NeuroCrypt: A Coercion-Resistant Implicit Memory Authentication System

Anonymous Authors
email@email.com
Institution
Country

ABSTRACT

Coercion or Rubber-Hose attacks are a difficult problem that have not been solved in cryptography. These attacks involve an adversary coercing a user to reveal the secret of an authentication system — bypassing all cryptographic measures. Previous research in countering such attacks has leveraged the implicit memory of the user to store authentication secrets, rendering it impossible for the user to voluntarily recollect the secret. Although such systems show promise, they are not practical for real-world usage. In our work, we set out to improve the practicality and efficacy of such a system.

We propose NeuroCrypt, an extended version of the Serial Interception and Sequence Learning (SISL) task that borrows concepts from cognitive psychology and Human Computer Interaction (HCI) to improve implicit retention while minimizing the cognitive load on the user. We explore the addition of stimuli from visual and auditory modalities to improve implicit learning through a study on 60 participants.

Our results suggest that employing stimuli from visual and auditory modalities contribute to a more practical implicit learning based authentication system with minimal cognitive burden on users. One important outcome is that auditory stimuli appear to be significantly more powerful during training (in terms of user attention) as well as retention. Further, we analyze and discuss the impact of the Authentication Threshold Value (σ_{ATV}) value in our system. To illustrate the impact of this value, we determine individual authentication success rates for the visual and auditory stimuli experiments for a range of σ_{ATV} (0.025 – 0.1).

CCS CONCEPTS

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

KEYWORDS

Do, Not, Us, This, Code, Put, the, Correct, Terms, for, Your, Paper

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Anonymous Authors. 2018. NeuroCrypt: A Coercion-Resistant Implicit Memory Authentication System. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The dominant user authentication system today is "knowledge-based" – a user is required to explicitly recollect a pre-set value, like a password, from their memory to validate their identity. Passwords are cumbersome [1], impose a high cognitive load on the user, and are prone to numerous attacks [4]. One major attack that is apparently insurmountable under these conditions is a rubber-hose or coercion attack [3]. In these attacks, an adversary forces the user to reveal the password through torture or blackmail instead of trying to break the encryption. One way of evading rubber-hose attacks is by ensuring that the user is never explicitly aware of their credentials but can still somehow prove possession of the same to the system. This is where we turn to cognitive psychology for an answer.

We use our implicit memory all the time for various repetitive tasks like brushing teeth, riding a bicycle, driving, etc. These are often motor actions that are carried out without explicit recollection or mental effort and are dependent on the cerebellum and the basal ganglia. In an authentication system incorporated with implicit memory, we are able to plant a secret key into the user's mind without them having conscious access to that key. Research that involves leveraging implicit memory to remember passwords has been conducted in the past for two main reasons – for its reduced cognitive burden on the user and the ability to overcome coercion attacks. Despite a substantial amount of interest, such systems remain impractical. Typical flaws include infeasible training duration and short retention periods which lead to multiple expensive "re-learning" sessions.

We propose NeuroCrypt, a coercion-resistant authentication system that uses an extended version of a serial interception sequence learning (SISL) [18] task that leverages implicit memory to store the authentication secret. The SISL task is designed like the popular arcade game – "Guitar Hero" [26] where users intercept falling cues by pressing keys on a keyboard. What is unknown to the user is that they are being trained to implicitly learn sequences that repeat in the form of falling cues in different columns. Although previous work in this task provided success in implicit sequence learning, the long training times and short retention periods made it an impractical solution to implement in the real world. NeuroCrypt attempts to improves the task in terms of both the rate and effectiveness of implicit learning by incorporating stimuli from the

visual and auditory modalities, carefully chosen from previous work in cognitive psychology. Our study involved two 'games' (visual and auditory). Each game comprises of three phases:

- **Training Phase** – Implicit learning of the passcode sequence
- **Authentication 1** – Testing of the user's retention and implicit learning of their trained sequence after 1 week
- **Authentication 2** – A repeat of Authentication 1 after another week has passed.

NeuroCrypt is a prototype for fall-back authentication in non-critical systems. We restrict our threat model to the usual one for rubber-hose attacks: an adversary may capture a trained user and coerce them to reveal everything they know. Then the adversary takes the authentication test and attempts to gain access to the system.

To evaluate the practicality and validate our proposed framework we conducted each of the games over three weeks to account for the three phases of the game. The visual game was conducted with 28 participants, and the auditory game with 32.

Contributions

We summarize our contributions as follows:

(1) We present NeuroCrypt, a prototype authentication scheme based on implicit learning and retention that serves as an extension to the system put forth by Bojinov et al. [3]. This is achieved by using human computer interaction and cognitive psychology techniques.
(2) We demonstrate the impact of computer-human interaction modalities such as 'vision' and 'audition', and how they impact implicit memory retention of sequences. Experiments with these modalities were conducted to help identify the best individual modality for the proposed authentication scheme. This will aid future research and help develop cross-modality enabled implicit memory-based authentication schemes with higher usability.
(3) We identify and perform systematic analysis on an Authentication Threshold Value (σ_{ATV}) for such a system. It is an important parameter which determines if a user's performance is satisfactory for successful authentication.
(4) Finally, we perform security and usability analysis of our proposed authentication system under the defined threat model in.

2 PRELIMINARIES AND BACKGROUND

2.1 Implicit Memory and Learning

In cognitive psychology, implicit learning and implicit memory refer to the unconscious effect that prior information processing may exert on subsequent behavior. Implicit memory is one of the two main types of long term memory in humans along with explicit (declarative) memory. It is hypothesized that the implicit memory system is primarily related to basal ganglia along with the cerebellum. Implicit memory helps humans carry out tasks without conscious awareness of the previous experiences. Some examples of activities that employ implicit memory include typing on a computer keyboard, brushing teeth and riding a motorcycle. Implicit learning is the process of learning without intention and awareness of what has been learned. This is considered to be one of the most important and complex cognitive processes for the acquisition of most motor, perceptual and cognitive skills. Implicit memory shows higher resistance to memory deficiencies and is more robust compared to explicit memory. Several tasks have been designed to showcase implicit learning in humans like the Serial Reaction Time task (SRT) [24], Artificial Grammar Learning task (AGL) [17] and Serial Interception Sequence Learning (SISL) [18, 22]. The evidence of learning in these tasks is proof of the implicit memory system. With intelligently designed systems, implicit learning can be leveraged for required information to be embedded in the human brain.

