Using EEG-Based BCI Devices to Subliminally Probe for Private Information

Mario Frank
University of California, Berkeley

Tiffany Hwu
University of California, Irvine

Sakshi Jain
LinkedIn Corporation

Robert T Knight
University of California, Berkeley

Ivan Martinovic
University of Oxford

Prateek Mittal
Princeton University

Daniele Perito
University of California, Berkeley

Ivo Sluganovic
University of Oxford

Dawn Song
University of California, Berkeley

ABSTRACT

EEG-based Brain-Computer-Interfaces are becoming available as consumer-grade devices, used in applications from gaming to learning programs with neuro-feedback loops. While enabling attractive applications, their proliferation introduces novel privacy concerns and security threats. One such example are attacks in which adversaries compromise EEG-based BCI devices and analyze the user's brain activity in order to infer private information such as their bank or area-of-living.

In this paper, we propose and analyze a more serious threat - a subliminal attack in which, given that the visual probing lasts for less than 13.3 milliseconds, the existence of any stimulus is below ones cognitive perception. We show that even under such limitation, the attacker can still analyze subliminal brain activity in response to the rapid visual stimuli and consequently infer private information about the user.

By running a proof-of-concept study with 27 participants, we experimentally evaluate the feasibility of subliminal attacks using EEG-based BCI devices. While not perfect, our results show that it is indeed feasible for attackers to subliminally learn probabilistic information about their victims.

1 INTRODUCTION

Brain-Computer Interface (BCI) devices are becoming increasingly popular for use in applications such as entertainment, accessibility, and cognitive enhancement [1]. A popular technology used in BCI for recording brain activity is Electroencephalography (EEG), which uses external scalp electrodes to capture fluctuations of the electrical potentials in the brain. The Emotiv device [2] is an example of low-cost commodity BCIs, intended for home usage with applications written by third-party developers and are available for download from application markets (see, e.g., [3]).

WPES'17, October 30, 2017, Dallas, TX, USA

ACM ISBN 978-1-4503-5175-1/17/10…$15.00
https://doi.org/10.1145/3139550.3139559

Martinovic et al. [4] recently emphasized that BCI devices may make the raw EEG signal available to potentially untrusted third-party applications. If such an application is malicious, it can in turn abuse the BCI device to infer private information about a victim, such as her/his preferred bank or area-of-living. The general idea of this attack is similar to a polygraph, where the interrogated person's physiological reactions are used to reason about his/her knowledge.

However, a fundamental limitation of the attacks proposed by Martinovic et al. is that they rely on supraliminal (consciously perceived) stimuli and are thus detectable. Based on these observations, we propose a *subliminal* attack that infers private information by probing the victim at a level below his/her cognitive perception. Similar to *subliminal advertising* (see, e.g., [5]), our key idea is to show the visual stimuli within the screen content that the user expects to see, but for a duration that is too short for conscious perception (several milliseconds), yet still sufficient to result in activation of certain parts of user's brain detectable by an attacker.

This is a challenging task. If the stimuli are shown too prominently, this increases the chance of the attack being detected. If, in contrast, the attacker hides the stimuli too well, the user's subliminal detection may not be sufficiently strong, reducing the probability of inferring relevant private information. Thus, the attacker must operate within this narrow regime of the user's input channel.

As the results of our experimental study with 27 participants show, such subliminal attack on users of EEG-based BCI devices are indeed feasible – attackers can make probabilistic inferences on the users' recognition of the person depicted in the visual stimuli in a manner that is concealed from the user.

2 BACKGROUND AND RELATED WORK

EEG-based BCI. Electroencephalography (EEG) monitors electrical activity at the scalp that corresponds to changes in ion concentrations of neurons in a functioning brain. EEG is widely used in a medical setting to monitor neurological diseases, such as epilepsy, to diagnosing possible brain deaths of comatose patients, communicate with patients who suffer from locked-in syndrome, or control of wheelchairs for handicapped [6]. In neuroscience research, EEG serves as a non-invasive, cost-effective method of measuring brain activity.

As an example of potential applications, EEG devices have successfully enabled users to spell letters using only EEG signals [7],

as well as to successfully perform *guilty-knowledge* tests, despite user's active efforts to conceal knowledge of some form [8].

