Exploring Effects of Auditory Stimuli on CAPTCHA Performance

Bruce Berg, Tyler Kaczmarek, Alfred Kobsa, and Gene Tsudik

University of California, Irvine
{bgberg, tkaczmar, akobsa, gtsudik} @uci.edu

Abstract. CAPTCHAs have been widely used as an anti-bot means for well over a decade. Unfortunately, they are often hard and annoying to use, and human errors have been blamed mainly on overly complex challenges, or poor challenge design. However, errors can also occur because of ambient sensory distractions, and performance impact of these distractions has not been thoroughly examined. The goal of our work is to explore the impact of auditory distractions on CAPTCHA performance. To this end, we conducted a comprehensive user study. Its results, discussed in this paper, show that various types of auditory stimuli impact performance differently. Generally, simple and less dynamic stimuli sometimes improve subject performance, while highly dynamic stimuli have a negative impact. This is troublesome since CAPTCHAs are often used to protect web sites offering tickets for limited-quantity events, that sell out very quickly, i.e., within seconds. In such settings, introduction of even a small delay can make the difference between obtaining tickets from the primary source, and being forced to use a secondary market. Our study was conducted in a fully automated experimental environment to foster uniform and scalable experiments. We discuss both benefits and limitations of unattended automated experiment paradigm.

1 Introduction

Completely **A**utomated **P**ublic **T**uring tests to tell **C**omputers and **H**umans **A**part (aka CAPTCHAs) are programs that generate and evaluate challenges that are easy solvable by a human, while hard to solve by software. CAPTCHAs have been used to prevent bot-based abuse of services for well over a decade [23]. They have become a fairly routine hurdle for users seeking to access online resources, such as: discussion forums, ticket sales, banking, and email account creation. Because of their widespread adoption, successful attacks, and pervasive dislike by users, most recent efforts in development have been invested into creating CAPTCHAs that are [3]: (1) usable: where humans are successful at least 90% of the time, (2) secure/robust: a state-of-the-art bot should not be successful more than 0.01% of the time, and (3) scalable: challenge are either automatically generated, or drawn a space that is too large to hard-code responses for each challenge. Consequently, CAPTCHA developers focused on text-based CAPTCHAs, i.e., those that present a jumbled alphanumeric code. This approach is popular since human users are quite good at identifying these alphanumeric codes in a distorted image, thus satisfying the usability requirement. Also, image segmentation and recovery known to be a hard problem for AI, satisfying the security requirement. Finally, such challenges can be randomly generated, satisfying the scalability requirement [8].

Gene Tsudik, Tyler Kaczmarek, Bruce Berg, Alfred Kobsa, Exploring Effects of Auditory Stimuli on CAPTCHA Performance, Proceedings of AsiaUSEC'20, Financial Cryptography and Data Security 2019 (FC). February 14, 2020 Kota Kinabalu, Sabah, Malaysia Springer, 2020.

However, not much attention has been paid to user's physical context while solving CAPTCHAs. Security-critical tasks, such as CAPTCHAs, are often performed in noisy environments. In many real-world settings, users are exposed to various sensory stimuli. Impact of such stimuli on performance and completion of security-critical tasks is not well understood. Any specific stimulus (e.g. police siren or fire alarm) can be incidental or malicious, i.e., introduced by the adversary that controls the environment. This threat is exacerbated and accentuated by the growth in popularity of Internet of Things (IoT) devices, particularly in contexts of "smart" homes or offices. As IoT devices become more common and more diverse, their eventual compromise becomes more realistic. One prominent example is the Mirai botnet [13] which used a huge number of infected smart cameras as zombies in a massive coordinated DDoS attack. A typical IoT-instrumented home environment, with "smart" lighting, sound and alarm systems (as well as appliances) represents a rich and attractive attack target for the adversary that aims to interfere with a user's physical environment in particular in order to inhibit successful CAPTCHA solving. We believe that this is especially relevant to some time-critical scenarios, such as web sites that sell limited numbers of coveted tickets for concerts, festivals, promotional airfares, etc. In these settings, a delay of just a few seconds can make a very big monetary difference.

Data Security 2019 (FC). February 14, 2020 Kota Kinabalu, Sabah, Malaysia Springer, 2020, In order to explore effects of attacks emanating from the user's physical environment we experimented with numerous subjects attempting to solve text-based CAPTCHAs in the presence of unexpected audio stimuli. We tested a total of 51 subjects in a fully unattended experimental setting. We initially hypothesized that introduction of audio stimuli would negatively impact subject task completion. While this was mostly confirmed, certain types of stimuli surprisingly demonstrated positive effects.

This paper is organized as follows: The next section describes the design and setup of our experiments are, followed by experimental results in Section 3. Next, we discuss the implications of the results and advantages of the unattended experimental environment. The paper concludes with directions for future work. Due to size limitations, we placed the following sections into the Appendix: (**A**) overview of related work and background material, (**B**) limitations of our study, and (**C**) ethical considerations.