2.2 Cognitive Load Theory

Cognitive load theory in cognitive psychology builds upon the information processing model that was accepted by the field in 1968 [2]. The essence of the model centers around the idea that information processing consists of three parts: sensory memory, working memory and learning memory. Sensory memory filters through and prioritizes sensory information that the human brain is fed during an event. This information is then passed on to the working memory where it is processed or removed. It is important to note that the working memory has a limit on how many chunks of information it can process at a given time - a feature that is the focus of the cognitive load theory. From the working memory, information is then stored in the long term memory. In his theory, John Sweller defines cognitive load as the amount of resources employed by the working memory [21]. Due to the limit of the working memory, any magnitude of information above the working memory's acceptable threshold does not cause learning. Therefore, during a training session, if a participant is bombarded with sensory information, learning is obstructed. A way to overcome this is to capitalize the working memory's ability to separately process information from different modalities.

2.3 Multimodal Processing

In cognitive psychology and Human-Computer Interaction, a modality is defined as a single independent channel of sensory input/output of a human. Examples of sensory modalities are vision (visual information), audition (auditory information) and tacticion (haptic information). When a system involves many of these modalities, it is said to be multimodal. A multimodal approach to training can be significantly effective as the working memory processes information from different modalities separately - ensuring reduced cognitive load and unimpeded learning [21]. Any human sense can be used as a computer to human modality. The most commonly used modalities are seeing and hearing as they are capable of transmitting information at a higher speed compared to other modalities – 250 to 300 [28] and 150 to 160 [27] words per minute respectively. The distribution of sensory information is carried on to implicit learning as well as reflected in studies that used a cross-modal approach to a SRT task [24]. The decision to combine visual stimuli with auditory stimuli to improve implicit learning was made due to the same reason.

2.4 Rubber-Hose Attack

Rubber-hose cryptanalysis [3] is a euphemism for the extraction of cryptographic secrets such as passwords or encrypted documents from someone forcefully by coercion or torture. Since a rubber hose is widely used in conveying systems for both pipeline and bends and in systems where a degree of natural flexibility is required, the term 'rubber-hose attacks' are used in place of the traditional term (i.e., a mathematical or cryptanalytic attack) because it signifies the similarity of the concept. Coercion here includes both psychological and physical torture. The attackers usually make threats involving severe personal consequences such as harsh legal penalties or inflicting violence upon friends and family members. Incentives to cooperate are usually plea bargains that involve clearing one's name of any criminal charges or erasure of some controversial records.

In most cryptographic systems, humans are the weakest link. Planning a direct attack on the security system's algorithm or protocols could be impractical and much more expensive as compared to simply targeting the people involved in its maintenance or usage. Thus, cryptographic systems are now being designed to ensure that human vulnerability is minimum. One such solution proposed was to employ a person's implicit memory to recall passwords. This meant that the users themselves would not be able to willingly access them since they had no conscious knowledge of them. Selecting and remembering secure passwords not only burdens the user in terms of cognitive load but also leaves them exposed to coercion attacks. An implicit memory-based authentication system eliminates the possibility of giving up the password under coercion while also minimising cognitive load. In our proposed implicit memory-based authentication system, we enhanced visual and auditory aspects of the original SISL task to analyze implicit learning and retention of the sequences.

3 RELATED WORK

Extensive research has been done on user authentication through biometrics, tokens, and passwords. This work vastly covers questions like "who are you?", "what do you have?", and "what do you know?". Biometric user authentication employs physiological and behavioural means of identification. Behavioural characteristics generally cover voiceprints, eye movement patterns, walking gait, etc., while physiological characteristics refer to fingerprints, iris recognition, face recognition, retinal scans, etc. [6, 14, 15, 19]. NeuroCrypt explores implicit learning; the system asks "what do you know without explicitly knowing it?" and comes under behavioural biometric authentication. Embedding sequences in implicit memory holds a huge advantage over other authentication methods due to relative ease and convenience. While learning a walking gait is difficult and changing one's retina is nearly impossible, embedding new sequences through a short training session in the implicit memory is relatively simple. Implicit learning systems promote segregation of confidential information since these systems sanction multiple authorization sequences per user.

Implicit Memory based Authentication Schemes

Weinshall and Kirkpatrick [11] leveraged knowledge about information recollection and storage in humans for developing a user authentication system. Participants were presented with a relatively large collection of same-sized pictures (100 – 200 arbitrarily taken from 20,000). The training phase was self-paced such that participants could go back and forth through the collection. For authentication, the participants were instructed to identify one image which was present during the training session from several sets of images. However, the proposed method employed explicit recognition of images and the model was not appropriate for deployment.

Denning et al. [12] proposed an implicit memory and priming effect based user authentication. This system uses associations between a pair of images – complete images of familiar objects and their degraded counterparts produced by using fragmented lines instead of continuous lines from the original images. Participants were initially presented with the original images and then the degraded images for authentication through a familiarisation task. Although small for an authentication scheme, a small priming effect was present in many images. However effective, the reliability and viability of the scheme depends highly on the identification and creation of images with sufficiently strong priming effect. Due to this crucial dependency, it would be improbable for the system to perform efficiently in case of a large number of users.

Castelluccia et al. [8] developed an authentication system through a Mooney images-based implicit learning approach. Mooney images are hard to label if the user hasn't been primed with the original image since they are low information two tone representations of the original images. Once a participant has been primed with the original gray-scale image from which the Mooney image is produced, recognition becomes a lot faster than otherwise. The shift from decoy distorted images to Mooney images is monumental for authentication since it triggers brain processes for implicit memory and recollection. During the priming session, the participants are presented with a Mooney image, the original image, and a label to describe the object in the images. This acquaints participants with the relation between the three vital pieces of information. For the authentication, the participants are presented with the primed and unprimed Mooney images. As the images come in a pseudo-random order, they are asked to write the label for each image and skip to the next image if unable to do so. The correctness and the recognition time for each participant is then noted down for analysis which proved successful retention.

Bojinov et al. [3] developed an implicit learning based authentication system which attempts to reduce disclosure of information through coercive attacks. The proposed scheme required participants to play a game and intercept objects falling at varying speed in four different columns until it reaches the sink at the bottom. This was performed by pressing keys on the keyboard that corresponded to the column on the screen where the circles were falling. Missing a circle or pressing the wrong key had no impact on the outcome. Authentication of the participants depended on a performance comparison of the trained and untrained sequences.

