Keystroke Biometrics: the User Perspective

Chee Meng Tey
DSO National Laboratories
tcheemen@dso.org.sg

Payas Gupta
New York University
payasgupta@nyu.edu

Kartik Muralidharan
Singapore Management University
kartikm.2010@smu.edu.sg

Debin Gao
Singapore Management University
dbgao@smu.edu.sg

ABSTRACT

Usability is an important aspect of security, because poor usability motivates users to find shortcuts that bypass the system. Existing studies on keystroke biometrics evaluate the usability issue in terms of the average false rejection rate (FRR). We show in this paper that such an approach underestimates the user impact in two ways. First, the FRR of keystroke biometrics changes for the worse under a range of common conditions such as background music, exercise and even game playing. In a user study involving 111 participants, the average penalties (increases) in FRR are 0.0360 and 0.0498, respectively, for two different classifiers. Second, presenting the FRR as an average obscures the fact that not everyone is suitable for keystroke biometrics deployment. For example, using a Monte Carlo simulation, we found that 30% of users would encounter an account lockout before their 50th authentication session (given a lockout policy of 3 attempts) if they are affected by external influences 50% of the time when authenticating.

Categories and Subject Descriptors

D.4.6 [**Software**]: OPERATING SYSTEMS—*Security and Protection, Authentication*; H.1.2 [**Information Systems**]: MODELS AND PRINCIPLES—*User / Machine Systems, Human factors*

Keywords

Authentication, Human Factors, Keystroke Biometrics

1. INTRODUCTION

Keystroke biometrics has a long history of research. Prior research has shown that typing patterns are unique to each individual and can be effectively used to identify and authenticate a user [1, 8, 15, 10, 9]. The advantages of keystroke biometrics include low cost (no requirement for specialised hardware) and transparency of use [10].

However, being a form of behavioral biometrics, keystroke biometrics has lower accuracy when compared to physical biometrics due to the inherent variation in human behavior. Such accuracy is usually evaluated using the false acceptance rate (FAR), the proportion of anomalous attempts wrongly classified as legitimate, and false rejection rate (FRR), the proportion of legitimate attempts wrongly classified to be anomalous. In terms of FAR, a recent work by Tey et al. showed that typing patterns can be imitated [14], increasing the FAR to unacceptable levels and thereby challenging the uniqueness assumption. For FRR, prior research has suggested that typing patterns are not resistant against external influences. For example, Cho et al. and Hwang et al. showed that with active cooperation from participants, typing patterns can be influenced by artificial rhythms and cues [3, 12]. Khanna et al. showed that typing speed decreases under negative emotions and increases for positive emotions [6]. Banerjee et al. suggested that changes in posture may also influence the typing pattern [2].

In this paper, we investigate the usability of keystroke biometrics under various external influencing factors and conditions. We identify a list of common conditions that may affect user typing patterns and investigate the extent to which the FRR is affected. Prior work has studied selected factors with the aim of either improving typing consistency [3, 12], or inferring information about the external factor from the typing pattern [6].

We conduct a user study involving 111 participants and measure the change in FRR using the Scaled Manhattan classifier [1] and a bioinformatics based classifier [11]. We find that the extent of influence depends on both the classifier and the external condition. The detailed analysis on typing under different conditions can be found in Section 5.

We analyze the FRR and determine using a Monte Carlo simulation the frequency of account lockout when keystroke biometrics is used as the authentication technique. Our results show that keystroke biometrics deployment requires an adjustment in the lockout policy. This however introduces an undesirable trade-off in the FAR. We also show that not everyone is suitable for keystroke biometrics and envision that future research looks into novel ways to identify users who are not unsuitable. We suggest a change in direction from finding classifiers that have decent performance for all users, to finding one that works well for a majority.

CODASPY'14, March 3–5, 2014, San Antonio, Texas, USA.

ACM 978-1-4503-2278-2/14/03 ...$15.00.
http://dx.doi.org/10.1145/2557547.2557573.

2. BACKGROUND

In this section, we present the background of keystroke biometrics systems by briefly describing the two classifiers used in our experiments. We discuss the reasons why we choose these two classifiers in Section 3.

Keystroke biometrics systems collect information about the typing patterns of users and compare it against reference models in its database in order to perform identification or authentication. In order to build the reference models, users need to provide a certain number of training samples. In the literature, the number of samples collected varies from 8 to 400 [7]. In this paper, we choose a training sample size of 100. Given a test sample, the reference model produces a score which is compared against the model threshold to determine if the test sample should be accepted or rejected.

In the rest of this paper, given a set of n training vectors each of length l, we use the notation t_{ij} to denote the elements in the training set, where $1 \leq i \leq n$ identifies the training vector and $1 \leq j \leq l$ identifies the element within each vector.

