Comparing and Fusing Different Sensor Modalities for Relay Attack Resistance in Zero-Interaction Authentication

Hien Thi Thu Truong*, Xiang Gao*, Babins Shrestha†, Nitesh Saxena†, N.Asokan‡ and Petteri Nurmi*
*University of Helsinki, Finland, Email: {htruong,xzgao,ptnurmi}@cs.helsinki.fi
†University of Alabama at Birmingham, USA, Email: {babins,saxena}@uab.edu
‡University of Helsinki and Aalto University, Finland, Email: asokan@acm.org

***Abstract*—Zero-Interaction Authentication (ZIA) refers to approaches that authenticate a user to a verifier (terminal) without any user interaction. Currently deployed ZIA solutions are predominantly based on the terminal detecting the proximity of the user's personal device, or a security token, by running an authentication protocol over a short-range wireless communication channel. Unfortunately, this simple approach is highly vulnerable to low-cost and practical *relay attacks* which completely offset the usability benefits of ZIA. The use of contextual information, gathered via on-board sensors, to detect the co-presence of the user and the verifier is a recently proposed mechanism to resist relay attacks.**

In this paper, we systematically investigate the performance of different sensor modalities for co-presence detection with respect to a standard Dolev-Yao adversary. *First*, using a common data collection framework run in realistic everyday settings, we compare the performance of four commonly available sensor modalities (WiFi, Bluetooth, GPS, and Audio) in resisting ZIA relay attacks, and find that WiFi is better than the rest. *Second*, we show that, compared to any single modality, fusing *multiple modalities* improves resilience against ZIA relay attacks while retaining a high level of usability. *Third*, we motivate the need for a stronger adversarial model to characterize an attacker who can compromise the integrity of context sensing itself. We show that in the presence of such a powerful attacker, each individual sensor modality offers very low security. Positively, the use of multiple sensor modalities improves security against such an attacker *if* the attacker cannot compromise multiple modalities simultaneously.

I. Introduction

In proximity-based "zero interaction authentication" (ZIA) [2] systems, a verifier device authenticates the presence of a prover device in physical proximity of the verifier while requiring *no additional interaction* by the user of the prover device. The zero interaction requirement is intended to improve usability of access control systems. For example, BlueProximity[1] allows a user to unlock the idle screen lock in her computer merely by physically approaching the computer while in possession of a mobile phone, previously paired with the computer, without having to perform any other action, such as typing in a password. Motivated by these usability considerations, there are many examples of ZIA systems, such as "Passive keyless entry and start" systems like "Keyless-Go"[2] PhoneAuth [3], and access control systems based on wearable devices [22].

Under the standard Dolev-Yao adversary model [5], an attacker is assumed to have complete control over the communication channel. In such a model, naïve ZIA schemes are vulnerable to *relay attacks* where a pair of colluding attackers relays messages between a legitimate user and verifier, thereby fooling the verifier into incorrectly concluding that the user is in close proximity. Relay attacks have been demonstrated to be practical for various short range wireless communication technologies like Bluetooth [14], [12], RFID [7] and NFC [8], making this vulnerability a serious threat.

The commonly proposed defense against such relay attacks, while preserving zero-interaction, is to use *distance bounding* techniques [1]. Distance bounding assumes that the prover and verifier share a security association. The prover is required to respond to a series of rapid-fire challenges from the verifier, which can then calculate a lower bound for the distance to the prover by measuring the elapsed time between sending a challenge and receiving a correct response. Distance bounding needs to be implemented at the lowest possible layer in the communication stack because even a small error in estimating processing time at the prover side can lead to large deviations in the distance bound. Therefore implementing distance bounding on commodity devices like ordinary smartphones might be a challenge.

An alternative approach is to leverage the fact that two co-present devices will "see" (almost) the same ambient environment. Modern computing devices are equipped with many "sensors" like microphones, wireless networking interfaces, global positioning system (GPS) receivers and so on. A device can extract information from such a sensor that is characteristic of that context. By having two mutually trusting devices exchange and compare context information, they can determine if they are co-present or not. This approach has recently been proposed for *single* sensor modalities, including WiFi [13], [23], audio [10], [19] Bluetooth and GPS [16].

Although these prior works constitute an important step towards addressing the hard problem of resisting relay attacks using off-the-shelf hardware, they leave several important questions unexplored, which we address in this paper. *First*, we compare the performance of different sensor modalities in resisting relay attacks against ZIA based on contextual co-presence. Although standalone evaluations of different modalities individually have been reported in prior work, they cannot be used for a fair comparison given that the data assessing each modality was collected in disparate settings. *Second*, we investigate whether the combination ("fusion") of multiple sensor modalities will perform better than using individual modalities in isolation. Prior work did not address this question. *Third*, we explore the question of finding the appropriate adversary model for ZIA based on contextual co-presence. While the Dolev-

[1]http://sourceforge.net/projects/blueproximity/
[2]http://techcenter.mercedes-benz.com/_en/keylessgo/detail.html

Yao model is sufficient for relay attacks in general, the use of contextual co-presence raises the possibility of an attacker who can subvert the integrity of context sensing (by faking signals in the context). Previous works have mainly considered resistance against false authentications in benign settings [10] or with respect to specific attack scenarios [23].

Contributions: This paper makes the following contributions:

1) We present the first "fair" comparison of four sensor modalities commonly available on commodity smartphones – audio, WiFi, Bluetooth and GPS – under the *same settings*. We show that the use of WiFi for contextual co-presence outperforms the other modalities in resisting relay attacks (Section V-C). Our analysis is based on a dataset collected from multiple users and devices, in a combination of predefined scenarios and everyday situations, using a common data collection framework we developed (Section III). We make the dataset and framework freely available for research purposes[3].
2) We demonstrate that fusing multiple modalities is effective: it can improve security, while maintaining a very similar level of usability as proximity-based ZIA mechanisms (Section V-C).
3) Using a simple model for adversaries who can compromise the integrity of context sensing, we show that individual modalities provide low security against such adversaries. Fusion can improve security *if* the adversary cannot compromise multiple sensor modalities simultaneously. Our results call for extensions of the Dolev-Yao model that incorporate integrity of context sensing as part of the model (Section VI).