Comparison with Bojinov et al. [3]

'Neuroscience Meets Cryptography' [3] proved that it is possible for humans to implicitly learn and retain long character sequences which can later be used as authentication credentials. The results were promising in showing that there was potential in leveraging

implicit memory to overcome rubber-hose attacks. The primary limitation of the system was the impracticality in deploying the system in a real-world scenario.

In our study, we extend the SISL task designed in [3] with the focus on bolstering performance and usability. We achieve this in multiple ways. Firstly, we integrate insights from the fields of Human Computer Interaction (HCI) to select certain stimuli that appeal to the Visual and Auditory modalities to improve implicit learning of the sequences. The effects of these stimuli on important metrics are tested in our detailed user studies. Secondly, we collect user demographic data to investigate usability and explicit recollection of the sequence. Finally, we design the selection of the Authentication Threshold Value (σ_{ATV}) that was left out in [3].

4 PROPOSED AUTHENTICATION SYSTEM

In this section, we present the specific construction of our NeuroCrypt authentication system. Our system prevents coercion attacks, as detailed later in threat model for the three different scenarios.

4.1 Overview

The authentication system that we are proposing in this paper is inspired by the Serial Interception Sequence Learning (SISL) [18, 22] task. The task is essentially a video game that facilitates implicit sequence learning. Very similar to the gameplay of the popular arcade game Guitar Hero [26], the subject is expected to intercept falling circular cues by pressing keys on the keyboard corresponding to the columns that they fall through. If the subject intercepts these circles when they reach the end line at the bottom of the screen, the game registers a hit. If the circle passes the end line without the subject pressing the corresponding key or if the subject presses the incorrect key, a miss is registered. No hit is registered if more than one key is pressed at the same time. The goal of the game is for the subject to accumulate the maximum number of hits they can. Their performance is measured using the hit-rate parameter which is the ratio of hits to attempts (the sum of hits and misses). The game displays multiple circles falling on the screen that enables the subject to get ready to press the next corresponding key. The speed of the circles falling is dictated by a difficulty-modulating algorithm that tries to maintain a hit-rate of around 70%.

The sequence-learning aspect of the game entails the repetition of a predetermined 30-item passcode sequence (the possible items being the keys that correspond to the six columns). The passcode sequence is repeated multiple times during the course of the training session and is separated by blocks of random noise that does not allow the subject to explicitly identify the sequence. The training sessions last for about 40 minutes and the subject is expected to have implicitly learnt the sequence to a sufficient level with minimal or no explicit retention of the sequence. This minimization of the explicit retention and increase in potential secret key sequences are achieved by employing several additional measures like the lack of repetition of items in the passcode sequence, the random noise and an addition of 2 extra columns (and thus possible items) compared to the original SISL task. Additionally, our proposed passcode sequence (30 items) is significantly more difficult to explicitly learn than the passcode sequence proposed in the original SISL task (12 items).

We have added other new visual and auditory features to the game that will help improve perception and reduce errors made by the subjects which will boost their implicit learning and retention of the sequence.[3]

In the authentication phase, the subject's performance is measured on their trained passcode sequence and two other randomly generated sequences. Superior performance in the passcode sequence over the random sequences leads to successful authentication.

4.2 Passcode Sequence Generation

The sequence generation algorithm is adopted from [3] so that direct comparisons between their SISL task and NeuroCrypt can be made later on. The passcode sequence is made up of 6 keys that correspond to the six columns in the game from the set $S = \{s, d, f, j, k, l\}$. Passcode sequences were generated from this set in a way such that they do not stand out to the user even when they are repeated many times during the training phase. To achieve this, passcode sequences would be made up of unique pairs that do not contain the same element. This was made possible by generating the sequence from the set of Euler cycles from a directed graph $G = (V, E)$ with each unique character in S as the vertices. This ensures that no character appears consecutively in the passcode sequence, The total number of possible passcode sequences will be the number of possible Euler cycles which can be calculated by using the BEST theorem [25]. The BEST theorem states that the number of Euler cycles represented by $ec(G)$ can be calculated by the given formula:

$$ec(G) = t_w(G) \prod_{v \in V} (deg(v) - 1)!$$

where $t_w(G)$ is the number of arborescences and $deg(v)$ is the indegree of a vertex v. Computing $ec(G)$ gives us approximately $2^{37.8}$. Hence the entropy of the passcode sequence is about 38 bits. We provide a detailed description of the process of passcode sequence generation in Algorithm 1.

Algorithm 1: Passcode Sequence Generation

```
Procedure generatePassSeq():
    passSeq = []; possibleItems = [s, d, f, j, k, l];    // possible passcode items
    foreach i ∈ {1, ··· , 30} do
        if i = 1 then
            randIndex := generateRandom(0, 5) // random integer in [0,..,5]
            randItem := possibleItems[randIndex] passSeq.append(randItem)
        end
        else
            Set randItem = 0;
            while (randItem == 0 OR passSeq[−1] == randItem) do
                randIndex := generateRandom(0, 5) randItem :=
                  possibleItems[randIndex]
            end
            passSeq.append(randItem)
        end
    end
    return passSeq;          // passSeq is returned after 30 items added.
```

4.3 The Phases of the Authentication System

The NeuroCrypt authentication system is composed of two phases – the **training phase** and the **authentication phase**. Both phases

have varying game sequences that were shown to the participants as visualised in Figure 1.

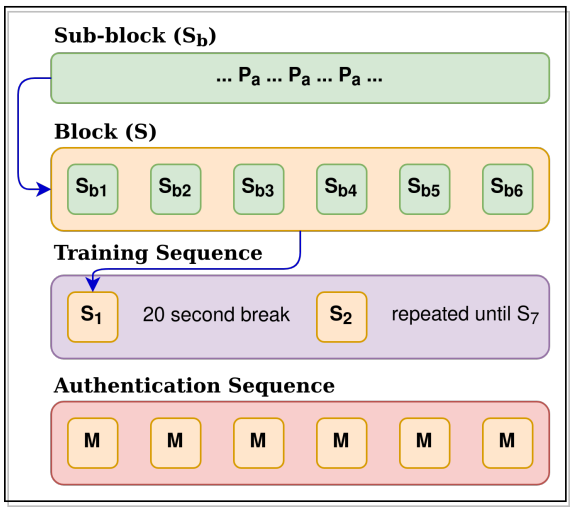

Figure 1: Schematic overview of the game sequences in NeuroCrypt. "P_a" is assigned passcode sequence for a specific user. "..." is random noise sequence. "P_1, P_2" are untrained valid passcode sequences; "M" is random permutation of (P_1, P_2, P_a).

