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

Tyler Kaczmarek, Alfred Kobsa, Gene Tsudik, and Robert Sy

tkaczmar@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…

Introduction

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

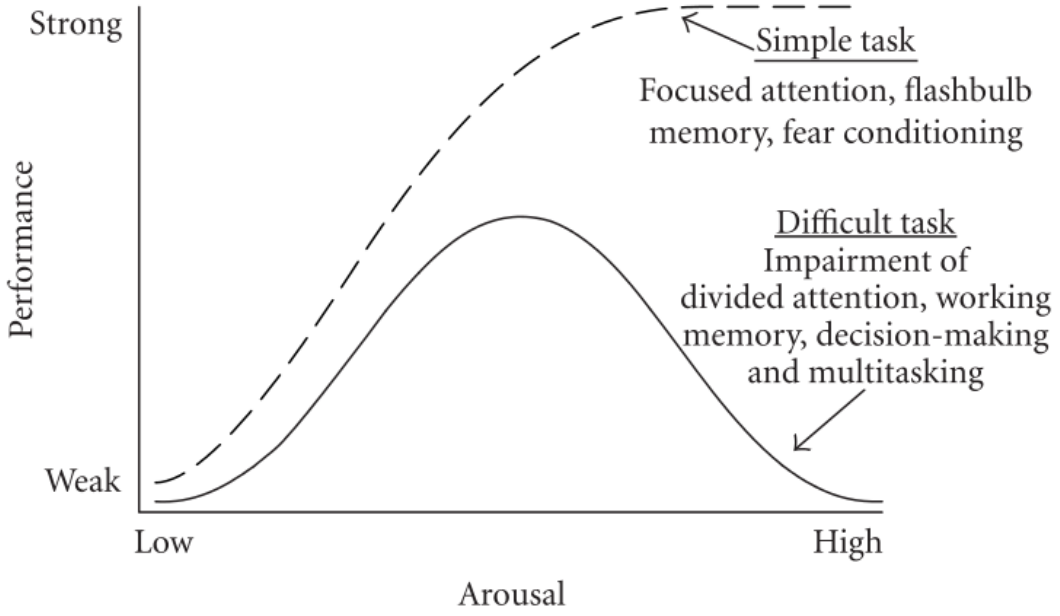
- Brief foray into Effect of Noise on Perception
- Previous User Studies
- Our setup
- The participants
- The experiment
- Results
- Discussion
- Lessons from our Design
- Moving forward

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

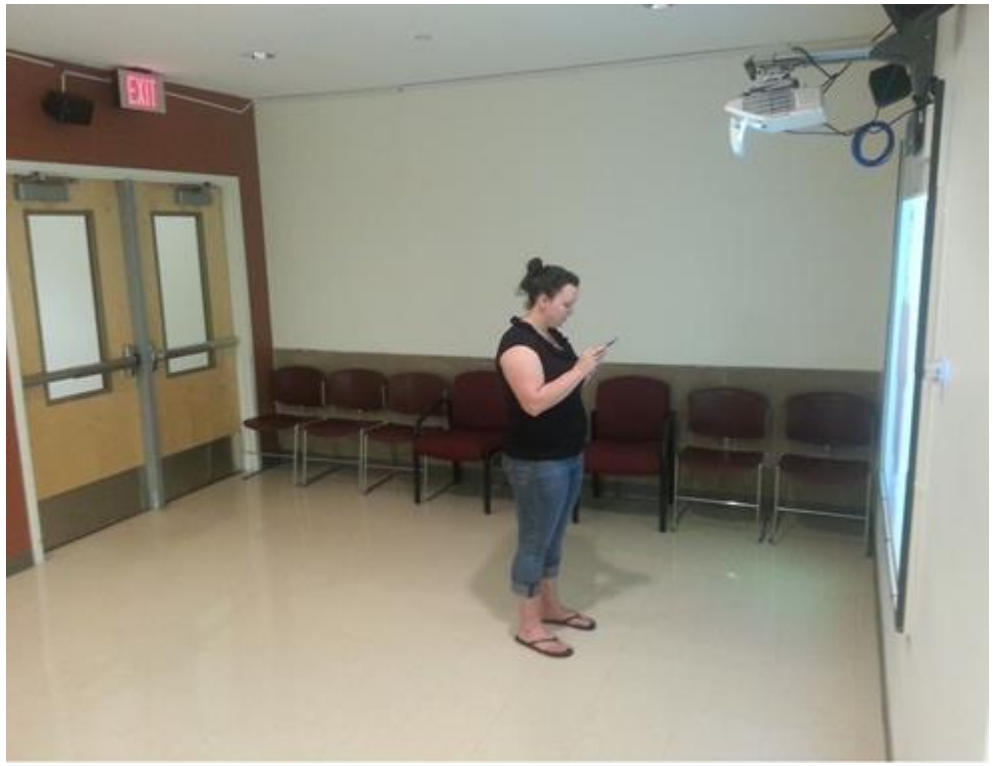
- Need at least 20-25 subjects for 5 different stimuli
- 125+ trials would be costly, time-consuming
- Solution : unattended experiment
- Looking at individual subjects preforming Bluetooth Pairing

