Prosody Training for Lowering Vishing Susceptibility

Short Paper

Abstract

This short paper reports a pilot test of prosody-focused training for voice phishing (vishing). While vishing is increasingly leveraging synthetic voices, most end-user security training remains text-centric (e.g., email phishing). We developed an audio-based prosody training intervention and evaluated user outcomes with Signal Detection Theory (SDT), separating sensitivity (d') from decision bias (criterion c). In a three-arm pilot ($N = 27$; AI-audio vs. human-audio vs. text-only), accuracy and sensitivity are stable from pre-test to post-test, but criterion shifts significantly toward a more conservative threshold ($\Delta c = +0.328$, $p < .001$), reducing both hit and false-alarm rates. Survey measures show increased perceived vulnerability and response efficacy. The results of the pilot study suggest that prosody-focused training can meaningfully shift decision thresholds even when sensitivity is unchanged.

Keywords: *Voice phishing, vishing, prosody, security training, protection motivation theory, signal detection theory, decision bias*

Introduction

Voice phishing (vishing) attacks exploit phone calls to induce victims to disclose sensitive information, transfer money, or bypass multi-factor authentication. Compared to email phishing, vishing is harder to filter automatically and may benefit from caller ID spoofing and synthetic or cloned voices. Synthetic speech spoofing and deepfake-speech detection are active areas in speech security research (Li et al., 2025), but end-user defenses must still support judgment under uncertainty and social pressure. Industry reports suggest continued growth in socially engineered attack vectors and increasing attacker sophistication (Verizon, 2024). More broadly, prior research on phishing and pretexting suggests that attackers succeed by exploiting human decision strategies and context cues (including personalized pretexts) rather than purely technical vulnerabilities (Dhamija et al., 2006; Downs et al., 2006; Jagatic et al., 2007; Workman, 2007). Prior work suggests that training can improve phishing avoidance, but also that training effects are heterogeneous and can shift behaviors beyond objective detection ability (Kumaraguru et al., 2010; Sheng et al., 2010).

Related voice-mediated training work suggests that audio interfaces can be used to teach detection strategies in other phishing modalities (Sharevski & Jachim, 2022). However, despite the salience of voice-based threats, mainstream security awareness programs still emphasize text-based cues and email-centric content. For vishing, a key open question is how user security awareness training impacts (i) discriminability between benign and malicious calls and (ii) decision thresholds that trade off false alarms (treating benign calls as suspicious) against misses (treating vishing calls as legitimate).

Therefore, we propose an audio-based training intervention that teaches prosodic cues (e.g., unnatural rhythm, mismatched affect, robotic cadence)and evaluates outcomes via Signal Detection Theory (SDT), separating sensitivity from decision bias (Green & Swets, 1966; Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999). We integrate SDT with Protection Motivation Theory (PMT) (Herath & Rao, 2009; Johnston & Warkentin, 2010; Rogers, 1983) and incorporate trust in AI explanations (Lee & See, 2004) to theorize when AI-assisted training amplifies (or dampens) reliance on prosodic cues. This short paper reports pilot evidence to refine the protocol for future research.

This short paper contributes: (1) a preliminary SDT-based evaluation of vishing training that separates sensitivity from criterion shifts, (2) a PMT mechanism model featuring prosodic-cue self-efficacy (PCSE) and AI-trust, and (3) evidence that decision bias can change even when sensitivity does not.

Theoretical Background and Hypotheses

IS research has long studied how security awareness training shapes security-compliant behaviors, highlighting rationality-based antecedents of policy compliance, the durability challenges of training effects, and

the role of feedback timing in shaping threat perceptions (Bulgurcu et al., 2010; Haeussinger & Kranz, 2013; Puhakainen & Siponen, 2010; Siponen & Vance, 2010; Yin et al., 2024). However, this literature primarily addresses text-based phishing. Vishing is attracting increased attention in IS literature (Ampel et al., 2026), and we extend this stream to voice phishing by focusing on prosodic cue utilization and separating changes in discriminability from changes in decision policy.

