Tiny Noise Can Make an EEG-Based Brain-Computer Interface Speller Output Anything

Xiao Zhang, Dongrui Wu, Lieyun Ding, Hanbin Luo, Chin-Teng Lin and Tzyy-Ping Jung

Abstract—An electroencephalogram (EEG) based brain-computer interface (BCI) speller allows a user to input text to a computer by thought. It is particularly useful to severely disabled individuals, e.g., amyotrophic lateral sclerosis patients, who have no other effective means of communication with another person or a computer. Most studies so far focused on making EEG-based BCI spellers faster and more reliable; however, few have considered their security. Here we show that P300 and steady-state visual evoked potential BCI spellers are very vulnerable, i.e., they can be severely attacked by adversarial perturbations, which are too tiny to be noticed when added to EEG signals, but can mislead the spellers to spell anything the attacker wants. The consequence could range from merely user frustration to severe misdiagnosis in clinical applications. We hope our research can attract more attention to the security of EEG-based BCI spellers, and more broadly, EEG-based BCIs, which has received little attention before.

I. INTRODUCTION

A brain-computer interface (BCI), which has been extensively used in neuroscience, neural engineering and clinical rehabilitation, is a communication pathway that allows people to interact with computers using brain signals directly [1]–[4]. There are many approaches to collecting signals from the brain. Electroencephalogram (EEG), usually measured from the scalp, may be the most popular one due to its simplicity and low cost [5].

An EEG-based BCI speller allows a user to input text to a computer by thought [6], [7]. It can help severely disabled individuals, e.g., amyotrophic lateral sclerosis patients, to communicate with computers or other people. There are mainly two categories of EEG-based BCI spellers, P300 spellers [6] and steady-state visual evoked potential (SSVEP) spellers [7], which elicit different EEG patterns, as illustrated in Fig. 1a.

A P300 speller, which uses P300 evoked potentials as its input signal [8], was first invented by Farwell and Donchin in 1988 [6] and further developed by many others [9]–[12]. P300

X. Zhang and D. Wu are with the Ministry of Education Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China.

L. Ding and H. Luo are with the School of Civil Engineering and Mechanics, Huazhong University of Science and Technology, Wuhan, China.

C-T Lin is with the Centre of Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia.

T-P Jung is with the Swartz Center for Computational Neuroscience, Institute for Neural Computation, University of California San Diego (UCSD), La Jolla, CA, USA. He is also with the Center for Advanced Neurological Engineering, Institute of Engineering in Medicine, UCSD, La Jolla, CA, USA.

Corresponding authors: D. Wu (drwu@hust.edu.cn) and L. Ding (dly@hust.edu.cn).
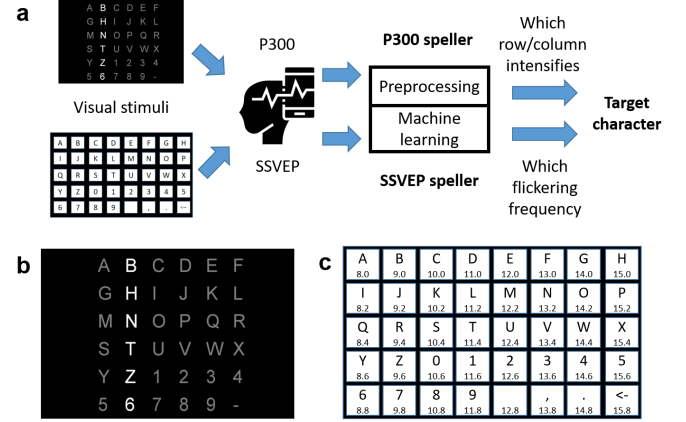

Fig. 1: **A P300 speller and an SSVEP speller. a**, Workflow of a P300 speller (top path) and an SSVEP speller (bottom path). For each speller, the user watches the stimulation interface, focusing on the character he/she wants to input, and EEG signals are recorded and analyzed by the speller. The P300 speller first identifies the row and the column that elicit the largest P300, and then outputs the character at their intersection. The SSVEP speller identifies the output character directly by matching the user's EEG oscillation frequency with the flickering frequency of each candidate character. **b**, Stimulation interface of a P300 speller, where the second column is intensified. **c**, Stimulation interface of an SSVEP speller. The number below each character indicates its flickering frequency (Hz).

is a positive deflection in voltage, typically appearing around 250 to 500 ms after a rare target stimulus occurs [13]. It is an endogenous potential linked to people's cognitive processes, such as information processing and decision making [14], [15]. The standard oddball paradigm is usually used to elicit P300, in which rare target stimuli are mixed with high-probability non-target ones. The P300 speller considered in this article uses a 6×6 character matrix, which consists of 26 letters and 10 other symbols, as shown in Fig. 1b. The user stares at the character he/she wants to input, while a row or column is rapidly intensified sequentially. The corresponding EEG signals are recorded and classified as a target (containing P300) or non-target (not containing P300) for each intensification. Then, the computer identifies the character at the intersection of the target row and the target column, which elicit the largest P300s, as the output. For reliable performance, each row and column may have to be intensified multiple times, which reduces the speed of the P300 speller.

Compared with the P300 speller, an SSVEP speller has the advantages of high information transfer rate (ITR), little user training, and some immunity to artifacts [16]–[18]. When the user stares at a visual target flickering at a specific frequency, usually between 3.5 Hz and 75 Hz, electrical signals of the same frequency, as well as its corresponding harmonics, can be observed from the EEG signals [16]. In an SSVEP speller, the pictures of different characters are flickering at different frequencies, so that a classifier can directly identify the output character from a large number of candidates by matching their flickering frequencies with the user's EEG oscillation frequency. Since all characters in an SSVEP speller are flickering simultaneously (in contrast to sequential intensification in a P300 speller), they can have much higher ITRs. The SSVEP speller considered in this article has 40 characters (Fig. 1c), whose stimulation frequencies are from 8 Hz to 15.8 Hz with 0.2 Hz increment [19].

Machine learning is used in BCI spellers to construct the classifiers. Most studies so far focused on making the BCI classifiers faster and more reliable; however, few have considered their security. It has been found in other application domains that adversarial examples [20], which are normal examples contaminated by deliberately designed tiny perturbations, can easily fool machine learning models. These perturbations are usually so small that they are indistinguishable to human eyes. Existing studies on adversarial examples focused largely on deep learning models for computer vision. For example, it was found that a picture of a panda, after adding a weak adversarial perturbation, can be misclassified as a gibbon by a deep learning classifier [21]. Kurakin *et al.* [22] found that printed photos of adversarial examples can degrade the performance of an ImageNet Inception classifier. Athalye *et al.* [23] 3D printed a turtle with an adversarial texture, which was classified as a riffle from almost every viewpoint. Recently, adversarial examples were also found in traditional machine learning models [24] and in many other application domains, e.g., speech recognition [25], text classification [26], malware identification [27], etc.

This article aims to expose a critical security concern in EEG-based BCI spellers, and more broadly, EEG-based BCIs, which has received little attention before. It shows that one can generate tiny adversarial EEG perturbation templates for target attacks for both P300 and SSVEP spellers, i.e., mislead the classification to any character the attacker wants, regardless of what the user intended character is. The consequence could range from merely user frustration to severe misdiagnosis in clinical applications [28]. We believe a new and more detailed understanding of how adversarial EEG perturbations affect BCI classification can inform the design of BCIs to defend such attacks.