2 Methodology

This section describes our experimental setup, procedures and subject parameters.
Apparatus: Our experimental setting was designed to allow for fully automated experiments with a wide range of sensory inputs. To accommodate this, we located the experiment in a dedicated office in the Psychology Department building of a large public university. The setup is comprised entirely of the following popular commercial-off-the-shelf (COTS) components: (1) Commodity Windows desktop computer with keyboard and mouse. (2) 19" Dell 1907FPc monitor, (3) Logitech C920 HD Webcam, and (4) Logitech Z200 Stereo Speaker System[1]. This experimental setup is supposed to mimic the typical environment where an average user might be presented with a CAPTCHA, i.e., an office.

[1] With the volume knob physically disabled.

Procedures: As mentioned earlier, the experimental environment was entirely unattended. An instructional PowerPoint presentation was used for subject instruction, instead of a live experimenter. This presentation was each subject's only source of information about the experiment. Actual experimenter involvement was limited to off-line activities: (1) periodic re-calibration of auditory stimuli, and (2) occasional repair or repositioning of some components that suffered minor damage or were moved throughout the study's lifetime. This unattended setup allowed the experiment to run without interruption 24/7/365. It was conducted over a 3-month period. The central goal was to measure performance of subjects attempting to solve as many CAPTCHAs as possible within a fixed timeframe. Subjects were expected to solve them continuously for 54 minutes. During this period, a subject was exposed to 4 rounds of 6 auditory stimuli. The control and stimuli were presented in a random order within each round, to mitigate any ordering effects on subject performance.

Why CAPTCHAs? We picked CAPTCHAs as the security-critical task for several reasons. First, CAPTCHAs do not require the subjects to enter any personally identifying information (PII) or secrets in order to solve them, and can be dynamically generated on the fly, allowing for the study of subject behavior across many different solution attempts. This is in contrast with other security-critical tasks, such as password entry. Second, solving CAPTCHAs is a fairly common task and it is reasonable to assume that all potential subjects are familiar with them, unlike infrequent tasks, e.g., Bluetooth pairing. Finally, the cognitive effort needed to solve CAPTCHAs (recognize-and-type) is higher than the simple comparison task in Bluetooth pairing, and is similar to recall-and-type tasks, such as password entry [21].

Phases: The experiment runs in four phases:

1. **Initial:** subject enters the office, sits down at a desktop computer and starts the instructional PowerPoint presentation. Duration: Negligible.
2. **Instruction:** subject is instructed in the nature of CAPTCHAs and the experimental procedure. Duration: 2-4 minutes
3. CAPTCHA: subject is presented with a random CAPTCHA. Upon submitting a solution, a new CAPTCHA is presented, regardless of the accuracy of the response. Subjects are exposed to the stimulus conditions for 24 rounds, each round lasting 2:15. Duration: 54 minutes.
4. Final: subject is taken to a survey page and asked to enter basic demographic information. Duration: 2-3 minutes

The entire experiment lasts between 58 and 61 minutes. Each subject's participation is recorded by the webcam and by screen-capturing software, to ensure compliance with the procedure. Since our objective is to assess overall impact of auditory stimuli on subject performance (and not performance degradation due to a surprise), the first 15 seconds of each stimulus condition were not used in data collection. This should accurately capture the enduring effect of the auditory stimuli, and ignore the spiking effect (i.e., surprise) on the attentional system due to the introduction of an unexpected stimulus [21].

2.1 CAPTCHA Generation

Since the study was concerned primarily with usability and less with robustness, we used text-based CAPTCHAs that follows the guidelines of [4] to create challenges that

are highly usable, and can be quicky solved in bulk. To facilitate this, a challenge generation algorithm was selected that created 5-character alphanumeric codes with thin occluding global lines, a small amount of global distortion and minimal local distortion of the characters. This yielded challenges that our subjects could easily and quickly solve in the baseline, i.e., Control case.

2.2 Stimuli Selection

The experiment consisted of two categories of auditory stimuli: (1) static with single volume level, and (2) dynamic, that changed volume throughout presentation.

Static sound stimuli were the sounds of: (1) crying baby, (2) babbling brook, and (3) human voice reading individual letters and digits in random order at a rate of two per second. (1) and (2) were chosen for their ecological significance as a source that needs attention, and a relaxing sound, respectively. The human voice stimulus was chosen to interfere with the task-specific cognitive processes used to solve CAPTCHAs. This is analogous to the Stroop effect, a phenomenon where subjects who attempt to read the written name of a color that is rendered in a different color (e.g., the word "red" written in blue ink) do so slower and in a more error-prone way than reading the same words in plain black ink [15]. Specific volumes of the three static stimuli were:

(1) Crying baby: 78 dB, (2) Babbling brook: 70 dB, and (3) Human voice: 75 dB

The two dynamic stimuli included: (1) randomly generated looming sounds, and (2) randomly ordered menagerie of natural, aversive sounds. The looming stimulus was an amplitude modulated tone that increased from nearly silent to 85 dB over 5 seconds. Its intensity curve is shown in Figure 1. Once the looming sound completed, it repeats at a different Left/Right speaker balance, selected randomly. This repeats continuously for the entire 2:15 minute stimulus window. The natural stimulus consisted of a randomly generated sequence of aversive sounds, which included: circular saw cutting wood, blaring vuvuzela, nails on a chalkboard, and spinning helicopter rotors. These sounds were played at a randomly selected volume from 75 to 88 dB. Each lasted for up to 2 seconds before changing to the next random sound.