To ground our research, we first frame the vishing problem through the lens of Protection Motivation Theory (PMT), which explains protective behavior in terms of threat appraisal (severity, vulnerability) and coping appraisal (response efficacy, self-efficacy, response costs) (Rogers, 1983; Siponen & Vance, 2010). In information security, PMT has been used to explain adoption of protective behaviors and responses to fear appeals (Anderson & Agarwal, 2010; Herath & Rao, 2009; Johnston & Warkentin, 2010). We adapt PMT to voice-based threats by introducing prosodic-cue self-efficacy (PCSE, the perceived ability to detect deceptive prosody in a voice call). Because vishing decisions can hinge on acoustic cues, PCSE should be a proximal mechanism linking training content to detection behavior. Accordingly, **H1** posits that audio-based training (AI-audio or human-audio) improves vishing detection relative to text-only training, since audio training provides direct prosodic exemplars that text descriptions cannot convey.

We further unpack H1 into modality-specific contrasts. **H1a** posits that AI-audio improves detection relative to text-only training, and **H1b** posits that human-audio improves detection relative to text-only training. Because the AI-audio and human-audio conditions deliver identical cue content, with only the narrator source differing, content-controlled equivalence predicts no systematic difference in detection outcomes between these arms (Ahuja et al., 2022; Ayaburi & Andoh-Baidoo, 2019). **H1c** therefore posits that AI-audio and human-audio training produce statistically equivalent changes in $\Delta d'$ and Δc, tested via two one-sided tests (TOST) with an equivalence bound of $|\delta| \leq 0.30\ d'$ units.

H2 posits that training effects operate through increased coping appraisal, specifically higher PCSE and response efficacy. Because threat appraisal is theorized to energize coping appraisal, **H3** posits that perceived severity and vulnerability increase PCSE.

PMT alone cannot answer the question of how training affects (i) discriminability between benign and malicious calls and (ii) decision thresholds that trade off false alarms (treating benign calls as suspicious) against misses (treating vishing calls as legitimate). We therefore also incorporate Signal Detection Theory (SDT), which characterizes performance in terms of separable sensitivity (d') and decision criterion (c) (Green & Swets, 1966; Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999). Training can increase d' (better discrimination), but can also shift c (a more liberal or conservative threshold) without improving discriminability. SDT is therefore useful for vishing training because it separates improved discriminability from changed decision policy. Criterion shifts should also have downstream costs. In vishing detection, a liberal shift may raise hit rates but also raise false alarms, increasing communication friction in routine calls. Therefore, **H5** posits that overly liberal shifts in criterion are associated with higher communication friction.

Finally, in the AI arm, trust in AI explanations (TAI) may moderate training effectiveness. Learners who trust AI-narrated cue explanations should engage more deeply with the prosodic training, amplifying both detection gains and PCSE (Lee & See, 2004). **H6** posits that in the AI-audio arm, post-training TAI positively moderates training effects on detection sensitivity ($\Delta d'$) and prosodic-cue self-efficacy (ΔPCSE).

Research Design

We use a randomized, controlled pre-test/post-test design with three conditions: (1) *AI-based audio training*, (2) *human-narrated audio training*, and (3) *text-only control.* Participants complete a pre-test consisting of short audio clips, receive training content aligned to their condition, and then complete a post-test with new clips. Post-test surveys measure PMT constructs, trust in AI explanations, and perceived communication friction. We show the design in Figure 1.

Procedure

Participants were recruited from a university population and completed (1) demographics and baseline measures, (2) an audio check (headphones/volume) to ensure stimuli audibility, (3) pre-test classification of