There have been some studies on adversarial attacks of time-series signals [25], [28]; however, they need to know the full time-series before computing the adversarial perturbations, which means these approaches are not causal and hence cannot be implemented in real-world applications. For example, to attack a voice command, previous approaches need to record the entire voice command first, and then design the perturbation. However, once the perturbation is obtained, the voice command has already been sent out (e.g., to a smart phone or Amazon Echo), so there is no chance to add the perturbation to the voice command to actually perform the attack.

What distinguishes the attack approaches in this article most from previous ones is that this article explicitly considers the causality in designing the perturbations. The adversarial perturbation template is constructed directly from the training set and then fixed. So, there is no need to know the test EEG trial and compute the perturbation specifically for it. The perturbation can be directly added to a test EEG trial as soon as it starts, and hence satisfies causality and can be implemented in practice. Thus, it calls for an urgent need to be aware of such attacks and defend them.

A closely related concept is universal adversarial perturbations [29], which can also be viewed as adversarial perturbation templates and have been used to attack deep learning models in image classification. In this article, we focus on the security of a traditional and most frequently used BCI pipeline, which consists of separate feature extraction and classification steps, whereas universal adversarial perturbations are usually designed for non-target attacks of end-to-end deep learning models.

II. RESULTS

A. Performance evaluation

We used two measures to evaluate the performance of a BCI speller. The first is classification accuracy, and the other is the ITR [30], which measures the speed of the speller:

$$\text{ITR} = \frac{1}{T}\left[\log_2 Q + R\log_2 R + (1-R)\log_2\frac{1-R}{Q-1}\right], \quad (1)$$

where T is the average time (minutes) spent to input a user character, Q the number of different characters (which was 36 in our P300 speller and 40 in the SSVEP speller), and R the classification accuracy. The unit of ITR is bits/min. When the classification accuracy is lower than random guess, i.e., $R \leq \frac{1}{Q}$, the ITR is directly set to 0.

To distinguish between the character the user wants to spell, and the one the attacker wants to mislead to, we denote the former *user character*, and the latter *attacker character*. Accordingly, *user score* and *user ITR* are used to describe the classification accuracy of user characters and the corresponding ITR, respectively. An *attacker score* is defined as the ratio that the perturbation template leads the speller to output an attacker character, and the corresponding *attacker ITR* is calculated by replacing R in equation (1) with the attacker score. A higher attacker score or attacker ITR represents a better target attack performance.

B. Security of the P300 speller

Dataset: We used a public P300 dataset (dataset II) introduced by by Rrusienski and Schalk [31]. It recorded 64-channel EEG signals from two subjects (A and B). The EEG data were sampled at 240 Hz, bandpass filtered to 0.1-40 Hz, then z-normalized for each channel. There were 85 training character trials and 100 test ones for each subject.

For each trial, a set of 12 random intensifications (six rows and six columns) were repeated 15 times (i.e., each row was intensified 15 times, and each column was also intensified 15 times). Each intensification lasted for 100 ms, after which the character matrix was blanked for 75 ms. So, it took $(100 + 75) \times 12 \times 15 = 31,500$ ms, or 31.5 s, to input a character. The spelling speed can be improved by using fewer repeats, e.g., 10 or 5; however, the spelling accuracy generally decreases with a smaller number of repeats.

The victim model: The victim model was a Riemannian geometry based approach, which won the Kaggle BCI challenge[1] in 2015. First, 16 xDAWN spatial filters [32], eight for the target trials and another eight for the non-target trials, were designed to filter all the trials. The template-signal covariance matrices of the EEG epochs were projected onto the tangent space of a Riemannian manifold [33]–[35], using Affine Invariant Riemannian Metric as its distance metric. Finally, we classified the feature vectors with a Logistic Regression model in the tangent space. The details can be found in the Supplementary Information. The model was trained with class-specific weights to accommodate class imbalance. All operations in these blocks are differentiable, so we re-implemented them using Tensorflow [36] to facilitate the gradient calculation.

To get the label (target or non-target) of an intensification, an epoch between 0-600 ms from the beginning of the intensification was extracted and fed into the victim model to calculate the target probability. Because each row and column was intensified multiple times, voting was performed for each trial to get the target row and target column, and hence the target character.

Baseline performance: The first part of Table I shows the baseline performance of the clean EEG data (without adding any perturbations). As the number of intensification repeats increased, the user score increased, indicating that the classification accuracy of the user characters increased. Meanwhile, the user ITR decreased, because the time needed to input each character significantly increased.

The second part of Table I shows the baseline performance when we added Gaussian noise to the raw EEG data, averaged over 10 runs. The Gaussian noise perturbations were preprocessed in the same way as the adversarial perturbations, by replacing $\widetilde{P}$ in equation (6) with Gaussian noise, so that they had the same energy. We use signal-to-perturbation ratio (SPR) to quantify the magnitude of the perturbation, which is also presented in the second part of Table I. Gaussian noise perturbations had almost no impact on the user score and the user ITR at all, not to mention forcing the P300 speller to output a specific attacker character. These results suggest that more sophisticated adversarial perturbations are needed to attack the P300 speller.

Performance under adversarial attacks: We added the adversarial perturbation template to the test EEG trials to validate whether it was effective in misleading the P300 speller. Fig. 2 shows the attacker scores of the 36 characters. The attacker can manipulate the P300 speller to spell whatever

character he/she wants, regardless of what the user intended character is, with higher than 90% average success rate.
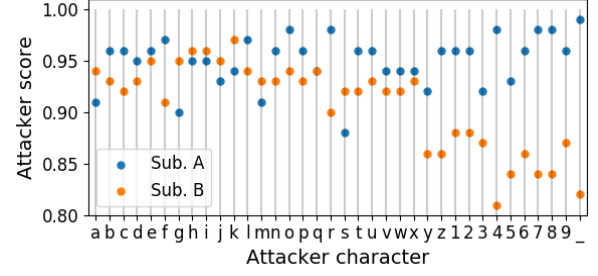

Fig. 2: **Attacker scores of manipulating the P300 speller to misclassify the 100 test character trials into a specific attacker character.** The P300 speller used 15 intensification repeats for each character.

The third part of Table I shows the average user scores and ITRs with different number of intensification repeats. The user scores and ITRs were close to zero, suggesting that the user almost cannot correctly input the character he/she wanted.

The fourth part of Table I shows the average attacker scores and ITRs with different numbers of intensification repeats. The attacker score increased with the number of intensification repeats, because more repeats increased the number of times that the attacker can inject the perturbation into the benign EEG trial.

To better quantify the magnitude of the perturbations, we also calculated two SPRs. The adversarial perturbation template was only added at some specific periods of the EEG trial, as shown in Fig. 3a, therefore we defined a *period SPR* to measure the SPR of the perturbed period, and also a *trial SPR* to measure the SPR of the entire trial. The last part of Table I shows these SPRs. They were higher than 20 dB, suggesting that the adversarial perturbation template may be undetectable when added to benign EEG trials.

Visualization of the adversarial perturbations: In addition to high attack performance, another requirement in adversarial attacks is that the perturbations should not be detected easily.

Fig. 3a shows a typical EEG trial before and after the adversarial perturbation on Subject A. For clarity, we only show channels F3, F4, Cz, P3 and P4, which evenly distribute on the scalp. One can barely distinguish the adversarial EEG trial from the original EEG trial.

A traditional way to visualize the P300 signal is to take the average of multiple P300 trials. We also took this approach to check if there was a noticeable difference between the average target (or non-target) trials, before and after perturbation. Fig. 3b shows the results from the Cz channel. One can hardly observe any differences. Fig. 3b also shows the spectrograms and topoplots of the difference between the average target EEG trial and the average non-target EEG trial. The original and adversarial spectrograms (or topoplots) show very similar energy distributions, and are hardly distinguishable by human eyes.