2.1 Scaled Manhattan classifier

The Scaled Manhattan classifier [1] requires a mean vector $\bar{\boldsymbol{x}}$ and an absolute deviation vector $\boldsymbol{d}$. Each element in $\bar{\boldsymbol{x}}$ and $\boldsymbol{d}$ is computed using the formulas:

$$\bar{x}_j = \frac{\sum_{i=1}^{n} t_{ij}}{n} \qquad d_j = \frac{\sum_{i=1}^{n} |t_{ij} - \bar{x}_j|}{n-1}$$

Given a test vector $\boldsymbol{y}$, the anomaly score s is computed using:

$$s = \sum_{j=1}^{l} \frac{|y_j - \bar{x}_j|}{d_j}$$

2.2 Bioinformatics based classifier

The bioinformatics based classifier [11] computes a motif vector using the following steps:

1. Computing a max vector $\boldsymbol{m}$

$$m_j = \max_i t_{ij}$$

2. Mapping each vector in the training set $\boldsymbol{t}$ to a normalised vector $\boldsymbol{u}$ such that:

$$u_{ij} = \frac{t_{ij}}{m_{ij}}$$

3. Mapping each normalised vector $\boldsymbol{u}$ to a bin vector $\boldsymbol{b}$ such that: b_{ij}:

$$b_{ij} = \begin{cases} 0 & \text{if } 0 < u_{ij} \leq 0.05 \\ 1 & \text{if } 0.05 < u_{ij} \leq 0.10 \\ \dots & \\ 19 & \text{if } 0.95 < u_{ij} \leq 1.00 \end{cases}$$

4. Computing a position specific scoring matrix P with 20 rows and l columns from the set of bin vectors such that each element p_{kj} of P is given by:

$$p_{kj} = \frac{\text{count}_i(b_{ij} = k-1)}{n}$$

where $1 \leq k \leq 20$ and $\text{count}_i(b_{ij} = k-1)$ counts for all i the number of times b_{ij} equals to $k-1$.

5. Computing a motif vector $\boldsymbol{f}$ such that:

$$f_j = \begin{cases} k & \text{if } \exists k : p_{kj} > 0.8 \\ -1 & \text{otherwise} \end{cases}$$

6. Counting the number of positive elements in $\boldsymbol{f}$. If it is 21 or less, a new max vector $\boldsymbol{m^*}$ is computed such that:

$$m_j^* = \begin{cases} m_j \times 1.1 & \text{if } f_j < 0 \\ m_j & \text{otherwise} \end{cases}$$

Steps 2 to 6 are then repeated, substituting $\boldsymbol{m^*}$ for $\boldsymbol{m}$, until the number of positive elements in $\boldsymbol{f}$ is more than 21. The last computed value of $\boldsymbol{m}$ is saved.

After obtaining the motif vector, to compute the anomaly score s for a test vector $\boldsymbol{y}$, a bin vector $\boldsymbol{y^*}$ is obtained following steps 2 and 3. The score s is computed by comparing the differences between $\boldsymbol{y^*}$ and $\boldsymbol{f}$ element by element as follows:

$$s = \sum_{j=1}^{l} D(y_j^*, f_j) \text{ where } D(\alpha, \beta) = \begin{cases} 0 & \text{if } \beta = -1 \\ 1 & \text{if } \beta \geq 0,\ \alpha = \beta \\ -1 & \text{if } \beta \geq 0,\ \alpha \neq \beta \end{cases}$$

2.3 Computation of threshold

Once an anomaly score is computed, a decision needs to be made to accept or reject the test sample. This requires a threshold parameter. For the Scaled Manhattan classifier, scores less than the threshold (close to the mean vector) are accepted. For the bioinformatics classifier, scores greater than the threshold (better match with motif vector) are accepted. In either case, the threshold defines a multidimensional boundary separating the acceptance space from the rejection space.

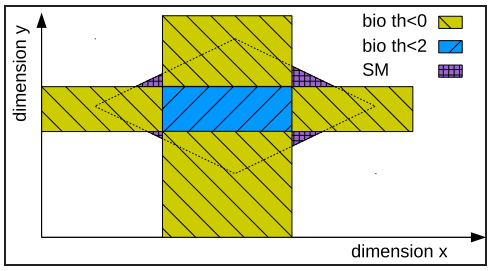


Figure 1: Comparison of classifier shapes

Figure 1 shows a possible arrangement of both classifiers that are trained on the same 2-d training set. Note that the centres of both shapes are not necessarily aligned. Any changes in typing pattern can be visualised as a shift on the diagram. Depending on the direction and magnitude of the shift, the resulting sample may be accepted by one classifier and rejected by the other. The Scaled Manhattan classifier is comparatively more vulnerable to shifts along the direction of the axes, while the bioinformatics based classifier is more vulnerable to shifts at 45° to the axes.

3. DESIGN CONSIDERATIONS AND APPROACH

High FRR results in usability problems because legitimate users may be unable to access the system or find it very

frustrating to do so. This results in security problems, because it tempts users to bypass the system in ways that may compromise security. For example, users may simply leave their screens unlocked if the authentication process is inconvenient. For this reason, it is important to identify usability issues. This section documents the considerations that affect our measurement of usability and our experimental approach.