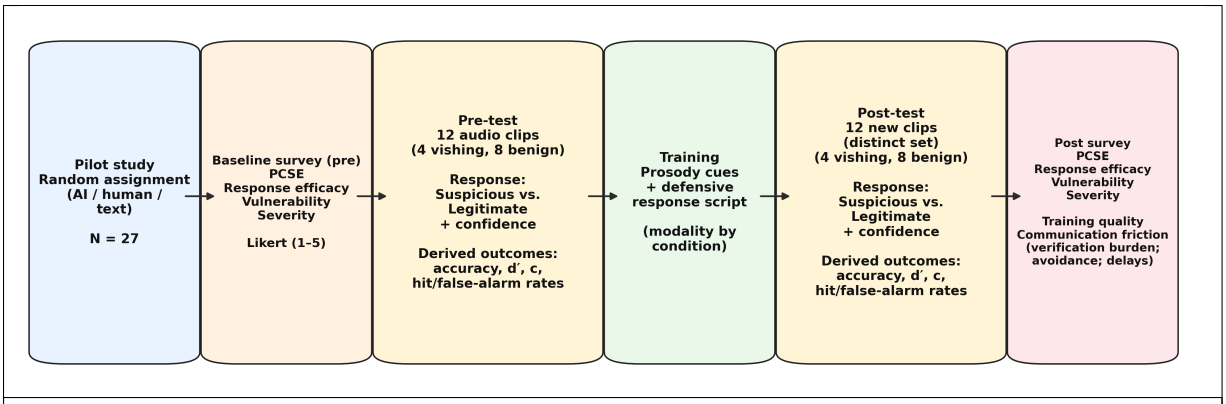

Figure 1. Pilot study design and measures.

audio clips (Suspicious vs. Legitimate) with confidence, (4) training content aligned to condition, (5) post-test classification with a counterbalanced clip set, and (6) survey measures and optional follow-up intent questions.

The intervention teaches *prosodic cues* relevant to vishing (e.g., unnatural cadence, mismatched affect, atypical stress patterns, inconsistent formality) and links those cues to defensive response scripts (pause, verify identity via known channels, and avoid disclosing credentials). The training contrasts malicious and legitimate call scripts across five prosodic cue families: (1) urgency and pressure cues, (2) requests for sensitive data, (3) tone and affect mismatches (robotic cadence, AI-like delivery), (4) caller ID spoofing indicators, and (5) scare tactics and threat escalation. The AI-audio and human-audio conditions deliver the same cue content in spoken form but differ in narrator source (synthetic vs. human). The text-only control receives the same content without audio examples.

Measures

Behavioral outcomes included detection accuracy, confidence, and SDT metrics (d', criterion c, hit rate, false-alarm rate) computed with standard log-linear corrections (Stanislaw & Todorov, 1999). Survey outcomes included PCSE, perceived severity, perceived vulnerability, and response efficacy (Herath & Rao, 2009; Rogers, 1983). Communication friction was operationalized as post-training verification behaviors and perceived burden (e.g., call-backs, delays, avoidance).

The main confirmatory analyses used cross-classified GLMMs and mixed-effects SDT models to separate training effects from stimulus difficulty, test PMT mediation and TAI moderation, and test equivalence via TOST (see §5 for full analytic detail).

Pilot Study

Sample and Procedure

The pilot included $N = 27$ participants (ages 18–24) recruited from a university population. Participants were randomly assigned to AI ($n = 10$), human ($n = 11$), or text ($n = 6$, lower because of missed attention checks) training. Table 1 summarizes participant characteristics by group. Each participant completed 12 pre-test trials (4 vishing, 8 benign) and 12 post-test trials (4 vishing, 8 benign). The pre-test and post-test used distinct clip sets. 24/27 participants passed an embedded attention check in the post-test survey. A duplicate-respondent flag appeared for one observation. Results were robust to excluding attention-check failures and the duplicate flag, with the caveat of reduced sample size.

Characteristic	AI	Human	Text	Total
N	10	11	6	27
Age 18–24	10 (100%)	11 (100%)	6 (100%)	27 (100%)
Gender				
Female	6 (60%)	10 (91%)	3 (50%)	19 (70%)
Male	4 (40%)	1 (9%)	3 (50%)	8 (30%)
Education				
High school	4 (40%)	4 (36%)	2 (33%)	10 (37%)
Some college	6 (60%)	7 (64%)	4 (67%)	17 (63%)
English proficiency				
Native	6 (60%)	8 (73%)	2 (33%)	16 (59%)
Advanced	4 (40%)	2 (18%)	3 (50%)	9 (33%)
Intermediate	0 (0%)	1 (9%)	0 (0%)	1 (4%)
Limited proficiency	0 (0%)	0 (0%)	1 (17%)	1 (4%)

Table 1. Participant characteristics by condition.