Training Phase. The generated 30 character-long passcode sequence (see Algorithm 1) is placed thrice intermittently in an 18 character random sequence with no consecutively repeating characters. This 108 character sequence will hereon be referred to as a sub-block.

A sub-block that is repeated 5 times will be called a block, which has $108 * 5 = 540$ characters. The participant is presented with 7 blocks with a short 20 second pause between each block. Thus, overall the participant will be presented with $7 * 540 = 3780$ characters which takes about 30 – 45 minutes to complete.

Authentication Phase. In this phase the participant is presented with the trained passcode sequence (that was repeatedly presented to the participant during the Training Phase) along with two untrained sequences which are members of the set of all possible passcode sequences. Two passcode sequences were chosen here to provide sufficient variation in sequences shown to the user while not costing a great amount of time taken to authenticate. A higher performance or hit-rate in the trained passcode sequence in relation to the untrained passcode sequences will validate the participant's identity. Let P_a be the trained passcode sequence and P_1 and P_2 be untrained sequences distinct from P_a, such that $P_1 \neq P_2$. Random permutation of the three sequences (P_a, P_1, P_2) are termed M and are presented to the participant six times.

Performance in a sequence i in M , $H(P_i)$ is defined as the hit-rate registered by the user in P_i. The system declares that authentication was successful if:

$$H(P_a) > avg(H(P_1), H(P_2)) + \sigma_{ATV}$$

where σ_{ATV} is large enough to minimize chance occurrences but small enough to prevent authentication failures and is called the authentication threshold. Analysis of σ_{ATV} is carried out in TODO.

4.4 Modalities Employed

NeuroCrypt differentiates itself from previous work done in the same field with the addition of visual and auditory stimuli as we aim to explore their effects on the rate of implicit learning and the subsequent retention period. We decided to leverage the visual and auditory modalities due to the plethora of work conducted in the field of cognitive psychology that tests the effectiveness of the two modalities when stimulated for learning.

Experiment Game 1 – Vision Modality: The Visual modality has always been prioritised when it comes to implicit learning through techniques like contextual cueing. Features such as spatial arrangement and colour contrast for visual distinction are a few of the features that improve learning by visual perception. The following were features added to the Game 1 to explore the benefits of visual stimuli on implicit learning:

- The color contrast of the cues and the background is maximized by making the cue black (#000000) and the background of the game screen white (#FFFFFF).
- Solid visual separators between the columns that distinctly map the columns to each hand and also provide spatial clarity which helps with faster cue-recognition when it is presented in the context of non-cue objects [3].
- Flashing of the visual cues that improve visual perception that in turn, boosts implicit sequence learning. [7].

Experiment Game 2 – Audition Modality: The Auditory modality has also been prioritised in work related to sequence learning due to the Auditory Scaffolding Hypothesis [9] which suggests that sound plays a significant role in perceiving and interpreting sequential information. The following features were added to the second Game to explore the benefits of auditory stimuli on implicit learning:

- A musical note is played on a key press by the participant to help provide the 'scaffolding' for greater learning of the sequence. To suppress the level of explicit information that could take place, the musical notes followed an unorthodox arrangement. 3 unique notes were mapped to 2 keys each, one key on each side ($\{s, d, f\} \rightarrow$ left and $\{j, k, l\} \rightarrow$ right). The musical notes are selected with the aim of minimizing perceptual grouping based on timbre and pitch [20].
- A mild white noise is played in the background while the participant is playing the game. It has been proven that background white noise improves performance in inattentive participants for memory tasks and reduces performance in attentive participants [23]. For this reason, the white noise volume is modulated over the course of a game. Volume is increased with falling performance and decreased with improving performance which is measured by the hit-rate of the user as measured by the system.

5 EXPERIMENT METHODOLOGY

This section discusses the fundamental research questions, an overview of the experiment, and the recruitment process.

Recruitment and Demographics. We recruited n = 60 participants for our study who committed to appearing for all three weeks of the study. Since we only contacted the student body of a university, our surveyed population comprised people between 17–27 years. Out of these, approximately 46.7% were male, 51.7% were female, and 1.6% were of other genders or preferred not to say. To prevent bias and preparation of any form from the participants, we did not reveal the aim or the details of the experiment and emphasized that their participation would be completely voluntary. The participants were informed that there would be a total of three sessions, each a week apart. They were to perform the experiment in person and had to report to the venue on the same day and time for three weeks. They were also asked to carry their own laptops and headphones. All Research Assistants were present throughout each session to resolve any doubts or logistical issues that came up during the experiments.

Ethics. Our study was reviewed and approved by our Institutional Review Board (IRB). Participants had to consent to take part and could drop out of the study at any time. Participants were given a monetary compensation at the end of the three weeks to convey our appreciation for their effort and time. Since participation in only one or two sessions is of no use, failure in appearing for any session resulted in no monetary compensation.

Limitations. Our study has a few limitations that may have hindered the accuracy of our results. Firstly, the session was conducted on personal laptops with multiple participants in a single room. This posed limitations in terms of distractions or diversions like technical difficulties involving hardware and internet connections as well as inter-personal distractions. We attempted to minimise both of these factors by instructing the participants on the exact hardware requirements for the study as well as personally maintaining silence in the room. Due to the weeks-long nature of the study, many participants (n = 7) failed to return for the follow-up authentication sessions. Such participants' data were removed from the collected dataset. Additionally, there was little diversity among the participants in the study as all of them ranged from ages of 18-27, with very similar levels of education. This is a limitation as we cannot confidently claim that the results are representative of the general population. Furthermore, the participants were all students who can be expected to possess high levels of familiarity of using a keyboard. Again, this is not representative of the general population. We attempted to control potential varying levels of keyboard familiarity in the study by maintaining a target hit-rate of 70% by modulating the speed of the game. Finally, students are also more competent at working on their laptops with distractions in the background – this also introuduces a bias since it is not representative of the general population.

