Usability of Security Critical Protocols Under Adversarial Noise

Tyler Kaczmarek
PhD Advancement to Candidacy Presentation
Computer Science Department
University of California Irvine

Outline

- Previous Work On Usability of Pairing techniques when exposed to adversarial stimuli
  - Introduction, Motivation and background
  - Unattended Experiment Design
  - Impact of Audio Noise on security-critical task performance [USEC 2015]
  - Impact of Visual Noise on security-critical task performance [Submitted to ASIACCS 2016]
- Future Work

Introduction, Motivation and Background

Introduction

- Personal wireless devices ubiquitous
- Used for Security-Critical tasks every day
  - PIN entry, Bluetooth pairing, CAPTCHA entry
- Extensive usability studies on ideal task techniques
- However…

Introduction Continued

We don't live In a sterile lab-like environment

- Distraction is everywhere
  - Audio, Visual, Olfactory, Tactile
- Does this induce failure?
- Does this impact task completion times?

Motivation

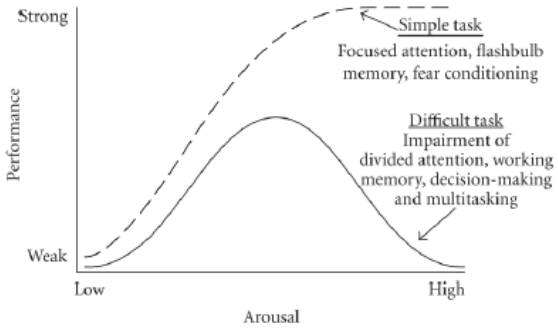
- Can an agent with control over the environment impact completion rates in security critical tasks?
  - Adversary increasing failure rates
  - Benefactor decreasing failure rates
- Can an agent with control over the environment impact completion speeds for security critical tasks?
- Do different stimuli cause different effects?
  - Difference in the sense stimulated?
  - Difference in the intensity of stimulation?

A Primer on Sensory Stimulation

- Mixed results
- Noise can have positive (Olmedo et al., 1977), negative (Beningus et al., 1975) or no effect (Harris, 1960)
- Related to subjects' overall arousal level
  - The type of noise (Koelega et al., 1990)
  - The complexity of the task (Hockey, 1970)

A Primer on Sensory Stimulation Continued

- Yerkes-Dodson Law (Cohen, 2011)
- Low levels of arousal error-prone
  - Sleepy, unengaged
- High levels of arousal error-prone
  - Overwhelmed, excessive stimulation
- In between is ideal
- Need to know arousal from task
  - Where do security-critical tasks fit?



Past User Studies of Security Critical Tasks

- Primarily aimed at most effective pairing method (Uzun et al., 2007) (Kobsa et al., 2009)
- "Short Authentication String" (SAS) protocols favored (Laur et al., 2005), (Kainda et al., 2009)
  - Subjects compare ~20 bits for equality
- Groups can complicate things (Kainda et al., 2010)
  - "Insecurity of conformity" (Nithyanand et al., 2010)
- Controlled, lab-like setting

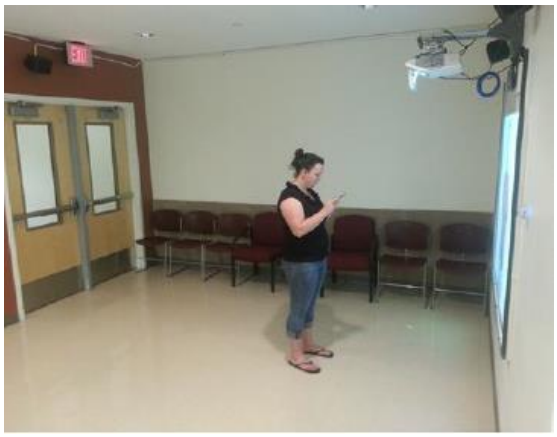
Unattended Experiment Design

The Setup

- Need 20-25 subjects per stimulus condition
  - Easily grows to hundreds of subjects
- Hundreds of trials costly, time consuming
- Solution: unattended setup
- Looking at subjects preforming Bluetooth Pairing
  - Use this setup for both visual and auditory experiments

