Effect of Data Size on Performance of Free-text Keystroke Authentication

Jiaju Huang, Daqing Hou, Stephanie Schuckers
Department of Electrical and Computer Engineering
Clarkson University, Potsdam NY
jiajhua, dhou, sschucke@clarkson.edu

Zhenhao Hou
Department of Computer Science
Brown University, Providence RI
zhenhao_hou@brown.edu

Abstract

Free-text keystroke authentication has been demonstrated to be a promising behavioral biometric. But unlike physiological traits such as fingerprints, in free-text keystroke authentication, there is no natural way to identify what makes a sample. It remains an open problem as to how much keystroke data are necessary for achieving acceptable authentication performance. Using public datasets and two existing algorithms, we conduct two experiments to investigate the effect of the reference profile size and test sample size on False Alarm Rate (FAR) and Imposter Pass Rate (IPR). We find that (1) larger reference profiles will drive down both IPR and FAR values, provided that the test samples are large enough, and (2) larger test samples have no obvious effect on IPR, regardless of the reference profile size. We discuss the practical implication of our findings.

1. Introduction

A biometric authentication system uses the physiological (fingerprints, face, hand geometry, iris) and/or behavioral (voice, signature, keystroke dynamics) traits of an individual to identify a person (identification) or to verify a claimed identity (verification) [3]. A typical biometric system operates in two distinct stages: the enrollment stage and the authentication stage. During enrollment, a user's biometric data (e.g., fingerprints) is acquired and processed to extract a feature set (e.g., minutiae points) that is stored in the database. The stored feature set, labelled with the user's identity, is sometimes referred to as a template [8]. In order to account for variations in the biometric data of a user, multiple templates corresponding to each user may be stored. During authentication, a user's biometric data is once again acquired and processed, and the extracted feature set is matched against the template(s) stored in the database in order to identify a previously enrolled individual or to validate a claimed identity. [8]

In this paper, we focus on free text, keystroke dynamics based user authentication [2]. During enrollment of keystroke dynamics authentication, a *reference profile* is created for each user by storing a certain amount of keystroke data in a database. A *test sample* is later collected from a user and matched against the reference profile in order to authenticate the user. Unlike modalities such as fingerprints, where there is a natural way to identify what makes a sample (e.g., a fingerprint impression), it is unclear how much keystroke data are needed by a reference profile and a test sample, respectively, and what is the effect of their size on authentication performance. This problem is related to, but different from, *template selection* [1, 8], a process by which prototype templates are chosen from a given set of samples. Some keystroke authentication algorithms, such as that of Leggett *et al*. [5], does not even require template selection. Gunetti and Picardi [2] is the only other study that briefly looks at this same problem, but their study is much less thorough, and covers fewer combinations, than ours.

To explore the relationship between the amounts of keystroke data [1] and authentication performance (False Alarm Rate, or FAR [2], and Imposter Pass Rate, or IPR), we conduct two experiments with two representative, state-of-art keystroke authentication algorithms [2] [5]. For both algorithms, while fixing other parameters such as threshold values, we systematically vary the reference profile size and the test sample size. We then analyze the relationship between data sizes and performance results. We find that (1) larger reference profiles will drive down both IPR and FAR values, provided that the test samples are large enough, and (2) larger test samples have no obvious effect on IPR, regardless of the reference profile size. We discuss the practical implication of our findings.

This is the first empirical study that explores the effect of data size on performance of *free-text* keystroke authentication. Killourhy and Maxion's study on keystroke dynamics error rates [4] is closely related. They investigate the

[1] We measure the "amounts of keystroke data" in terms of the number of characters that a user has typed.

[2] In contrast, some researchers prefer to referring to FAR as "False Acceptance Rate," which corresponds to IPR, or "Imposter Pass Rate," in this paper.

effects of six factors on authentication performance, including "training amount," which is the number of repetitions of typing the same given password (*.tie5Roanl*) in a user's profile. The key difference is that their study is about keystroke authentication based on *static*, rather than *free* text, so the issue of test sample size does not apply as in their work, the test sample is fixed. While they draw essentially the same conclusion regarding the effect of larger reference profiles, understandably, their results do not include the interaction between reference profiles and test samples.