II. Background

ZIA: Figure 1 shows the system model for ZIA based on contextual co-presence. A ZIA scheme involves a user U who intends to authenticate to a verifier terminal T (e.g., a PC, car or gate) using a device D (e.g., a phone or smart key). U does not explicitly take part in the authentication process other than by approaching T while carrying D. ZIA is triggered by the devices sensing each other over a short-range wireless communication channel like Bluetooth. T will authenticate U by running a standard challenge-response based entity authentication protocol with D over the proximity communication channel. D and T pre-share a key K, which allows D to authenticate to T in the entity authentication protocol.

Standard Adversary Model: The goal of the adversary A against the ZIA protocol is to fool T into concluding that U is nearby and thus needs access to T even when U is far away (and not intending to authenticate). A possesses standard Dolev-Yao capabilities [5]: it has complete control of the communication channel over which the authentication protocol between T and D is run but does not have physical possession of D nor is able to compromise either D or T. However, we allow A to be in close physical proximity of D as well as T, even when D and T are far apart and U does not intend for D to authenticate to T. A could take the form of a "ghost-and-leech" [12] duo (A_d, A_t) such that A_d is physically close to D and A_t is physically close to T, and A_d and A_t communicate over a high-bandwidth connection. Such an adversary pair can completely compromise the security of ordinary ZIA schemes by simply initiating a protocol session between D and T, relaying messages (e.g., the challenge and response) between them, leading T to conclude that D is in proximity[4].

[3]http://se-sy.org/projects/coco

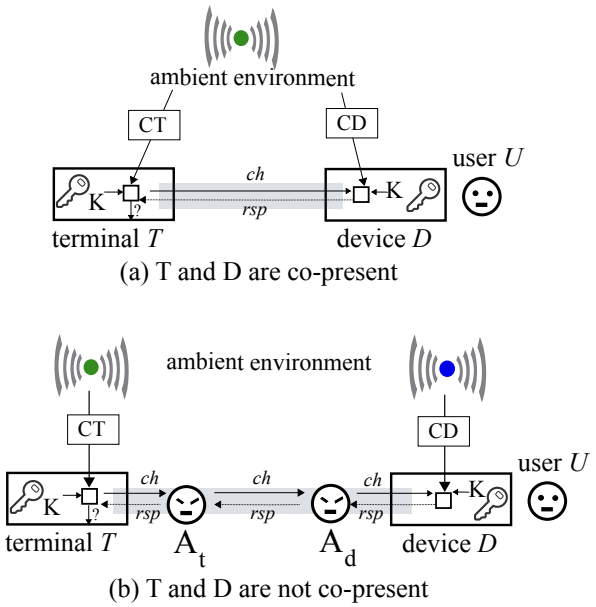

Fig. 1: System Model for ZIA with Contextual Co-presence.

ZIA Enhanced with Contextual Co-Presence: The contextual co-presence approach to ZIA aims to prevent such a relay attack as follows. When D sends a ZIA trigger to T it responds with a challenge ch. D and T then initiate context sensing for a fixed duration t. D appends ch to the sensed context information CD and computes an authenticated encryption of the result using key K to create the response rsp, which is sent to T. In the meantime, T finishes sensing its own context CT and compares it with CD extracted from rsp. T can conclude that D is in proximity if CT and CD are sufficiently similar. Note that context sensing is not run continually, but only when an authentication request takes place, implying a minimal energy overhead from the inclusion of sensing. When multiple (n) sensor modalities are used, CD and CT are vectors of the form $CD = CD_1, CD_2, ..., CD_n$, $CT = CT_1, CT_2, ..., CT_n$. T compares each CT_i with received CD_i in making the co-presence decision. In such a ZIA scheme enhanced with the contextual defense, A still can not manipulate the authentication protocol between D and T. However, as we will later explore in Section VI, A may undermine the integrity of context sensing.

III. Data Collector

We developed a data collection framework as an application installed on user devices. Our goal in developing this application is to have an easy-to-use, non-intrusive tool that

[4]We have developed a proof-of-concept to demonstrate the feasibility of such a relay attack against Blueproximity (a practical ZIA instantiation) using off-the-shelf hardware/software. Due to lack of space, the details of this implementation are not reported here.

allows potentially a large set of users to collect co-presence ground truth data. We also wanted a tool that can be easily re-purposed to conduct real-world and controlled experiments. Existing context sensing frameworks such as SensorDrone[5] are not suitable because they are designed for data collection on individual devices, making it cumbersome to collect co-presence data from multiple devices. Concretely we aimed for the following characteristics:

- A framework with a plug-in mechanism that allows later addition of new sensor modalities;
- The possibility for a user to indicate whether two devices are co-present or not by providing input on only one of them.
- A balance between collecting ample data without imposing excessive battery consumption while still letting the user to temporarily disable data collection.

A. Design and Usage

Figure 2 depicts the architecture of the data collector. It consists of the back-end synchronization server, the pair of clients, and the communication between server and clients.

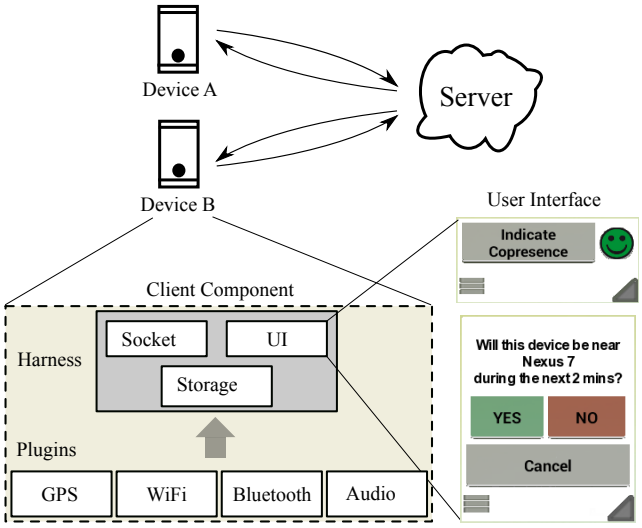

Fig. 2: Data collector architecture.

Server facilitates the "binding" of two devices of the same user. It provides a communication channel between a pair of bound devices for forwarding control messages to synchronize data collection. It also stores the collected data samples.