5.1 Experiment Overview

The experiment was conducted twice – one with visual stimuli and the other with auditory stimuli. We referred to the different iterations of the game as the Visual Game and the Audio Game, respectively. We recruited 28 participants for the visual game and 32 participants for the audio game. Assignment of participants to the two games was completely random. This was done to eliminate any kind of bias from both the participants and the research assistants.

- **First Session:** This session was the Training Session where participants were trained with a system-assigned passcode sequence which was covertly embedded in their game sequence. Since participants were being familiarised with the game and their sequences for the first time, this session lasted for about 35 – 45 minutes.
- **Explicit Retention Test:** After the participants were done with the game, we checked for any explicit retention of the pass code sequence through a questionnaire. They were presented with five different videos of sequences from the system and asked to rate them on a scale of 1 to 10 on the basis of familiarity with those sequences. Among the five videos, one of them showed the system-assigned sequence pertaining to that user and the other four were randomly generated sequences. They were also asked to fill in information about their linguistic background, past video gaming experience, familiarity with musical instruments, etc. to gauge a better idea about the diversity of the participants and the factors that may impact their performance. This test successfully helped us assess any explicit recognition of the trained sequences in the participants. It also provided information about the participants' background and cognitive abilities.
- **Second and Third Sessions**: These two sessions were used for authenticating the participants and were called Auth One and Auth Two respectively. In these sessions, we studied the implicit retention of the trained sequences in the trained participants. These sessions took only about 10 minutes for each participant.

6 RESULTS AND ANALYSIS

In this section, we discuss the experimental results obtained from our study in 6.1, the analysis of the Authentication Threshold Value σ_{ATV} in 6.2 and the results from explicit recognition tests in 6.3.

6.1 Experimental Results

Now we discuss the results obtained from the two experiments conducted over three weeks each. We analyse these results with respect to general performance gains, block-wise performance and the overall performance advantages. These aspects of the results were chosen to verify evidence of implicit learning and retention which in turn demonstrate the feasibility and promise of NeuroCrypt.

General Hit-Rate Comparison in Training Session

Figure 2 shows that there is a general increase in hit-rate when the user is on the passcode sequence compared to the general noise across the blocks in both the audio and visual games. This is proof of learning (established to be implicit in 6.3) taking place during the training session.

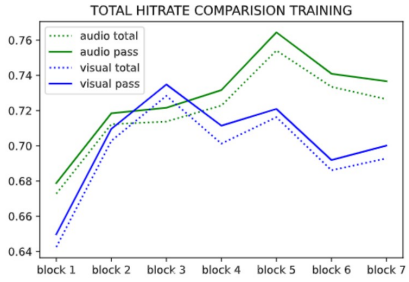

Figure 2: Graph that show the hit-rate (% in Y-axis) for the Visual and Audio games over the passcode and noise training game sequence

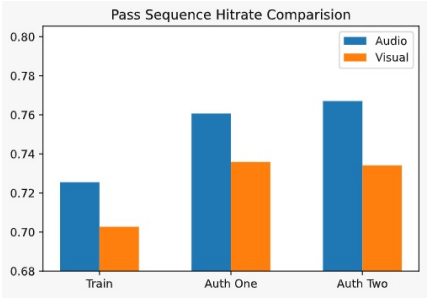

Figure 3: Graph that show the passcode sequence-specific hit-rate (% in Y-axis) for the Visual and Audio games in the three sessions.

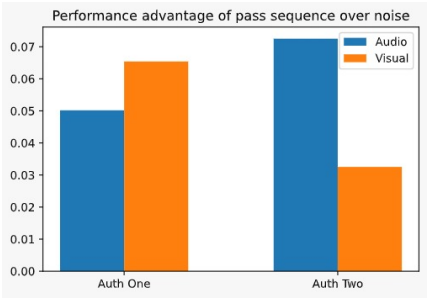

Figure 4: Graphs that show the Blockwise Performance (hit-rate (%) in Y-axis) for the Visual and Audio games' training session

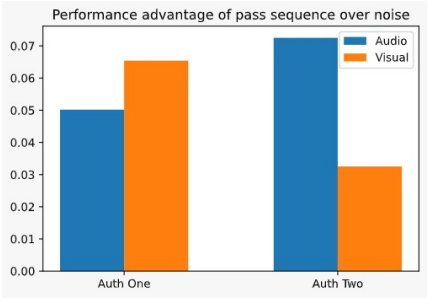

Figure 5: Graph that shows the Authentication performance (hit-rate (%) in Y-axis) advantages for the Visual and Audio games

Passcode Sequence Hit-Rate in Visual and Audio Games

Results from Figure 3 show the observed performance of the participants in terms of their hit-rate on their passcode sequence. Both games' data reflect an increase in performance after training which provides evidence that the passcode sequence learning has taken place. The increase in hit-rate between the training and the Auth One sessions are around the same for both games at around 3 – 3.5%. It is also worth noting that the average hit-rate across all sessions for the Audio game was around 3% higher than that of the Visual game. This could hint at the superior efficacy of audio stimuli at keeping the users focused and engaged compared to the visual stimuli.

Training Blockwise Performance Advantage

Figure 4 shows the training advantages for the participants in the audio and visual games during the training session. Training advantage can be defined as the gain in hit-rate of the participant during the passcode sequence compared to their hit-rate during noise.

The Audio Training Advantage follows an expected pattern from Block 2. The direction of the line is inverted between each subsequent block. This can be attributed to the effect of white noise stimuli. When the participants are not doing well during the game, the white noise increases in the background and instigates the users to focus and get a higher hit-rate. The positive effect of white noise on a participant's motor skills were verified here.

For the visual game, however, there was no observable pattern in Figure 4. The most unexpected feature of the graph is the steep decline in the training advantage between block 4 and block 5. The expected result here would be a general increase in the training advantage over the course of all blocks due to the lack of dynamic white noise. The steep decline between blocks 4 and 5 could be attributed to a possible drop in attention around the 25 – 35 minute mark that inhibits the participant to perform in the expected manner. The verification of the possible attention drops in the game should be studied in future work.

Authentication Performance Advantage

Figure 5 sheds light on a couple of insights from the performance advantage for both Audio and Visual games. Firstly, Auth1 performance advantage in the visual game was higher than for the audio game but we see a reversal of that in Auth2. Secondly, the Performance advantage in the audio game was more pronounced in the second authentication. This is a key insight as it could mean that audition is a better modality for implicit retention of long sequences over longer periods of time.

