1.18.0), which transmitted the EEG signals via Lab Streaming Layer (LSL) [8], and all data were recorded using LabRecorder (version 1.16.5).

BCI Paradigm and Task: We employed an auditory ERP-based BCI paradigm with six Dutch words as stimuli. Each word was associated with a corresponding Dutch sentence that served as a cue. All audio files were presented at a sampling rate of 44.1 kHz, and exclusively via headphones (Sony MDR-7506).

To simulate spatially separated sound sources, virtual auditory positions were generated using Reaper (version 7.36) in combination with the dearVR MICRO plugin (version 1.0.3). Six fixed virtual positions were defined in the horizontal plane around the participant, covering the full 360° with equal angular spacing (see Figure 1(a)). In the spatial mode, the six words were presented from the six virtual positions. In the mono mode, the six words were presented without spatial separation.

The experiment consisted of a familiarization phase prior to the experimental sessions which included offline and online tasks. The familiarization phase was conducted during the first session, and EEG was not recorded during this phase. It included three tasks.

First, in the position familiarization task, the participant was introduced to the six virtual auditory source positions. The purpose of this task was to help the participant get used to the simulated spatial locations. The participant pressed keys 1 to 6 to play sounds from the corresponding virtual positions. This allowed him to actively explore and perceive the spatial separation between the different sound sources.

Second, in the sentence completion task, the participant learned the mapping between each sentence cue and its corresponding target word. After a sentence cue was presented, the participant was asked to select the correct target word from the six-word set. This ensured that the participant had memorized the cue–target associations before proceeding to the formal experimental sessions. The sentence-word pairs used are listed in Table 1.

Third, in the position identification task, the participant indicated the perceived spatial position of the target word after hearing a sentence cue and a six-word sequence. The purpose was to verify whether the participant had established a reliable association between sentence cues, target words, and their corresponding virtual positions.

After completing the familiarization phase, the participant proceeded to the formal offline and online tasks. The

offline and online tasks were organized into runs, where each run consisted of six trials. In each trial, the participant attended one of the six target words which had been cued by the sentence at the trial start.

During the **offline task**, the participant passively performed the auditory task. The EEG data collected during the offline runs were used to train a block Toeplitz linear discriminant analysis (btLDA) classifier [9], which was used in later online runs to decode the ERP responses and provide feedback.

At the beginning of each trial, the sentence cue was presented. This was followed by a silent interval, with the combined duration of the sentence and silence fixed to 6 s. Then, a fixation cross was displayed on the screen before the stimulation sequence started 2 s later.

The sequence consisted of 15 iterations of the six-word set, resulting in a total of 90 word presentations per trial. Within each iteration, the order of the six words was pseudo-randomized. To avoid temporal predictability, successive words were separated by at least three in-between words. In the spatial condition, each word was assigned to one spatial position within a trial. This relation was changed pseudo-randomly for the next trial.

After the fixation cross disappeared, a 3 s inter-trial pause was presented before the next trial began. The participant completed several offline runs across different sessions. Both the stimulus onset asynchrony (SOA) and the audio presentation mode (spatial or mono) differed between sessions. (see Figure 2).

In the **online task**, each trial included the sentence cue and an auditory stimulus presentation sequence as in the offline task, but in addition also a self-introspection rating phase, and a subsequent feedback display (see Figure 1(b)). Specifically, the fixation cross disappeared 2 s after the onset of the last word stimulus had been played, and a 5 s countdown started. During this period, the participant used a continuous mechanical slider to rate his perceived performance for that trial, the self-introspection score. The subsequent feedback phase lasted 3 s, during which both the self-introspection score and the true system score were displayed on the screen. The next trial started after the feedback phase.

Across the online sessions, task difficulty was manipulated by varying the SOA and the auditory presentation mode. Three SOA levels were tested: 350 ms, 250 ms, and 200 ms. The average duration of individual word stimuli was approximately 280 ms. For SOAs shorter than the word duration, successive auditory stimuli overlapped in time, which increased attentional demands.

To ensure consistency within each session, the six-word set remained unchanged across trials. Between sessions, predefined replacement words could be introduced if noticeable performance imbalances between target words had been observed in the previous session. This procedure prevented persistent ceiling effects while maintaining stable stimulus conditions within a session.