3.1 Feedback vs without feedback

In practical authentication scenarios, users know if the last login attempt succeeded or failed. The focus of this paper is to investigate the FRR after various user activities and under various conditions. To provide a more realistic evaluation, we decide to collect the data in our experiments with feedback.

We are interested to know the effects on FRR when feedback is provided. After enrolment, we plan for 2 sessions, one with feedback and one without. Comparing the FRR of these sessions allows us to determine if feedback affects the FRR, in which case existing evaluation methods may require changes to take this into account. The outcome is presented in Section 5.

3.2 Choice of conditions

There are a large number of possible activities and conditions that could potentially induce changes of FRR in authentication. In this section, we describe those that are chosen and the rationale.

3.2.1 Background music and noise

Background music occurs fairly commonly in certain workplaces or during celebratory periods. Certain users may prefer to work while listening to music. Certain environments may also be noisier in various circumstances. Prior research by Cho et al. and Hwang et al. showed that if a rhythmic music is played and users are instructed to actively follow the rhythm when typing, there is a change in their typing pattern [3, 12].

In our case, users are under passive influence. It is possible that a similar though reduced effect exists. For this experiment, we choose to play the fast paced 'The Dark Knight Rises' soundtrack by Hans Zimmer.

3.2.2 Low lighting and time pressure

Low lighting conditions arise in meeting rooms darkened for presentation or voluntary blackout events such as Earth Hour. For users who need to orientate their fingers by looking at the keyboard but who are not using a backlit keyboard, it may slowdown their typing.

Time pressure is a very common emotional condition of modern life that may affect the typing pattern. Prior research by Khanna et al. [6] showed that it is possible to infer changes in the emotion from the typing pattern. We ask in this paper whether the change in FRR is significant when there is a change in emotion due to time pressure. However, the inducement of time pressure needs to be done without causing the participants distress or causing them to type haphazardly. We display a ticking timer and inform the participants to complete their typing before the timer runs down, while offering a reward for the participant with the lowest FRR. Refer to Section 4.3 for the rationale of this reward.

The results from the low lighting and time pressure experiments also allow us to answer this question: during an emergency in a critical infrastructure environment, can an operator working under emergency backup lights and time pressure unlock his control station?

3.2.3 Plaster

A plaster applied around a finger is a common remedy for minor ailments. Although its occurrence may be rare, once applied, any effect on the typing pattern persists until the removal of plaster or the recovery of the underlying ailment. It is therefore interesting to evaluate its effect because existing evaluation methods are likely to exclude it.

3.2.4 Standing

We include standing because of past experience where it was observed to occur. Such a scenario may also arise in field work where standard office amenities are not expected to be replicated. It is known that posture affects typing pattern [2]. Our focus in this paper is the quantitative change.

3.2.5 Change of keyboard

Our emphasis is on the quantitative aspect of the change in FRR when there is a change of keyboard. This scenario is interesting because the outcome determines whether the keystroke biometrics database needs to be trained for different input devices. BYOD (Bring Your Own Device) developments, shared workstations, and even helpdesk technicians who may need login access to user devices are examples where a quantitative analysis is useful.

We include in our experiments only participants who are laptop users. Our experiments measure the change in FRR when they switch from the laptop keyboard to a standard external keyboard or an ergonomic keyboard.

3.2.6 Computer gaming and physical exercise

Both computer gaming and physical exercise are very common activities. Both induce emotional changes and possibly physical muscle strain. For our experiments, we look for participants who already have existing gaming or exercise habits. The experimental structure is designed to fit in with their routine rather than having the researchers dictate to them what they should do. For exercise in particular, we need to emphasise this point to the participants to avoid increasing the chance of injury.

For the gaming and exercise conditions, there is a possibility that the effect is long lasting and persists well beyond the end of the activity. We include an experiment to measure the FRR change after different rest periods.

3.3 Choice of userid and password

We choose as password the string `ths.ouR2` which had been shown to be more difficult to imitate [14]. This password also fits the complexity criteria in user environments and is in our opinion not atypical of actual user passwords. We choose `user1024` as the userid. We acknowledge the great diversity of userids in use. However, we have the constraint of choosing only one userid for all participants to avoid introducing the effect of typing difficulty into the sample data. We also need to avoid choosing a userid (e.g., based on a common name) that may be more familiar to some participants than others. The string `user1024` represents a good compromise in our opinion.

3.4 Choice of classifier

We compute the results using two classifiers. We choose the Scaled Manhattan classifier by Araujo [1] because it was identified as the best classifier (in terms of accuracy) in a survey by Killourhy and Maxion [7]. We choose as the second classifier, the bioinformatics based classifier by Revett [11], which was identified (in another survey) by Banerjee [2] as a classifier with good accuracy.

For experiments involving feedback, it is unrealistic to provide feedback from two classifiers at the same time. We choose arbitrarily to provide feedback for only the Scaled Manhattan classifier.