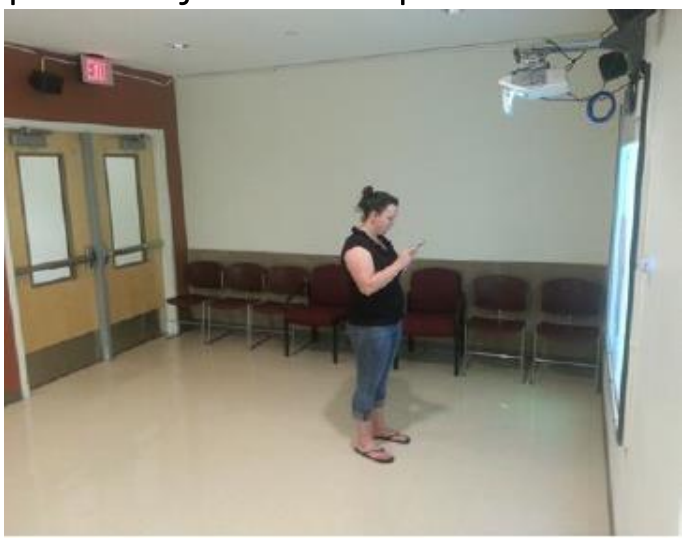
The Setup – Subject's Perspective

- Set up in Comp Sci building on campus
- Potential subjects followed posted advertisements
- Location: low-traffic public alcove
- Equipment: Smartboard, projector system, 4 Phillips HUE SmartBlubs and 4 speakers



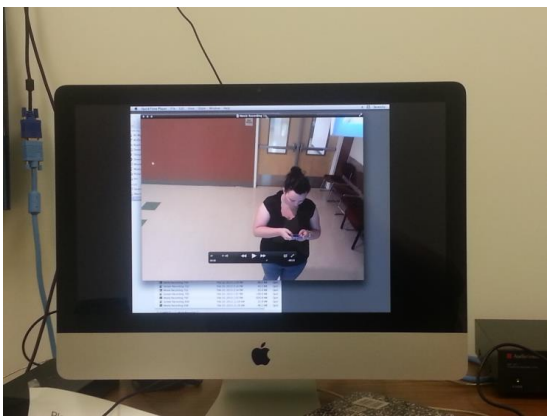
Experiment environment, side view

The Setup – Subject' Perspective Close Up



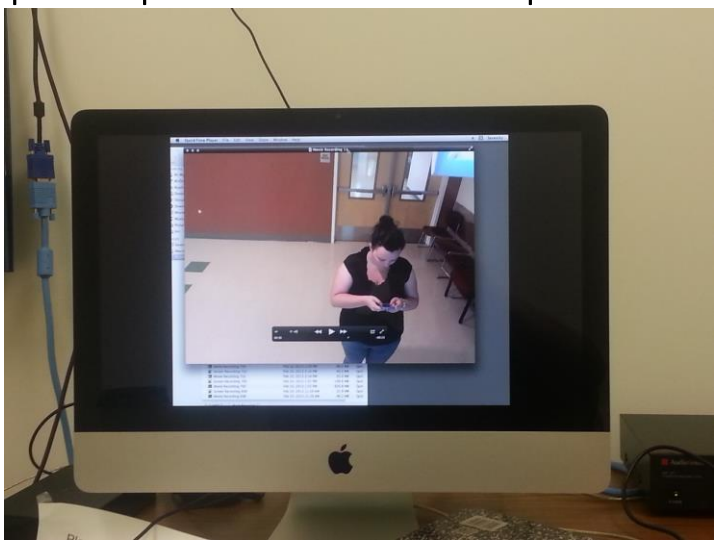
The Setup – Experimenter's Perspective

- Webcam recorded subjects
- Experimenters review after the fact
  - Used to confirm gender, correct participation, etc
- No real-time experimenter participation
- Experiment ran 24/7/365



Example Video Recording

The Setup – Experimenter's Perspective Close Up



The Setup – Data Collection

- After task completion
  - Subject filled out short survey
  - Subject given promised reward ($5 Amazon gift card)