Subliminal stimulation. While the majority of sensory stimulation that we are exposed to in everyday lives is sufficiently intensive to be consciously perceived (**supraliminal**), in certain situations a stimulus can also be made **subliminal** if its intensity is carefully controlled. For instance, if a visual stimuli is shown sufficiently briefly, individuals might not be consciously aware of it, but research has shown that subliminal stimuli measurably impact one's behavior, for instance by influencing the choice of consumer brands [9].

Researchers agree on the existence of *perceptual threshold*, an intensity that defines whether one will cognitively perceived a stimulation or not, but determining specific values for any given stimulus type is not straightforward. For instance, neuroscience literature speaks of 10 ms to 55 ms as a suggested presentation time range for a stimulus to be subliminal (a good overview of designing experiments with subliminal stimuli can be found in [10]), but it is also known that this duration not only varies significantly among individuals, but also changes from day to day for the same individual. As a result, most definitions of perceptual threshold focus on levels of stimulation that result in stimulus being undetected in a certain percentage of times it was presented to a user.

Recognizing faces for authentication. The human ability to recognize and remember faces over extended periods of time has been utilized as a method of authentication in which the user identifies familiar faces within a grid of images [11]. The idea of Passfaces was further extended by the use of commodity BCI devices [12]. Instead of manually choosing the correct faces out of a grid of images, the authors used eye trackers and considered a 0.5 second fixation on a face as a selection of it. Among other applications of facial recognition, being deployed in various real-world scenarios, Facebook requires one to identify their friends in tagged photos for security verification [13]. Consequently, any information about faces familiar to a user should be considered vulnerable private information.

Subliminal face recognition. Existing neuroscientific work on the subliminal perception of human faces shows that ERPs in response to unpleasant facial expressions have a higher positive amplitude than pleasant expressions. Furthermore, this effect shows even through very fast unmasked subliminal presentations of stimuli, at 1ms [14]. Although in our experiment several subjects had noticed the stimuli, research shows that presentation times as short as 1 ms could still reveal enough information in their EEG signal to extract desired information about faces [15].

3 SYSTEM AND ADVERSARY MODEL

The system model consists of the user, the computer, and the BCI-device. The user uses the BCI device with various applications from the third-party developer platform, actively supporting setting up the device and calibrating it.

The adversary is an application developer who's goal is to obtain private user information by exploiting a BCI-device's API to access the raw EEG signal recorded during use of a malicious application. This could be any of the scenarios proposed in [4] such as guessing the banking provider, PINs, or month of birth. As an example, a repressive regime can try to identify which users are familiar with some of the key persons of the underground opposition.

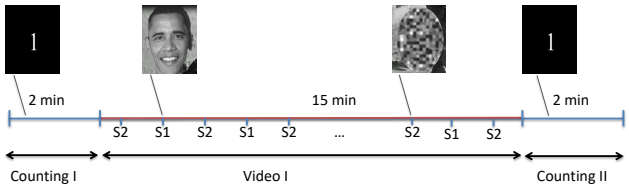

Figure 1: The experimental protocol is divided into 3 sub-experiments: Counting I, Video I and Counting II. The embedded visual stimuli S_j are depicted above the timeline.

The adversary can, for instance, modify a benign game or video viewer by inserting malicious code and inconspicuously upload the application to an online marketplace. This allows him to collect the EEG signal recorded while the user is exposed to the different images displayed on the screen, and to deduce private information under the assumption that e.g. the stimuli known to the user trigger the strongest EEG response.

4 EXPERIMENTS

Due to many factors that can negatively affect the outcome, experimentally investigating the feasibility of the proposed subliminal side-channel is a challenging task. For instance, the equipment used could be sub-optimal, the video used to hide the attack could have many still images where it is hard to hide a stimulus, the secret that is being attacked could be too complex, and so on. Therefore, our experiment is designed to investigate whether the attack is indeed feasible in a basic scenario, instead of starting off with sophisticated variants. We thus make design decisions that minimize the chance that the attack fails due to factors that we can control. For instance, we use a good EEG device and a video with flickering artifacts that helps hide the attack.