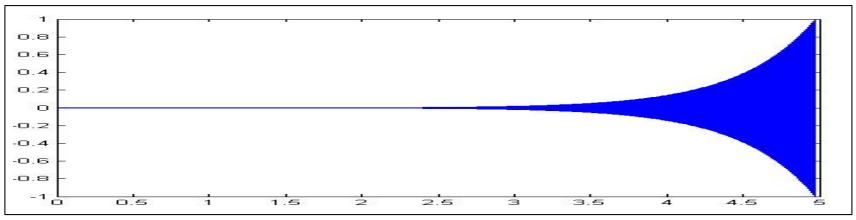

Fig. 1. Looming Sound Intensity Function

Even the highest stimuli volume (88 dB) is well within the *safe range*, as defined by US Occupational Safety & Health Administration (OSHA) guidelines.[2]. Clearly, an

[2] OSHA requires all employers to implement a Hearing Conservation Program where workers are exposed to a time-weighted average noise level of 90 dB or higher over an 8 hour work shift. Our noise levels were for a much lower duration, and only the very loudest was within the regulated range. See: `https://www.osha.gov/SLTC/noisehearingconservation/`

adversary that controls the victim's environment would not be subjected to any such ethical guidelines, and could thus use much louder stimuli.

2.3 Psychophysical Description of Stimuli

The chosen stimuli have the potential to produce different effects. Except for the babbling brook, selection of the sounds was guided by the intent to elicit a negative emotional response and increased level of general arousal. It is reasonable to expect a negative impact of these sounds on task performance. However, any capture of an individual's attention by an aversive stimulus is likely to be momentary, occurring primarily when the stimulus is first introduced. In cognitive science, attention is conceptualized as a limited resource. Probably for good reason, the greatest demand on attention is in response to a change in the environment. Once an assessment is made that a stimulus does not require a response, adaptation to the stimulus from a foreground target into a background context proceeds relatively rapidly as attention is redistributed to other demands. Although an aversive sound may remain aversive throughout its presentation, its capacity to disrupt performance on a complex task might rapidly fade after onset. This could serve to sharpen an individual's focus for the task at hand [22].

However, the auditory attentional system is not nearly as adept at dealing with many rapid changes in the environment that occur in quick succession [1]. Dynamic synthetic sounds can be designed to attract attention resources without being aversive. To the human auditory attention system, a looming sound is not easily classified as a single, non-threatening change in the environment. Instead, it embodies a context of continuous, approaching and potentially threatening change. This unclassifiable context "tricks" the system into a state of sustained engagement, and can deplete the subject's attentional resources. Because of this phenomenon, we suspect that highly dynamic sounds have the greatest impact on subject performance.

2.4 Initial Hypotheses

Our initial intuitive hypothesis was that introduction of unexpected auditory stimuli while solving CAPTCHAs would have negatively impact subject performance. We expected two outcomes, as compared to a distraction-free (Control) setting:

[H1]: Higher error rates, and

[H2]: Longer completion times in successful cases

We hypothesized this because, although mixed results were observed in [2] for Bluetooth pairing, solving CAPTCHAs is a more difficult cognitive task (requires more attention) even in the distraction-free (Control) case [22].

2.5 Recruitment

Recruitment was handled through the human subjects lab pool of Psychology Department at a large public university. A brief description of the study was posted on an online bulletin, and undergraduate students were allowed to sign up for the experiment and were compensated with course credit. Not surprisingly, the subject pool was dominated by college-age (18-25) individuals and the gender split was somewhat uneven: 35 female (69%) and 16 male subjects (31%).

Table 1. Subject Failure Rates

Stimulus	#Successful Entries	#Unsuccessful Entries	Failure Rate	Odds Ratio wrt Control	p
None (Control)	6413	616	0.088	-	-
Baby	6074	1544	0.203	2.31	< 0.001
Brook	6332	574	0.083	0.901	0.090
Looming	5039	719	0.125	1.483	< 0.001
Natural	5787	723	0.111	1.299	< 0.001
Voice	4582	697	0.132	1.581	< 0.001
Total	34227	4873	0.125	-	-

Table 2. Avg Times (sec) for Successful Solutions

Stimulus	Mean Time	Standard Deviation	DF wrt Control	t-value wrt Control	p	Cohen's D
None (Control)	4.621	3.771	-	-	-	-
Baby	4.520	5.267	12485	0.016	0.986	0.022
Brook	3.472	5.100	11743	15.026	< 0.001	0.400
Looming	6.092	2.212	11450	17.373	< 0.001	0.323
Natural	5.909	4.751	12198	18.505	< 0.001	0.300
Voice	6.480	6.985	10993	18.07	< 0.001	0.331

3 Results

This section discusses the results, starting with data cleaning and proceeding to subject task completion effects.

Data Cleaning: A total of 58 subjects took part in the study. However, 7 of them were non-compliant with the experimental procedure, and prematurely quit the experiment. Since this behavior was captured by the recording software, all data from these subjects was discarded.

Task Failure Rate: As Table 1 shows, every audio stimulus – except for brook – had a substantial, statistically significant impact on subject failure rates. Furthermore, each of these was shown by their Odds ratios to have a large effect size. Thus, the impact on failure rates, though seemingly small, is a large proportional increase in failures when subjects are exposed to any stimulus, with the most impactful stimulus (crying baby) more than **doubling** subject failure rates. Interestingly, there was no direct correlation between dynamicity of the stimulus and its impact on failure rates, as the Brain Arousal Model would suggest [22]. This opens up a potential attack vector for the adversary that controls the auditory environment, as discussed in Section 4.