Session Design: The experiment comprised six sessions conducted on different days. The exact number of

Table 1: **Dutch sentence–word pairs.** During a trial, the participant attended one target word, which was associated with and cued by a specific sentence. All stimuli were presented in Dutch; English translations are provided for reference only.

Sentence (Dutch)	Sentence (English)	Target (Dutch)	Target (English)
Ik mix een milkshake in de ____	I mix a milkshake in the ____	Blender	Blender
We eten aardappelen, vlees en ____	We eat potatoes, meat, and ____	Groente	Vegetable
Op de kermis vind ik een grote ____	At the fair, I find a large ____	Knuffel	Stuffed toy
Ik zie mijn reflectie in de ____	I see my reflection in the ____	Spiegel	Mirror
Om zijn nek draagt de zakenman een ____	Around his neck, the businessman wears a ____	Stropdas	Tie
Het veld wordt omgeploegd met een ____	The field is plowed with a ____	Tractor	Tractor

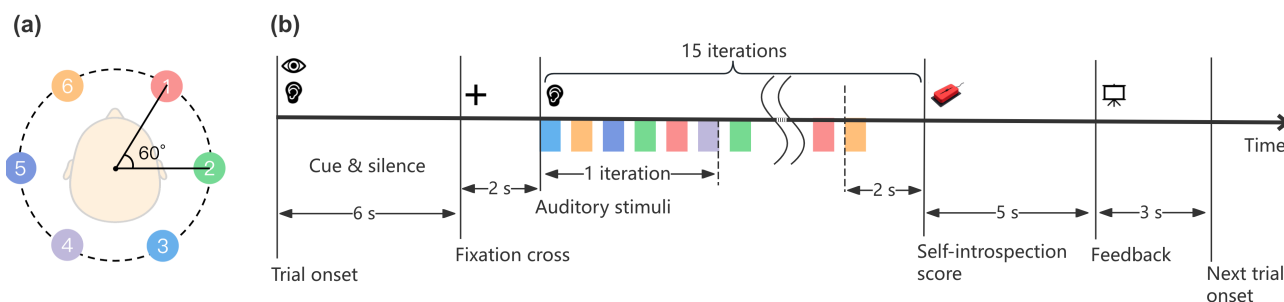


Figure 1: **Experimental setup and online trial timing.** (a) Arrangement of the six virtual auditory speaker locations, equally distributed around the participant in the horizontal plane with an angular separation of 60°. (b) Schematic illustration of one online trial. The structure was identical to the offline trial up to stimulus presentation, followed by a self-introspection rating and a feedback display.

offline and online runs in each session is shown in Figure 2. Rather than predefining a fixed number of runs, the aim was to collect as many trial-wise scores as possible within each session while keeping the total session duration feasible for the participant. Each session comprised offline and online tasks, except for session 5, which included online tasks only. In sessions 1–4 and session 6, EEG data were first collected during offline tasks for classifier training. The trained classifier was subsequently used in the online tasks of the same session. Session 2 contained two online phases. After the initial spatial online runs, the presentation mode was switched to mono. The classifier trained on the spatial data was directly reused for the mono runs, allowing us to assess within-session transfer without additional recalibration. Session 5 consisted of online runs only, for which the classifier trained in the previous session was reused.

Across sessions, task difficulty was manipulated by varying the SOA and the auditory presentation mode. Earlier sessions used longer SOAs and spatial separation, whereas later sessions introduced shorter SOAs and mono presentation to increase task demands. These changes were intentionally exploratory in this pilot study. They served to identify a parameter range that remained challenging while still allowing stable task performance.

The stimulus set remained constant within each session. Sessions 1–3 used the default word set (Blender, Groente, Knuffel, Spiegel, Stropdas, Tractor). In Sessions 4–6, the word “Groente” was replaced by “Stoeprand” to balance performance across target words.

Self-Introspection Score: The self-introspection score was obtained directly from the participant using a continuous mechanical slider interface described above. After each trial, the participant rated how they expected the

BCI system to evaluate their performance of that trial. The slider value was recorded within the range of $[-1, 1]$, enabling direct comparison with the system score.

The participant was instructed that a value of -1 , corresponding to the leftmost slider position, indicated “I think the system will give a very negative score”. The value 0 , positioned in the middle, indicated “I think the system will give a neutral score” and 1 , corresponding to the rightmost slider position, indicated “I think the system will give a very positive score”. Based on their subjective judgment of the current trial, the participant then used the slider to provide their introspection score prior to receiving the system score as feedback.