6.2 Impact of Authentication Threshold Value

The Authentication threshold value (σ_{ATV}) is the pre-set difference that a participant should display in performance in the passcode sequence over the untrained sequence, hereby referred to as noise.

The σ_{ATV} value should be set such that it is high enough to disallow random better performance in the passcode sequence and low enough to not miss out on a trained user with relatively lower implicit retention of the passcode sequence.

Since implicit retention tends to degrade over time, it is clear that σ_{ATV} will have to reduce with time from the training session to maintain the scheme's efficacy.

Authentication Threshold Value (σ_{ATV}) Estimation. Let the difference between overall hit-rate in passcode sequence (H_p) and hit-rate in noise sequence (H_r) of a participant be $d = H_p - H_r$. The d value of all participants will help us estimate the σ_{ATV} value. All our participants were trained, so the σ_{ATV} value has to be higher than the lowest d for 100% authentication of trained users i.e $\sigma_{ATV} \leq$ lowest d value. A zero or negative d value suggests no implicit retention of the sequence. We calculate the authentication success rates (ASR) for the audio game and the visual game for different σ_{ATV} values in Table 1. ASR is defined as the percentage of trained users who are correctly authenticated by the system.
From Table 1, the key takeaways are:

- As the σ_{ATV} values increase for both games in both authentication sessions, the corresponding ASRs decrease.
- The ASRs in the second Authentication Session for the Audio game are consistently higher than its counterparts in the first Authentication Session. Additionally, the Audio ASRs for the second Authentication Session are also higher than its Visual counterparts in the same session. This translates to the fact that more users are able to authenticate in the Audio game compared to the number of users who authenticated in the same game a week ago as well as the number of users who authenticated in the same session in the Visual game. This implies that the Audio stimuli induced stronger implicit retention within the participants.
- The d values' variances for the Visual games (Authentication 1: 0.0116, Authentication 2: 0.0111) is more than twice the variances for Audio games (Authentication 1: 0.0042, Authentication 2: 0.0051). These values suggest that the Performance Advantages for the participants in the Audio games were a lot more grouped together than the users in the Visual games. This suggests that calculating the optimal σ_{ATV} value for the authentication system that verifies the user's identity will be easier in the audio game.
- As most participants in the study were in the same age group, sigma value declaration based on this study alone might not be representative of other age groups.

6.3 Explicit Recognition Results

Explicit memory refers to a conscious and intentional recollection of events or information. Explicit recognition of information or events indicates a conscious retention of the data. Recall that explicit memory is the phenomenon that we are trying to suppress in our authentication system to avoid rubber-hose attacks. To measure participants' explicit knowledge, they were presented with a questionnaire that asked the participants to rate four untrained and their trained sequence based on familiarity. The higher selection of untrained sequences suggests that explicit recognition of the trained sequences was low in the trained participants. This suggests

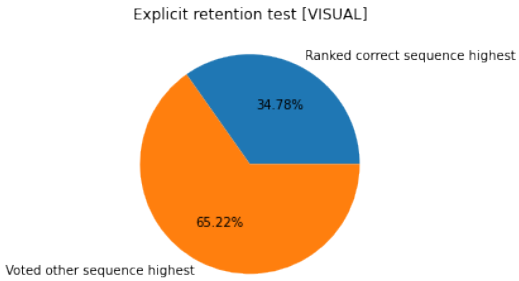

Figure 6: Proportion of participants who selected the trained sequence the highest and otherwise in the Visual Game

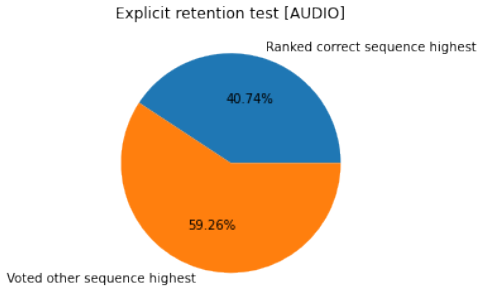

Figure 7: Proportion of participants who selected the trained sequence the highest and otherwise in the Auditory Game

that the training and test sessions were successfully structured to reduce explicit retention and heighten implicit learning.
40.74% of the participants rated the trained sequence as most familiar for Game 2 involving audition modalities (Figure 7), while 34.78% participants rated the chose their trained sequence as the most familiar for Game 1 incorporating visual modalities (Figure 6). The higher selection of trained sequences for Game 1 can be attributed to the Auditory Scaffolding Hypothesis [10]. This suggests that sound provides stronger support required by organisms require to interpret and process sequential information as compared to the visual modality. The auditory inputs given with every key press during Game 2 possibly provided the participants with the scaffolding needed for processing and identifying the trained sequences. This resulted in a higher explicit recognition of the trained sequences in participants who played Game 2 with audition modalities than those who played the Game 1.

7 SECURITY AND USAABILITY ANALYSIS

In this section, a security and privacy analysis of our framework is carried out. The significance of a secure and usable authentication system cannot be overstated as it is the gateway that grants users access to corresponding information by verifying their identity.

Table 1: NeuroCrypt success rates (%) for trained users on different Authentication Threshold (σ) values for Auth One and Auth Two.

σ	**Authentication 1** (1 week after training)		**Authentication 2** (2 weeks after training)	
	Audio Success Rate	Visual Success Rate	Audio Success Rate	Visual Success Rate
0.025	56.25	57.14	65.62	50.00
0.050	43.75	42.85	46.87	42.85
0.075	34.37	25.00	37.50	35.71
0.1	18.75	21.42	34.37	32.14

Authentication systems are based on four fundamental pieces of information: something the user is, something the user has, something the user knows, and recently proposed, someone the user knows [5]. If the user of the system can provide proof in some or all of these areas, they are admitted into the system. NeuroCrypt aims to provide a means for fall-back authentication systems while replacing traditional explicit knowledge-based authentication systems. Here, the piece of information that will grant access is something the user knows without knowing that they know it. This critical deviation from traditional knowledge-based authentication systems makes NeuroCrypt a more favourable candidate for fall-back authentication. The two main aspects where our system is more favourable are in terms of security attacks (mainly coercion attacks) and usability [12].