---

[1]https://www.kaggle.com/c/inria-bci-challenge

TABLE I: **P300 speller attack results. Before attack: Baselines on clean EEG data (without adding any perturbations) and Gaussian-noise-perturbed EEG data, and the corresponding SPRs (dB). After attack: Average user/attacker scores/ITRs of the 36 attacker characters in target attacks, and the corresponding period and trial SPRs (dB).**

| Subject | Number of repeats | Before attack | | | | | After attack | | | | Period SPR | Trial SPR |
| | | Clean | | Gaussian noise | | | User | | Attacker | | | |
		Score	ITR	Score	ITR	SPR	Score	ITR	Score	ITR		
A	5	0.64	13.07	0.65	13.40	20.8	0.072	0.248	0.825	19.8	20.8	25.8
	10	0.85	10.62	0.84	10.40	21.0	0.049	0.052	0.900	11.7	21.0	25.9
	15	0.91	8.03	0.92	8.19	21.0	0.040	0.021	0.950	8.7	21.0	25.9
B	5	0.79	18.41	0.79	18.41	25.2	0.107	0.578	0.713	15.6	25.2	30.2
	10	0.91	11.96	0.89	11.50	25.5	0.061	0.093	0.860	10.9	25.5	30.4
	15	0.93	8.35	0.91	8.03	25.6	0.049	0.034	0.907	8.0	25.6	30.5

C. Security of the SSVEP Speller

Dataset: The dataset was first introduced by Wang *et al.* [19] as a benchmark dataset for SSVEP-based BCIs. The 64-channel signals were recorded from 35 subjects using an extended 10-20 system. During the experiments, the subjects were facing a monitor, in which a 5×8 character matrix was flickering. Different flickering frequencies were assigned to the 40 characters respectively, ranging from 8 Hz to 15.8 Hz with 0.2 Hz increment, as shown in Fig. 1c. Six blocks of EEG signals were recorded from each subject, each with 40 trials, corresponding to the 40 target characters. Each trial was downsampled to 250 Hz and lasted 6 seconds, including 0.5 s before stimulus onset, 5 s for stimulation, and 0.5 s after stimulus offset.

Chen *et al.* [37] showed that an SSVEP at the stimulation frequency and its harmonics usually starts to be evoked with a delay around 130-140 ms; hence, we extracted EEG signals between [0.13, 1.38] s after the stimulus onset as the input to the victim model. Nine channels over the occipital and parietal areas (Pz, POz, PO3, PO4, PO5, PO6, Oz, O1 and O2) were chosen. The signals were bandpass filtered to 7-90 Hz with a fourth-order butterworth filter.

The victim model: Extracting the frequency information of SSVEPs is an essential step in recognizing the stimulation frequency, and hence the user character. A natural solution is to utilize fast Fourier transform to estimate the spectrum, so that the energy peaks can be matched to the stimulation frequency; however, canonical correlation analysis (CCA) was recently shown to be more promising in identifying the stimulation frequency [37], [38]. Thus, CCA-based frequency recognition was used in the victim model.

CCA is a statistic approach which can be used to extract the underlying correlation between two multi-channel time series [39]. Its main idea is to find a linear combination of channels for each time series, so that their correlation is maximized. When applied to SSVEP spellers, CCA is utilized to calculate the maximum correlation between the input EEG signals and a standard reference signal, which consists of the sinusoidal signal of a stimulation frequency and its $(N_q - 1)$ harmonics ($N_q = 5$ in our case).

Mathematically, let $X \in \mathbb{R}^{N_e \times N_s}$ denote an EEG trial with N_e channels and N_s samples, and Y_f a standard reference

signal of stimulation frequency f. The (c, n)-th entry of Y_f is:

$$Y_f(c, n) = \begin{cases} \sin\left((c+1)\pi \frac{f}{f_s} n\right), & c \text{ is odd} \\ \cos\left(c\pi \frac{f}{f_s} n\right), & c \text{ is even} \end{cases}, \quad (2)$$

where f_s is the sampling rate, $1 \le c \le 2N_q$, and $1 \le n \le N_s$. To calculate the maximum correlation coefficient $\rho(X, Y_f)$, X and Y_f are first z-normalized, and then $\rho(X, Y_f)$ is computed as the square root of the largest eigenvalue of matrix

$$S(X, Y_f) = (XX^T)^{-1} X Y_f^T (Y_f Y_f^T)^{-1} Y_f X^T, \quad (3)$$

i.e.,

$$\rho(X, Y_f) = \sqrt{\lambda_{\max}\left(S(X, Y_f)\right)}. \quad (4)$$

More detailed derivations can be found in the Supplementary Information.

Let $F = \{f_i\}_{i=1}^K$ be the set of K candidate stimulation frequencies ($K = 40$ in our case). Then, the SSVEP speller outputs the character corresponding to the following stimulation frequency:

$$f^* = \arg\max_{f \in F} \rho(X, Y_f). \quad (5)$$

Baseline performance: Among the 35 subjects, eight with the best baseline performances (shown in the first part of Table II) were used in our experiments.

Because SSVEPs are highly susceptible to periodic noise, we evaluated the robustness of the victim model to both Gaussian noise and sinusoidal noise of a random frequency chosen from 40 stimulation frequencies, and a random phase chosen from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$. The SPRs were all set to 25 dB, so that the energy of the Gaussian noise and periodic noise was comparable to that of the adversarial perturbation templates. The second and third parts of Table II show the results on these noisy data, averaged over 10 runs, respectively. The victim model was almost completely immune to the Gaussian noise. The periodic noise degraded the model performance more than the Gaussian noise.