Task Completion Times: Table 2 shows average completion times for successful CAPTCHA completions under each stimulus. Results illustrate that all stimuli (except crying baby) have a statistically significant departure from the mean ($p < 0.001$) after applying a conservative Bonferroni correction to account for 5 pairwise comparisons to Control. However, while the looming, natural and voice stimuli have a negative effect on subject performance and slow down subject task completion, brook has a positive effect and lower average task completion times. Also, although these effects appear to be highly pronounced due to their significance, their effect size is small, with Cohen's D values ranging from 0.300 to 0.400. Implications of these impacts on task completion times are discussed in Section 4.

Table 4 shows a one-way analysis of variance (ANOVA) evaluation of differences in means of each stimulus, excluding Control. There is a significant difference ($p < 0.0001$) in completion times across different stimuli. Furthermore, Bartlett' test for homogeneity of variances was performed over each stimulus, again excluding Control. Bartlett's test rejected the null hypothesis that all distributions of completion times have the same variance ($\chi^2 = 5521.543,\ p < 0.0001$). These results assert that different stimuli influence subject task performance differently. This suggests that there are different aspects to the specific stimulus that can be altered to impact performance differently. Implications are discussed in the next section.

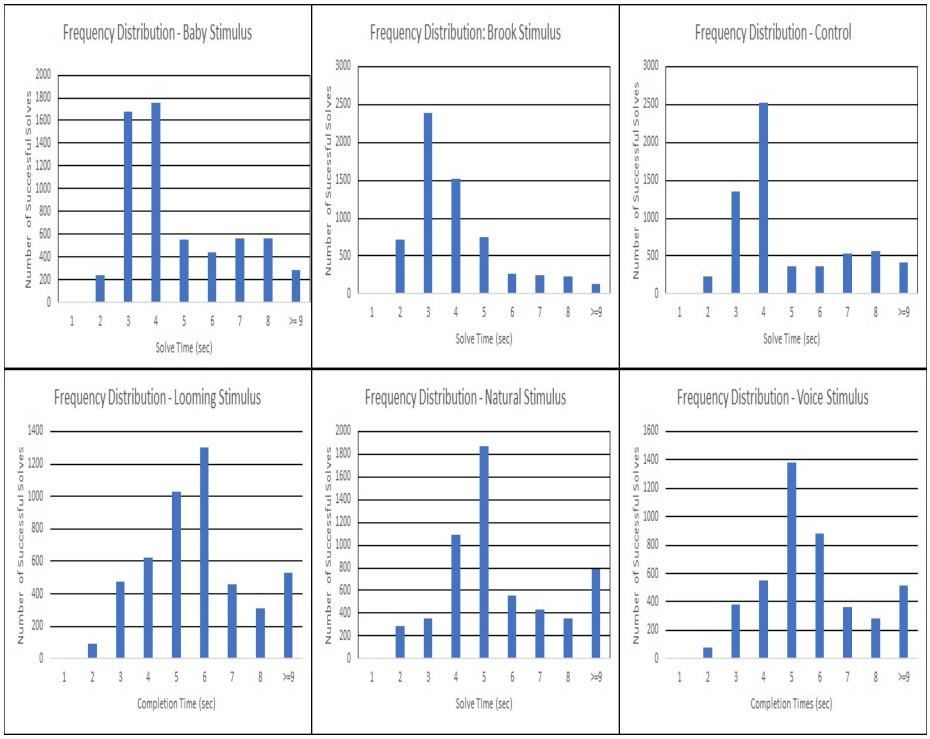

Fig. 2. Frequency Distribution of All Stimuli

Figures 3 show frequency distributions of response times by stimulus. They are similar to exponentially modified Gaussian distributions, consistent with reaction time distributions [25]. This is somewhat expected, since subjects were instructed to solve CAPTCHAs as quickly and as accurately as they could. Alhough this correlation can help future studies into the cognitive task of completing text-based CAPTCHAs, it is out of the scope of this paper.

We note that the stimuli with the greatest impact on subject completion times have much heavier tails than other distributions. These correspond to the highly dynamic stimuli which also negatively impact subject failure rates. In particular, voice stands out because it is a task-specific stimulus; its exaggerated effect on subject performance is discussed below.

4 Discussion of Observed Effects

As results show, subjects solving CAPTCHAs are not uniformly impacted by different stimuli. We observed both positive and negative effects. More dynamic or task-specific stimuli (such as looming, voice and natural) negatively impact subject performance, while the simplest static stimulus (brook) had a positive effect. Interestingly, crying baby had a substantial negative effect on subject failure rates, though it did not significantly influence subject completion times.