7.1 Bane of Knowledge Based Passwords

Knowledge-based authentication systems traditionally depend upon users' explicit recollection of some information. They either use simpler passwords multiple times which are not secure enough to reliably defend against dictionary attacks or use stronger passwords which impose a cognitive burden on the user and decrease memorability [16] [13] [12]. Thus, authentication schemes based upon explicit memory suffer from a tension between security and usability [12]. Social engineering attacks involve the exploitation of human vulnerabilities rather than technical ones as the human factor is often considered the weakest link in a security system. In a study conducted on the limitations of human memory with respect to password usage, 72% of the participants reported forgetting or mixing up passwords [16]. An important factor affecting this was the number of unique passwords owned by the users. It was recommended to have no more than five different passwords as trying to recall more led to risks of forgetting a password at least once a month [16].

7.2 Basic Threat Model

As part of our security analysis, we highlight three cases that could compromise an authentication system and evaluate the efficacy of our system during such a case.

Case 1 – Oscar obstructs some users like Alice and coerces them into revealing as much information as they can. Assuming Oscar is able to make his way into the secured facility, the basic coercion threat model only allows him one chance to fool Bob. Let us say that the training procedure has successfully embedded a predicate p in Alice's brain. Oscar has intercepted u number of users and subjected each to q number of trials, his probability of success at best is $\frac{qu}{|\sum|}$. ($|\sum|$ = total number of possible sequences with 30 characters, i.e., 248 billion). Since each authentication test takes approximately 5 minutes, we assume an upper bound of $q = 105$ trials per user because this means spending about 500,000 minutes on authentication tests. A year has around 525,600 minutes, implying that even in the best case scenario, Oscar will have to subject users to authentication tests non-stop for a year. This has 2 drawbacks:

- Constant subjection to a separate authentication test will interfere with the already learnt sequence by Alice, rendering her useless to the adversary i.e., Oscar.
- Prolonged absence of Alice will alert the authorities and lead to revocation of her credentials.

It might not be possible to adequately determine attacker success rates. However, let us assume Oscar is able to perform the described operation with $u = 100$ users. Even then, his success probability would be:

$$\frac{qu}{|\sum|} = 10^5 \times \frac{100}{|\sum|} = 2^{-16} \tag{1}$$

β in Case 1 is only 0.0015% signifying the extremely low probability of Oscar being successful.

Case 2 – Oscar himself takes the authentication test by showing up at the secure facility where it is being conducted. Oscar is only allowed one chance to authenticate correctly; otherwise, he could memorize at least one of the sequences by coercing users and taking the test repeatedly. In the system proposed by Bojinov et al. [3], if Oscar is allowed to authenticate more than once, he could try to memorise part of the authentication sequence and will have a $\frac{1}{3}$ probability of passing the authentication test and fooling Bob. In NeuroCrypt, we try to better this adversarial scenario by having the lengths of the passcode sequences for all users differ by a value of ± 2 so the sequence could be 28, 29, 30, 31 or 32 characters long. Even if Oscar is aware of the length of the sequence and where the breaks in the blocks are for the user whose authentication test he is taking, it gives him a $(1/3)^5 = 0.0041$ chance of success. *Note: The data collected in this paper was before this modification and hence this will be implemented in the next set of experiments we run.* β in Case 2 is only 0.41% signifying the low probability of Oscar winning. However, if the σ value is low for the user whom Oscar is impersonating, his success rate could be higher as he will need to have a moderate to lower average hit-rate to authenticate into the system.

Oscar could also deliberately degrade his performance across 2 blocks and exhibit an artificial performance increase for one block in favour of the trained sequence. He could iterate between

coercing users and taking the test himself to record all 3 30-character sequences and then offline subject the trained user, Alice, to those 3 sequences. This would help Oscar successfully determine the password and he could then train himself on that sequence. Oscar is then guaranteed success at the next authentication trial. Although this is a difficult move to pull off because of the requirement of memorizing an approximately 30-character sequence at the speed the game is played, it is not impossible. Additionally, here Oscar will show an obvious performance gap (a significant difference in hit-rate for a particular sequence as compared to the other sequences during authentication) for the other 2 sequences before passing the authentication test. An observable performance gap between the pass sequence and the random sequences is an indication of the system being under attack. This also applies to a potential attack where Oscar is actually an expert player at all 3 sequences but deliberately degrades his performance for the non-pass sequences to fool Bob.

Case 3 – Oscar tries to recreate the sequence. Another potential attack to the system could include Oscar profiling Alice's knowledge and using fragments to reconstruct the sequence or at least a sufficient part of it to meet the authentication threshold. An experiment has been conducted by Bojinov et al. [3] on assessing whether trigrams could be useful in this regard. The participants went through the same training phases. During the test, participants performed a sequence constructed to provide each of the 150 possible trigrams exactly 10 times by constructing ten different 150-trial units that each contained all possible trigrams in varying order. Performance on each trigram was measured by percent correct as a function of the current response and two responses prior. The percent correct on each trigram was individually calculated. The trigrams for which the participants showed high learning would be on top. Thus, if there had been sufficient implicit learning at the trigram level, the untrained ones would be towards the bottom for the individuals. However, the 34 participants averaged 73.9% correct (SE 1.2%) for trigrams from the trained sequence and 73.2% correct (SE 1.1%) for the rest displaying that the difference was unreliable.

Directions for future research could include conducting experiments to determine the exact number of character subsets that can be used to recreate the sequence. A limitation here would be the exponential increase in the number of fragments to be tested with increase in the number of characters per subset. However, if determined correctly, this will help in developing more robust and complex sequences as passwords.

8 CONCLUSION AND FUTURE WORK

We introduce NeuroCrypt, an implicit memory based authentication system which defends against coercion attacks by leveraging visual and auditory concepts borrowed from cognitive psychology. The selected visual and auditory stimuli promised to improve implicit learning while minimising the cognitive load on the user. We perform real-world experiments (n=60) and analyse implicit retention by measuring user performance in trained passcode sequences. Then, we analyse the effect of the audio and visual stimuli, observing that audio stimuli are more effective in keeping users focused and engaged, and are also superior to visual stimuli in terms of passcode retention over long time periods (2 weeks). Visual stimuli produce stronger implicit learning in a shorter training period but also result in lower retention rates over time. Analysis of user performance in passcode and noise sequences, d, revealed that the variance of d was much higher in for participants who played the audio game over the visual game. This indicates that the use of audio stimuli provides a more robust approach to ensuring security (through the selection of a more reliable authentication threshold value).