System Score: For each trial, EEG responses were processed using a btLDA classifier. Each trial consisted of 15 iterations of a six-word stimulus set. For every stimulus, the classifier produced an output, which was transformed into a probability using a sigmoid function.

For each of the six words, probabilities were averaged across the 15 iterations present in a trial. The true *target word* for that trial was defined by the cueing sentence, with each sentence corresponding to exactly one word (see Table 1), irrespective of whether it obtained the highest mean classifier score. Among the remaining five words, the *best non-target* was defined as the non-target word with the highest mean probability. This comparison between true *target* and *best non-target* focuses on the most difficult classification case, as the best non-target represents the strongest competitor to the target.

To quantify how well the system distinguished the target from competing stimuli, the area under the receiver operating characteristic curve (AUC) was computed using the 15 probabilities of the target word and the 15 probabilities of the best non-target word.

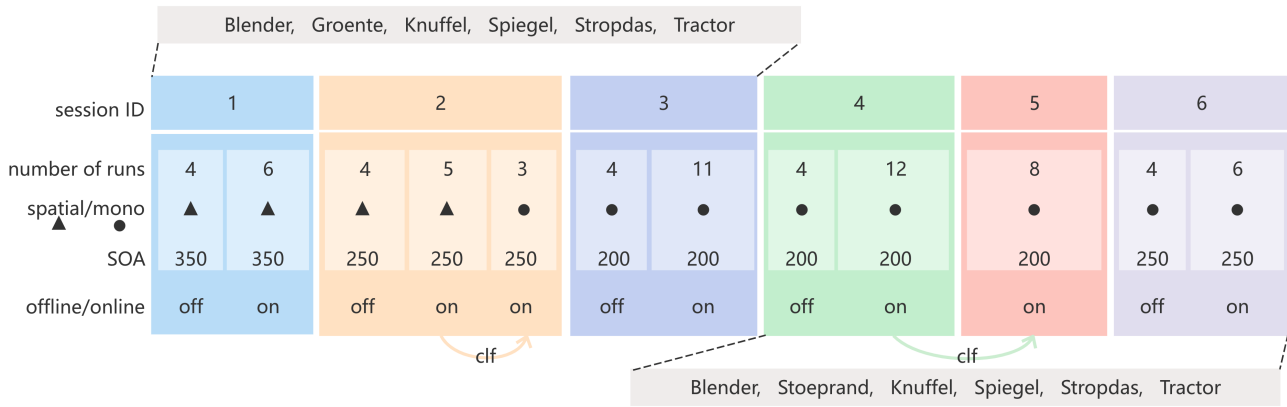


Figure 2: Overview of the experimental protocol. The experiment consisted of six sessions with different presentation modes (spatial or mono), numbers of runs, and SOAs. Triangles (▲) indicate spatial presentation and circles (●) indicate mono presentation. Labels “off” denote offline runs used for classifier training, whereas “on” denote online runs. Sessions 1–2 used SOAs of 350 ms and 250 ms, sessions 3–5 used an SOA of 200 ms, and session 6 used an SOA of 250 ms. The first three sessions used the word set *Blender, Groente, Knuffel, Spiegel, Stropdas, Tractor*, while in the last three sessions *Groente* was replaced by *Stoeprand*.

The resulting AUC value was linearly rescaled to the range $[-1, 1]$. That is, for trial t :

$$\text{SystemScore}_t = 2 \times \text{AUC}_t - 1 \quad (1)$$

With this transformation, -1 represents reversed discrimination, 0 corresponds to chance level, and 1 indicates perfect separation. The relationship between the self-introspection score and the system score is summarized in Figure 3.

EEG Data Processing and Feature Extraction: EEG data were recorded and stored in XDF format. Offline data were used for classifier training, and online data were processed in real time using the trained classifier.

A subset of 32 EEG channels was used for analysis (Fp1, Fp2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, TP9, CP5, CP1, CP2, CP6, TP10, P7, P3, Pz, P4, P8, PO9, O1, Oz, O2, PO10). The continuous EEG data were first epoched from -0.65 s to 1.35 s relative to stimulus onset. These extended epochs were band-pass filtered between 0.5 and 20 Hz, downsampled to 100 Hz to reduce computational load. Finally, the epochs were cropped to the analysis window from -0.35 s to 1.05 s relative to the stimulus onset.