Panel A: Behavioral SDT outcomes						
Metric	**Pre**	**Post**	Δ	p	d_z	**Notes**
Accuracy	0.808	0.814	+0.006	.847	0.04	
Sensitivity (d')	1.652	1.619	-0.033	.850	-0.04	
Criterion (c)	-0.149	0.179	+0.328	<.001	0.74	95% CI [0.15, 0.50]
Hit rate (log-linear)	0.810	0.707	-0.103	.021	-0.47	
False-alarm rate (log-linear)	0.261	0.197	-0.064	.040	-0.42	
"Vishing" label rate	0.449	0.360	-0.089	.0036	-0.62	
Type-2 AUC (metacognitive)	0.590	0.727	+0.137	.039	0.47	
Panel B: Survey outcomes (Likert 1–5)						
PCSE	3.815	4.046	+0.231	.124	0.31	α/ω pre|post: .76/.77|.81/.82
Response efficacy	4.000	4.296	+0.296	.028	0.45	α/ω pre|post: .55/.65|.63/.74
Vulnerability	2.469	3.432	+0.963	<.001	1.04	α/ω pre|post: .78/.79|.64/.67
Severity	3.790	3.827	+0.037	.762	0.06	α/ω pre|post: .55/.62|.84/.86

Table 2. Pilot outcomes ($N = 27$).
Notes: Δ =Post−Pre; p = paired t-test; d_z = within-subject effect size; d' = sensitivity; c = criterion (higher is more conservative); α = Cronbach's alpha; ω = McDonald's omega.

Pilot Results: SDT Outcomes

Table 2 (Panel A) reports our pilot study results. Overall accuracy and sensitivity (d') are stable from pre-test to post-test, yet d' remains well above chance at both phases (pre: $t(26) = 12.67$, $p < .001$; post: $t(26) = 8.30$, $p < .001$), confirming that participants were actively discriminating throughout. However, criterion shifted significantly from $c = -0.149$ (pre) to $c = 0.179$ (post; $\Delta c = +0.328$, 95% CI [0.15, 0.50], $t(26) = 3.82$, $p = .001$, $d_z = 0.74$), indicating a more conservative threshold (less vishing labeling). Notably, pre-training criterion was significantly liberal ($c = -0.149$, $t(26) = -2.66$, $p = .013$) and post-training criterion was significantly conservative ($c = +0.179$, $t(26) = 2.41$, $p = .023$). Both hit rate and false-alarm rate decreased ($p < .05$), consistent with a meaningful shift in decision policy rather than improved discriminability. Non-parametric Wilcoxon tests corroborate the criterion shift ($p = .002$) and reduced vishing labeling ($p = .0056$), and robustness checks excluding attention-check failures. The duplicate-flagged response retain a positive criterion shift ($\Delta c \in [0.245, 0.314]$, $p \leq .011$). Training also selectively increased confidence on correct vishing detections (hit confidence: $M_{pre} = 5.90$, $M_{post} = 6.28$, $p = .045$, $d_z = 0.41$) without a significant change in confidence on correct benign responses ($p = .203$), indicating targeted certainty gains on the security-critical response type.

Exploratory between-arm tests do not support H1 in the pilot. Per-condition $\Delta d'$ estimates are: AI -0.21, human $+0.01$, text $+0.29$; the text control had the numerically largest d' improvement (albeit with underpowered, noisy estimates rather than a true advantage). For H1 (pooled audio vs. text), audio (AI+human) vs. text suggests no differential change in $\Delta d'$ (Welch $p = .351$; $g = -0.30$) or Δaccuracy (Welch $p = .298$; $g = -0.32$). Breaking H1 into modality-specific contrasts, AI vs. text is not significant for $\Delta d'$ (Welch $p = .205$; $g = -0.59$) and human vs. text is not significant ($p = .686$; $g = -0.16$). For H1c, AI vs. human suggests no significant difference in $\Delta d'$ (Welch $p = .600$; $g = -0.22$) or Δc (Welch $p = .562$; $g = 0.25$), consistent with the equivalence prediction; however, TOST with bound $|\delta| \leq 0.30$ does not yet establish equivalence (TOST $p = .43$ for $\Delta d'$, $p = .17$ for Δc), as the pilot is underpowered for this comparison ($n = 6$ in the text arm; $n = 10/11$ in audio arms). A full study with $N \approx 200$ is needed for a meaningful TOST. Exploratory confidence analyses show a significant between-condition difference in mean confidence change (ANOVA $F(2, 24) = 3.42$, $p = .049$): the Human arm showed near-zero confidence inflation ($\Delta = +0.17$) paired with the largest Type-2 AUC gain ($0.53\ to 0.76$). In contrast, the Text arm showed the largest increase in confidence ($\Delta = +0.54$) but the weakest calibration improvement. These results suggest that human-narrated audio fosters calibration, while text-only training may inflate confidence without commensurate accuracy gains. These patterns are directional only, given small per-condition samples.