Performance under adversarial attacks: We generated 40 adversarial perturbation templates, each forcing the SSVEP speller to output a specific character. Fig. 4 shows their attacker scores. For six of the eight subjects, their output character can be manipulated to any character the attacker
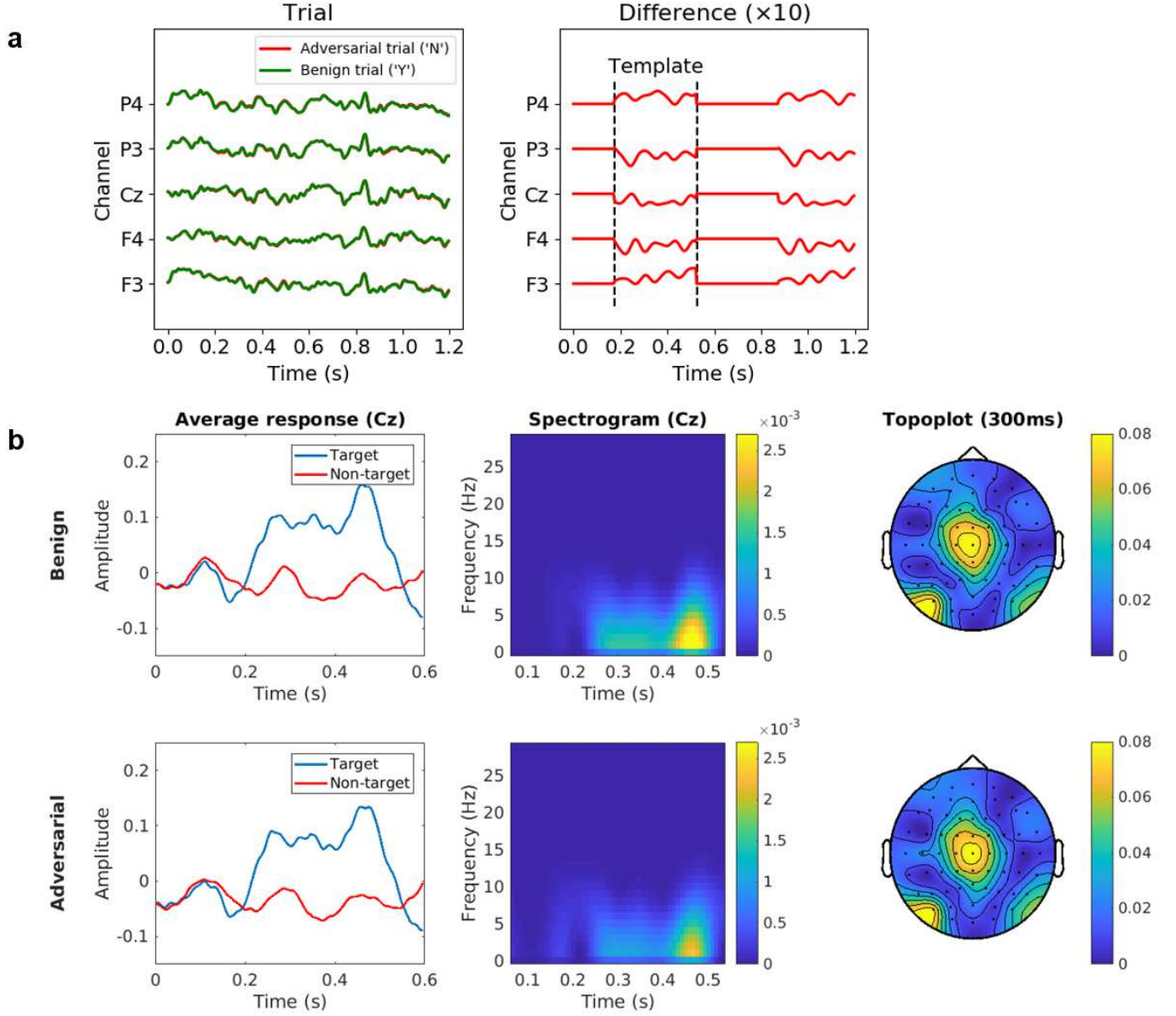
Fig. 3: **Visualization of the adversarial perturbations for Subject A. a**, EEG trials before and after adversarial perturbation, and the difference (magnified ten times) between the adversarial trial and the benign trial. The non-zero part of the difference is the adversarial perturbation template, which is added to a benign EEG trial according to the attacker character. The adversarial perturbation led the P300 speller to misclassify letter *Y* into *N*. **b**, left column: average of $100 \times 15 \times 2 = 3,000$ target trials and average of $100 \times 15 \times 10 = 15,000$ non-target trials in channel Cz, for benign and adversarial trials; middle column: spectrogram of the difference between the average target trial and the average non-target trial in channel Cz, for benign and adversarial trials; right column: topoplot of the difference between the average target trial and the average non-target trial, for benign and adversarial trials.

wanted, at 70%-100% success rate. Interestingly, due to individual differences, Subjects 3 and 25 showed some resistance to adversarial perturbation templates.

The fifth and sixth parts of Table II show the averaged user and attacker performances, respectively. The adversarial perturbation templates were very effective on most subjects (except Subjects 3 and 25), reducing both the user scores and the user ITRs to almost zero, i.e., the user almost cannot correctly input any character he/she wanted. The attacker

scores for five subjects were close to one, i.e., the attacker was able to force the SSVEP speller to output any character he/she wanted. The SPRs were all around 25 dB, comparable to the SPRs for random noise.

Visualization of the adversarial perturbations: This subsection shows the characteristics of the adversarial perturbation templates, and verifies their imperceptibility to some widely-used approaches for evaluating the quality of SSVEPs.

Fig. 5a shows the EEG signals before and after adversarial

TABLE II: **SSVEP speller attack results. Before attack: Baselines on clean data (without adding any perturbations), Gaussian-noise-perturbed EEG data and periodic-noise-perturbed EEG data. After attack: Average user/attacker scores/ITRs of 40 attacker characters in target attacks, and the corresponding SPRs (dB).**

| Subject	Before attack							After attack				
	Clean		Gaussian Noise		Periodic Noise		SPR	User		Attacker		SPR
	Score	ITR	Score	ITR	Score	ITR		Score	ITR	Score	ITR	
3	0.88	182.5	0.88	181.6	0.73	135.6	25.0	0.44	61.1	0.58	93.3	25.3
4	0.90	186.9	0.90	187.1	0.71	127.8	25.0	0.07	2.3	0.95	210.1	25.7
12	0.90	188.8	0.90	188.0	0.82	160.7	25.0	0.26	26.6	0.75	139.5	25.5
22	0.82	160.0	0.79	150.2	0.77	146.5	25.0	0.11	6.1	0.91	191.0	25.1
25	0.90	189.1	0.89	184.1	0.87	178.1	25.0	0.78	148.2	0.17	13.8	26.7
26	0.90	187.8	0.88	180.3	0.60	98.8	25.0	0.03	0.1	1.00	229.9	24.8
32	0.87	176.9	0.87	179.6	0.64	110.3	25.0	0.03	0.0	1.00	231.4	24.9
34	0.80	154.7	0.79	151.8	0.47	71.3	25.0	0.03	0.0	1.00	231.2	25.9
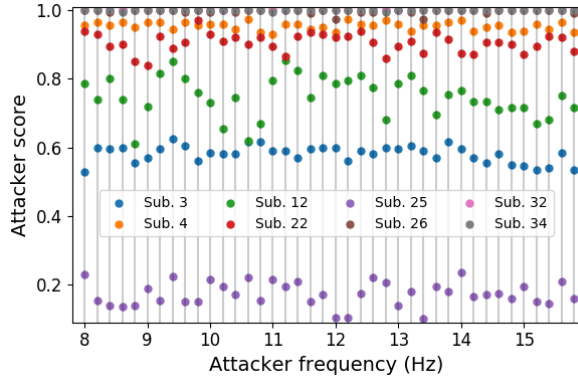
Fig. 4: **Attacker scores of manipulating the SSVEP speller to misclassify the $5 \times 40 = 200$ test character trials into a specific attacker frequency (character).**
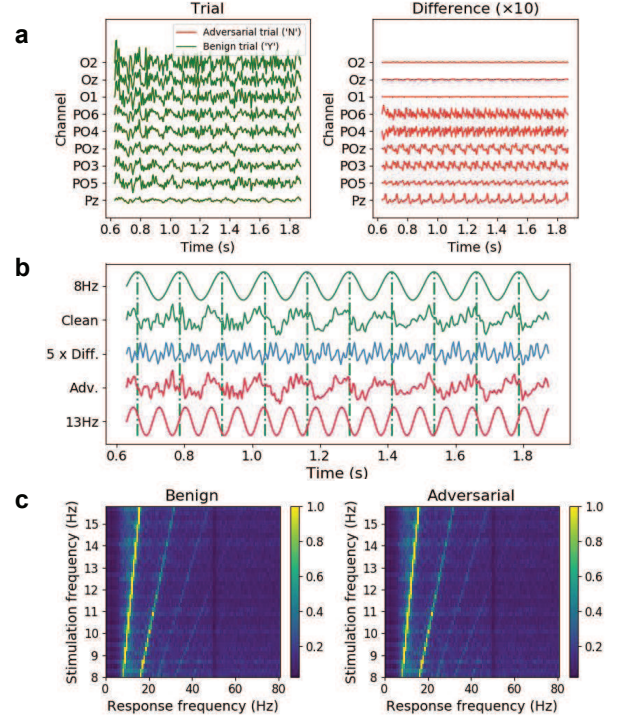