For feature extraction, the post-stimulus interval was divided into consecutive time windows of 75 ms duration, resulting in 14 temporal segments covering the interval from 0 to 1.05 s. Within each window, EEG samples were averaged for each channel. The resulting values across all channels and time windows were concatenated to form the feature vector for classification with btLDA.

Classifier Updating Strategy: For sessions that included an offline phase (Sessions 1–4 and Session 6), an initial btLDA classifier was trained using the offline data collected at the beginning of the session. In Session 5, no offline data were collected. Instead, the classifier trained at the end of Session 4 was reused as the initial model. To enable adaptation during online operation, a sliding-window update strategy was applied. The classifier was

continuously trained on the most recent six runs ($3,240$ trials) using a first-in-first-out (FIFO) scheme. Each trial was decoded before being added to the training window, ensuring that no trial contributed to its own prediction. At the start of the online task, newly acquired trials were added until six runs were available; thereafter, the window was updated after each trial by removing the oldest and adding the newest. In session 5, initial trials were decoded using the classifier from session 4, and subsequent trials were decoded using models that had been updated with previously completed trials from session 5.

RESULTS

Trial-wise Relationship Between Self-Introspection and System Scores: We examined trial-wise correlations separately within each session (see Figure 4). Session 1 showed a significant positive correlation ($r = .536$, $p < .001$), and session 3 showed a similar significant positive correlation ($r = .483$, $p < .001$). In session 2, correlations were computed separately for spatial and mono trials: both spatial trials ($r = .501$, $p = .005$), and mono trials ($r = .618$, $p = .006$) showed a significant positive correlation. In session 4, the positive association remained but was weaker ($r = .244$, $p = .039$). In session 5, no significant correlation was observed ($r = -.075$, $p = .615$ (n.s.)). In session 6, a significant positive correlation re-emerged ($r = .377$, $p = .023$). Overall, these results show that the strength of the relationship varied across sessions, likely influenced by changes in task configuration.

When pooling trials from all sessions, the correlation remained significantly positive ($r = 0.481$, $p < 0.001$), indicating that the participant’s self-introspection scores generally co-varied with the system scores.

Word-Level Classification Performance and Stimulus Adjustment: To examine class separability across target words, confusion matrices were computed separately for sessions 1–3 and sessions 4–6 (see Figure 5).

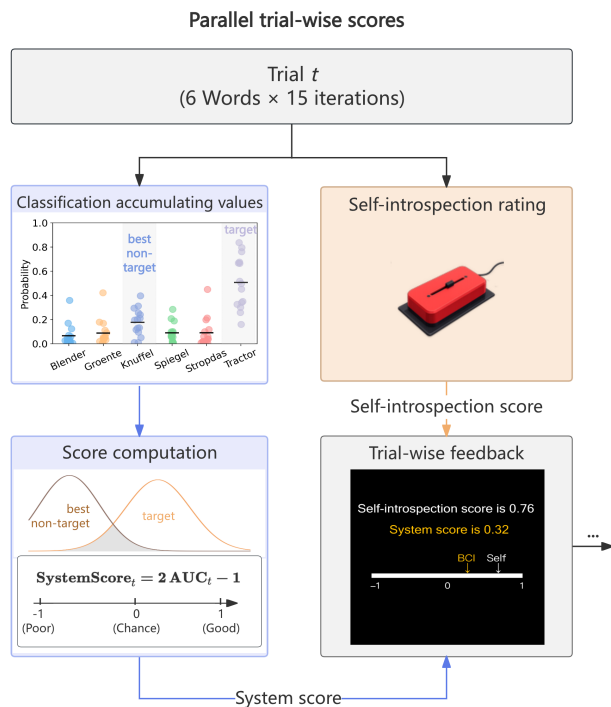


Figure 3: **Trial-wise self-introspection and system scores.** For each trial, classifier outputs were used to compute an AUC value, which was linearly rescaled into the system score. Participants also provided a self-introspection score after stimulus presentation. Both scores were presented as trial-wise feedback.