Table 3. Yerkes-Dodson Relationship Between Sensory Arousal Levels & Performance

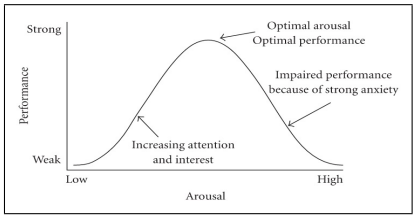

Table 4. One-Way ANOVA Between Stimulus Completion Time Distributions

Source of Variation	Sum of Squares	Degrees of Freedom	Variance	F	p
Between Groups	41601.39	4	10400.349	412.340	<0.0001
Within Groups	676183.75	26809	25.22		
Total	717785.15	26813			

The above is mostly consistent with the Yerkes-Dodson Law, which, states that a subject's overall level of sensory arousal is a determining factor in their performance at any task. At a low level of arousal, a subject is uninterested, and unengaged with the task at hand, and thus does not perform optimally. Similarly, an overstimulated subject is likely to have attention split between the arousing stimuli and the task at hand; thus performance suffers. However, there is a middle ground where a subject's overall arousal level allows being engaged with, yet not overwhelmed by, the task, thus yielding optimal performance. This relationship between sensory arousal and performance generally follows an upside-down U-shaped curve, as shown in Table 3 [7]. We now consider the implications of beneficial and negative observed effects.

Beneficial Effects: Only the babbling brook stimulus had a positive impact on subject failure rates and completion times.

Intuitively, our subjects were not highly engaged with the assigned task. Their general level of sensory arousal was similar to that of performing any boring/routine security-critical task. Because of this low level of initial engagement, the Yerkes-Dodson Law implies that introduction of additional stimulation can improve task performance. In our case, this resulted in increased speed of correct CAPTCHA completion under the babbling brook stimulus. This simple and static (yet relaxing) stimulus served to pique subject arousal without overwhelming their attentional resources.

The above illustrates the fine line between optimal sensory arousal and overstimulation. While our subjects might not have been sufficiently engaged with the task at hand, results imply that cognitive resources required to successfully solve CAPTCHAs as quickly as possible left little additional room for stimulation before the subject became overstimulated. However, this beneficial effect suggests that there must be a range of stimulation that can reliably improve performance. Thus, there could be a way for benign actors to incorporate sensory stimulation into security-critical tasks (such as CAPTCHAs) to push subjects along the Yerkes-Dodson curve towards a more beneficial level of sensory arousal, yielding better performance.

Negative Effects: Several types of auditory stimuli negatively impacted subjects' successful completion. However, collected data shows that this impact is not consistent across all stimuli. The negative effect may be tied to certain features of a particular stimulus. Instances of significant degradation in subject success rates were linked to dynamic sound stimuli, more than static ones. However, this comes with the noted exception of crying baby. While static, it had by far the greatest negative impact on subject failure rates. This could be related to the ecological significance of the sound of a crying baby. In turn, it might be that highly dynamic or aversive stimuli (e.g., Natural or

Looming) are not necessarily the most effective adversarial stimuli, despite what the Yerkes-Dodson model asserts. Instead, ecologically-significant stimuli such as crying baby could be crafted for a specific victim population.

Negative impact on subject task completion rates under these conditions could pave the way for the adversary who controls the ambient soundscape. Through the use of specifically-crafted sounds with shifting intensity levels (or high ecological significance), the adversary could force a user into failing CAPTCHAs as a denial-of-service (DoS) attack. Moreover, not being limited by any ethical boundaries, the adversary can increase the volume far beyond OSHA-recommended safe levels. This would allow creation of even more dynamic stimuli and could push performance degradation beyond the doubling of errors we observed with the crying baby stimulus. Also, more dynamic stimuli impacted completion speed of successful subjects, slowing them down. The one-way ANOVA analysis we performed on stimuli distributions implies that different stimuli impact completion speeds differently. Furthermore, voice was the stimulus with the greatest impact on subject completion times. This is noteworthy because the task itself revolves around visual interpretation of letters and numbers.

It is reasonable to assume that subjects are confounded by the sensory crossfire of listening to random letters and numbers being read aloud while they try to read and write random letters and numbers. This is analogous to the Stroop effect, and implies that some features of the specific stimuli impact completion speeds differently [15]. The adversary can use the knowledge of the specific task to construct an optimal interfering stimulus.

The real threat of negative effects occurs when they are combined. CAPTCHAs are often used as a defense against the abuse of bots at point-of-sale of limited-quantity time-sensitive services, such as event tickets or travel flash sales. These limited commodities typically sell out completely, within seconds of availability [11]. Therefore, even a single CAPTCHA failure or a second-long delay, can cause a victim to totally miss out on a potentially important (to them) opportunity.

5 Unattended Setup Analysis

Advantages: The primary goal of our study was **not** to assess accuracy of the unattended experimental setup. However, results from the Control case are analogous to the attended experiment in [4] which used short alphanumeric CAPTCHAs with 1-px. global lines. Results obtained in the Control case for our experiment: mean solving time of 4.62 seconds and accuracy of 0.912 for a 5 character code are consistent with predictions in [4] for the same type of CAPTCHAs. This reinforces equivalence between unattended and attended experimental paradigms.

In general, unattended setups are very well-suited for completing rote, repetitive tasks, such as solving numerous CAPTCHAs. Since subject performance appears to be in-line in both paradigms, an unattended setup saves person-hours that are otherwise spent on logistics of scheduling and physically attending experiments. Moreover, there is no burden on the subject to adhere to a particular schedule, or a limited time-window, since the experiment can run 24/7/365. Furthermore, although it was not done in this case, the unattended paradigm allows for seamless, identical replication in multiple locations simultaneously, which is impossible in an attended manner. Finally, this paradigm entirely avoids experimenter bias: since no one is present during the experiment, there is no way to taint data collection by experimenter's actions.