Pilot Results: Survey Outcomes

We assessed internal consistency and related diagnostics using the full item set. Across multi-item scales, Cronbach's α ranged from .55 to .84 and McDonald's ω ranged from .62 to .86. Post-training scales showed acceptable reliability for perceived training quality ($\alpha = .79$, $\omega = .80$) and modest reliability for confidence-friction ($\alpha = .60, \omega = .63$). For post-training PMT constructs, HTMT ratios ranged from .34 to .79 (max=.79, below the conventional 0.85 threshold) and inter-construct correlations were moderate (max $|r| = .59$), providing preliminary discriminant validity evidence. As preliminary nomological evidence, PCSE was positively associated with response efficacy post-training ($r = .51$, $p = .006$).

Survey measures showed a large increase in perceived vulnerability ($\Delta = 0.963$, $p < .001$, $d_z = 1.04$) and a moderate increase in response efficacy ($\Delta = 0.296$, $p = .028$, $d_z = 0.45$). PCSE increased modestly ($\Delta = 0.231$, $p = .124$, $d_z = 0.31$), while perceived severity did not change. Post-training perceived training quality was high ($M = 4.35$, 1–5). The communication-friction proxy (post-only; 1–5) averaged $M = 4.07$ ($SD = 0.68$). Table 2 (Panel B) summarizes paired tests and internal consistency. Directionally, PCSE gains were largest in the text ($\Delta = +0.46$) and human ($\Delta = +0.39$) conditions relative to AI ($\Delta = -0.08$). At the same time, perceived vulnerability increased most in the AI condition ($\Delta = +1.20$ vs. human: $\Delta = +0.79$, text: $\Delta = +0.89$); these cross-condition patterns are illustrative only given small per-condition samples ($n = 6$–11) and will be tested formally in the full study.

The pilot primarily shifted *decision policy* rather than discriminability. Criterion moved from liberal to conservative while d' remained stable. Although survey outcomes show increases in vulnerability and response efficacy, participants labeled fewer calls as vishing. In SDT terms, the participants required stronger evidence to apply a suspicious label. In PMT terms, higher coping appraisal may reduce anxiety-driven overreaction and support more deliberate use of diagnostic cues. Pre-training false alarms carried high mean confidence ($M = 5.39/7$), suggesting initial over-vigilance was genuine rather than indiscriminate. These results are consistent with the friction mechanism in which confident false alarms impose real social costs on routine callers. Additional exploratory analyses confirm that larger conservative criterion shifts are associated with lower reported friction (Pearson $r = -.45$, Spearman $r_s = -.43$, $p_s = .025$), directly motivating H5 and the main study's friction modeling.

Discussion and Contributions

Our pilot study provided two interesting insights. First, prosody-focused instruction can shift *decision policy* even when discriminability is unchanged. Further, SDT makes this tradeoff explicit by connecting criterion shifts to downstream costs (false alarms as friction) and risks (misses as attack success). Second, additional exploratory analyses supported improved metacognitive calibration after training. The trial-level confidence correctness correlation increased from pre $r = .18$ to post $r = .27$ (both $p < .001$) and Type-2

ROC AUC (metacognitive sensitivity independent of accuracy) increased significantly from 0.590 to 0.727 ($\Delta = +0.137$, $t(21) = 2.21$, $p = .039$, $d_z = 0.47$). These results suggest that participants became reliably better at knowing when they were correct (i.e., not merely better at shifting their response threshold).

In the AI arm ($n = 10$), post-training trust in the AI narrator (TAI) is significantly correlated with both $\Delta d'$ ($r = .70$, $p = .024$) and ΔPCSE ($r = .72$, $p = .019$). These results suggest that learners who trusted the AI-delivered cue explanations showed larger sensitivity and self-efficacy gains. A marginal negative trend between pre-training TAI and $\Delta d'$ ($r = -.54$, $p = .104$) is consistent with an automation-complacency mechanism. Higher baseline AI trust may attenuate effortful engagement with the prosodic examples, while skeptical learners process them more actively. Changes in sensitivity and PCSE also exhibit a positive trend ($\Delta d'$ vs. ΔPCSE: $r = .60$, Spearman $r_s = .68$, $p_s = .030$). These AI-arm patterns provide pilot support for H6 ($n = 10$). Between-arm TAI moderation (condition $\times$ TAI interaction on $\Delta d'$ and Δc) was not significant in the pilot ($p > .37$), as expected given the available sample size. Table 3 summarizes the results of our pilot study by hypothesis, direction, and evidence.