4.1 Test Population and Setup

After obtaining approval from the Institutional Review Board, 29 undergraduate and graduate students (21 males and 8 females) in the Computer Science department were recruited to participate in our experiment, 2 of which had unusable data due to recording problems. All subjects were self-screened for neurological disorders and metal implants which could potentially interfere with recording. Prior to the experiment, subjects were informed of the basic EEG procedures, but not yet informed of the subliminal nature of the stimuli. The participants signed informed consent and received compensation in the form of a $40 gift card. The experiment took 90 minutes total for each user, including setup time. This was the main limiting factor for population size. ActiveTwo BioSemi equipment [16] was used for the collection of EEG data. Participants were measured and fitted with a tight cap, and 64 Ag/AgCl electrodes were attached to the cap with conducting gel. All electrodes were then attached to a low-noise DC coupled post-amplifier, with a sampling rate of 1024 Hz. All stimuli were presented in a dim room on a CRT monitor (75Hz refresh rate) using presentation software [17].

4.2 Experimental Protocol

After the setup described above, the participants were instructed to remain relaxed for the entire duration of the experiments and . other interaction with the participants was kept as short and concise as possible. The experiment consisted of three parts: two repetitions of the counting task, and watching the video, as shown in Figure 1.

Counting I and II. In a version of the standard task used to calibrate BCI devices, the participant was presented with a randomly permuted sequence of numbers from 0 to 10. Each number except 1 appeared exactly 16 times. The digit 1 could appear anywhere between 14-18 times, chosen uniformly at random. The participant was asked to count the number of occurrences of the number 1. Each stimulus lasted for 250 ms, and pauses between stimuli were randomly chosen to be between 250 ms and 375 ms long. At the end of this step of the experiment, the participants were asked for their count to check for correctness. This part of the experiment lasted for about 2 minutes. It was carried out in the beginning of the experiment (Counting I) and at the end (Counting II).

Video I. In this phase, the participant was instructed to watch a 15 minute long black and white video extracted from Charlie Chaplin's "The Gold Rush" (1925). They were asked to pay attention to the plot of the video to make sure they concentrated on watching the video through its entire duration. Two kinds of stimuli (S1 and S2) were used, one with a black and white portrait of Barack Obama (S1) and the other being a blurred image of a human face (S2). We choose these stimuli in order to make sure that every subject was familiar with S1 and would not recognize S2. Given that the our experiments were conducted at a US university at the time when Barack Obama was the acting president, we can safely assume that we know the correct answer for all participants.

A stimulus was shown every 5 seconds, making a total of 180 stimuli over 15 minutes. Every 4^{th} stimulus was S1 and was displayed at the top right corner of the image frame. The position of S2 rotated along the remaining three corners. Each stimulus was shown for 13.3 ms. The limiting factor of this time was the screen refresh rate (75Hz).

Recognition survey. As we are ultimately interested in understanding the feasibility of carrying out the attack subconsciously, all participants were asked at the end of the experiment if they noticed anything odd in the video. If they negated, no further questions were asked, otherwise they were asked for details of what they saw. We categorize their answers as follows: participant recognized nothing, participant saw something, participant saw a face, participant saw "Barack Obama".

5 DATA ANALYSIS

The raw data consists of wave signals from a number of different electrodes. For preprocessing, we first divide the signal into epochs, each epoch ranging from 200 ms prior to 1000 ms after every stimulus onset. Each such epoch is associated with the respective stimulus that triggers it. For each epoch, we then calculate the mean of the first 200 ms to get a baseline and subtract this baseline from the entire epoch. We reduce the high frequency noise by passing the signals through a low pass filter with a pass band of [0.35, 0.4] in normalized frequency units and applying a median filter of size 4.

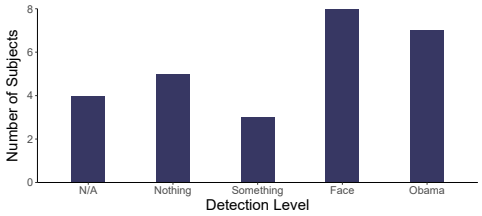

Figure 2: Aggregated levels of participants' detection of the visual stimuli hidden in the video. While the stimulus did not remain hidden from all participants, the stimulation was indeed subliminal, with only 7 participants able to correctly detect the stimuli that was repeatedly shown throughout the duration of the 20 minute long video.