Client records and uploads sensor data, and provides the UI via which the user indicates co-presence ground truth. The client software framework consists of a harness with common functionality (communication with Server, UI, etc.) and a plugin interface for integrating sensing modules for different sensor modalities. Clients use the communication channel via Server to synchronize sensing.

Usage: A data collection user begins by binding two devices with the help of Server. Once the devices are bound, they maintain an open connection to Server. A user can provide co-presence ground truth via the UI shown in Figure 2. It can be launched either manually (at any time) by the user activating the "Indicate co-presence" button or periodically (once every 30 minutes by default; the frequency is configurable by the user). The UI is launchable only if the peer device (and hence Server) are reachable. When the user indicates ground truth in one device, sensing is initiated on both devices. The resulting sensor data (collection of samples) and the ground truth are sent to Server for persistent storage with a unique sequence number. The data is then deleted on devices as well.

B. Sensor Data

We currently have plugins for the following sensor modalities commonly available on modern smartphones:

GPS Raw Data: We record the identifiers of visible GPS satellites and the "signal strength" for each of them in the form of signal-to-noise ratio (SNR). The identifier is the "pseudo-random noise code" (PRN) which is an integer ($1 \ldots 32$). Each data sample consists of multiple records taken at the rate of one every second over a two minute period. Each record contains the set of identifiers and SNRs observed at that instant. The SNR ranges from 0 to 100. Where a location fix is available, we record longitude, latitude, altitude and accuracy.

WiFi: For each visible WiFi access point (AP), we record the list of link-layer addresses (BSSID) and the associated received signal strength indicators (RSSI), supported capabilities and the frequency of the WiFi channel advertised by that AP. RSSI ranges from -100 to -20 dBm.

Bluetooth: For each visible Bluetooth device, we record the identifier (BDADDR) and received signal strength indicator (RSSI). RSSI ranges from -100 to -20 dBm.

Audio: Ambient audio is recorded in standard PCM format (wav file) without compression. Each PCM wave is sampled in 44100Hz with 16-bit encoding. Because raw audio is sensitive, by default, we do not store raw audio on Server. Instead, we extract certain features (as described in Section IV). Users however have the option of changing this default to let their client(s) upload raw audio to Server.

IV. Experimental Setup

We carried out an extensive empirical investigation of the effectiveness of different sensor modalities, both individually and in different combinations, in strengthening ZIA solutions against relay attacks. In this section, we describe the data that we collected, and the features used in our analyses.

A. Data Collection

Everyday Dataset: Using our data collector framework, five testers collected data for 15 days in mid 2013. Hardware variations across devices are well-known to cause significant changes in sensor measurements. To ensure robustness of results with respect to device variations, we collected data using tablets and phones from different manufacturers and with different models. We gave no specific instructions to the testers about what scenarios or locations in which they should collect data. Consequently, the resulting dataset is *uncontrolled*, consisting of data collected in various everyday settings and locations (e.g., university campus, labs, libraries, cafeteria,

[5]http://www.sensorcon.com/sensordrone/

home, streets), Data collection was done in two different cities: Birmingham, Alabama, USA and Helsinki, Finland. This dataset contains 2303 samples, of which 1140 samples (49.5%) are from co-present devices and 1163 (50.5%) from non co-present devices. Each sample contains data from sensor modalities available at the time on the respective devices (2117 with audio, 1600 with Bluetooth, 782 with GPS and 2269 with WiFi). For each sample, we scan all available sensors simultaneously: 2 minutes for GPS scanning, 10 scans for WiFi (about 30 seconds), 10 seconds for recording ambient audio, and 10 scans for Bluetooth (up to 12 seconds for each scan).

Ethical considerations: The data collection was performed by persons giving explicit consent. The data is released for research purposes in anonymized form. The anonymization was carried out by (a) replacing each device identifier with its SHA-1 value, (b) replacing the pair of GPS co-ordinates from the two devices in a sample by the great-circle distance between the pair, and (c) raw audio data is replaced by relevant features as discussed below.

B. Features

We investigated various possible features that can be extracted from the data in different sensor modalities, finally settling on the most promising features as discussed below.

1) Features for Bluetooth, WiFi, GPS: For all sensors involving radio-frequency (RF) emissions, we studied a common set of features. Let a sample from an RF sensor modality be of the form (m, s) where m is an identifier of a sensed device and s is the associated signal strength. Let S_a and S_b denote the set of records sensed by a pair of bound devices A and B, and let n_a and n_b denote the number of different beacons (i.e., WiFi access points, satellites or Bluetooth devices) observed by devices a and b. We define the following sets:

$S_a = \{(m_i^{(a)}, s_i^{(a)}) \mid i \in \mathbb{Z}_{n_a-1}\}$.
$S_b = \{(m_i^{(b)}, s_i^{(b)}) \mid i \in \mathbb{Z}_{n_b-1}\}$.
$S_a^{(m)} = \{m \ \forall (m,s) \in S_a\}$, $S_b^{(m)} = \{m \ \forall (m,s) \in S_b\}$.
$S_\cap = \{(m, s^{(a)}, s^{(b)}) \ \forall m | (m, s^{(a)}) \in S_a, (m, s^{(b)}) \in S_b\}$.
$S_\cup = S_\cap \cup \{(m, s^{(a)}, \theta) \ \forall m | (m, s^{(a)}) \in S_a, m \notin S_b^{(m)}\}$
$\cup \{(m, \theta, s^{(b)}) \ \forall m | (m, s^{(b)}) \in S_b, m \notin S_a^{(m)}\}$,
θ is modality-specific (see below).
$S_\cap^{(m)} = \{m \ \forall m | (m, s^{(a)}, s^{(b)}) \in S_\cap\}$.
$S_\cup^{(m)} = \{m \ \forall m | (m, s^{(a)}, s^{(b)}) \in S_\cup\}$.
$L_a^{(s)} = \{s^a | (m, s^{(a)}, s^{(b)}) \in S_\cap\}$.
$L_b^{(s)} = \{s^b | (m, s^{(a)}, s^{(b)}) \in S_\cap\}$.

$S_\cap$ consists of devices seen by both A and B; $S_\cup$ represents all devices seen by A or B with θ filled in as the "signal strength" for devices that are *not* seen by either device. The features of interest are as follows (the first five have been used in prior work such as [6], [23], [13]).