Fig. 5: **SSVEP speller attack results. a**, left column: EEG trials before and after adversarial perturbation, for Subject 26; right column: the difference (adversarial perturbation) between the adversarial EEG trial and the benign EEG trial for Subject 26, magnified by ten times to it them visible. The adversarial perturbation led the SSVEP speller to misclassify letter Y into N. **b**, detailed signal analysis for Channel POz of Subject 26. The clean signal was the average of all six trials of 8 Hz stimulation frequency, and the adversarial trial was the average of the same trials with δ_{13Hz} added. Standard 8 Hz and 13 Hz sinusoidal signals are shown as references. The green dot-dashed lines mark the 8 Hz periodicity. **c**, Normalized spectra of SSVEPs for 40 stimulation frequencies, averaged over all the chosen channels and all 40 subjects.

perturbations, along with the magnified difference. The SSVEP speller misclassified the user character, which was supposed to be Y (8.6 Hz), into N (13.2 Hz). Human eyes can barely recognize the difference between the benign and the adversarial EEG trials. After being magnified by 10 times, the perturbation looks periodical, which can modify the user frequency to the attacker frequency.

We compared the clean and adversarial EEG signals with standard sinusoidal signals in Fig. 5b, using Subject 26 as an example. We took the average of the clean temporal waveforms of 8 Hz SSVEPs from Channel POz, and did the same for their adversarial signals with δ_{13Hz} added (which forced the SSVEP speller to output the character of 13 Hz stimulation frequency). We chose Channel POz because the adversarial perturbation on this channel had one of the largest amplitudes, as shown in Fig. 5a. Fig. 5b shows that both clean and adversarial EEG signals were synchronized with the standard 8 Hz sinusoidal signal, indicated by the green dot-dashed lines. Comparing the 13 Hz sinusoidal signal with the magnified difference, the synchronization can also be observed, suggesting that the adversarial perturbation template introduced a frequency component matching the attacker character, which was imperceptible to human eyes but powerful enough to mislead the SSVEP speller.

Fig. 5c shows the spectrum analysis of SSVEPs for 40 stimulation frequencies. We averaged the spectra of the benign EEG signals of the same stimulation frequency from all 40 subjects and all chosen channels, so that background activities can be suppressed. The left column of Fig. 5c, for

benign trials, clearly shows that the visual stimulus, flickering at a stimulation frequency, can evoke SSVEP of the same frequency and its harmonics. The right column of Fig. 5c shows the same property of adversarial trials, whose attacker character was randomly chosen and fixed for each stimulation frequency. We cannot observe noticeable differences between the two columns in Fig. 5c, demonstrating the challenge in detecting the adversarial perturbation templates.

III. DISCUSSION

This article aims to expose a critical security concern in EEG-based BCI spellers, and more broadly, EEG-based BCIs, which has received little attention before. It shows that one can generate adversarial EEG perturbation templates for target attacks for both P300 and SSVEP spellers, i.e., deliberately-designed tiny perturbations can manipulate an EEG-based BCI speller to output anything the attacker wants with high success rate, demonstrating the vulnerability of BCI spellers. We should emphasize that these attacks are not specific to the victim models used in this article. They may also be used to attack many other classifiers in BCIs with little modification.

It should be noted that the current approaches have two limitations: (a) they require some subject-specific EEG trials to construct the adversarial perturbation template; and, (b) they need to know the exact timing of the stimulus to achieve the best attack performance. They could be more dangerous if these limitations are resolved.

The first limitation may be alleviated by utilizing the transferability of adversarial examples, which was one of the most dangerous properties of adversarial examples. It was first discovered by Szegedy *et al.* [20] in 2014 and further investigated by many others [24], [40]–[42]. The transferability means that adversarial examples generated from one model can also be used to attack another model, which may have a completely different architecture and/or be trained from a different dataset. Thus, it may be possible to construct the adversarial perturbation template from some existing subjects and then apply it to a new subject.

The second limitation is that the attacker needs to know the precise time synchronization between adversarial perturbation templates and EEG signals. To study how the synchronization time delay affects the attack performance, we show the relationship between the user/attacker scores and the time delay in adding the perturbation template in Fig. 6. For the P300 speller, when the time delay increased, the user scores increased rapidly and the attacker score decreased rapidly, suggesting that the adversarial template was sensitive to the synchronization. The SSVEP speller was more robust to the time delay. These results may suggest that hiding the time synchronization information can help defend adversarial attacks in EEG-based BCI spellers. However, attacks insensitive to the synchronization may also be possible. For example, the idea of *adversarial patch* [43], which is a tiny picture patch that can mislead the classifier when added anywhere to a large picture to be classified, may be used to increase the robustness to the synchronization time delay. Thus, defending the attackers may not be an easy task.
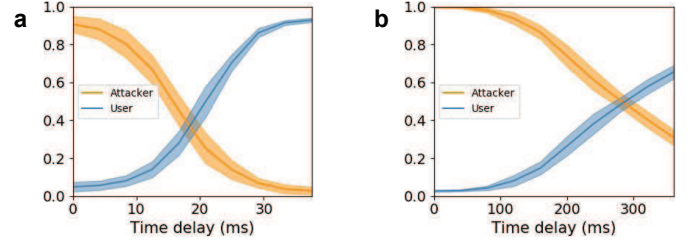

Fig. 6: **User and attacker scores with respect to the synchronization time delay.** The curve represents the mean of all attacker characters, and the shadow the standard deviation. **a**, scores for the P300 speller, where 100 test trials for Subject A were perturbed to be misclassified as each of the 36 attacker characters. **b**, scores for the SSVEP speller, where $5 \times 40 = 200$ test trials for Subject 26 were perturbed to be misclassified as each of the 40 attacker characters.

Finally, we need to emphasize again that the goal of this study is not to damage EEG-based BCIs. We aim to demonstrate that serious adversarial attacks to EEG-based BCIs are possible, and hence expose a critical security concern, which received little attention before. Our future research will develop strategies to defend such attacks. Meanwhile, we hope our study can attract more researchers' attention to the security of EEG-based BCIs.

IV. METHODS

A. Attack the P300 Speller

The main idea to construct the adversarial perturbation template was to find a universal perturbation which leads the P300 classifier to classify non-target epochs into target ones. The approach was to get the directions pointing from non-target epochs to the decision boundary of the victim model, and then sum up these directions as the universal perturbation. These directions can be identified by simply calculating the gradients of the loss with respect to the input non-target EEG epochs, assuming the decision boundary is linear. Though the victim model includes nonlinear operations, the attack approach still worked surprisingly well.

Let X be an EEG trial, y its label (0 for non-target, and 1 for target), f the victim model which gives the label probability for each input X, $J(X, y, f)$ the loss function (cross-entropy loss in our case), and D_{NT} the dataset containing all non-target epochs in the training set. Then, the overall direction can be computed as:

$$\widetilde{P} = \sum_{(X,y) \in D_{NT}} \frac{\nabla_X J(X, 1-y, f)}{\|\nabla_X J(X, 1-y, f)\|_F}. \quad (6)$$

After obtaining $\widetilde{P}$, we filtered it by a fourth-order Butterworth bandpass filter of $[0.1, 15]$ Hz, extracted the first 350ms signal, and then normalized it in each channel so that the L2 norm is 1. Denote the result as $\widehat{P}$. Then, the adversarial perturbation P was computed as:

$$P = \epsilon \cdot \widehat{P}, \quad (7)$$

where ϵ is a constant controlling the energy of the perturbation ($\epsilon = 0.5$ in our experiments).