IRB Clearance

- Fully cleared with Institutional Review Board as “Exempt”
- Limited sound volume
- Do not use any subject secrets

The Setup – Subject's Perspective

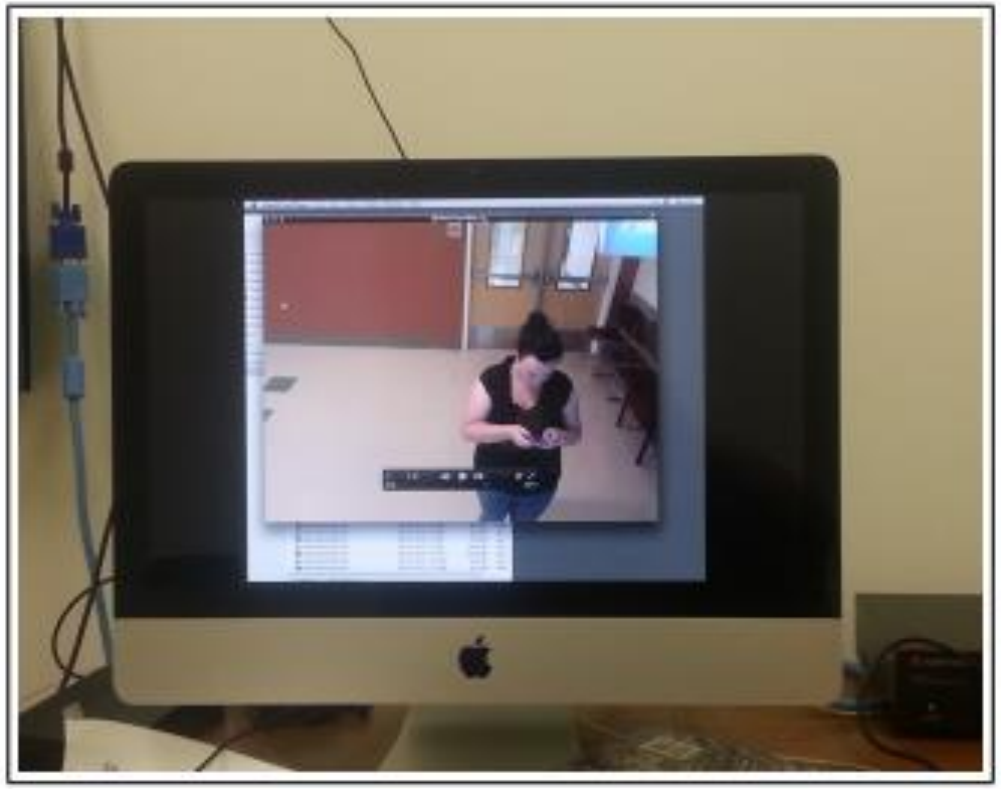
- Set up in Comp Sci building on campus
- Potential subjects followed posted advertisements
- Led to a low-traffic public alcove
- Had a Smartboard, projector system and 4 speakers