Figure 5(a) shows the pooled confusion matrix for sessions 1–3. In these sessions, the word *Groente* showed noticeably lower classification performance compared to the other words, as reflected by reduced diagonal counts and increased confusion with competing words. This imbalance was consistent with the participant’s subjective report indicating that this word was more difficult to attend to. Based on both classifier performance and participant feedback, *Groente* was replaced by *Stoeprand* in sessions 4–6. Figure 5(b) presents the pooled confusion matrix for sessions 4–6 after the replacement. Following this adjustment, classification performance across words became more balanced.

DISCUSSION

This pilot study investigated whether a single participant could estimate his own performance in an auditory BCI task, a process we refer to as self-introspection. Across six sessions, we compared subjective self-introspection scores with system scores obtained from a btLDA classifier. The multi-session design allowed us to assess whether this relationship remained consistent across sessions and under different task difficulty levels. The main objective was to evaluate the feasibility of measuring subjective performance awareness within a multi-session auditory BCI paradigm.

Overall, the results suggest that self-introspection can be meaningfully assessed in this paradigm. Across sessions, we observed a consistent positive relationship between

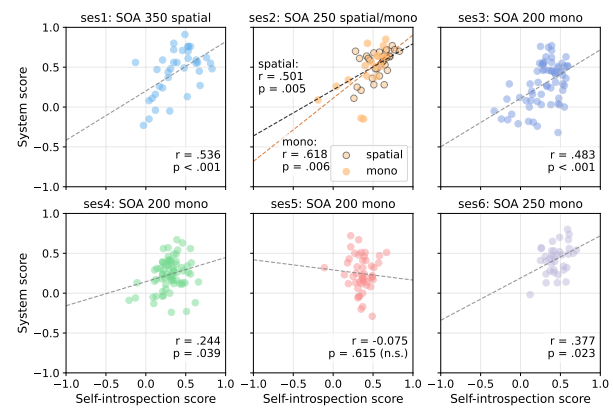


Figure 4: **Session-specific trial-wise correlations between self-introspection scores and system scores.** Each panel represents one session. In session 2, spatial and mono trials are analyzed separately and shown with separate regression lines; spatial data points are marked with black outlines. Correlation coefficients and corresponding significance levels are indicated within each subplot.

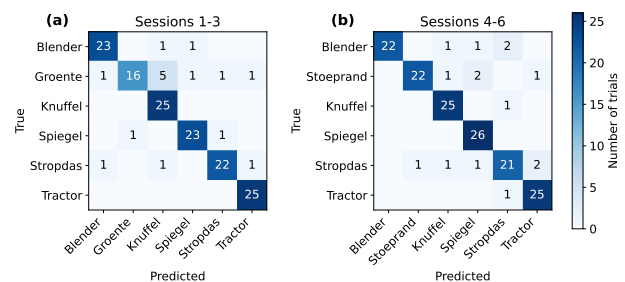


Figure 5: **Confusion matrices pooled across sessions.** (a) Sessions 1–3 using the original word set including *Groente*. (b) Sessions 4–6 after replacing *Groente* with *Stoeprand*. Color intensity reflects the number of trials.

the self-introspection scores provided by the participant and the system scores. Although the strength of this relationship varied across sessions, the overall pooled correlation indicates that the participant was generally able to anticipate how the BCI system would evaluate their performance. These findings provide initial support for the idea that users may possess a degree of internal awareness of their BCI performance.

At the same time, the present study involved only a single participant, and therefore the observed relationship cannot yet be generalized. Consequently, the present findings should be interpreted primarily as evidence that the paradigm is technically and experimentally feasible. Future studies involving a larger participant group will be necessary to determine whether the observed relationship generalizes across individuals.

The pilot phase revealed several aspects that can be improved in future studies. In the present study, SOA and presentation mode (spatial or mono) were varied within sessions to explore task feasibility under different difficulty levels. However, for studies targeting learning effects, a stable task configuration is preferable, as frequent changes may introduce variability that interferes

with learning. Future experiments should therefore keep parameters constant within and ideally across sessions. A key challenge is to select an appropriate initial difficulty level that prevents ceiling effects while still allowing performance improvements over time.

Another consideration concerns the interpretation of the system score derived from the AUC metric. Although it provides a continuous measure of separability, its meaning at the single-trial level is not always straightforward. In Figure 6, each point represents one trial, with the horizontal axis showing the scaled system score. Circles indicate trials with correct decoding (target word has the highest mean classifier score), whereas crosses indicate incorrect decoding.