To mislead the P300 speller, one only needs to tamper with some specific signal periods according to the onset of the target stimuli. Because in a practical P300 speller the same row or column is never intensified successively, the perturbation template can last more than one intensification period. In our experiments, the template lasted $2 \times 175 = 350$ ms, i.e., two intensification periods.

Fig. 7 illustrates the attack procedure. The benign EEG trial would output character *7*, since the last row and the third column of the character matrix have the highest P300 probability, and their intersection is *7*. However, after applying the perturbation template, the trial outputs the character *Z*, because the fifth row and the second column have the highest P300 probability. Interestingly, the adversarial template acts like random noise when it is not synchronized with an intensification onset. As shown in Fig. 7, the last 175 ms of the template does not influence the classification of the corresponding intensification.

B. Attack the SSVEP Speller

There are two difficulties in attacking the victim model of the SSVEP speller. First, the victim model is not fixed, as the parameters of CCA vary in different EEG trials. Second, unlike the P300 speller whose base victim model only needs to classify the input into two classes, there are many more classes in the SSVEP speller. These make adversarial attacks of the SSVEP speller much more challenging.

The purpose was to generate the adversarial perturbation template $\delta_{\widehat{f}} \in \mathbb{R}^{N_e \times N_s}$, which can lead the SSVEP speller to output the attacker character of stimulation frequency $\widehat{f}$. For each user, we used the first block $\mathcal{D} = \{X_i\}_{i=1}^{N}$ to craft $\delta_{\widehat{f}}$, and the remaining five blocks to evaluate its attack performance.

According to the victim model, $\delta_{\widehat{f}}$ should be able to maximize $\rho(X + \delta_{\widehat{f}}, Y_{\widehat{f}})$ in equation (4), such that

$$\arg\max_{f \in F} \rho(X + \delta_{\widehat{f}}, Y_f) = \widehat{f}. \tag{8}$$

In other words, $\delta_{\widehat{f}}$ can be crafted by solving

$$\max_{\delta_{\widehat{f}}} \sum_{X \in \mathcal{D}} \lambda_{\max}(S(X + \delta_{\widehat{f}}, Y_{\widehat{f}})), \tag{9}$$

where $S(X, Y)$ is defined in equation (3).

Since $S(X + \delta_{\widehat{f}}, Y_f)$ is not symmetric, it is difficult to calculate the derivatives of its largest eigenvalue, resulting in challenges in optimization. Because of the fact that the largest eigenvalue is always no smaller than the average of all eigenvalues:

$$\lambda_{\max}(S(X + \delta_{\widehat{f}}, Y_f)) \geq \frac{1}{N_e} \sum_{j} \lambda_j(S(X + \delta_{\widehat{f}}, Y_f))$$
$$= \frac{1}{N_e} \mathrm{tr}(S(X + \delta_{\widehat{f}}, Y_f)), \tag{10}$$

instead of solving equation (9) directly, we can maximize its lower bound to reduce the optimization difficulty:

$$\max_{\delta_{\widehat{f}}} \sum_{X \in \mathcal{D}} \mathrm{tr}(S(X + \delta_{\widehat{f}}, Y_{\widehat{f}})). \tag{11}$$

Because the effective frequency band of SSVEP signals is 7-90 Hz, we introduced a new variable $\mathbf{r}_{\widehat{f}}$ so that

$$\delta_{\widehat{f}} = \mathrm{filt}(\mathbf{r}_{\widehat{f}}), \tag{12}$$

where $\mathrm{filt}(\cdot)$ means retaining only the 7-90 Hz effective signal frequency components. As a result, we can ensure the integrity of the adversarial template during signal filtering. In addition, we added $\alpha \cdot \|\delta_{\widehat{f}}\|_F$ to penalize the energy of the perturbation, where α is the penalty coefficient.

Finally, the problem becomes:

$$\min_{\mathbf{r}_{\widehat{f}}} -\sum_{X \in \mathcal{D}} \mathrm{tr}(S(X + \mathrm{filt}(\mathbf{r}_{\widehat{f}}), Y_{\widehat{f}})) + \alpha \cdot \|\mathrm{filt}(\mathbf{r}_{\widehat{f}})\|_F. \tag{13}$$

Gradient descent was used to update $\mathbf{r}_{\widehat{f}}$, and the iteration stopped when the SPR was lower than a threshold, which was set to 25dB in our experiments.

REFERENCES

[1] B. Graimann, B. Allison, and G. Pfurtscheller, *Brain-computer interfaces: A gentle introduction*. Berlin, Heidelberg: Springer, 2009, pp. 1–27.

[2] C. T. Lin, Y. T. Liu, S. L. Wu, Z. Cao, Y. K. Wang, C. S. Huang, J. T. King, S. A. Chen, S. W. Lu, and C. H. Chuang, "EEG-based brain-computer interfaces: A novel neurotechnology and computational intelligence method," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 3, no. 4, pp. 16–26, Oct. 2017.

[3] H. He and D. Wu, "Transfer learning for brain-computer interfaces: A Euclidean space data alignment approach," *IEEE Trans. on Biomedical Engineering*, vol. 67, no. 2, pp. 399–410, 2020.

[4] R. Chavarriaga and J. d. R. Millán, "Learning from EEG error-related potentials in noninvasive brain-computer interfaces," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 18, no. 4, pp. 381–388, 2010.

[5] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, Feb. 2012.

[6] L. A. Farwell and E. Donchin, "Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalography and Clinical Neurophysiology*, vol. 70, no. 6, pp. 510–523, Dec. 1988.

[7] X. Chen, Y. Wang, M. Nakanishi, X. Gao, T.-P. Jung, and S. Gao, "High-speed spelling with a noninvasive brain-computer interface," *Proc. National Academy of Sciences*, vol. 112, no. 44, pp. E6058–E6067, 2015.

[8] S. Sutton, M. Braren, J. Zubin, and E. R. John, "Evoked-potential correlates of stimulus uncertainty," *Science*, vol. 150, no. 3700, pp. 1187–1188, Nov. 1965.

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

[10] P. Meinicke, M. Kaper, F. Hoppe, M. Heumann, and H. Ritter, "Improving transfer rates in brain computer interfacing: A case study," in *Proc. Advances in Neural Information Processing Systems*, BC, Canada, Dec. 2003, pp. 1131–1138.

[11] N. Xu, X. Gao, B. Hong, X. Miao, S. Gao, and F. Yang, "BCI competition 2003-data set IIb: Enhancing P300 wave detection using ICA-based subspace projections for BCI applications," *IEEE Trans. on Biomedical Engineering*, vol. 51, no. 6, pp. 1067–1072, Jun. 2004.

[12] C. Guan, M. Thulasidas, and J. Wu, "High performance P300 speller for brain-computer interface," in *Proc. IEEE Int'l Workshop on Biomedical Circuits and Systems*, Singapore, Dec. 2004, pp. S3–5.

[13] J. Polich, "Updating P300: An integrative theory of P3a and P3b," *Clinical Neurophysiology*, vol. 118, no. 10, pp. 2128–2148, Oct. 2007.

[14] R. M. Chapman and H. R. Bragdon, "Evoked responses to numerical and non-numerical visual stimuli while problem solving," *Nature*, vol. 203, no. 4950, p. 1155, Sep. 1964.

[15] S. Sutton, P. Tueting, J. Zubin, and E. R. John, "Information delivery and the sensory evoked potential," *Science*, vol. 155, no. 3768, pp. 1436–1439, Mar. 1967.