There is much that remains to be established with NeuroCrypt. Firstly, we would like to perform a direct comparison with [3] to investigate the difference in performance with and without the integrated visual and auditory stimuli. Secondly, retraining frequencies for the system needs to be studied. This would involve exploration of the deterioration in implicit retention of the passcode sequences over time in the trained users. An ideal authentication system would re-train the user before their performance advantage falls below σ_{ATV} while keeping usability in mind. Additionally, we also want to explore the addition of stimuli from the haptic modality while also studying the performance effects of a combination of stimuli from the visual, audio and haptic modalities. Finally, we also intend on expanding the demographics of the study by age and background in order to analyse how the optimal σ_{ATV} may vary across populations.

ACKNOWLEDGMENTS

REFERENCES

[1] Angela Sasse Anne Adams. 1999. Users are not the Enemy. *Commun. ACM* 42, 12 (1999), 41–46.

[2] Richard C Atkinson and Richard M Shiffrin. 1968. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*. Vol. 2. Elsevier, 89–195.

[3] Hristo Bojinov, Daniel Sanchez, Paul Reber, Dan Boneh, and Patrick Lincoln. 2014. Neuroscience meets cryptography: Crypto primitives secure against rubber hose attacks. *Commun. ACM* 57, 5 (2014), 110–118.

[4] Joseph Bonneau, Cormac Herley, Paul C Van Oorschot, and Frank Stajano. 2012. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *2012 IEEE Symposium on Security and Privacy*. IEEE, IEEE, 553–567.

[5] John Brainard, Ari Juels, Ronald L Rivest, Michael Szydlo, and Moti Yung. 2006. Fourth-factor authentication: somebody you know. In *Proceedings of the 13th ACM conference on Computer and communications security*. 168–178.

[6] Christoph Bregler. 1997. Learning and recognizing human dynamics in video sequences. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 568–574.

[7] Jean-Pierre Bresciani, Franziska Dammeier, and Marc O Ernst. 2008. Tri-modal integration of visual, tactile and auditory signals for the perception of sequences of events. *Brain research bulletin* 75, 6 (2008), 753–760.

[8] Claude Castelluccia, Markus Dürmuth, Maximilian Golla, and Fatma Deniz. 2017. Towards implicit visual memory-based authentication. In *Network and Distributed System Security Symposium (NDSS)*.

[9] Christopher Conway, David Pisoni, and William Kronenberger. 2009. The Importance of Sound for Cognitive Sequencing Abilities: The Auditory Scaffolding Hypothesis. *Current directions in psychological science* 18 (10 2009), 275–279. https://doi.org/10.1111/j.1467-8721.2009.01651.x

[10] Christopher M Conway, David B Pisoni, and William G Kronenberger. 2009. The importance of sound for cognitive sequencing abilities: The auditory scaffolding hypothesis. *Current directions in psychological science* 18, 5 (2009), 275–279.

[11] S. Kirkpatrick D. Weinshall. 2004. Passwords You'll Never Forget, But Can't Recall. *ACM SIGCHI Extended Abstracts on Human Factors in Computing Systems* (2004), 1399–1402.

[12] Tamara Denning, Kevin Bowers, Marten Van Dijk, and Ari Juels. 2011. Exploring implicit memory for painless password recovery. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2615–2618.

[13] Payas Gupta. 2013. Exploiting Human Factors in User Authentication. (2013), 170.

[14] Amit Kale, Aravind Sundaresan, AN Rajagopalan, Naresh P Cuntoor, Amit K Roy-Chowdhury, Volker Kruger, and Rama Chellappa. 2004. Identification of humans using gait. *IEEE Transactions on image processing* 13, 9 (2004), 1163–1173.
[15] Fabian Monrose, Michael K Reiter, and Susanne Wetzel. 2002. Password hardening based on keystroke dynamics. *International journal of Information security* 1, 2 (2002), 69–83.
[16] Denise Ranghetti Pilar, Antonio Jaeger, Carlos FA Gomes, and Lilian Milnitsky Stein. 2012. Passwords usage and human memory limitations: A survey across age and educational background. *PloS one* 7, 12 (2012), e51067.
[17] Emmanuel M Pothos. 2007. Theories of artificial grammar learning. *Psychological bulletin* 133, 2 (2007), 227.
[18] Paul J Reber, Daniel J Sanchez, and Eric W Gobel. 2011. Models of sequential learning. In *Proceedings of the Eighth International Conference on Complex Systems (this volume), NECSI.*
[19] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. 2000. Speaker verification using adapted Gaussian mixture models. *Digital signal processing* 10, 1-3 (2000), 19–41.
[20] Jennifer C Romano Bergstrom, James H Howard Jr, and Darlene V Howard. 2012. Enhanced implicit sequence learning in college-age video game players and musicians. *Applied Cognitive Psychology* 26, 1 (2012), 91–96.
[21] Natalie Ruiz. 2011. *Cognitive load measurement in multimodal interfaces.* Ph. D. Dissertation. University of New South Wales, Sydney, Australia.
[22] Werner Sævland and Elisabeth Norman. 2016. Studying different tasks of implicit learning across multiple test sessions conducted on the web. *Frontiers in psychology* 7 (2016), 808.
[23] Göran BW Söderlund, Sverker Sikström, Jan M Loftesnes, and Edmund J Sonuga-Barke. 2010. The effects of background white noise on memory performance in inattentive school children. *Behavioral and brain functions* 6, 1 (2010), 1–10.
[24] Philipp Taesler, Julia Jablonowski, Qiufang Fu, and Michael Rose. 2019. Modeling implicit learning in a cross-modal audio-visual serial reaction time task. *Cognitive Systems Research* 54 (2019), 154–164.
[25] T van Aardenne-Ehrenfest and NG de Bruijn. 2009. Circuits and trees in oriented linear graphs. In *Classic papers in combinatorics.* Springer, 149–163.
[26] By Wikipedians. [n. d.]. *Guitar hero series.* PediaPress.
[27] James R Williams. 1998. Guidelines for the use of multimedia in instruction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 42. SAGE Publications Sage CA: Los Angeles, CA, 1447–1451.
[28] Martina Ziefle. 1998. Effects of display resolution on visual performance. *Human factors* 40, 4 (1998), 554–568.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009