1) Jaccard distance: $1 - \frac{|S_\cap^{(m)}|}{|S_\cup^{(m)}|}$
2) Mean of Hamming distance: $\frac{\sum_{k=1}^{|S_\cup|} |s_k^{(a)} - s_k^{(b)}|}{|S_\cup|}$
3) Euclidean distance: $\sqrt{\sum_{k=1}^{|S_\cup|} (s_k^{(a)} - s_k^{(b)})^2}$
4) Mean exponential of difference: $\frac{\sum_{k=1}^{|S_\cup|} e^{|s_k^{(a)} - s_k^{(b)}|}}{|S_\cup|}$
5) Sum of squared of ranks: $\sum_{k=1}^{|S_\cap|} (r_k^{(a)} - r_k^{(b)})^2$ where, $r_k^{(a)}$ (respectively $r_k^{(b)}$) is the rank of $s_k^{(a)}$ ($s_k^{(b)}$) in the set L_a (L_b) sorted in ascending order.
6) Subset count: $\sum_{i=1}^{T} f_i$. Here T is the scanning time (seconds)
$f_i = 1$ if $S_{a_i}^{(m)} \neq \emptyset$, $S_{b_i}^{(m)} \neq \emptyset$,
$(S_{a_i}^{(m)} \subseteq S_{b_i}^{(m)}$ *or* $S_{a_i}^{(m)} \supseteq S_{b_i}^{(m)})$
$f_i = 0$ otherwise. S_{a_i}, S_{b_i} are the set of records by A and B respectively at the i^{th} second

WiFi: Features 1-5 are used. Since we do multiple scans in each sample, we use the mean value of RSSI for a BSSID (m) from all of the scans as the signal strength (s) value. θ is -100.

Bluetooth: Features 1,3 are used with BDADDR as identifier (m) and average RSSI as signal strength (s). θ is -100.

GPS: All features are used with PRN as identifier (m) and mean SNR as signal strength (s). θ is 0.

Note that feature 6 is used only for GPS. This is because the set of satellites visible to a device varies greatly depending on the sensitivity of GPS hardware. Thus, one device may see a subset of the satellites seen by the another co-present device. In such cases, metrics like Jaccard distance perform poorly whereas the subset count could perform better. When GPS co-ordinates are available for A and B in a sample, we also use the orthodromic distance [9] as a feature.

2) Features for Audio: We consider two features proposed by Halevi et al. [10], which were found to be the most robust among all algorithms tested: Schurmann and Sigg [19], SoundSense [15], and Shazam audio fingerprinting [24]. The other features either required careful synchronization between the two audio samples or were highly sensitive to variations in the microphone characteristics of the devices. The two features that we consider are defined as follows:

- Max cross correlation:
  $M_{corr}(a,b) = Max(cross\ correlation(X_a, X_b))$
- Time frequency distance:
  $D(a,b) = \sqrt{(D_{c,time}(a,b))^2 + (D_{d,freq}(a,b))^2}$
  where, $D_{c,time}(a,b) = 1 - M_{corr}$, $D_{d,freq}(a,b) = ||FFT(X_a) - FFT(X_b)||$ is the Euclidean norm of the distance.

Here X_a and X_b denote the raw (16-bit PCM) audio signals recorded by A and B and FFT(X_a), FFT(X_a) denotes the Fast Fourier Transforms of the corresponding signals.

V. Analysis and Results

A. Analysis Methodology and Metrics

We treat contextual co-presence detection as a classification task. All our experiments have been performed using ten-fold cross-validation and Multiboost [25] as the classification algorithm. In all experiments, decisions trees (J48 Graft) are used as the weak learners. From each experiment, we record the 2x2 confusion matrix, containing the number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN).

The classification performance of contextual co-presence detection directly influences both the security and usability

of the underlying ZIA mechanism. In particular, the security of the system is determined by FP rate as it indicates the probability of T concluding that D (and hence U) is co-present erroneously. Usability, on the other hand, is represented by the FN rate as it determines the probability of T not being able to authenticate U even though U is co-present. In addition to evaluating the FP and FN rates, we consider two metrics for the overall classification performance: F-measure and the Matthews' correlation coefficient (MCC).

The F-measure (Fm) uses precision ($\frac{TP}{TP+FP}$) and recall ($\frac{TP}{TP+FN}$) for each class. $Fm_i = 2 \cdot \frac{precision_i \cdot recall_i}{precision_i + recall_i}$, $Fm = \frac{\sum_{i=1}^{c} w_i \cdot Fm_i}{\sum_{i=1}^{c} w_i}$, where i is the class index, $w_i = n_i/N$ with n_i being the number of samples of the i^{th} class and N being the total number of samples, c is the number of classes.

MCC is an approximate statistical measure for deciding whether the prediction is significantly more correlated with the data than a random guess. MCC is related to chi-square statistic for a 2x2 contingency table: $|MCC| = \sqrt{\frac{\chi^2}{n}}$. It can be calculated directly from the confusion matrix as: $|MCC| = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$. It takes values between -1 and +1, with +1 representing perfect prediction, and -1 total disagreement between prediction and ground truth while 0 represents no better than random guess.

B. Effect of Time Budget

Although we collected data for two minutes in each sample, the realistic time budget for ZIA is much smaller (typically 5-15 seconds) due to usability reasons (e.g., being able to unlock a terminal or a door quickly). To see the effect of sampling time on the performance of classification, we consider the performance with different time budgets. For a time budget of n seconds, we only consider the sensor data recorded by the device in a sample within the first n seconds. Table I shows the results for the uncontrolled dataset for different time budgets. Although the overall performance is reasonable with a 5-second limit (FN=8.95%; FP=7.14%, Fm=0.921, MCC=0.841), data was often missing different sensor modalities: among 2303 instances, 80% is without GPS data, 37% without WiFi data, 40% without Bluetooth and 8% without Audio. With a 10-second budget the performance is significantly better than with a 5-second budget as more data is captured by sensors, but it flattens out thereafter. We will use a time budget of 10-second for all analyses in this paper from now on.