4. EXPERIMENT

In this section, we describe our experimental setup given the considerations discussed in Section 3. The basic idea of the experiments is to engage the participants in certain activities and conditions, and to measure the FRR of their authentication attempts. The whole study is divided into three sessions: s1, s2, and s3. s1 is conducted to build an anomaly dataset for training the classifier. s2 and s3 are conducted in a lab environment. s2 involves experiments that are unlikely to have any long-term effects on typing patterns while s3 involves experiments that *may* have long-term effects. Figure 2 shows the various stages of s2 and s3. Note that s2 and s3 are conducted on different days.

4.1 Participants and Setup

We recruit a total of 111 students in our tests. We receive IRB approval from our university, and compensate the participants for completion of the entire set of tasks. Table 4.1 summarises the participant demographics of our study.

Experiment	Male	Female
Background Noise	11	7
Low Light	16	7
Time Pressure	10	8
Plaster	9	8
Standing	14	4
Ergonomic KB	4	4
Standard External KB	6	3
Exercise with 45 min rest	14	4
Exercise with 15 min rest	11	5
Heavy Gamer with 45 min rest	16	4
Heavy Gamer with 15 min rest	11	7
Light Gamer with 45 min rest	9	9
Light Gamer with 15 min rest	9	12

Table 1: Demographics

4.2 s1: Anomaly dataset collection

s1 collects typing samples for the purpose of building an anomaly dataset. Each participant contributes 5 samples (without feedback of the classifier acceptance) for use as anomaly data for other participants. Each of them is required to type in a userid and password (provided by us) via our web interface. Participants are paid $4 for s1. Figure 3 shows the interface used.

This part of the study is designed to be conducted online. However, as some participants were unable to access the web interface due to technical glitches (e.g., browser compatibility), they completed s1 in a lab environment prior to the start of s2. For these users, their typing patterns are excluded from the anomaly data set so as to ensure everyone uses a largely similar set of anomaly data. Out of 111 participants, 60 contributed 5 samples each for the anomaly data. Each of these participants' anomaly data-set contains 59×5 samples. The remaining 51 participants' anomaly data-set contains 60×5 samples.

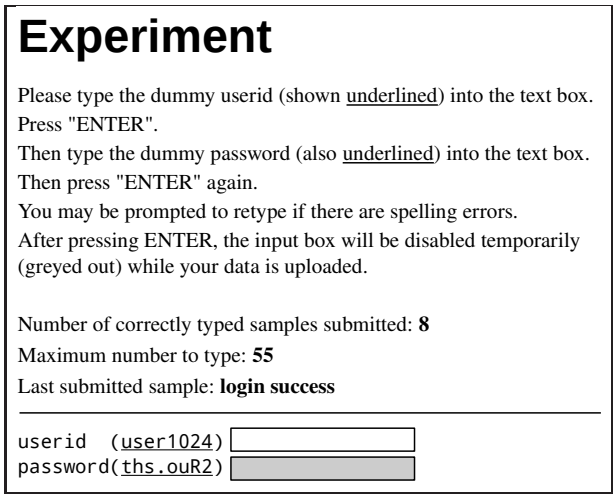


Figure 3: User Interface for data collection

4.3 s2: Experiments with short-term effects

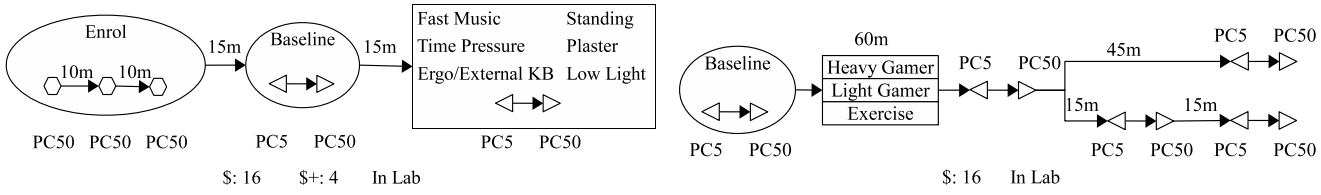
s2 is designed to investigate the effectiveness of keystroke biometrics under certain scenarios that we consider to have *short-term* effects. It is conducted in a lab environment post s1.

Participants are first required to answer a questionnaire capturing their emotional and physical state. This is followed by the creation of the training set whereby participants type 50 samples of the userid and password pair thrice with a gap of 10 minutes between each period. As discussed in Section 3, we design the session such that each collection period is short. We use the first 100 samples to train the classifier and the next 50 to calculate the FRR under no-feedback conditions, where there is no indication whether prior authentication attempts were successful.

Next, after a rest of 15 minutes, participants are asked to type 55 samples with feedback. We use the submitted samples to calculate the FRR, with feedback, for that user. This FRR is compared with the FRR (under no-feedback conditions) to observe the impact of feedback (if any) on a participant's FRR. We show the result of this evaluation in Section 5.

Finally, after another 15 minutes of rest, participants are asked to submit 55 samples under the following conditions: listening to fast music, with a plaster on the right index finger, using an ergonomic keyboard, using a standard external keyboard, under time pressure, under low light, and while standing (see Section 3.2).

