An Unattended Study of Users Performing Security Critical Tasks Under Adversarial Noise

Tyler Kaczmarek, Alfred Kobsa, Gene Tsudik, and Robert Sytkaczmar@uci.edu
UC Irvine
Introduction

• Personal wireless devices are ubiquitous
• Used for security-critical tasks every day
  • PIN entry, Bluetooth Pairing, CAPTCHA entry
• Extensive usability studies on ideal techniques
• But wait...
We don’t live in a sterile lab-like environment

• Distraction is everywhere
• Audio, visual, olfactory, tactile
• Does this cause failure?
• Does this slow us down?
• Can intentional distraction wreak havoc?
Motivation

• Can an agent with environmental control impact user success rates in security critical tasks?
  • Adversary increasing failure rate?
  • Benefactor decreasing failure rate?

• Can an agent with environmental control impact user completion speeds in security critical tasks?

• Do different sounds cause different effects?
  • Based on volume?
  • Based on sound type?
Contributions

• First study of effects of auditory noise on Security Task completion
• First unattended study
  • 147 subjects, 5 stimuli
Overview

• Brief foray into Effect of Noise on Perception
• Previous User Studies
• Our setup
• The participants
• The experiment
• Results
• Discussion
• Lessons from our Design
• Moving forward
A Look at Distraction

• Mixed results
• Auditory noise can have positive, negative or no effect
• Related to subjects’ overall sensory arousal
  • The type of noise
  • The complexity of the task
A Look at Distraction cont’d

- Yerkes-Dodson Law
  - Low sensory arousal levels can be error-prone
    - Sleepy, unengaged
  - High sensory arousal levels can be error-prone
    - Sensory overload
- In between is ideal
- Where do security-critical tasks fit?
User Studies of Security Critical Tasks

• Primarily aimed at most effective pairing method
• “Short Authentication String” (SAS) protocols favored
  • Subjects compare ~20 bit strings for equality
• Groups can complicate things
  • “insecurity of conformity”
  • We focus on individuals
• Controlled, lab-like setting
The Setup

• Need at least 20-25 subjects for 5 different stimuli
• 125+ trials would be costly, time-consuming
• Solution: unattended experiment
• Looking at individual subjects performing Bluetooth Pairing
IRB Clearance

• Fully cleared with Institutional Review Board as “Exempt”
• Limited sound volume
• Do not use any subject secrets
The Setup – Subject’s Perspective

• Set up in Comp Sci building on campus
• Potential subjects followed posted advertisements
• Led to a low-traffic public alcove
• Had a Smartboard, projector system and 4 speakers

Experiment environment, side view
The Setup – Experimenter’s Perspective

• Webcam recorded subjects
• Experimenters review after the fact
  • Used to confirm single subjects, gender etc.
• No active experimenter participation
• Experiment ran 24/7 for several months

Example Video Recording
The Subjects

• 147 total subjects
• Volunteers around Engineering / Comp Sci section of campus
• 94% “college aged” (18-29), 6% older (30+)
• 69% male, 31% female
• Vast majority of devices were Smartphones
The Experiment – Subject Task

• Subject interacts with recorded “proxy experimenter”
  • Proxy reads off single instruction set
  • No live monitoring or assistance is given

• Subject asked to pair personal device with ours via Bluetooth
  • 2 minute time window to pair

Experiment environment, back view
The Experiment - Stimuli

• During the pairing process either:
  • Nothing happens (Control case)
  • Recording of crying baby played
  • Recording of helicopter played
  • Recording of hammering played
  • Recording of table-saw played

Subject Pairing Devices
The Experiment – Sound Parameters

• Sounds were played at safe high volume
  • From 69 dB to 80 dB
    • Below OSHA threshold of 85 dB
  • Lower volumes less arousing
  • Higher volumes potentially dangerous

• Unrealistic limitation
  • Adversary can be unethical
The Experiment – Data Collection

• After pairing:
  • Subject filled out short survey
  • Subject given promised reward ($5 Amazon card)
Results – Data Cleaning

• Several cases purged
  • Subjects using old Flip phones (10)
  • Subjects in groups (29)
  • Subjects with obvious hearing impairment (0)
## Results – Raw Failure Rates

### Failure Rate by Stimulus (single trial)

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Successful Subjects</th>
<th>Unsuccessful Subjects</th>
<th>Failure Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (control)</td>
<td>27</td>
<td>13</td>
<td>0.34</td>
</tr>
<tr>
<td>Baby</td>
<td>23</td>
<td>1</td>
<td>0.04</td>
</tr>
<tr>
<td>Hammer</td>
<td>33</td>
<td>3</td>
<td>0.08</td>
</tr>
<tr>
<td>Helicopter</td>
<td>24</td>
<td>1</td>
<td>0.04</td>
</tr>
<tr>
<td>Saw</td>
<td>20</td>
<td>2</td>
<td>0.09</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>127</strong></td>
<td><strong>20</strong></td>
<td><strong>0.14</strong></td>
</tr>
</tbody>
</table>

### Failure Rate by Stimulus (multiple trials)

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Successful Subjects</th>
<th>Unsuccessful Subjects</th>
<th>Failure Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (control)</td>
<td>28</td>
<td>13</td>
<td>0.32</td>
</tr>
<tr>
<td>Baby</td>
<td>24</td>
<td>1</td>
<td>0.04</td>
</tr>
<tr>
<td>Hammer</td>
<td>34</td>
<td>3</td>
<td>0.08</td>
</tr>
<tr>
<td>Helicopter</td>
<td>24</td>
<td>1</td>
<td>0.04</td>
</tr>
<tr>
<td>Saw</td>
<td>20</td>
<td>2</td>
<td>0.09</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>130</strong></td>
<td><strong>20</strong></td>
<td><strong>0.13</strong></td>
</tr>
</tbody>
</table>
## Results – Analysis of Failure Rates