Hypothesis	**Dir.**	**Evidence (p)**
H1. Audio-based training (AI-audio or human-audio) improves vishing detection relative to text-only training.	Opp.	.351
H1a. AI-audio improves vishing detection relative to text-only training.	Opp.	.205
H1b. Human-audio improves vishing detection relative to text-only training.	Opp.	.686
H2. Training effects operate through increased coping appraisal, specifically higher PCSE and response efficacy.	Partial	.124/.028
H3. Perceived severity and vulnerability increase PCSE.	Mixed	.189/.320
H1c. AI-audio and human-audio produce equivalent changes in $\Delta d'$ and Δc (TOST, $\lvert\delta\rvert \leq 0.30$).	Consist.	.600/.562*
H5. Overly liberal shifts in criterion are associated with higher communication friction.	Yes	.018
H6. in the AI-audio arm, post-training TAI positively moderates effects on $\Delta d'$ and ΔPCSE.	Yes	.024/.019

Table 3. Hypotheses summary. *Dir.=pilot direction vs. hypothesis; p-order follows wording; H5 uses Δc. *Standard Welch p; TOST not yet established (TOST $p = .43/.17$); pilot underpowered for equivalence test.*

Limitations and Ethics

The pilot is small ($N = 27$) and uses a student sample, limiting generalizability and between-condition inference. Several survey items require reworking to ensure we achieve high reliability and validity. Similarly, pre-training response efficacy and severity scales showed below-threshold internal consistency ($\alpha = .55$) and will need to be revised for the main study. Voice-based deception research also requires careful debriefing and safeguards to prevent the transmission of harmful attack strategies. Our training emphasizes defensive heuristics and recommends verification behaviors aligned with organizational best practices.

Conclusion

Voice phishing defense requires training that accounts for audio cues and the security-usability tradeoffs inherent in suspicion decisions. Our pilot suggests that training can meaningfully shift decision thresholds even when sensitivity is unchanged, underscoring the value of SDT for evaluating vishing training. The shift from a significantly liberal pre-training criterion ($c = -0.149$, $p = .013$) to a significantly conservative post-training criterion ($c = +0.179$, $p = .023$) is precisely the regime in which communication friction would be expected to emerge. The main study will test this link and estimate the effects of training on sensitivity, criterion, and friction.

Planned Next Steps: Building on the pilot, the full study will follow our protocol with a randomized, controlled pre-test/post-test design comparing AI-audio training, human-audio training, and text-only control. To better detect policy shifts and downstream effects, we will target $N \approx 200$ participants to sup-

port between-condition inference and mediation tests. Detecting small-to-moderate within-subject effects ($d \approx 0.30$) requires approximately 90 participants; between-condition effects require approximately 175 per arm, motivating a target of $N \approx 200$ refined via pilot-derived variance components. Robustness checks will include preregistered exclusions, randomization balance checks, stimulus-equating diagnostics, and sensitivity analyses across outcome definitions and link functions.

Stimuli will be constrained to AI-generated male voices with neutral accents to reduce speaker variability, and we will conduct a separate norming study to build matched A/B clip sets and counterbalance set order.

Post-test surveys will measure PMT constructs, PCSE, TAI, and a refined multi-item friction scale alongside behavioral proxies (response time, verification intent). Analyses will use cross-classified GLMMs and mixed-effects SDT models for trial-level outcomes, multilevel SEM for mediation and moderation, and TOST (bound $|\delta| \leq 0.30$) for H1c. TAI will be tested as a within-arm moderator of AI-condition sensitivity and PCSE gains.

References

Ahuja, V. K., Rosser, H. K., Grover, A., & Hale, M. (2022). Phish finders: Improving cybersecurity training tools using citizen science [ICIS 2022 Proceedings]. *International Conference on Information Systems (ICIS)*. Retrieved March 4, 2026, from http://aisel.aisnet.org/icis2022/security/security/1

Ampel, B. M., Samtani, S., & Chen, H. (2026). Automatically detecting voice phishing: A large audio model approach [Forthcoming]. *MIS Quarterly*.