Participants are encouraged to type the samples consistently and are in fact provided with a monetary incentive of $4 for doing so. The reason is that in practice, there is a natural penalty (account-lockout, repeated attempts) that discourages haphazard typing. This is missing in an experimental setting. The $4 incentive compensates for this.



(a) s2 (b) s3

Figure 2: Experimental structure — PCn: Type userid and password n times, ○: No feedback to user on the acceptance of their typing pattern (by the classifiers), ◁: Users are asked for their perception before providing the feedback, ▷: Feedback is provided to user, $: Amount paid for session, $+: amount paid as bonus for meeting certain criteria.

There is a possible risk of overcompensation that mask a worse FRR increase. However, the incentive provides certainty of the lower bound of the FRR increase.

At the end of s2, participants answer a questionnaire on the task experience. Note that samples with any typographical errors are not collected, but the count and order of such samples are recorded. Participants are required to retype whenever they make a typographical error.

4.4 s3: Experiments with long-term effects

The aim of s3 is to investigate the effectiveness of keystroke biometrics under scenarios that we consider to have relatively *long-term* effects. Similar to s2, participants first answer a questionnaire stating their emotional and physical states. We then collect the baseline typing pattern (55 samples) of each participant.

After the collection of baseline samples, participants are asked to perform the assigned activity (either playing a computer game or going to the gym). Each activity lasts for 60 minutes following which participants return to the lab to type 55 more samples as in s2.

After this sample collection, participants are divided into two groups. The first group is allowed to rest for 45 minutes following which 55 more typing-samples are collected. This is done to observe the effect of a long rest duration on the accuracy of their typing. Users are not allowed to use their smartphones or laptops during this rest period. They are, however, allowed to read magazines and other periodicals (provided by us) to pass time. This is to ensure that no typing is performed during the rest period, thereby minimising any undesirable influence on subsequent sample collection.

The second group is allowed to rest only for 15 minutes after which they are asked to type 55 samples. This cycle of a 15 minute rest followed by typing 55 samples is then repeated. This is done to observe the effect of intermittent rest on the accuracy of their typing. As in the first group, the participants do not perform any other typing. At the end of s3, participants answer a questionnaire on the task experience.

5. RESULTS

In this section, we present the results of our experiments. We omit the results for the experiments involving long term and intermittent rest at the end of s3. This is due to both the brevity of space and the inconclusive results for those experiments.

In our analysis, the significance level is computed using a paired student's t-test. Paired t-test is used because it indicates (subject to a confidence level) whether there is a difference in the FRR with and without the environmental influences. Our null hypothesis is that there are no changes in FRR. The alternative hypothesis is that there are increases in FRR. A single-tailed test is used because we assume any changes in typing pattern increase the FRR. Subsequent results (see Figure 7) support this. The significance level follows a standard text book classification [5], where the descriptions 'highly significant', 'significant' and 'weakly significant' are used to describe p-values in the ranges of [0, 0.01), [0.01, 0.05), and [0.05, 0.10), respectively. The criteria for pairing is based on the user.

5.1 Participant fatigue and feedback

In this section, we address concerns relating to participant fatigue and the provision of feedback that may affect the validity of our results.

5.1.1 *Number of samples collected*

First, we want to know whether the number of samples typed (55) is overwhelming and causes those typed later to be significantly different from those typed earlier. Figure 4 shows the relation between the average FRR for all participants and the sample index for the s2 baseline experiment.

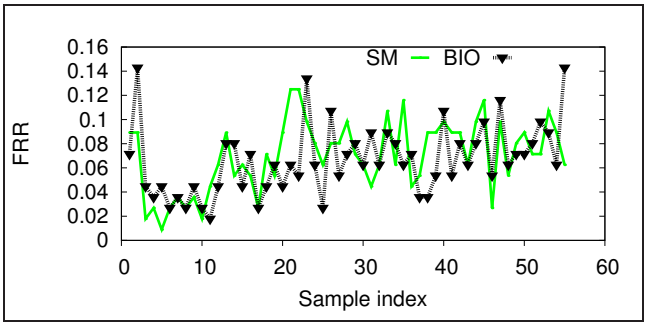


Figure 4: Effect of number of samples collected on the FRR

The coefficient of correlation between FRR and sample index are 0.4339 and 0.3994 for the Scaled Manhattan and bioinformatics classifier, respectively. This means that there is statistical support for the intuition that latter samples tend to have higher FRR than earlier samples. The slope of the regression lines through the two sets of data are, however, just 0.0007885 and 0.0007228, respectively. Therefore, the effect is not overwhelming. The time taken to collect 55

typing samples range from 3.5 min to 12.2 min, with an average of 5.3 min. The typing time is therefore short relative to the rest period.

5.1.2 Feedback vs no feedback

We compare for all participants the FRR of the samples collected without feedback with their session baseline which is collected with feedback. For the Scaled Manhattan classifier, the mean FRR of samples collected with feedback is 0.01106 higher than that without feedback. However, the difference is not statistically significant. On the other hand, the difference for the bioinformatics based classifier is 0.01427 and is highly significant.