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

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

- First study of effects of auditory noise on completion of security tasks
- First unattended study of this nature
  - 147 subjects, 5 stimulus conditions
- Presented at USEC 2015

The Subjects

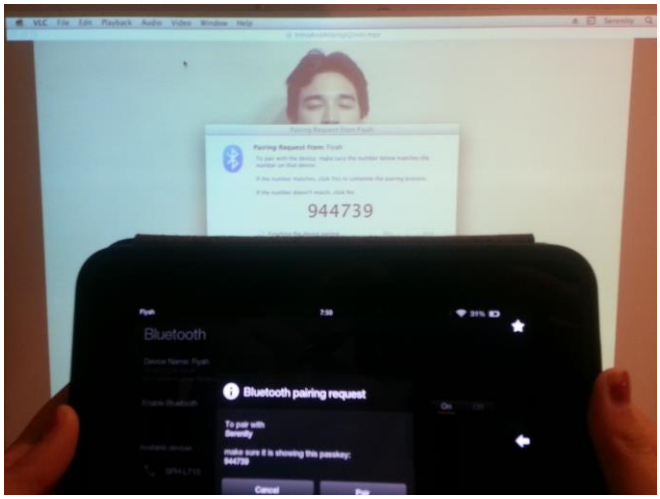
- 147 total subjects
- Volunteers around Engineering / Comp Sci section of campus
- 94% college-age (18-29), 6% older (30+)
- 69% male, 31% female
- Overwhelmingly used smartphones

Experiment – Subject Task

- Subject Interacts with recorded "proxy experimenter"
  - Proxy reads off instruction set
  - No live monitoring or assistance
- Subject asked to pair personal device with ours via Bluetooth
  - Given 2 minute time window

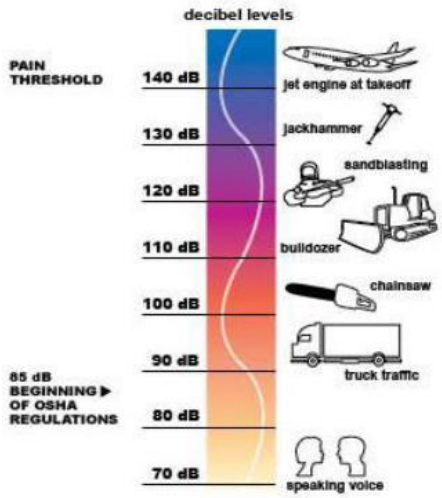
Experiment - Stimuli

- During the Pairing process either:
  - Nothing happens (control)
  - Recording of crying baby plays
  - Recording of flying helicopter plays
  - Recording of hammering plays
  - Recording of a circular saw plays



Experiment – Stimuli Parameters

- Sounds played at a 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



Results – Data Cleaning

- Several Cases Purged
  - Subjects using flip phones
  - Subjects in groups
  - Subjects with obvious hearing impairment

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

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	DF 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 fewer 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

Lights, Camera, Action! Studying Effects of Visual Distractions on Completion of Security Tasks

Lights, Camera, Action! Studying Effects of Visual Distractions on Completion of Security Tasks

- User study of effects of visual noise on completion of security tasks
- Uses similar setup to study on auditory noise
  - 169 subjects, 7 conditions
- In submission at ASIACCS 2016

The Subjects

- 169 subjects
- Volunteers from Comp Sci / Engineering part of campus
- 74% male, 26% female
- 95% college-age (18-29) 5% older (30+)
- Overwhelmingly used smartphones

The Experiment – Subject Task

- Identical task to audio experiment
- Subject Interacts with recorded "proxy experimenter"
  - Proxy reads off instruction set
  - No live monitoring or assistance
- Subject asked to pair personal device with ours via Bluetooth
  - Given 2 minute time window