Limitations: As mentioned earlier, some subjects were non-compliant and their data was discarded. This occurred despite clear instructions (during the initial phase) that CAPTCHAs had to be solved continuously for 54 minutes. Non-compliance is a basic limitation of the unattended setup: no one can enforce the rules in real-time[3].

Our setup did not capture fine-grained data about subjects' awareness of the stimuli. In the video recordings of some subjects, there is some evidence of them noticing the stimuli in obvious ways, such as making verbal remarks, or turning their heads towards the speakers. However, there is no firm evidence that shows any subject's failure to notice a given stimulus. Such information would be crucial for development of a realistic adversarial model.

The nattended setup might be both appropriate and useful for assessment of task performance, completion of questionnaires or any study that has subjects act in a fixed manner. However, it is not well-suited for adaptive data collection, e.g., what may be obtained in a loosely-structured interview. Also, since there is no on-site real-time interaction, every subject has an identical experience, which can cause the loss of corner-case data.

6 Conclusions & Future Work

As IoT-enabled sensory environments become more common, the threat of having to complete security-critical tasks in an adversary-controlled environment increases. This trend motivates studying the impact of external stimuli on performance of such tasks. Research described in this paper sheds some light on the impact of sensory stimulation on performance of security-critical tasks. However, there remain numerous outstanding issues and directions for future work:

• Our results highlight the threat of realistic distributed adversary that aims to induce extra errors and/or longer task completion. While this may not be seen as dire, due to the nature of CAPTCHAs, it opens up a worrisome attack vector for cognitively similar tasks. Notably, many systems implementing two-factor authentication use a similar challenge format to CAPTCHAs, with the distinction that challenges are sent to the user in plain text, instead of a distorted image. Replication of a similar experiment using more security-sensitive task (e.g., two-factor authentication) would point to a more obvious security threat. This would outline practical security concerns for emergent IoT-rich environments where the auditory environment could become adversary-owned.

• It is unclear whether our results can be generalized to non-text CAPTCHAs. Many popular CAPTCHA implementations utilize photographic images, such as Google's ReCAPTCHA, which asks users to identify numbers in pictures of address signs, or objects within regions of a picture (e.g. all regions of a large image that contain a car) [24]. Since recognition of objects within images is a different cognitive task than "deciphering" distorted text, it would be worthwhile to see if effects of unexpected auditory stimuli could be replicated with other CAPTCHA types.

• Finally, we intend to further explore the space of sensory stimuli's impact on performance of security-critical tasks. We aim to create a general framework of the Yerkes-Dodson relationship between sensory stimulation and user performance of arbitrary

[3] Although it would have been possible to detect non-compliance automatically, e.g., via an inactivity timeout, non-compliant subject data would still be discarded

security-critical tasks. This framework would be instrumental in both detailing the potential threats of a hostile "smart" sensory environment and describing a set of best-practice for service providers that want to optimize usability for required security challenges.

References

1. BENIGNUS, V. A., OTTO, D. A., AND KNELSON, J. H. Effect of low-frequency random noises on performance of a numeric monitoring task. *Perceptual and motor skills 40*, 1 (1975), 231–239.
2. BERG, B. G., KACZMAREK, T., KOBSA, A., AND TSUDIK, G. An exploration of the effects of sensory stimuli on the completion of security tasks. *IEEE Security & Privacy 15*, 6 (2017), 52–60.
3. BURSZTEIN, E., BETHARD, S., FABRY, C., MITCHELL, J. C., AND JURAFSKY, D. How good are humans at solving captchas? a large scale evaluation. In *Security and Privacy (SP), 2010 IEEE Symposium on* (2010), IEEE, pp. 399–413.
4. BURSZTEIN, E., MOSCICKI, A., FABRY, C., BETHARD, S., MITCHELL, J. C., AND JURAFSKY, D. Easy does it: more usable captchas. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (2014), ACM, pp. 2637–2646.
5. CHANG, R., AND SHMATIKOV, V. Formal analysis of authentication in bluetooth device pairing. *FCS-ARSPA07* (2007), 45.
6. CHELLAPILLA, K., LARSON, K., SIMARD, P., AND CZERWINSKI, M. Designing human friendly human interaction proofs (hips). In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2005), ACM, pp. 711–720.
7. COHEN, R. A. Yerkes–Dodson law. In *Encyclopedia of clinical neuropsychology*. Springer, 2011, pp. 2737–2738.
8. EL AHMAD, A. S., YAN, J., AND NG, W.-Y. Captcha design: Color, usability, and security. *IEEE Internet Computing 16*, 2 (2012), 44–51.
9. HARRIS, W. *Stress and Perception: The Effects of Intense Noise Stimulation and Noxious Stimulation upon Perceptual Performance*. Ph.D. thesis, University of Southern California, 1960.
10. HOCKEY, G. R. J. Effect of loud noise on attentional selectivity. *The Quarterly Journal of Experimental Psychology 22*, 1 (1970), 28–36.
11. KAISER, E., AND FENG, W.-C. Helping ticketmaster: Changing the economics of ticket robots with geographic proof-of-work. In *INFOCOM IEEE Conference on Computer Communications Workshops, 2010* (2010), IEEE, pp. 1–6.
12. KHALIL, A., ABDALLAH, S., AHMED, S., AND HAJJDIAB, H. Script familiarity and its effect on captcha usability: An experiment with arab participants. *International Journal of Web Portals (IJWP) 4*, 2 (2012), 74–87.
13. KOLIAS, C., KAMBOURAKIS, G., STAVROU, A., AND VOAS, J. Ddos in the iot: Mirai and other botnets. *Computer 50*, 7 (2017), 80–84.
14. LAZEM, S., AND GRACANIN, D. Social traps in second life. In *2010 Second International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)* (Mar. 2010), pp. 133–140.
15. MACLEOD, C. M. Half a century of research on the stroop effect: an integrative review. *Psychological bulletin 109*, 2 (1991), 163.
16. MAROTTA, V., AND ACQUISTI, A. Online distractions, website blockers, and economic productivity: A randomized field experiment. *Preliminary Draft* (2017).