If participants ignored the feedback, there should not be any significant differences. The results therefore suggest that users are not ignoring feedback and, surprisingly, are affected by it in a negative way. A possible explanation is that when users are trying hard to make sure they are typing "correctly", their typing pattern actually becomes slightly different from their enrolment pattern. This factor is not accounted for in existing literature and justifies our decision to collect typing samples with feedback.

5.2 Change of input devices

Table 2 shows the change in FRR when the input device is changed from the laptop keyboard to an external keyboard or an ergonomic keyboard. For both cases, the changes are large and highly significant. The Scaled Manhattan classifier is more adversely affected compared to the bioinformatics based classifier. Most people are unfamiliar with ergonomic keyboards and that has resulted in an extremely high FRR.

Exp	N	SM-B	SM-D	BIO-B	BIO-D
Ext. KB	9	0.1010	0.1697	0.0747	0.0889
Ergo. KB	8	0.0818	0.7500	0.0841	0.6182

Table 2: Effect of device factors on typing pattern. N: number of participants. SM-B: Baseline FRR (Scaled Manhattan). SM-D: Increase in FRR (Scaled Manhattan). BIO-B: Baseline FRR (Bioinformatics). BIO-D: Increase in FRR (Bioinformatics). All changes are highly significant.

It is clear from our results that patterns obtained using one keyboard should not be used to authenticate users when typing on a different type of keyboard. Given that in both home and corporate environments, owning multiple devices such as both a PC and a laptop is a common scenario, the results raise the question of how keystroke biometrics deployment should handle multiple input devices. A straightforward solution is to enrol the user multiple times, once for each device. However, it remains unclear whether a user can maintain a consistent typing pattern for each device when one is used only occasionally. We do not explore this issue in this paper and leave it as future work.

5.3 Effect of various conditions on FRR

Figure 5 summarises the effect of the conditions discussed in Section 3. We can see that other than the fast music experiment under the Scaled Manhattan classifier, all other conditions exhibit varying degrees of statistical significance. Different classifiers and groups of participants are affected differently by these conditions. The standing condition, plaster, and exercise are among the factors making the larger differences, although such changes in FRR are not as big as those observed with the change in input devices.

The result for fast music is interesting. The change in FRR for the Scaled Manhattan classifier is minor and statistically insignificant. This contrasts with the one observed for the bioinformatics classifier. An explanation can be made based on a shape based interpretation of these two classifiers (see Section 2). Any changes in typing pattern corresponds to a multi-dimensional movement. It is therefore possible to move beyond the enclosure of one classifier region while remaining in the other classifier region.

Participants in the fast music experiment report that they type faster. Referring to Figure 1, this corresponds to a diagonal movement towards the approximate direction of the origin in a 2-d setting. The exact direction depends on the relative extent of movement in each dimension. This means that for the fast music experiments, the participants' typing pattern moved out of the bioinformatics region, but remained in the Scaled Manhattan region.

6. LOCKOUT AND ANNOYANCE

Even without keystroke biometrics, there is an FRR due to typing when users remember their passwords correctly, but type them wrongly. It is not a serious problem in practice because if a user makes typing errors on the initial tries, he can slow down and ensure the subsequent attempt is correct. With keystroke biometrics, such a strategy is no longer workable. If a user slows down, he increases the FRR due to keystroke biometrics. If he does not slow down, he increases the FRR due to typing mistakes. Implementing keystroke biometrics therefore increases the overall FRR of the system by more than just the keystroke biometrics FRR.

In this subsection, we investigate what the impact of this increase would be, in terms of both the likelihood of account lockout and the user reaction. To estimate the effect of the overall FRR on the user experience, we conducted a Monte Carlo simulation. We divide the users into groups based on the conditions shown in Figure 5. The simulation first picks a group, then a user within that group, both randomly. The overall FRR is computed for 2 cases: (a) baseline only, without any influence of the selected condition and (b) baseline + condition. The overall FRR takes into account both typing errors and keystroke biometrics error. The simulation computes, for each login session, the number of tries required to login.

Figure 6 shows the cumulative percentage of users who will encounter a lockout after a certain number of login sessions. For example, approximately 30% of users will encounter a lockout before their 50th authentication session if the lockout policy is 3 attempts and if, in 50% of the trials, they are under the influence of one of the conditions of Figure 5.

Figure 6 shows that keystroke biometrics is not for everyone. Even with a lockout policy of 5, approximately 10% of users will find keystroke biometrics very frustrating to use. This suggests that finding classifiers that work for all users is a difficult task. A possible alternative is to find classifiers that work well for the majority. The unsuitable minority requires an alternative login mechanism.