TABLE I: Overall performance vs. time budget

Time Budget (s)	5	8	10	12	15
%FN	8.95	2.19	1.67	1.40	1.49
%FP	7.14	2.67	1.98	2.15	2.15
MCC	0.841	0.951	0.966	0.964	0.964
Fm	0.921	0.976	0.983	0.982	0.982

C. Performance of Single and Multiple Modalities

Next we focus on investigating the effectiveness of single modality co-presence detection, and on assessing the potential improvements provided by the fusion of multiple context modalities. The results of this investigation are shown in Table II. For a given sensor modality, we only consider samples that have data from that sensor. To facilitate comparison, we study the fusion of modalities for the same set of samples in each case. Among individual modalities (column 2) WiFi performs best (Fm = 0.989, MCC = 0.978) and GPS worst (Fm = 0.776, MCC = 0.550). Bluetooth and Audio exhibit similar performance with the former (Fm = 0.885, MCC = 0.773) slightly better than the later (Fm = 0.857, MCC = 0.715).

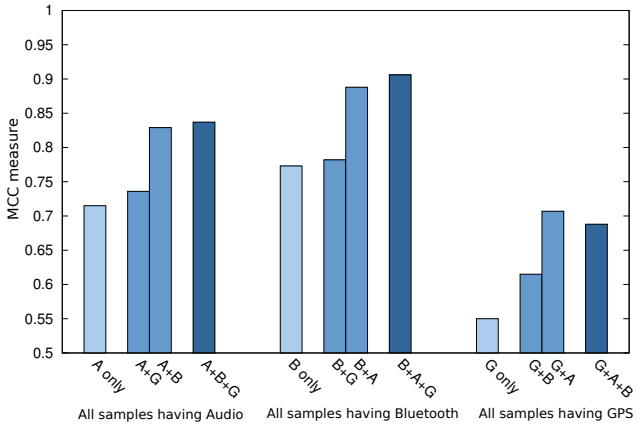

Fig. 3: MCC comparison for three modalities Audio (A) - Bluetooth (B) - GPS (G).

The results for Bluetooth, audio and GPS clearly demonstrate that relying solely on any single one of these modalities is not sufficient for satisfying the usability and security requirements of ZIA. Moreover, from Figure 3 we can observe that the performance of these modalities improves when they are fused with another modality.

To see if the performance of an individual modality varied greatly depending on the sampled values, we analyzed the performance separately for samples with values in different ranges ("bands"). Tables III shows the results. A band consists of those samples where the records from both devices fall in the range corresponding to that band (e.g., there were 551 samples in which both devices saw only one other Bluetooth device). Several conclusions can be drawn from the table. First, the performance is significantly worse in some bands (e.g., "< 2" for Bluetooth). In a practical ZIA implementation, samples falling in such bands can be filtered out when evaluating contextual co-presence. Second, the performance of GPS naturally improves when more satellites are visible – but within our 10s time budget, GPS performs poorly because the vast majority of the samples contain only one visible satellite.

D. Small-scale Controlled Dataset

To assess the robustness of the results with respect to common sources of noise in sensor measurements, such as variations in device placement (pocket vs bag) and variations in the characteristics of the ambient environment (noisy vs quiet), we supplemented the everyday dataset with a limited dataset collected from predefined settings. This *controlled dataset* was collected in order to determine if there was any potential systematic bias as to how our testers collected the data in

TABLE II: Individual modalities vs Fusion of modalities; (A) Audio, (B) Bluetooth, (G) GPS, (W) WiFi

	All samples containing Audio (sample size = 2117)							
	A only	A+B	A+G	A+W	A+B+G	A+B+W	A+G+W	A+B+G+W
FN(%)	19.9	12.49	20.41	1.52	12.59	1.52	1.73	1.62
FP(%)	9.28	5.21	7.07	1.59	4.33	1.77	1.59	1.77
MCC	0.715	0.829	0.736	0.969	0.837	0.967	0.967	0.966
Fm	0.857	0.914	0.866	0.984	0.918	0.983	0.983	0.983
	All samples containing Bluetooth (sample size = 1600)							
	B only	B+A	B+G	B+W	B+A+G	B+A+W	B+G+W	B+A+G+W
FN(%)	15.54	7.64	18.25	0.74	6.78	0.49	0.49	0.37
FP(%)	7.35	3.55	4.18	1.27	2.66	1.01	1.14	1.01
MCC	0.773	0.888	0.782	0.980	0.906	0.985	0.984	0.986
Fm	0.885	0.944	0.886	0.990	0.952	0.992	0.992	0.993
	All samples containing GPS (sample size = 782)							
	G only	G+A	G+B	G+W	G+A+B	G+A+W	G+B+W	G+A+B+W
FN(%)	23.6	14.89	25.28	1.97	18.54	1.69	2.53	1.97
FP(%)	21.36	14.32	13.85	3.52	12.91	3.99	3.52	3.76
MCC	0.55	0.707	0.615	0.944	0.688	0.941	0.938	0.941
Fm	0.776	0.854	0.808	0.972	0.845	0.971	0.969	0.971
	All samples containing WiFi (sample size = 2269)							
	W only	W+A	W+B	W+G	W+A+B	W+A+G	W+B+G	W+A+B+G
FN(%)	0.36	0.27	0.45	0.45	0.18	0.18	0.27	0.18
FP(%)	1.83	1.83	1.83	1.83	1.83	1.83	1.92	1.83
MCC	0.978	0.979	0.977	0.977	0.980	0.980	0.978	0.980
Fm	0.989	0.989	0.989	0.989	0.990	0.990	0.989	0.990

TABLE III: Performance for different modality bands, #IDs = number of device IDs seen and N: number of samples.

Audio						**Bluetooth**						**GPS**					
RMS[a]	N	%FN	%FP	MCC	Fm	#IDs	N	%FN	%FP	MCC	Fm	#IDs	N	%FN	%FP	MCC	Fm
≤ 500	919	11.14	4.09	0.855	0.933	< 2	551	38.12	4.01	0.642	0.826	≥ 1	757	27.12	21.84	0.511	0.757
rest	1198	23.89	16.51	0.594	0.795	rest	1049	10.34	0.72	0.883	0.941	≥ 5	314	19.85	15.30	0.647	0.828
												≥ 10	39	6.25	4.35	0.894	0.949

[a]RMS refers to audio signal's root mean square level.