We observed that some trials were decoded correctly despite AUC values below 0.5 (scaled < 0), while others were incorrect despite AUC values above 0.5 (scaled > 0). This apparent inconsistency arises because this AUC-like metric quantifies the global separability between the target and the strongest competing non-target across repetitions within a trial. It does not directly reflect whether the target achieved the highest *mean* score among all six words in that trial. As a result, the binary decoding outcome and the continuous AUC-based system score are not necessarily aligned. A discrepancy becomes problematic near the decision boundary.

This mismatch may influence how participants interpret the feedback. In particular, if the feedback does not consistently reflect the decoding outcome, it may affect both the participant's ability to anticipate the system's evaluation and the learning process during repeated BCI use.

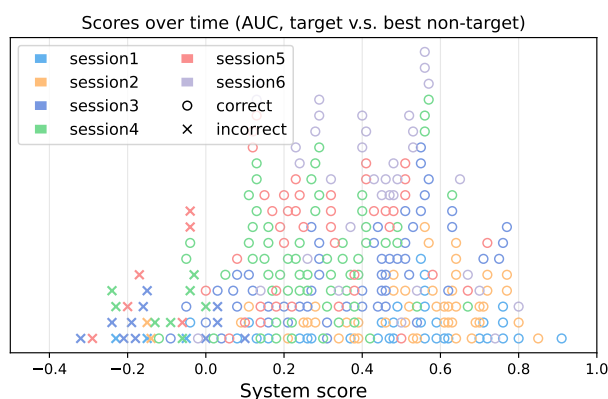


Figure 6: Distribution of trial-wise system scores across sessions. Each point represents one trial. The horizontal axis shows the scaled system score derived from the AUC ($2 \times \text{AUC} - 1$). Circles indicate correctly decoded trials, whereas crosses indicate incorrectly classified trials. Colors correspond to different sessions.

CONCLUSION

The present pilot study demonstrates the feasibility of measuring self-introspection in an auditory BCI task and suggests that a user may possess an internal sense of their BCI performance. While the results are preliminary

due to the single-participant design, the study provides a proof of concept for integrating subjective performance awareness into auditory neurofeedback BCI paradigms. Future work with more participants and improved experimental parameters are required to determine whether self-introspection can reliably predict BCI performance and potentially support adaptive training strategies.

ACKNOWLEDGMENTS

This project has received funding from the European Union's research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101118964, and from the Dutch Research Council (grant 024.005.022).

REFERENCES

- [1] P. Séguin et al., "The challenge of controlling an auditory BCI in the case of severe motor disability," *Journal of NeuroEngineering and Rehabilitation*, vol. 21, p. 9, Jan. 2024.
- [2] M. Musso et al., "Aphasia recovery by language training using a brain-computer interface: A proof-of-concept study," *Brain Communications*, vol. 4, no. 1, fcac008, Jan. 2022.
- [3] M. Schreuder, B. Blankertz, and M. Tangermann, "A New Auditory Multi-Class Brain-Computer Interface Paradigm: Spatial Hearing as an Informative Cue," *PLOS ONE*, vol. 5, no. 4, e9813, Apr. 2010.
- [4] M. Schreuder, T. Rost, and M. Tangermann, "Listen, You are Writing! Speeding up Online Spelling with a Dynamic Auditory BCI," *Frontiers in Neuroscience*, vol. 5, Oct. 14, 2011.
- [5] C. Jeunet, B. N'Kaoua, and F. Lotte, "Advances in user-training for mental-imagery-based BCI control: Psychological and cognitive factors and their neural correlates," *Progress in Brain Research*, vol. 228, pp. 3–35, 2016.
- [6] F. Lotte et al., "A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031 005, Jun. 2018.
- [7] S. Kojima et al., "Influence of pitch modulation on event-related potentials elicited by Dutch word stimuli in a brain-computer interface language rehabilitation task," *Journal of Neural Engineering*, vol. 22, no. 3, p. 036 010, Jun. 1, 2025.
- [8] C. Kothe et al., "The lab streaming layer for synchronized multimodal recording," *Imaging Neuroscience*, vol. 3, IMAG.a.136, 2025, Open Access.
- [9] J. Sosulski and M. Tangermann, "Introducing block-Toeplitz covariance matrices to remaster linear discriminant analysis for event-related potential brain-computer interfaces," *Journal of Neural Engineering*, vol. 19, no. 6, Nov. 7, 2022.