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

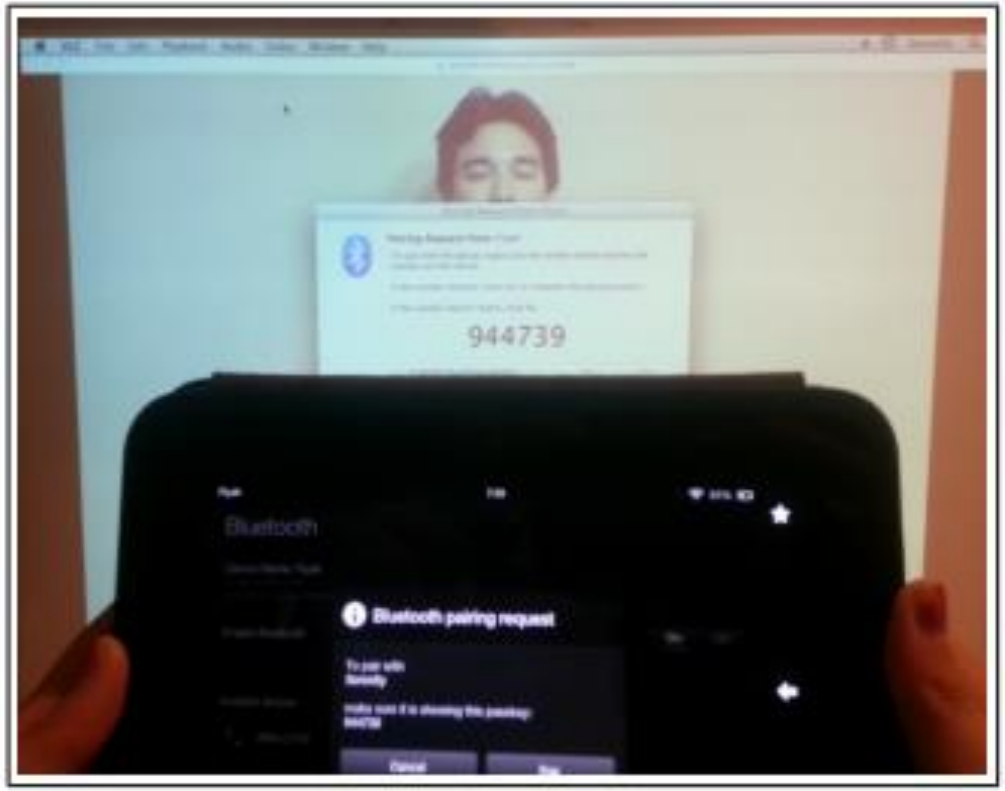
"proxy experimenter"
  - Proxy reads off single instruction set
  - No live monitoring or assistance is given
- Subject asked to pair personal device

with ours via Bluetooth
  - 2 minute time window to pair

Experiment environment, back view

The Experiment - Stimuli

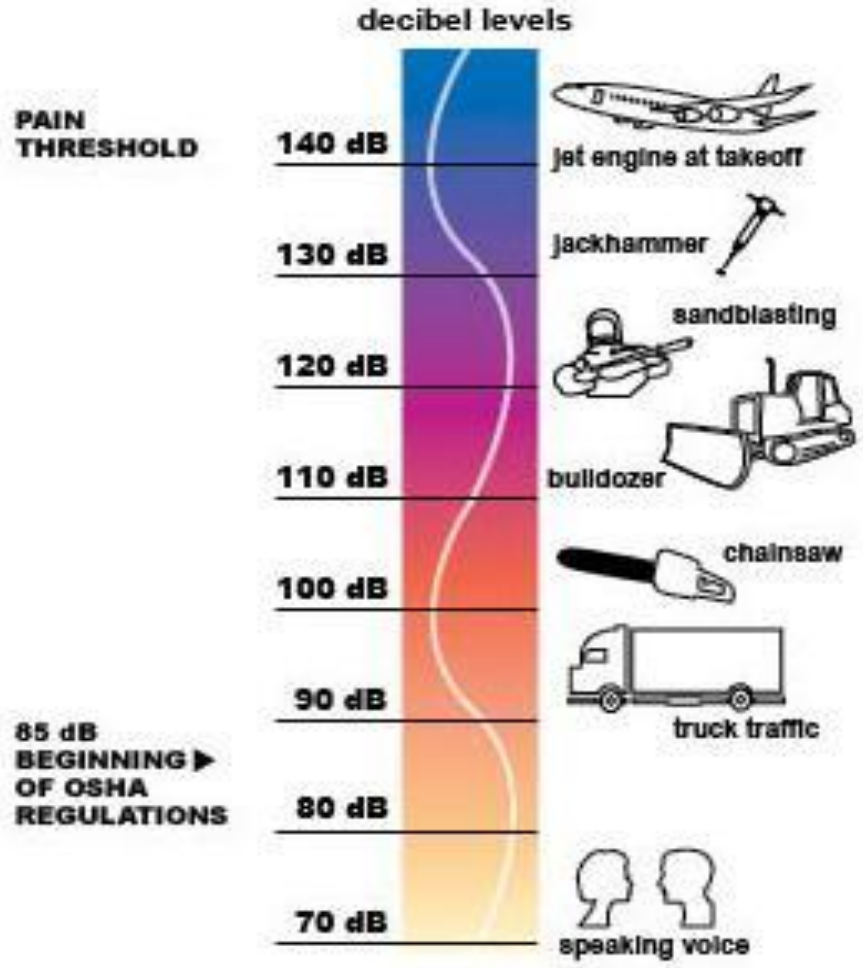
- During the pairing process either:
  - Nothing happens (Control case)
  - Recording of crying baby played
  - Recording of helicopter played
  - Recording of hammering played
  - Recording of table-saw played