TABLE IV: Controlled setting (sample sizes in brackets)

	Single modality				All modalities			
	%FN	%FP	MCC	Fm	%FN	%FP	MCC	Fm
Audio(74)	18.18	16.67	0.644	0.825	4.55	3.33	0.917	0.960
Bluetooth(94)	4.44	2.04	0.936	0.968	4.44	0	0.958	0.979
GPS(37)	18.18	26.67	0.552	0.784	4.55	0	0.946	0.973
WiFi(88)	4.44	2.33	0.932	0.966	4.44	2.33	0.932	0.966

the uncontrolled dataset. The controlled dataset, contains 94 samples (44 from co-present devices and 50 from non co-present devices) which were collected by two users. All were taken in noisy environments (in crowded areas and noisy streets). In each sample, one device was within an enclosure (pocket or backpack) while the other was exposed (e.g., in the user's hands).

Table IV shows the performance of the classification in controlled dataset for different sensor modalities (single, and all together). The results do not indicate any clear systematic difference between the two datasets in terms of the classification performance, suggesting that generally the evaluated context sensing mechanisms are robust across variations in environmental characteristics and in device placements.

The performance of WiFi exceeds other modalities, providing near perfect results for the uncontrolled dataset. One possible reason is that in most of the samples in this dataset, the two devices are either very close or very far from other. This is reasonable since our focus is on preventing relay attacks where the common case is for the attacker to attempt relaying when the two legitimate devices are far apart. However, it is reasonable to ask whether the FP rate of WiFi will remain as high when the non co-present devices are much closer to each other. To investigate this issue, we conducted another small-scale controlled experiment where we collected data from four devices. Pairs of devices were placed in two offices that were approximately 15 meters apart, and 100 samples containing all sensor modalities were recorded for a duration of two hours, in which 50% is from the co-present pair and 50% from the non co-present pair. The results show that (a) WiFi performance degrades slightly with FP% rising from 1.83% to 7.14% and (b) the fusion of multiple sensor modalities does improve the FP rate (to 4.76%) compared to using WiFi alone.

E. Effect of Personalized Training Model

So far, we used data from all users to create a common user-independent model. A natural question is whether a user-specific model would perform better. To see this, we separated

TABLE V: Analysis of personalized model for individual users, blanks indicate insufficient data.

Modalities	User1					User 2					User 3				
	N	%FN	%FP	MCC	Fm	N	%FN	%FP	MCC	Fm	N	%FN	%FP	MCC	Fm
Personalized Model: Trained and tested with personal data															
Audio	494	0.76	0.85	0.984	0.992	228	21.55	18.58	0.599	0.799	209	6.88	18.37	0.737	0.905
Bluetooth	435	0.77	0	0.99	0.995	198	3	4.08	0.929	0.965	133	-	-	-	-
GPS	52	31.58	15.15	0.539	0.787	20	-	-	-	-	59	-	-	-	-
WiFi	496	0.76	0	0.992	0.996	229	0.86	0.88	0.983	0.991	219	1.25	1.67	0.966	0.986
All	496	0.76	0	0.992	0.996	229	0.86	0.88	0.983	0.991	220	0.63	3.33	0.966	0.986
Common Model: Trained with all data and tested with personal data															
All	496	0	0	1	1	229	0	2.65	0.974	0.987	220	0	3.33	0.977	0.991

data by individuals and used them to train "personalized" models. Note that a personalized model is trained using data from only two devices, whereas the common model was computed using data from multiple pairs of devices. Accordingly, the user-specific evaluation also assesses the robustness of our results hardware variations. Table V summarizes the results for three users (uncontrolled data set) with the most data. Since a personalized model is more cumbersome (it would require each user to train the model), it has to be significantly better than the common model to justify its use.

F. Summary

We showed that WiFi is the most effective sensor modality for resisting relay attacks against ZIA schemes based on contextual co-presence detection. We also showed that for all combinations of sensor modalities, fusing all available modalities will improve security (low false positives) of such ZIA schemes while retaining the high level of usability (low false negatives) characteristic of ZIA.

VI. Adversarial Analysis

So far, we assumed the Dolev-Yao [5] adversary model. However, the Dolev-Yao model is intended for analyzing traditional communication protocols. Attacks against the integrity of context sensing are known. For example, Tippenhauer et al [21] showed how to defeat WiFi-based positioning systems with inexpensive equipment. Our proof-of-concept attack against BlueProximity was based on changing the Bluetooth device address on the Bluetooth controller on a PC. It is not difficult to imagine an attacker capable of generating same dominant sound near a pair of devices in two different locations. All this demonstrate the need for a stronger adversary model that would cover the capability for interfering with context sensing.

Prior work on contextual co-presence largely limited their security analysis to benign failures only [10]. The occasional exceptions involved testing resistance against a few types of attacks interfering with context sensing [23]. In contrast, we argue that there is a need for a precise but realistic formulation of a contextual adversary without having to spell out specific attacks. Once such an adversary model is defined, different contextual co-presence schemes can be analyzed with respect to such an adversary.

A Model for a Context Adversary: Faking contextual information may require conspicuous equipment (like fake access points) or actions (like playing loud music). Observe that D is usually carried by the human user U whereas T may be unattended. We therefore postulate that A_t, the attacker near T, can more easily interfere with the context sensing of T undetected than can A_d with D. Furthermore, we assume that it is infeasible for an attacker to *suppress* existing context signals. Therefore, one way to characterize the context attacker is as follows:

- A_d can perfectly measure the context information that D would sense,
- A_t can fool T into sensing any context information it chooses; Specifically A_t can receive context information from A_d, reproduce it perfectly near T; and
- A_t (A_d) cannot suppress any other ambient context information from being sensed by T (D).

While this is still a very powerful attacker, analyzing our features for classification with respect to such an attacker may give some insights into the relative security of different sensor modalities.

Analysis: For RF-based sensors, the context adversary as defined above can be modelled by replacing S_b with $S_a \cup \{(m,s)\ \forall (m,s) \in S_b, m \notin S_a^{(m)}\}$. For audio, since raw audio data is additive, the adversary can be modelled replacing X_b by $X_a + X_b$. To estimate the effect of such an adversary, we took the following approach. We used our uncontrolled dataset with ten-fold validation. Training is done using the nine folds of the dataset as before. But the test data was transformed as described above to model the effect of the context adversary.