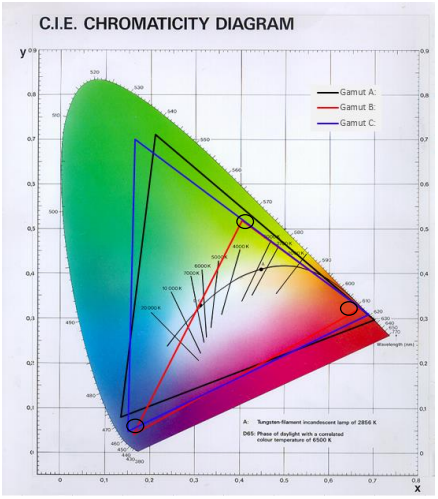
The Experiment - Stimuli

- During the Pairing process either:
  - Nothing happens (control)
  - Solid Red
  - Flickering Red
  - Solid Blue
  - Flickering Blue
  - Solid Yellow-Green
  - Flickering Yellow-Green

A subject pairing under the solid red condition

The Experiment – Stimuli Parameters

- Colors Chosen from CIE color space with emotive impact in mind (Naz and Epps, 2004)
  - Red, CCV: X = 0.674, Y = 0.322
    positive arousing connotations
  - Blue, CCV: X = 0.168, Y = 0.041
    positive relaxing connotations
  - Yellow-Green, CCV: X = 0.408, Y = 0.517
    negative arousing connotations



Results – Data Cleaning

- Group participation not discarded for 1st participant
- Otherwise identical to Data Cleaning for previous study
- Potentially Color-Blind subjects retained
  - Task does not require subjects to distinguish between colors
  - Post-hoc ANOVA analysis reinforces this decision

Results – Raw Failure Rate

Stimulus	Successful Subjects	Unsuccessful Subjects	Failure Rate
None (control)	32	15	0.32
Solid Red	11	9	0.45
Flickering Red	9	11	0.55
Solid Blue	14	6	0.30
Flickering Blue	8	12	0.60
Solid YG	10	12	0.54
Flickering YG	7	13	0.65
Total	91	78	0.46

Subject Failure Rates Under Each Stimulus

Results – Analysis of Failure Rates

Stimulus	Total Pairings	Failure Rate	Wald Statistic	Nuisance Parameter	*p*
None (control)	47	0.32	--	--	--
Solid Red	20	0.45	1.02	0.88	0.17
Flickering Red	20	0.55	2.58	0.91	0.04
Solid Blue	20	0.30	2.71	0.89	0.49
Flickering Blue	20	0.60			0.03
Solid YG	22	0.54			0.06
Flickering YG	20	0.65	2.05	0.84	0.01

Control and Stimulus

Stimulus	Odds Ratio WRT Control	95% Confidence Interval WRT Control
None (control)	--	--
Solid Red	1.7	0.60-5.11
Flickering Red	2.61	0.89-.7.63
Solid Blue	0.91	0.29-2.85
Flickering Blue	3.2	1.08-9.47
Solid YG	1.79	0.91-7.24
Flickering YG	3.96	1.31-11.6

Odds Ratio and 95% Confidence Interval Between Control and Stimulus

Results – More Analysis of Failure Rates

- Barnard's Exact Test shows significant increase in some failure rates
  - Only for "flickering" conditions
  - Dynamic changes more stimulating than static changes (Koelega et al., 1990)
- higher failure rates with noise mean
  - ***over***stimulated
  - Cannot focus on task effectively
- Worse performance than control case
- Negligible difference between genders
- Negligible difference between individuals and first member in group

Results – Analysis of Completion Times

Stimulus	Mean Time	Standard Deviation	DF wrt Control	T-value WRT control	*p*
None (control)	34.5	11.93	-	-	-
Solid Red	87.81	24.56	41	9.56	< 0.01
Flickering Red	90.44	15.62	39	11.59	< 0.01
Solid Blue	106.36	17.39	44	16.32	< 0.01
Flickering Blue	91.25	24.11	38	9.61	< 0.01
Solid YG	90.3	19.08	40	11.1	< 0.01
Flickering YG	90.29	19.06	37	10.01	< 0.01

Subject Completion Times Under Each Stimulus

Results – Analysis Of Variance (ANOVA) on Completion Times

	Sum of Squares	Degrees of Freedom	Mean Square	*F*	*p*
Between Groups	2964.28	5	592.86	1.466	0.217
Within Groups	21440.33	53	404.535		
Total	24404.61	58			

Results of One-Way ANOVA Test

Results – More Analysis of Completion Times

- Significant difference in every case
- ANOVA shows that the differences are the same
  - Each stimulus effects completion time in the same way
- Negligible difference between genders
- Negligible difference between individuals and first member of a group