[16] F. Beverina, G. Palmas, S. Silvoni, F. Piccione, S. Giove *et al.*, "User adaptive BCIs: SSVEP and P300 based interfaces." *PsychNology Journal*, vol. 1, no. 4, pp. 331–354, Dec. 2003.
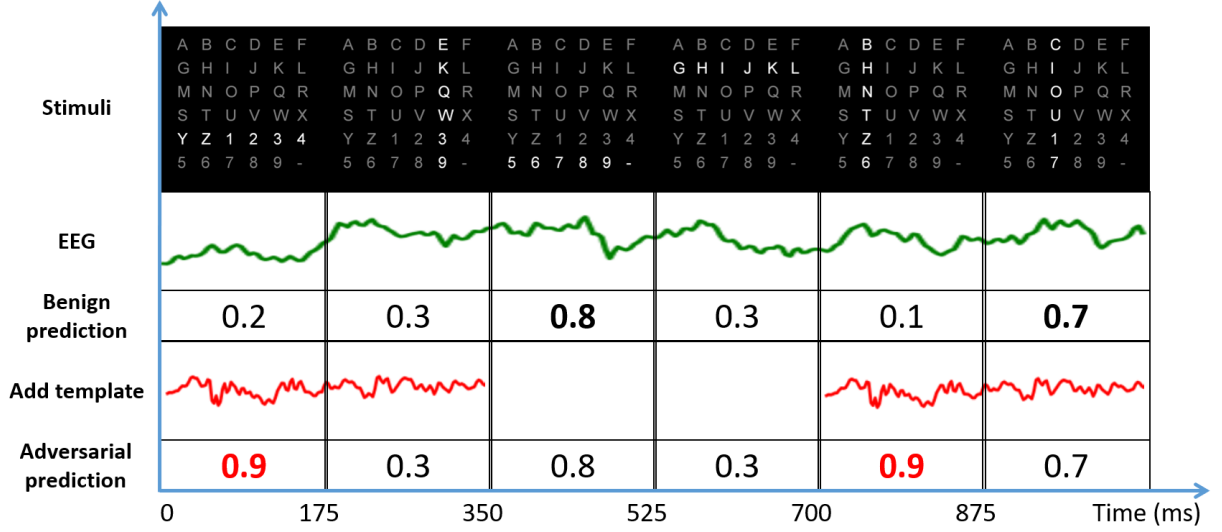
Fig. 7: **Illustration of the attack procedure.** The attacker character is *Z*, whereas the user character is *7*. For the benign EEG trial, the P300 speller can correctly identify that P300 is elicited by the intensifications of the last row and the third column. To mislead the P300 speller, adversarial perturbation template is added during the periods of 0-350ms and 700-1050ms, so that the fifth row and the second column are believed to elicit P300 with the highest probability. The added adversarial perturbation templates do not influence the results of the second and the last stimuli, because their corresponding periods are out of synchronization with the templates. As a result, the P300 speller misclassifies the perturbed trial to attacker character *Z*.

[17] Y. Wang, X. Gao, B. Hong, C. Jia, and S. Gao, "Brain-computer interfaces based on visual evoked potentials," *IEEE Engineering in Medicine and Biology Magazine*, vol. 27, no. 5, pp. 64–71, Sep. 2008.
[18] F.-B. Vialatte, M. Maurice, J. Dauwels, and A. Cichocki, "Steady-state visually evoked potentials: Focus on essential paradigms and future perspectives," *Progress in Neurobiology*, vol. 90, no. 4, pp. 418–438, Apr. 2010.
[19] Y. Wang, X. Chen, X. Gao, and S. Gao, "A benchmark dataset for SSVEP-based brain-computer interfaces." *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1746–1752, Oct. 2017.
[20] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int'l Conf. on Learning Representations*, Banff, Canada, Apr. 2014.
[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int'l Conf. on Learning Representations*, San Diego, CA, May 2015.
[22] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. Int'l Conf. on Learning Representations*, Toulon, France, Apr. 2017.
[23] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proc. 35th Int'l Conf. on Machine Learning*, Stockholm, Sweden, Jul. 2018, pp. 284–293.
[24] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," *CoRR*, vol. abs/1605.07277, 2016. [Online]. Available: https://arxiv.org/abs/1605.07277
[25] N. Carlini and D. A. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE Symposium on Security and Privacy*, San Francisco, CA, May 2018, pp. 1–7.
[26] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," *CoRR*, vol. abs/1707.07328, 2017. [Online]. Available: https://arxiv.org/abs/1707.07328
[27] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, "Adversarial perturbations against deep neural networks for malware classification," *CoRR*, vol. abs/1606.04435, 2016. [Online]. Available: https://arxiv.org/abs/1606.04435
[28] X. Zhang and D. Wu, "On the vulnerability of CNN classifiers in EEG-based BCIs," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 5, pp. 814–825, May 2019.
[29] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, Jul. 2017, pp. 1765–1773.
[30] J. R. Wolpaw, H. Ramoser, D. J. McFarland, and G. Pfurtscheller, "EEG-based communication: Improved accuracy by response verification," *IEEE Trans. on Rehabilitation Engineering*, vol. 6, no. 3, pp. 326–333, Sep. 1998.
[31] D. Rrusienski and G. Schalk, "Documentation Wadsworth BCI dataset (P300 evoked potentials)," 2004. [Online]. Available: http://www.bbci.de/competition/iii/
[32] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "xDAWN algorithm to enhance evoked potentials: Application to brain-computer interface," *IEEE Trans. on Biomedical Engineering*, vol. 56, no. 8, pp. 2035–2043, Aug. 2009.
[33] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain-computer interface classification by Riemannian geometry," *IEEE Trans. on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, Apr. 2012.
[34] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Classification of covariance matrices using a Riemannian-based kernel for BCI applications," *Neurocomputing*, vol. 112, pp. 172–178, Jul. 2013.
[35] F. Yger, M. Berar, and F. Lotte, "Riemannian approaches in brain-computer interfaces: A review," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1753–1762, Oct. 2017.
[36] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symposium on Operating Systems Design and Implementation*, Savannah, GA, Nov. 2016, pp. 265–283.
[37] X. Chen, Y. Wang, S. Gao, T.-P. Jung, and X. Gao, "Filter bank canonical correlation analysis for implementing a high-speed SSVEP-based brain–computer interface," *Journal of Neural Engineering*, vol. 12, no. 4, p. 046008, Aug. 2015.
[38] Z. Lin, C. Zhang, W. Wu, and X. Gao, "Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs," *IEEE Trans. on Biomedical Engineering*, vol. 53, no. 12, pp. 2610–2614, Nov. 2006.
[39] H. Akaike, "Canonical correlation analysis of time series and the use of an information criterion," in *Mathematics in Science and Engineering*. Elsevier, 1976, vol. 126, pp. 27–96.
[40] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *CoRR*, vol. abs/1611.02770, 2016. [Online]. Available: http://arxiv.org/abs/1611.02770

[41] F. Tramr, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "The space of transferable adversarial examples," *CoRR*, vol. abs/1704.03453v2, 2017. [Online]. Available: http://arXiv.org/abs/1704.03453v2

[42] L. Wu, Z. Zhu, C. Tai, and W. E, "Understanding and enhancing the transferability of adversarial examples," *CoRR*, vol. abs/1802.09707, 2018. [Online]. Available: http://arXiv.org/abs/1802.09707

[43] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *CoRR*, vol. abs/1712.09665, 2017. [Online]. Available: http://arxiv.org/abs/1712.09665

[44] M. R. Hestenes, "Multiplier and gradient methods," *Journal of Optimization Theory and Applications*, vol. 4, no. 5, pp. 303–320, 1969.