The results for WiFi, Bluetooth and audio are shown in Table VI. (We did not include GPS in this analysis because GPS performed poorly to begin with and spoofing GPS is likely to be harder than the other modalities. Nevertheless, we expect the adversary model to hold for GPS as well and is likely to yield similar results.) The first and the third row show the performance of individual and multiple sensor modalities in the presence of the context attacker. All individual modalities are insecure with respect to such an attacker. *If* we can assume that the attacker is capable of compromising only one sensor modality at a time, the use of multiple modalities restores security in the case of audio and Bluetooth, thanks to the effect of WiFi. In the case of WiFi itself, the fusion of the other modalities results in only a modest increase in security. The second row of Table VI shows the difference in false positive rate with respect to the same modalities in the absence of the attacker. False positive rate of Bluetooth and Audio

TABLE VI: Performance in adversarial setting

Modalities	Audio				Bluetooth				WiFi			
	FN(%)	FP(%)	MCC	Fm	FN(%)	FP(%)	MCC	Fm	FN(%)	FP(%)	MCC	Fm
Single modality	16.14	**100**	-0.298	0	15.17	**99.11**	-0.268	0.281	0.45	**75.17**	0.365	0.556
Difference from Table. II	-16.77	+91.23	-0.905	-0.857	-0.37	+91.76	-1.041	-0.604	+0.09	+73.34	-0.613	-0.433
Fused of multi-modalities	1.75	**3.01**	0.952	0.976	0.37	**1.22**	0.984	0.992	0.45	**65.8**	0.444	0.625

has comparable increases (+91.76% and +91.23% respectively) while the increase in WiFi is a more modest 73.34%, implying that although the powerful context attacker is very successful across the board, WiFi performs somewhat better than the other modalities against such an attacker.

VII. Related Work

Relay Attack Resilience: Distance bounding [1] techniques are commonly used to avoid relay attacks [11], [18]. As we saw in Section I, it may not be realistic for commodity devices.

Halevi et al. [10] developed techniques using ambient audio for co-presence detection. Their experiments were done using identical device models rather than using different device models as in our work. Our results show that in such divergent scenarios, their techniques perform less well than in [10]. Nevertheless, their techniques are the best among different audio techniques we tested for ambient audio.

Narayanan et al. [17] studied the use of various modalities for private proximity detection and concluded that on WiFi broadcast packets and WiFi access point IDs are likely to perform best. Our systematic experiments confirm that WiFi access point IDs perform well. We ruled out the use of WiFi broadcast packets because they are not accessible to applications in ordinary smartphone platforms.

Krumm et al. [13] proposed "NearMe" which uses WiFi similarity features for proximity detection. They built a model using data collected in an office building environment and tested in a cafeteria environment. They conjecture that their approach generalizes well to other settings. Our analysis with the uncontrolled dataset collected from diverse environments confirms that WiFi access point and signal strength information works well for general settings.

Czeskis et al. [4] proposed "secret handshakes" to avoid ghost-and-leech attack by limiting the context where the contactless card communicates with the reader. They used only accelerometer data as contextual information.

Pairing Using Contextual Information: Secure pairing using contextual information is a harder problem in that it requires the two devices being paired to extract sufficient entropy (e.g., 128 bits) from the context to serve as a cryptographic key. In contrast, contextual co-location determination does not require secrecy for the context information. There has been significant work in secure device pairing using contextual information such as WiFi or audio. Schurmann et al. [19] presented an approach that uses binary fingerprints from ambient audio to establish a secure channel between two co-present devices. Varshavsky et al. [23] presented Amigo to authenticate co-present devices using various features extracted from the WiFi environment. All such previous work on pairing has focused on a single sensor modality. In contrast, we consider the use of multiple modalities simultaneously.

VIII. Summary and Discussion

In this paper, we addressed the issue of using different sensor modalities for co-presence detection to be used in applications that need ZIA. To the best of our knowledge, our work is the first that fairly compares performance of different modalities and shows that the use of multiple modalities can improve security of co-presence detection (relay attack resilience) without significantly degrading its usability. We believe that our insights can help improve the design of ZIA schemes.

Energy Consumption: One potential concern in adding context sensing to ZIA is the effect on battery consumption. However, this is not a serious problem for the following reasons. First, context sensing is triggered only when an authentication request happens by D coming in proximity of T, *and* T is locked. Second, a pre-requisite for context sensing is an authenticated trigger sent from D to T, which precludes denial of service attacks. Third, although a relay attacker could cause repeated triggers, this can be resisted by system design, such as introducing exponential back off after a small number of failed authentication attempts.

Limitations: Our data collection and analysis is targeted for evaluating contextual co-presence techniques for the particular use case of ZIA. As such, our results (such as the effectiveness of WiFi as a sensor modality) may not generalize to other applications of contextual co-presence detection. We let users decide what constitutes co-presence, which is a reasonable approach for evaluating ZIA but may not be so for other applications. On the positive side, our data collection framework can be easily adapted to collect data for different scenarios, such as indicating ground truth in terms of the exact distance rather than as a boolean value.

The accuracy required for co-presence detection varies from application to application. Apart from the small-scale controlled experiment in Section V-D, we did not focus on estimating the exact granularity of co-presence in terms of distance. Other work, such as NearMe [13], suggest that co-presence can be determined up to a distance of 20m.

It is possible that co-presence determination is inconclusive, for example because sufficient sensor data is not available in a given situation. In such a case, depending on the application, it may be reasonable to trade off usability for security by relaxing the "zero interaction" requirement, e.g., by asking the user U to confirm co-presence on device D (recall that D and T share a security association).

Extensions: It is natural to consider the use of other forms of sensor modalities. Indeed, in a recent paper [20], we

investigated the use of sensor modalities that represent the *physical* ambient environment. We are expanding the data collection to include more users so that we can resolve the question of whether personalized models are more effective for contextual co-presence detection than common, offline models. We are incorporating our model into Blueproximity and plan to conduct a user study to evaluate its usability. Our adversarial analysis is intended as a first step – the question of how to characterize a context adversary formally yet realistically remains open.