Subject Pairing Devices

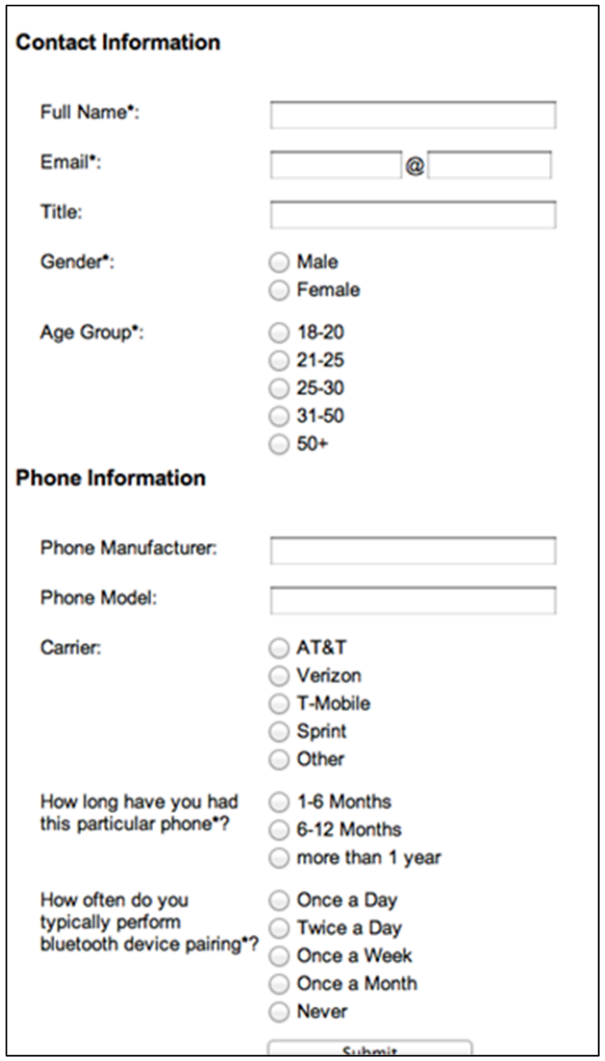
The Experiment – Sound Parameters

- 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)



Contact Information

Full Name*:

Email*: @

Title:

Gender*: Male, Female

Age Group*: 18-20, 21-25, 25-30, 31-50, 50+

Phone Information

Phone Manufacturer:

Phone Model:

Carrier: AT&T, Verizon, T-Mobile, Sprint, Other

How long have you had this particular phone*? 1-6 Months, 6-12 Months, more than 1 year

How often do you typically perform bluetooth device pairing*? Once a Day, Twice a Day, Once a Week, Once a Month, Never

Submit

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

Stimulus	Successful Subjects	Unsuccessful Subjects	Failure Rate
None (control)	27	13	0.34
Baby	23	1	0.04
Hammer	33	3	0.08
Helicopter	24	1	0.04
Saw	20	2	0.09
Total	127	20	0.14

Failure Rate by Stimulus (single trial)

Stimulus	Successful Subjects	Unsuccessful Subjects	Failure Rate
None (control)	28	13	0.32
Baby	24	1	0.04
Hammer	34	3	0.08
Helicopter	24	1	0.04
Saw	20	2	0.09
Total	130	20	0.13

Failure Rate by Stimulus (multiple trials)

Results – Analysis of Failure Rates

Stimulus	Total Pairings	Failure Rate	Wald Statistic	Nuisance Parameter	*p*
None (control)	40	0.34	--	--	--
Baby	24	0.04	2.65	0.95	0.03
Hammer	36	0.08	2.58	0.91	0.01
Helicopter	26	0.04	2.71	0.89	0.01
Saw	22	0.09	2.05	0.84	0.03

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

Stimulus	Odds Ratio WRT Control	95% Confidence Interval WRT Control
None(control)	--	--
Baby	0.09	0.01 – 0.74
Hammer	0.18	0.04 – 0.73
Helicopter	0.09	0.01 – 0.71
Saw	0.20	0.04 – 1.02

Odds Ratio and 95% Confidence Interval Between Control and Stimulus

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 ***over***stimulated
  - Narrowed focus
- Better performance than under-stimulated control case
- Negligible difference between genders

Results – Analysis of Completion Times

Stimulus	Mean Time	Standard Deviation	DoF WRT Control	*t*-value WRT Control	*P*
None (control)	34.41	13.78	--	--	--
Baby	31.13	10.06	63	0.97	0.35
Hammer	28.82	9.76	74	1.84	0.07
Helicopter	31.33	13.13	63	0.81	0.39
Saw	38.45	17.15	60	0.90	0.38

Pairwise *t*-test on completion times between control and stimulus

Stimulus	Cohen's d WRT Control	95% CI WRT Control
None (control)	--	--
Baby	0.27	-4.00 - 4.29
Hammer	0.47	-3.80 - 3.66
Helicopter	0.23	-4.04 - 5.48
Saw	-0.27	-4.54 - 6.89

Cohen's d and 95% Confidence Ratios between Stimuli and Control

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?