The rest of the paper is organized as follows. Section 2 describes the two algorithms, the dataset used, and the experimental settings. Sections 3 and 4 present and analyze the results from the two experiments, respectively. Finally, Section 6 summarizes the results of this work and provides future directions for research.

2. Experimental Setting

In this section, we describe the two algorithms [2] [5] that we use in our experiments, the dataset, and how we process the data to create a reference profile and a test sample for the experiments.

2.1. Experiment 1: Gunetti-Picardi's Algorithm

For our first experiment, we implemented Gunetti-Picardi's user verification algorithm (cf. Section 5.2 in [2]). Briefly, given a test sample X and a user A, Gunetti-Picardi's verification algorithm verifies whether X comes from A, as follows:

1. For each legal user U in the system, calculate two metrics $m(U)$ and $md(U, X)$ defined as follows:

   $m(U)$: the average distance between all the samples in U's profile.

   $md(U, X)$: the average distance between all the samples in U's profile and the test sample X.

2. If there exists any other legal user B of the system, to whom the test sample X is closer than to the verified user A, that is, $md(B, X) \leq md(A, X)$, then reject X for user A. Otherwise, X is closest to A in the system; conduct a further test in the next step to ensure that X is *sufficiently close* to A's profile samples.

3. Furthermore, if test sample X is closer to the core of user A's profile samples than the average distance between themselves ($m(A)$). That is, if $md(A, X) \leq m(A)$ is true, then accept X. Otherwise, $md(A, X) > m(A)$ holds, and check the next condition.

4. Finally, accept X, if $md(A, X)$ is closer to $m(A)$ than to any other $md(B, X)$ computed by the system, i.e.:

   $md(A, X) - m(A) < md(B, X) - md(A, X)$

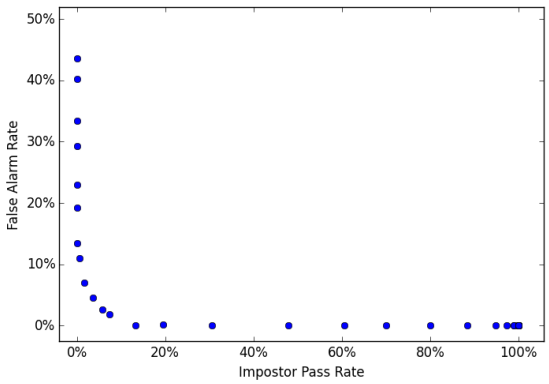

Figure 1. ROC curve for user 4989044. Each of the 346 combinations of threshold and standard deviation yields an FAR and IPR that is represented by a single data point.

The distance between typing samples $S1$ and $S2$ is measured in terms of n-graphs (n consecutive keystrokes typed by a user), using a measurement called *"A" measure* [2]. The "A" measure between $S1$ and $S2$ is defined as follows, in terms of the n-graphs they share and a constant value t:

$A_{t,n}(S1, S2)$ = 1 − (number of *similar* n-graphs shared by $S1$ and $S2$) / (total number of shared n-graphs)

A constant t is needed for defining n-graph similarity. More specifically, let $G_{S1,d1}$ and $G_{S2,d2}$ be the same n-graph occurring in typing samples $S1$ and $S2$, with durations $d1$ and $d2$, respectively. We say that $G_{S1,d1}$ and $G_{S2,d2}$ are *similar* if $1 < max(d1, d2)/min(d1, d2) \leq t$, where t is some constant greater than 1.

In our implementation, we use digraph durations and trigraph durations as features and choose the same value for t as in [2] ($t = 1.25$). We sum up the A_2 measure and A_3 measure as the final distance between two samples.