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Total Pairings</th>
<th>Failure Rate</th>
<th>Wald Statistic</th>
<th>Nuisance Parameter</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (control)</td>
<td>40</td>
<td>0.34</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Baby</td>
<td>24</td>
<td>0.04</td>
<td>2.65</td>
<td>0.95</td>
<td>0.03</td>
</tr>
<tr>
<td>Hammer</td>
<td>36</td>
<td>0.08</td>
<td>2.58</td>
<td>0.91</td>
<td>0.01</td>
</tr>
<tr>
<td>Helicopter</td>
<td>26</td>
<td>0.04</td>
<td>2.71</td>
<td>0.89</td>
<td>0.01</td>
</tr>
<tr>
<td>Saw</td>
<td>22</td>
<td>0.09</td>
<td>2.05</td>
<td>0.84</td>
<td>0.03</td>
</tr>
</tbody>
</table>

### Odds Ratio WRT Control and 95% Confidence Interval WRT Control

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Odds Ratio WRT Control</th>
<th>95% Confidence Interval WRT Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (control)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Baby</td>
<td>0.09</td>
<td>0.01 – 0.74</td>
</tr>
<tr>
<td>Hammer</td>
<td>0.18</td>
<td>0.04 – 0.73</td>
</tr>
<tr>
<td>Helicopter</td>
<td>0.09</td>
<td>0.01 – 0.71</td>
</tr>
<tr>
<td>Saw</td>
<td>0.20</td>
<td>0.04 – 1.02</td>
</tr>
</tbody>
</table>

Barnard’s Exact Test on Failure rates Between Control and Stimulus

An Unattended Study of Users Performing Security Critical Tasks Under Adversarial Noise
Results – More Analysis of Failure Rates

• Barnard’s Exact Test shows significant reduction in failure rates
• Lowered failure rates with noise mean
  • aroused
  • But not overstimulated
  • Narrowed focus
• Better performance than under-stimulated control case
• Negligible difference between genders
## Results – Analysis of Completion Times

### Pairwise $t$-test on completion times between control and stimulus

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Mean Time</th>
<th>Standard Deviation</th>
<th>DoF WRT Control</th>
<th>$t$-value WRT Control</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (control)</td>
<td>34.41</td>
<td>13.78</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Baby</td>
<td>31.13</td>
<td>10.06</td>
<td>63</td>
<td>0.97</td>
<td>0.35</td>
</tr>
<tr>
<td>Hammer</td>
<td>28.82</td>
<td>9.76</td>
<td>74</td>
<td>1.84</td>
<td>0.07</td>
</tr>
<tr>
<td>Helicopter</td>
<td>31.33</td>
<td>13.13</td>
<td>63</td>
<td>0.81</td>
<td>0.39</td>
</tr>
<tr>
<td>Saw</td>
<td>38.45</td>
<td>17.15</td>
<td>60</td>
<td>0.90</td>
<td>0.38</td>
</tr>
</tbody>
</table>

### Cohen’s $d$ and 95% Confidence Ratios between Stimuli and Control

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Cohen’s $d$ WRT Control</th>
<th>95% CI WRT Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (control)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Baby</td>
<td>0.27</td>
<td>-4.00 - 4.29</td>
</tr>
<tr>
<td>Hammer</td>
<td>0.47</td>
<td>-3.80 - 3.66</td>
</tr>
<tr>
<td>Helicopter</td>
<td>0.23</td>
<td>-4.04 - 5.48</td>
</tr>
<tr>
<td>Saw</td>
<td>-0.27</td>
<td>-4.54 - 6.89</td>
</tr>
</tbody>
</table>

---

An Unattended Study of Users Performing Security Critical Tasks Under Adversarial Noise

23
Results – More Analysis of Completion Times

• insignificant difference in every case

• Hammering *approaches* significant difference
  • How is Hammering different?
  • Baby crying: organic, continuous sound
  • Helicopter: mechanical, continuous sound
  • Saw: mechanical, continuous sound
  • Hammering: mechanical, discrete sound
  • Evidence not strong enough for conjecture

• Negligible difference between genders
Discussion

• Why less errors?
• Bluetooth pairing is quick, simple task
• Low levels of sensory arousal in control
• Audio noise puts subjects in “sweet spot”
  • Gets above lower arousal threshold
  • Doesn’t put over high arousal threshold
Discussion

• So, phones should screech during Bluetooth pairing?
  • No, results only show facilitation by *some* noise over *no* noise
  • Overstimulation can occur
  • The top-end threshold of arousal is unknown
  • Results suggest malicious shattering of silence as ineffective
Lessons Learned

• Single instruction set doesn’t cover all knowledge levels
  • No verbose explanation for unsure subject

• Subjects like to act in groups
  • Explicit prevention is desirable, but hard
  • Unattended nature lends to filtering out after the fact
Moving Forward – Improving the Process

• More representative subjects
  • College-aged people more tech-savvy
  • Familiarity can skew true error rate

• More security-critical task
  • Setup was clearly contrived
  • No motivation for security of device

• More complex task
  • Bluetooth paring too easy?
  • Complicated task may induce more arousal
Moving Forward – Different Experiments

• Stimulation threshold
  • When do mistakes start?

• Visual
  • Sight is our dominant sense
  • Easier to over stimulate?

• Combined sensory input
  • Multiple sources – more stimulation
  • Is there a “sweet spot” where errors start?
Questions?