APPENDIX A
THE VICTIM MODEL OF THE P300 SPELLER

The details of the victim model of the P300 speller are introduced.

A. xDAWN spatial filters

The original xDAWN filter [32] was designed for P300 evoked potentials by enhancing the target response with respect to the non-target response. We used a generalized version in our experiments, which was implemented in *pyRiemann*[2].

More specifically, let $\mathbf{D} = \{(X_i, y_i)\}_{i=1}^N$ be the training set, where $X_i \in \mathbb{R}^{N_e \times N_s}$ is the i-th mean-centered EEG epoch (N_e is the number of channels, and N_s the number of time domain samples), and $y_i \in \{0, 1\}$ its corresponding label (0 for *non-target*, and 1 for *target*). The average epoch $\overline{X}_c$, $c \in \{0, 1\}$, is first calculated. Spatial filters $U_c \in \mathbb{R}^{N_f \times N_e}$ were then designed to maximize the signal to signal-plus-noise ratio for each class:

$$U_c = \arg\max_U \frac{\mathrm{tr}\left(U\overline{X}_c \overline{X}_c^T U^T\right)}{\mathrm{tr}\left(U X_{all} X_{all}^T U^T\right)}, \qquad (14)$$

where N_f is the number of filters ($N_f = 8$ was used in our experiments), X_{all} is obtained by concatenating all EEG epochs in $\mathbf{D}$ along the channels, and tr is the trace of a matrix. Generalized eigenvalue decomposition can be used to solve equation (14).

After obtaining the filters for both classes, the concatenated spatial filters $U = [U_0; U_1]$ can be used to filter each EEG epoch:

$$\widetilde{X}_i = U X_i. \qquad (15)$$

B. Tangent space projection

Covariance matrices of the EEG trials are widely-used in BCIs. However, they lie on a Riemannian manifold of Symmetric Positive Definite (SPD) matrices, and hence cannot be directly used by Euclidean space classifiers, such as Logistic Regression and Support Vector Machines. To solve this problem, the covariance matrices are projected onto the Euclidean tangent space of a reference SPD matrix, and then the vectorized features are used by Euclidean space classifiers.

More specifically, we first calculate the augmented covariance matrix C_i for each $\widetilde{X}_i$:

$$C_i = \begin{bmatrix} ZZ^T, & Z\widetilde{X}_i^T \\ \widetilde{X}_i Z^T, & \widetilde{X}_i \widetilde{X}_i^T \end{bmatrix}, \qquad (16)$$

[2]https://pyriemann.readthedocs.io/en/latest/index.html

where $Z = [U\overline{X}_0; U\overline{X}_1]$. Then, C_i is projected onto the tangent space of the reference SPD matrix C_f, which is the geometric mean of $\{C_i\}_{i=1}^N$, i.e.,

$$C_f = \arg\min_C \left(\sum_{i=1}^N \delta(C, C_i)^2 \right), \qquad (17)$$

where $\delta(C_A, C_B)$ is the Affine Invariant Riemannian Metric distance:

$$\delta(C_A, C_B) = \left\| \mathrm{logm}\left(C_A^{-1/2} C_B C_A^{-1/2}\right) \right\|. \qquad (18)$$

The vectorized features are:

$$\mathbf{s}_i = \mathrm{upper}\left(\mathrm{logm}\left(C_f^{-1/2} C_i C_f^{-1/2}\right)\right), \qquad (19)$$

where $\mathrm{upper}(\cdot)$ vectorizes the upper triangular part of a symmetric matrix. A weight of $\sqrt{2}$ is applied to the off-diagonal elements, and a weight of 1 to the rest, during the vectorization. $\mathbf{s}_i$ can then be fed into any Euclidean space classifier.

APPENDIX B
CANONICAL CORRELATION ANALYSIS (CCA)

This section introduces CCA, which can be used to extract the underlying correlation between two time series.

A. Problem setup

Let $X \in \mathbb{R}^{C_1 \times N}$ and $Y \in \mathbb{R}^{C_2 \times N}$ be two multi-channel time series, where C_1 and C_2 represent the number of channels, and N the number of time domain samples. X and Y are z-normalized in each channel.

The main idea of CCA is to find a pair of canonical variables, denoted as $\mathbf{a} \in \mathbb{R}^{C_1 \times 1}$ and $\mathbf{b} \in \mathbb{R}^{C_2 \times 1}$, for X and Y respectively, so that the correlation coefficient ρ between $\mathbf{a}^T X$ and $\mathbf{b}^T Y$ can be maximized. The problem can be mathematically formulated as:

$$\max_{\mathbf{a}, \mathbf{b}} \frac{\mathbf{a}^T XY^T \mathbf{b}}{\sqrt{\mathbf{a}^T XX^T \mathbf{a}} \sqrt{\mathbf{b}^T YY^T \mathbf{b}}}, \qquad (20)$$

which can be re-expressed as:

$$\max_{\mathbf{a}, \mathbf{b}} \quad \mathbf{a}^T XY^T \mathbf{b}, \qquad (21)$$
$$\text{s.t.} \quad \mathbf{a}^T XX^T \mathbf{a} = 1, \mathbf{b}^T YY^T \mathbf{b} = 1.$$

B. Solution of CCA

There are several approaches to solve equation (21). Here we introduce the Lagrange multiplier method [44].

Denote AB^T by S_{AB}. Then, equation (21) can be rewritten as:

$$\max_{\mathbf{a}, \mathbf{b}} \quad \mathbf{a}^T S_{XY} \mathbf{b}, \qquad (22)$$
$$\text{s.t.} \quad \mathbf{a}^T S_{XX}^T \mathbf{a} = 1, \mathbf{b}^T S_{YY}^T \mathbf{b} = 1.$$

According to the Lagrange multiplier method, equation (22) is equivalent to $\max_{\mathbf{a}, \mathbf{b}, \lambda, \theta} J(\mathbf{a}, \mathbf{b}, \lambda, \theta)$, where:

$$J(\mathbf{a}, \mathbf{b}, \lambda, \theta) = \mathbf{a}^T S_{XY} \mathbf{b} - \frac{\lambda}{2}(\mathbf{a}^T S_{XX} \mathbf{a} - 1) - \frac{\theta}{2}(\mathbf{b}^T S_{YY} \mathbf{b} - 1).$$
$$(23)$$

By setting the first partial derivatives to zero, i.e.,

$$\nabla_{\mathbf{a}}J = S_{XY}\mathbf{b} - \lambda \cdot S_{XX}\mathbf{a} = 0, \qquad (24)$$

$$\nabla_{\mathbf{b}}J = S_{YX}\mathbf{a} - \theta \cdot S_{YY}\mathbf{b} = 0, \qquad (25)$$

$$\frac{\partial J}{\partial \lambda} = -\frac{1}{2}(\mathbf{a}^T S_{XX}\mathbf{a} - 1) = 0, \qquad (26)$$

$$\frac{\partial J}{\partial \theta} = -\frac{1}{2}(\mathbf{a}^T S_{YY}\mathbf{b} - 1) = 0, \qquad (27)$$

we have

$$\lambda = \theta = \mathbf{a}^T S_{XY}\mathbf{b}. \qquad (28)$$

It should be noted that equation (28) is also the definition of the correlation coefficient ρ.

According to equations (24) and (25), we have:

$$S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{YX}\mathbf{a} = \lambda^2 \mathbf{a} = \rho^2 \mathbf{a}, \qquad (29)$$

which implies that ρ^2 equals the largest eigenvalue of $S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{YX}$, and $\mathbf{a}$ is the corresponding eigenvector.

$\mathbf{b}$ can be obtained in a similar way.