Figure 6 also shows that deployment of keystroke biometrics must come with an adjustment in lockout policy, but at a cost in security. Keystroke biometrics is considered as a security mechanism to enhance the security of weak password [10]. However, when the password is weak and the

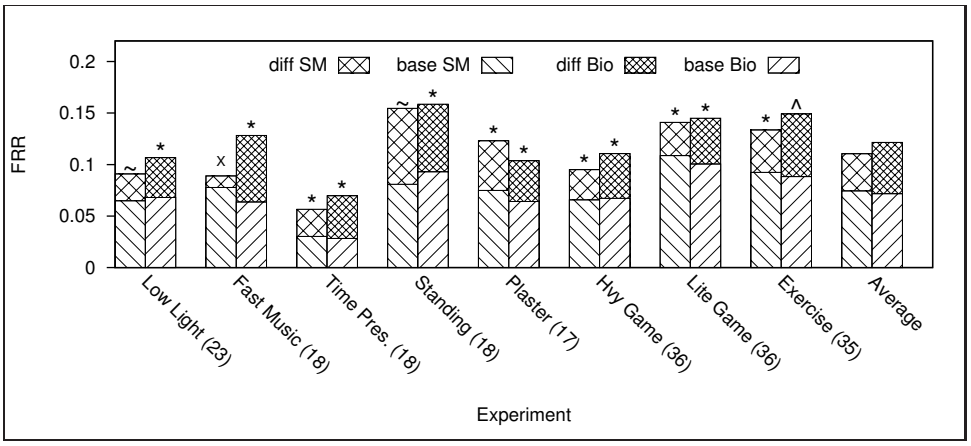


Figure 5: Effect of various conditions on typing pattern (baseline and increase in FRR). Numbers in brackets indicate the number of participants. ∧: highly significant. *: significant. ∼: weakly significant. x: not significant.

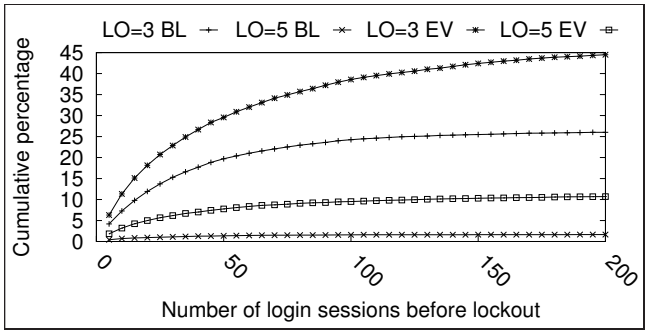


Figure 6: Number of authentication sessions before user gets lockout. LO: Lockout policy. BL: Baseline (no environmental factors). EV: Environmental factors in effect 50% of the time.

lockout is adjusted to 5, the added security may be modest. Table 3 shows the chance that an attacker can break into a system if he already knows the password. For an FAR value of 0.05, the overall FAR is 0.22622 if the attacker is allowed to make 5 attempts. In other words, approximately 1 in 4 such attempts will succeed.

	FAR			
Lockout Policy	0.001	0.01	0.02	0.05
1	0.00100	0.01000	0.02000	0.05000
2	0.00200	0.01990	0.03960	0.09750
3	0.00300	0.02970	0.05881	0.14263
4	0.00399	0.03940	0.07763	0.18549
5	0.00499	0.04901	0.09608	0.22622

Table 3: Chance of attacker success before lockout (detection).

Other than lockout, we are interested in possible user annoyance. We asked the participants for their reactions if, on average, they need to type their credentials multiple times before they can login. Table 4 summarises the responses.

Figure 7 shows the cumulative percentage of users who will encounter a login session in which they succeed but find very annoying. For example, more than 40% of users will

	2 tries	3 tries	4 tries	5 tries
Okay	37.62	5.94	0.99	0.00
Mildly annoying	51.49	51.49	16.83	3.96
Very annoying	10.89	42.57	82.18	96.04

Table 4: User response (%) to the number of tries require for authentication. E.g. 10.89% of users are very annoyed if they need 2 tries to login.

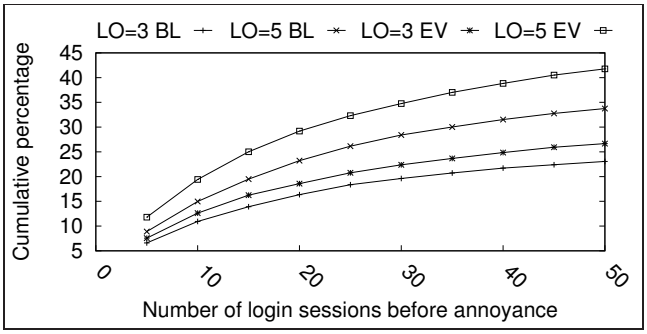


Figure 7: Number of authentication sessions before user encounters a very annoying one. LO: Lockout policy. BL: Baseline (no environmental factors). EV: Environmental factors in effect 50% of the time.

encounter an annoying session before their 50th authentication session if the lockout policy is 5 attempts and if, in 50% of the trials, they are under the influence of the conditions in Figure 5. The occurrence of annoyance is lower when the lockout policy is 3 because users get locked out before they have a chance to succeed (but feel annoyed). This shows that even after solving the lockout problem, keystroke biometrics still has to contend with user frustration over the multiple login attempts required.