17. OLLESCH, H., HEINEKEN, E., AND SCHULTE, F. P. Physical or virtual presence of the experimenter: Psychological online-experiments in different settings. *International Journal of Internet Science 1*, 1 (2006), 71–81.
18. OLMEDO, E. L., AND KIRK, R. E. Maintenance of vigilance by non-task-related stimulation in the monitoring environment. *Perceptual and motor skills 44*, 3 (1977), 715–723.
19. O'MALLEY, J. J., AND POPLAWSKY, A. Noise-induced arousal and breadth of attention. *Perceptual and motor skills 33*, 3 (1971), 887–890.
20. RIVA, G., TERUZZI, T., AND ANOLLI, L. The use of the internet in psychological research: comparison of online and offline questionnaires. *CyberPsychology & Behavior 6*, 1 (2003), 73–80.
21. ROGERS, R. D., AND MONSELL, S. Costs of a predictible switch between simple cognitive tasks. *Journal of experimental psychology: General 124*, 2 (1995), 207.
22. SÖDERLUND, G., ET AL. Positive effects of noise on cognitive performance: Explaining the moderate brain arousal model. In *The 9th Congress of the International Comisssion on the Biological Effects of Noise* (2008), Leibniz Gemeinschaft, pp. 378–386.
23. VON AHN, L., BLUM, M., HOPPER, N. J., AND LANGFORD, J. Captcha: Using hard ai problems for security. In *International Conference on the Theory and Applications of Cryptographic Techniques* (2003), Springer, pp. 294–311.
24. VON AHN, L., MAURER, B., MCMILLEN, C., ABRAHAM, D., AND BLUM, M. recaptcha: Human-based character recognition via web security measures. *Science 321*, 5895 (2008), 1465–1468.
25. WHELAN, R. Effective analysis of reaction time data. *The Psychological Record 58*, 3 (2008), 475–482.

A A: Background & Related Work

This section overviews related work in automated experiments, and human-assisted security methods.We also provide psychological background theory related to effects of sensory arousal on subject task performance.

A.1 Automated Experiments

There has been a prior study focusing on effects of visual and auditory stimuli on completion of a specific security-critical task – Bluetooth pairing [2]. It showed that introduction of unexpected stimuli has a spectrum of beneficial and detrimental effects on subject performance. That initial result motivates a more thorough examination of the space of security-critical tasks, since Bluetooth pairing is a very simple (and infrequent) cognitive task that only requires a single button press to confirm matching codes [5].

Some prior work focused on evaluating virtually-attended remote experiments and unattended online surveys. in comparison with those conducted in the traditional lab setting. Ollesch et al.[17] collected psychometric data in a physically attended experimental lab setting and its virtually attended remote counterpart. No significant differences were found. This is further reinforced by Riva et al.[20] who compared data collected from unattended online, and attended offline, questionnaires. Finally, Lazem and Gracanin [14] replicated two classical social psychology experiments where both the participants and the experimenter were represented by avatars in Second Life[4], instead of being physically co-present. Here too, no significant differences were observed.

[4] See `secondlife.com`

Finally, Marotta and Acquisti explored the impact of access to potentially distracting websites, particularly social media sites, on MTurk users performing a variety of tasks[16]. They found that restricting access to tdistracting sites increased user productivity by 8 tasks an hour. This is orthogonal to our exploration, as [16] is dealing with the users' self-control and the work presented here is focused on externally-presented auditory stimuli.

A.2 User Studies of Text-Based CAPTCHAs

Given ubiquity of CAPTCHAs, it is surprising that only a few usability studies have been conducted.

Chellapilla et al. [6] performed the first usability evaluation of CAPTCHAs, by examining character-based CAPTCHAs and evaluating Robustness/Usability tradeoffs. Results showed that sophisticated segmentation algorithms can violate robustness goals of popular, currently deployed text-based CAPTCHAs. However, service providers are hesitant to switch to more difficult CAPTCHAs for fear of low user acceptability.

Bursztein et al. [3] conducted a large-scale evaluation of user performance with several CAPTCHA schemes. Performance varied widely from scheme to scheme, with user's success rates ranging from 91% to 70%. This contradicted self-reported statistics, e.g., from Ebay, which claimed a 98% successful completion rate. Audio-only CAPTCHAs were found to be extremely difficult for most users, with success rates as low as 35%. This motivates guidelines for user-friendly text-based, and the need for further study of audio-only, CAPTCHAs.