Classification. The goal of the attacker is to train a classifier which identifies if the stimulus is relevant for the user. From a technical perspective, we want to evaluate whether the classifier can extract sufficient information from the recorded EEG signal in order to determine one of the three different types of brain activity: 1) unknown face, 2) a face that the user subliminally recognizes, or 3) a plain video sequence without any subliminal stimulation.

Classification Setting. In our setting, each epoch is one observation. Each epoch corresponds to a single stimulus and contains the signals from all the EEG channels for a time period of [signal - 200ms, signal + 1000 ms]. If C denotes the number of channels being used and f denotes the sampling frequency, we have $(1000 + 200)f = S$ measurements per channel per epoch. We group the signals from all the channels for each epoch into a feature vector of dimensionality $K = C \times S$.

In the testing phase, the classifier is provided with a set of fresh observations $\mathbf{x_i}$ for which it must output label predictions $\mathbf{y_i}$. In other words, the classifier must predict for each epoch, if the corresponding stimulus shown to the participant is relevant or not.

We use the boosted logistic regression classifier (BLR) which consists of a set of $M \in \mathbb{N}$ individual classifiers f_m with $m \in \{1, \ldots, M\}$ that all output individual classifier scores. In our analysis, we only use those channels that are located along the z-axis, parietal, and occipital areas of the scalp, where P300 ERPs are usually the strongest. In particular those channels are: 'Fz', 'Cz', 'Pz', 'P3', 'P4', 'PO7', 'PO8', 'Oz'.

6 EVALUATION

In this section, we evaluate the feasibility of subliminally probing for private user information by running several experiments, each representing a different scenario that an attacker might attempt.

6.1 Stimulus Subliminality

As described in Section 2, the level of cognitive perception of a visual stimulus depends on a number of factors which vary significantly between scenarios, environments, and individuals. Given the fact that the same stimulus is repeatedly shown to participants in our experiment, we expect some of them to be able to detect it, even if we shorten the stimulus duration in comparison to what is often reported in relevant literature (10 ms to 55 ms).

The results of asking participants if anything seemed "strange" while watching the video are given in Figure 2. The subjects are

Figure 3: Detecting subliminal responses based on supraliminal training (Numbers on Faces). For each user, the classifier outputs which type of stimulus results in recognition, the three possible candidates being: Obama, Unknown face, and random periods with no subliminal stimuli (Blank). Outputs are correct for the majority (18 out of 27) of participants (Figure a), irrespective of their level of stimuli detection (Figure b), showing that such an attack is indeed feasible.

divided into five different groups, representing the different recognition levels of the users. As expected, the subliminality of the stimulus varied for different users. While a total of 7 participants were indeed able to recognize an image of Barack Obama, 5 users did not notice anything unusual, while further 3 of them only "saw something". Despite being detected by a subset of users, our stimuli remained completely hidden from some users. This supports the hypothesis that private information could be subliminally probed using EEG based BCI devices, especially if the attacker can adapt to a specific victim by gradually increasing the stimulus duration in order to ensure that it remains undetected.

6.2 Subliminal Probing

We now evaluate the attack that subliminally probes for user's private information.

Setup. During training, the classifier was provided with data from Counting I. For testing, we extract all epochs triggered by hidden images of Barack Obama, all epochs triggered by the unknown face, and equally many epochs taken from random frames where the video was not manipulated. We let the classifier output a score for each epoch of this dataset. Recall that, based on the training data used, this score outputs the classifier's belief that the user has 'counted' the respective stimulus. Even though the user did not actively count the target stimulus (she should not even realize that it is on the screen), the classifier is searching for the same artifacts in the EEG signal.

As in the counting experiment, the final output of the classifier is the candidate stimulus that gets the highest average classifier score. This time, there are three possible outcomes. Since we assume that all participants recognize an image of Barack Obama, we can compare the classifier output against this ground truth.

Results. We show the classifier output split by different levels of user awareness in Figure 3. For 18 users the classifier outputs the correct answer. For 5 users, BLR predicted 'Unknown face' and for 4 users BLR predicted 'Blank'. From a machine learning perspective, it appears that the attack works, as the classifier is able to distinguish a relevant stimulus from irrelevant stimuli. The *reduction in guessing entropy* is expectedly smaller than in previously reported supraliminal attacks (and our baseline); however, it still *equals a high 20.84%*. This is an important result, which shows that attackers could indeed carefully design their visual stimuli such that they remain subliminal, and still probabilistically reduce the entropy of guessing relevant private information using EEG-based BCI devices.