ACKNOWLEDGMENTS

This work was partially supported by a Google Faculty Research Award, a donation from Nokia and US NSF grant CNS-1201927. Xiang Gao and Petteri Nurmi were supported by TEKES as part of the Internet of Things and Data to Intelligence programs, respectively, of DIGILE (Finnish Strategic Centre for Science, Technology and Innovation in the field of ICT and digital business). We thank the testers who participated in the data collection. Cooper Filby contributed to the demonstration of relay attacks against Blueproximity. Tzipora Halevi, Dominik Schurmann, Ngu Nguyen and Haipan Guo helped us analyze the use of different algorithms for comparing the audio data.

REFERENCES

[1] S. Brands and D. Chaum. Distance-bounding protocols. In *Workshop on the theory and application of cryptographic techniques on Advances in cryptology*, EUROCRYPT '93, pages 344–359. Springer-Verlag New York, Inc., 1994.

[2] M. D. Corner and B. D. Noble. Zero-interaction authentication. In *Proc. 8th annual international conference on Mobile computing and networking*, MobiCom '02, pages 1–11, New York, NY, USA, 2002. ACM.

[3] A. Czeskis, M. Dietz, T. Kohno, D. Wallach, and D. Balfanz. Strengthening user authentication through opportunistic cryptographic identity assertions. In *Proc. 2012 ACM conference on Computer and communications security*, CCS '12, pages 404–414, New York, NY, USA, 2012. ACM.

[4] A. Czeskis, K. Koscher, J. R. Smith, and T. Kohno. RFIDs and Secret Handshakes: Defending Against Ghost-and-leech Attacks and Unauthorized Reads with Context-aware Communications. In *Proc. 15th ACM conference on Computer and communications security*, CCS '08, pages 479–490, New York, NY, USA, 2008. ACM.

[5] D. Dolev and A. C.-C. Yao. On the security of public key protocols. *IEEE Transactions on Information Theory*, 29(2):198–207, 1983.

[6] O. Dousse, J. Eberle, and M. Mertens. Place learning via direct WiFi fingerprint clustering. *2012 IEEE 13th International Conference on Mobile Data Management*, 0:282–287, 2012.

[7] A. Francillon, B. Danev, and S. Čapkun. Relay attacks on passive keyless entry and start systems in modern cars. In *Proc. Network and Distributed System Security Symposium (NDSS)*, 2011.

[8] L. Francis, G. Hancke, K. Mayes, and K. Markantonakis. Practical NFC Peer-to-peer Relay Attack Using Mobile Phones. In *Proc. 6th international conference on Radio frequency identification: security and privacy issues*, RFIDSec'10, pages 35–49, Berlin, Heidelberg, 2010. Springer-Verlag.

[9] W. Gellert, S. Gottwald, and M. Hellwich. *The VNR concise encyclopedia of mathematics*. Van Nostrand Reinhold New York, 2nd edition, 1989.

[10] T. Halevi, D. Ma, N. Saxena, and T. Xiang. Secure Proximity Detection for NFC Devices Based on Ambient Sensor Data. In *Proc. 17th European Symposium on Research in Computer Security (ESORICS)*, volume 7459 of *Lecture Notes in Computer Science*. Springer, 2012.

[11] G. Hancke and M. Kuhn. An RFID distance bounding protocol. In *Security and Privacy for Emerging Areas in Communications Networks, 2005. SecureComm 2005.*, pages 67–73, 2005.

[12] Z. Kfir and A. Wool. Picking virtual pockets using relay attacks on contactless smartcard. In *Proc. First International Conference on Security and Privacy for Emerging Areas in Communications Networks*, SECURECOMM '05, pages 47–58, Washington, DC, USA, 2005. IEEE Computer Society.

[13] J. Krumm and K. Hinckley. The NearMe Wireless Proximity Server. In *In Proc. Ubiquitous Computing (UbiComp)*, 2004.

[14] A. Levi, E. Cetintas, M. Aydos, C. Koc, and M. Caglayan. Relay Attacks on Bluetooth Authentication and Solutions. In *Computer and Information Sciences (ISCIS)*, 2004.

[15] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell. SoundSense: Scalable Sound Sensing for People-Centric Applications on Mobile Phones. In *Proc. 7th international conference on Mobile systems, applications, and services*, pages 165–178, 2009.

[16] D. Ma, N. Saxena, T. Xiang, and Y. Zhu. Location-Aware and Safer Cards: Enhancing RFID Security and Privacy via Location Sensing. *IEEE Trans. Dependable Secur. Comput.*, 10(2):57–69, Mar. 2013.

[17] A. Narayanan, N. Thiagarajan, M. Lakhani, M. Hamburg, and D. Boneh. Location Privacy via Private Proximity Testing. In *Proc. Network and Distributed System Security Symposium (NDSS)*, 2011.

[18] J. Reid, J. M. G. Nieto, T. Tang, and B. Senadji. Detecting Relay Attacks with Timing-based Protocols. In *Proc. 2nd ACM symposium on Information, computer and communications security*, ASIACCS '07, pages 204–213, New York, NY, USA, 2007. ACM.

[19] D. Schurmann and S. Sigg. Secure communication based on ambient audio. *IEEE Transactions on Mobile Computing*, 12(2):358–370, Feb. 2013.

[20] B. Shrestha, N. Saxena, H. T. T. Truong, and N. Asokan. Drone to the rescue: Relay-resilient authentication using ambient multi-sensing. In *Proc. Eighteenth International Conference on Financial Cryptography and Data Security 2014*, pages –, 2014. (to appear).

[21] N. O. Tippenhauer, K. B. Rasmussen, C. Pöpper, and S. Čapkun. Attacks on Public WLAN-based Positioning Systems. In *Proc. 7th international conference on Mobile systems, applications, and services*, MobiSys '09, pages 29–40, New York, NY, USA, 2009. ACM.

[22] B. Tognazzini. The apple iwatch. Blog posting in AskTOG: Interaction Design Solutions for the Real World, Feb 2013. http://asktog.com/atc/apple-iwatch/.

[23] A. Varshavsky, A. Scannell, A. LaMarca, and E. De Lara. Amigo: Proximity-based authentication of mobile devices. In *Proc. 9th International Conference on Ubiquitous Computing (UbiComp)*, pages 253–270. Springer, 2007.

[24] A. Wang. The Shazam music recognition service. *Communications of the ACM*, 49:44–48, 2006.

[25] G. I. Webb. Multiboosting: A technique for combining boosting and wagging. *Mach. Learn.*, 40(2):159–196, Aug. 2000.