Yan and El Ahmed [8] examine what makes CAPTCHAs usable, and non-intrusive. Color is identified as the primary culprit in intrusiveness, as clashing schema can interfere with presentation of the site itself. Furthermore, coloring a CAPTCHA lowers robustness, since it gives an easy target for segmentation, i.e., separating the image by color. Surprisingly, inclusion of color in a CAPTCHA is clamed to be a benefit for both usability and robustness if done correctly. However, what constitutes correct color usage is left as an open problem.

Khalil et al. examine the impact of alphabet familiarity on CAPTCHA performance using different character sets [12]. Familiarity with the alphabet used to construct a text-based CAPTCHA does not impact error rates. However, users' satisfaction is positively correlated with their familiarity level with the alphabet being used.

Burszstein et al. [4] paramaterized CAPTCHA features to find the most usable combination. This was done with particular focus on low-security CAPTCHAs that could sacrifice robustness and allow bots to achieve $> 0.01\%$ success rate. Subjects were found to prefer CAPTCHAs composed of English-language words with positive connotations (such as "cutest") with simple global distortions, and very few intersection or occluding lines. The study concluded with a candidate CAPTCHA design that showed a 95.4% success rate.

To date, there has been no evaluation of user performance with CAPTCHAs in a noisy environment.

A.3 Effects of Sensory Stimulation

Sensory stimulation has variable impact on task performance. This is due to many factors, including the subject's current level of arousal. The Yerkes-Dodson Law stipulates an inverse quadratic relationship between arousal and task performance [7]. It implies that, across all contributing stimulants, subjects who are either at a very low – or very high – level of arousal are unlikely to perform well, and there exists an optimal level of arousal for correct task completion.

An extension to this law is the notion that completion of less complex tasks that produce lower levels of initial arousal in subjects benefits from inclusion of external stimuli with low

to medium arousal. At the same time, completion of complex tasks that produce a high level of initial arousal suffers from inclusion of external stimuli. Hockey [10] and Benignus et al. [1] classified this causal relationship by defining task complexity as a function of the task's event rate (i.e., how many subtasks must be completed in a given time-frame) and the number of sources that originate these subtasks. External stimulation can serve to sharpen the focus of a subject at a low arousal level, improving task performance [18]. Conversely, it can overload subjects that are already at a high level of arousal, and induce errors in task completion [9].

O'Malley and Poplawsky [19] argued that sensory noise affects behavioral selectivity. Specifically, while a consistent positive or negative effect on task completion may not occur, a consistent negative effect was observed for tasks that require subjects to react to signals on their periphery. Meanwhile, a consistent positive effect on task completion was observed for tasks that require subjects to react to signals in the center of their field of attention. This leads the authors to claim that sensory stimulation has the effect of narrowing the subject's area of attention.

B B: Study Shortcomings

This section discusses some shortcomings of the study.

Homogeneous Subjects: Our subject group was comprised of young and tech-savvy college students. This is a consequence of the experiment's location and recruitment methods. Replication of this experiment in a non-academic setting would be useful. However, recruiting an appropriately diverse set of subjects is still difficult, even in a public setting. Ideal venues might be stadiums, concert halls, fairgrounds or shopping malls. Unfortunately, deployment of the unattended setup in such public locations is logistically infeasible. Since such public areas are already full of other sensory stimuli, reliable adjustment of subjects' arousal level in a consistent manner would be very hard. Furthermore, it would be very difficult to secure expensive experimental equipment.

Synthetic Environment: Even though we attempted to provide a realistic environment for CAPTCHAs, our setup was obviously a contrived, artificial and controlled space. Typically, people encounter CAPTCHAs while using their own devices from their own homes or offices. As such, it would be intuitive to conduct a study remotely over the Internet. However, this would introduce many compounding and potentially dangerous variables. First, there would be no way of knowing ahead of time the exact nature of the potential subjects' auditory environment. This could lead to complications ranging from the trivial nullification of collected data (e.g., if subject's audio-out is muted) all the way to potential hurting subject's auditory faculties (e.g., in-ear headphones turned to a dangerously high volume).

This further complicates measurement of any effects of auditory stimuli, as it becomes unclear if any two subjects encounter the stimuli the same way. For example, a subject using headphones at a high volume is going to have a drastically different experience than a subject using speakers at a low volume. These differences will confound the actual impact of the stimuli, making it extremely difficult to quantify any meaningful effect on task performance. Because of the need of homogeneity in presentation of the stimuli, it is easy to see how such an online experiment would be ineffective in practice.

C C: Ethical Consideration

Experiments described in this paper were fully authorized by the Institutional Review Board (IRB) of the university, well before the study. The level of review was: Exempt, Category II. Further IRB-related details are available upon request. No sensitive data was harvested during the experiments and minimal identifying information was retained. In particular:

1. No names, addresses, phone numbers or other identifying information was collected from the participants.

2. Although email addresses were solicited in order to confirm participation, they were erased very soon thereafter.
3. Video recordings of the experiments were kept for study integrity purposes. However, they were erased before the IRB expiration time.

Finally, with regard to safety, sound levels were maintained at between 70 and 88 dB, which is (especially, for only 2:15 minutes) generally considered safe.