7. LIMITATIONS

We discuss some limitations of our experiments in this section. In both sessions s2 and s3, we ask participants to abstain from exercise and gaming for the prior 24 hours, to avoid introducing any possible influence from these activi-

ties. We are dependent on the participants to observe these restrictions in good faith.

In our experiments, we collect the typing samples via a web based interface using participants' own laptop. The latter is because we want to ensure they are using a familiar keyboard layout. However, this also implies that the input devices come in a variety of OS and hardware. In particular, the timing granularity varies. For Windows XP systems, this can be up to 16 ms. In the literature, the timing granularity varies from 0.2 ms [7], 1 ms [1] and 10 ms [13, 4]. To ensure timings are collected accurately, we also exclude laptops which are overloaded with too many computationally intensive background processes.

In all our experiments, a single password is used. That password is chosen to be difficult to type, which unfortunately increases the chance of typing it wrongly, and in turn increases the FRR. If participants are to use an easier password, the number of typing errors should decrease. However, we believe our chosen password is more representative of those mandated by corporate security policies.

8. CONCLUSION

We find that various environmental factors such as playing computer games, exercise, lighting conditions, fast music, emotional pressure, and even the application of a plaster on one finger result in varying increase of FRR. Different users are also affected differently. Certain users are shown to be unsuitable for keystroke biometrics. This suggests that instead of a research direction that attempts to find classifiers that work for all users, a more practical approach may be to find users who have low baseline FRR and are tolerant towards environmental changes. This implies a hybrid approach where keystroke biometrics is implemented for only a subset of users, and is complemented by an alternative authentication mechanism.

Acknowledgment

This work was supported in part by Singapore Management University grant 13-C220-SMU-003.

9. REFERENCES

[1] L. Araujo, L. Sucupira, Jr., M. Lizarraga, L. Ling, and J. Yabu-Uti. User authentication through typing biometrics features. *Trans. Sig. Proc.*, 53(2):851–855, Feb. 2005.

[2] S. Banerjee and D. Woodard. Biometric Authentication and Identification using Keystroke Dynamics: A Survey. *Journal of Pattern Recognition Research*, 7:116–139, 2012.

[3] S. Cho, C. Han, D. H. Han, and H. il Kim. Web based keystroke dynamics identity verification using neural network. *Journal of Organizational Computing and Electronic Commerce*, 10:295–307, 2000.

[4] S. Haider, A. Abbas, and A. Zaidi. A multi-technique approach for user identification through keystroke dynamics. In *IEEE International Conference on Systems, Man and Cybernetics*, SMC 2000, pages 1336–1341, 2000.

[5] G. Keller and B. Warrack. *Statistics for management and economics*. Number v. 1 in International student edition: Wadsworth. Thomson/Brooks/Cole, 2003.

[6] P. Khanna and M.Sasikumar. Article: Recognising emotions from keyboard stroke pattern. *International Journal of Computer Applications*, 11(9):1–5, December 2010. Published By Foundation of Computer Science.

[7] K. Killourhy and R. Maxion. Comparing anomaly-detection algorithms for keystroke dynamics. In *Dependable Systems Networks, 2009. DSN '09. IEEE/IFIP International Conference on*, pages 125 –134, Jul 2009.

[8] K. S. Killourhy. *A Scientific Understanding of Keystroke Dynamics*. Dissertation, Carnegie Mellon University, 2012.

[9] F. Monrose and A. D. Rubin. Keystroke dynamics as a biometric for authentication. *Future Gener. Comput. Syst.*, 16(4):351–359, Feb 2000.

[10] A. Peacock, X. Ke, and M. Wilkerson. Typing Patterns: A Key to User Identification. *IEEE Security and Privacy*, 2(5):40–47, Sept. 2004.

[11] K. Revett. A bioinformatics based approach to user authentication via keystroke dynamics. *International Journal of Control, Automation and Systems*, 7:7–15, 2009.

[12] S. seob Hwang, H. joo Lee, and S. Cho. Improving authentication accuracy using artificial rhythms and cues for keystroke dynamics-based authentication. *Expert Systems with Applications*, 36(7):10649 – 10656, 2009.

[13] D. X. Song, D. Wagner, and X. Tian. Timing analysis of keystrokes and timing attacks on SSH. In *Proceedings of the 10th conference on USENIX Security Symposium - Volume 10*, SSYM'01, pages 25–25, Berkeley, CA, USA, 2001. USENIX Association.

[14] C. M. Tey, P. Gupta, and D. Gao. I can be You: Questioning the use of Keystroke Dynamics as Biometrics. In *Proceedings of NDSS*, San Diego, CA, Feb 2013.

[15] D. Umphress and G. Williams. Identity verification through keyboard characteristics. *International Journal of Man-Machine Studies*, 23(3):263–273, Sept. 1985.