Anderson, C. L., & Agarwal, R. (2010). Practicing safe computing: A multimethod empirical examination of home computer user security behavioral intentions. *MIS Quarterly*, *34*(3), 613–643. https://doi.org/10.2307/25750694

Ayaburi, E., & Andoh-Baidoo, F. K. (2019). Understanding phishing susceptibility: An integrated model of cue-utilization and habits [ICIS 2019 Proceedings (Short Paper)]. *International Conference on Information Systems (ICIS)*. Retrieved March 4, 2026, from http://aisel.aisnet.org/icis2019/cyber_security_privacy_ethics_IS/cyber_security_privacy/43

Bulgurcu, B., Cavusoglu, H., & Benbasat, I. (2010). Information security policy compliance: An empirical study of rationality-based beliefs and information security awareness. *MIS Quarterly*, *34*(3), 523–548. https://doi.org/10.2307/25750690

Dhamija, R., Tygar, J. D., & Hearst, M. (2006). Why phishing works. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 581–590. https://doi.org/10.1145/1124772.1124861

Downs, J. S., Holbrook, M. B., & Cranor, L. F. (2006). Decision strategies and susceptibility to phishing. *Proceedings of the Second Symposium on Usable Privacy and Security (SOUPS '06)*, 79. https://doi.org/10.1145/1143120.1143131

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Haeussinger, F., & Kranz, J. (2013). Information security awareness: Its antecedents and mediating effects on security compliant behavior [ICIS 2013 Proceedings]. *International Conference on Information Systems (ICIS)*. Retrieved March 4, 2026, from http://aisel.aisnet.org/icis2013/proceedings/SecurityOfIS/9

Herath, T., & Rao, H. R. (2009). Encouraging information security behaviors in organizations: Role of penalties, pressures and perceived effectiveness. *Decision Support Systems*, *47*(2), 154–165. https://doi.org/10.1016/j.dss.2009.02.005

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM*, *50*(10), 94–100. https://doi.org/10.1145/1290958.1290968

Johnston, A. C., & Warkentin, M. (2010). Fear appeals and information security behaviors: An empirical study. *MIS Quarterly*, *34*(3), 549–566. https://doi.org/10.2307/25750691

Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., & Hong, J. (2010). Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology*, *10*(2), 7:1–7:31. https://doi.org/10.1145/1754393.1754396

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Li, M., Ahmadiadli, Y., & Zhang, X.-P. (2025). A survey on speech deepfake detection. *ACM Computing Surveys*, *57*(7), 1–38. https://doi.org/10.1145/3714458

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum Associates.

Puhakainen, P., & Siponen, M. (2010). Improving employees' compliance through information systems security training: An action research study. *MIS Quarterly*, *34*(4), 757–778. https://doi.org/10.2307/25750704

Rogers, R. W. (1983). Cognitive and physiological processes in fear appeals and attitude change: A revised theory of protection motivation. In J. T. Cacioppo & R. E. Petty (Eds.), *Social psychophysiology: A sourcebook* (pp. 153–176). Guilford Press.

Sharevski, F., & Jachim, P. (2022). "alexa, what's a phishing email?": Training users to spot phishing emails using a voice assistant. *EURASIP Journal on Information Security*, (1). https://doi.org/10.1186/s13635-022-00133-w

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L. F., & Downs, J. (2010). Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI '10)*, 373–382. https://doi.org/10.1145/1753326.1753383

Siponen, M., & Vance, A. (2010). Neutralization: New insights into the problem of employee information systems security policy violations. *MIS Quarterly*, *34*(3), 487–502. https://doi.org/10.2307/25750688

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Verizon. (2024). *2024 Data Breach Investigations Report (DBIR)*. Verizon.

Workman, M. (2007). Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security. *Journal of the American Society for Information Science and Technology*, *59*(4), 662–674. https://doi.org/10.1002/asi.20779

Yin, D., Mullarkey, M. T., de Vreede, G.-J., & Limayem, M. (2024). Timing of feedback after phishing simulations: Evidence from a randomized field experiment [ICIS 2024 Proceedings]. *International Conference on Information Systems (ICIS)*.