The attack works almost independently of the extent to which the victims realize that the video has been manipulated and in each recognition group, the classifier found the correct answer for the majority of users.

7 CONCLUSION

This work examined the feasibility of subliminal attacks on users of EEG-based brain-computer interfaces (BCIs). By running a series of experiments with 27 subjects, we find that our attack is able to detect brain responses to subliminal stimulation with accuracy that is comparable to the results previously reported for supraliminal attacks, even when the classifier is trained on a different type of brain responses than the ones that are being probed for. Consequently, by carefully designing the visual stimuli, an attacker can reduce the entropy of guessing user's private information by more than 20%, while at the same time achieving that the victim remains unaware of being probed.

As a first attempt to perform subliminal probing, our experiments have been carried out in a controlled setting to demonstrate their feasibility and exclude other factors that might impede success. However, with the recent improvements of measurement performance and the reduction of prices, the pervasiveness of EEG-based BCI devices in our daily lives is likely to increase. Consequently, this paper makes an important step towards raising the awareness about the possibility of some of the attacks to even happen below the level of victim's conscious perception.

REFERENCES

[1] SmartBrain Technologies, "http://www.smartbraintech.com/store/pc/all-attention-brain-exercisers-c9.htm."
[2] Emotiv Systems, "www.emotiv.com."
[3] Emotiv's App Store, "http://www.emotiv.com/store/app.php."
[4] I. Martinovic, D. Davies, M. Frank, D. Perito, T. Ros, and D. Song, "On the feasibility of side-channel attacks with brain-computer interfaces," in *21st USENIX Security Symposium*. USENIX Association, Aug 2012.
[5] J. C. Karremans, W. Stroebe, and J. Claus, "Beyond vicary's fantasies: The impact of subliminal priming and brand choice," *Journal of Experimental Social Psychology*, vol. 42, no. 6, pp. 792 – 798, 2006.
[6] T. Carlson and J. Del R. Millan, "Brain-controlled wheelchairs: A robotic architecture," *IEEE Robotics and Automation Magazine*, vol. 20, no. 1, pp. 65–73, 2013.
[7] U. Hoffmann, G. Garcia, J.-M. Vesin, K. Diserens, and T. Ebrahimi, "A boosting approach to P300 detection with application to brain-computer interfaces," in *2nd International IEEE EMBS Conference on Neural Engineering*, 2005, pp. 97 –100.
[8] L. a. Farwell and S. S. Smith, "Using brain MERMER testing to detect knowledge despite efforts to conceal." *Journal of forensic sciences*, vol. 46, no. 1, pp. 135–43, 2001.
[9] J. C. Karremans, W. Stroebe, and J. Claus, "Beyond Vicary's fantasies: The impact of subliminal priming and brand choice," *Journal of Experimental Social Psychology*, vol. 42, no. 6, pp. 792–798, 2006.
[10] Nick Epley, "Laboratory manual: Science or science fiction? investigating the possibility (and plausibility) of subliminal persuasion."
[11] S. Brostoff and M. Sasse, "Are passfaces more usable than passwords? a field trial investigation," in *People and Computers XIV - Usability or Else!*, S. McDonald, Y. Waern, and G. Cockton, Eds. Springer London, 2000, pp. 405–424.
[12] P. Dunphy, A. Fitch, and P. Olivier, "Gaze-contingent passwords at the atm," in *4th Conference on Communication by Gaze Interaction (COGAIN)*, 2008.
[13] "http://www.facebook.com/help/search/?q=security+verification."
[14] E. Bernat, S. Bunce, and H. Shevrin, "Event-related brain potentials differentiate positive and negative mood adjectives during both supraliminal and subliminal visual processing," *International Journal of Psychophysiology*, vol. 42, no. 1, pp. 11 – 34, 2001.
[15] B. J. Liddell, L. M. Williams, J. Rathjen, H. Shevrin, and E. Gordon, "A temporal dissociation of subliminal versus supraliminal fear perception: An event-related potential study," *J. Cognitive Neuroscience*, vol. 16, no. 3, pp. 479–486, Apr. 2004.
[16] BioSemi, "www.biosemi.com."
[17] Neurobehavioral Systems, Inc., "http://www.neurobs.com/."