Discussion

- Why more errors for flickering conditions?
  - Dynamic stimulus more arousing than static one
  - Moves subject sensory arousal to far right of Yerkes-Dodson curve
- Why so much slower?
  - Sight is dominant sense (Eimer, 2004)
  - Stimulus in same sense as task

Future Work

Future Work – CAPTCHA performance

- Bluetooth pairing has issues as a security critical task
  - Infrequent task
  - Very simple
  - Little cognitive load
- Moving forward: examine impact of noise on CAPTCHA solving
  - Larger cognitive load
  - Frequent task
  - More familiar to average user
  - No personal secrets

Future Work – Internet of Things(IoT)

- Cyber physical spaces are becoming ubiquitous
- Verification of embedded devices rely on attestation techniques
  - Developed by experts for use by experts
  - No usability studies of these techniques exist
  - Extremely limited UI on devices
- Need to conduct user studies on performance on these tasks
  - Evaluate flaws in current protocols
  - Develop Usable protocols for attestation of IoT devices

References

- Ronald A Cohen. 2011. Yerkes-Dodson Law. In *Encyclopedia of clinical neuropsychology.* Springer 2737-2738
- Martin Eimer. 2004. Multisensory integration: how visual experience shapes special perception. *Current biology* 14, 3 (2004), R115-R117.
- William Harris. 1960. *Stress and Perception: The Effects of Intense Noise Stimulation and Noxious Stimulation upon Perceptual Performance.* Ph.D. Thesis. University of Southern California
- G. R. J. Hockey. 1970. Effect of loud noise on attentional selectivity. *The Quarterly Journal of Experimental Psychology* 22, 1 (1970) 28-36
- Ronald Kainda, Ivan Flechais, and A. W Roscoe 2009. Usability and security of out-of-band channels in secure device pairing protocols. In *Proceedings of the 5th Symposium on Usable Privacy and Security*(2009), 11:1-11:12
- Ronald Kainda, Ivan Flechais, and A. W Roscoe 2010. Two heads are better than one: security and usability of device associations in group scenarios. In *Proceedings of the Sixth Symposium on Usable Security (SOUPS '10).* 5:1-5:13
- Alfred Kobsa, Rahim Sonawalla, Gene Tsudik, Ersin Uzun and Yang Wang. 2009. Serial hook-ups: a comparative usability study of secure device pairing methods. *Proceedings of the 5th Symposium on Usable Privacy and Security* (2009), 10:1 – 10:12.
- Harry S Koelega, Jan-Albert Brinkman, Bert Zwep, and Marinus N Verbaten. 1990. Dynamic vs static stimuli in their effect on visual vigilance performance. *Perceptual and motor skills*70, 3 (1990), 823-831
- Sven Laur, N. Asokan, and Kaisa Nyberg. 2005. Efficient Mutual Data Authentication Using Manually Authenticated Strings. Cryptology ePrint Archive, Report 2005/424. (2005)
- Kaya Naz and Helena Epps. 2004. Relationship between color and emotion: A study of college students. *College Student J* 38, 3 (2004), 396x
- Rishab Nithyanand, Nitesh Saxena, Gene Tsudik and Ersin Uzun. 2010. Groupthink: usability of secure group association for wireless devices. *Proceedings of the 12th ACM international conference on Ubiquitous computing* (2010), 331-340
- Esteban L. Olmedo and Roger E. Kirk. 1977. Maintenance of vigilance by non-task related stimulation in the monitoring environment. *Perceptual and motor skills* 44, 3 (1997) 715-723
- John J. O'malley and Alex Poplawskyy. 1971. Noise-induced arousal and breadth of attention. *Perceptual and motor skills*33, 3 (1971), 887-890.
- Ersin Uzun, Kristiina Karvonen, and N. Asokan. 2007. Usability Analysis of Secure Pairing Methods. In *Financial Cryptography and Data Security*, Sven Dietrich and Rachna Dhamija (Eds.) Lecture Notes in Computer Science, Vol. 4886. Springer Berlin Heidelberg, 307-324

Questions?