2.2. Experiment 2: "Zone-of-Acceptance"

The "zone-of-acceptance" algorithm by Leggett *et al.* [5] is based on the assumption that the latency times for all occurrences of a digraph in the reference profile follow a normal distribution $\mathcal{N}(\mu, \sigma)$. If the average latency time for a digraph in the test sample falls between the *acceptance region* ($[\mu - d * \sigma, \mu + d * \sigma]$), the digraph is then considered *accepted*; otherwise, it is *rejected*.

The similarity score for a test sample is in turn defined as the ratio of the number of accepted digraphs over the total number of digraphs appearing in the test sample. The test sample is accepted if its similarity score is greater than or equal to a set threshold. In their study, Leggett *et al.* set $d = 0.5$ and an acceptance threshold of 0.6, *for all users* [5]. Instead, we are able to calibrate a different pair of d and threshold values for *each user*.

We calibrate the d for standard deviations and accep-

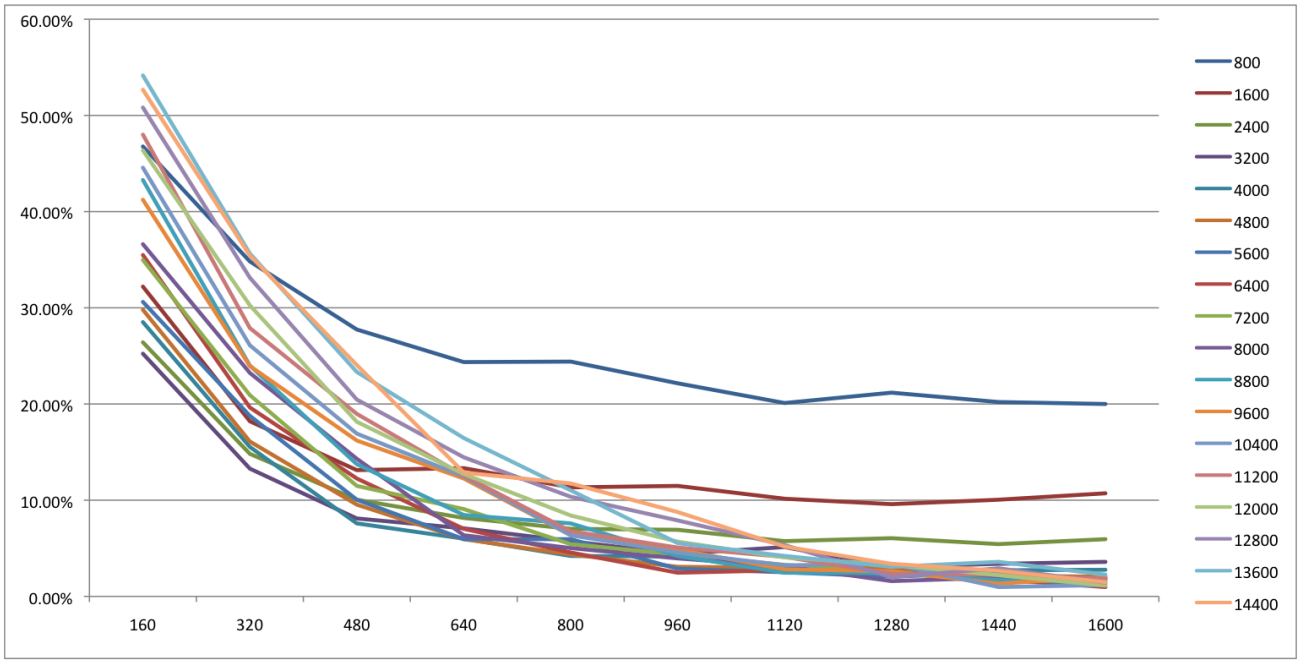

Figure 2. Effect of Test Sample Size on False Alarm Rate (FAR). X Axis Represents Test Sample Sizes (number of keystrokes). Y Axis Represents FAR Values. Each Line Represents the Results of Testing Reference Profiles of the Same Size, Ranging from 800 to 14,400 Keystrokes, against Test Samples of Different Sizes, Ranging from 160 to 1,600 Keystrokes.

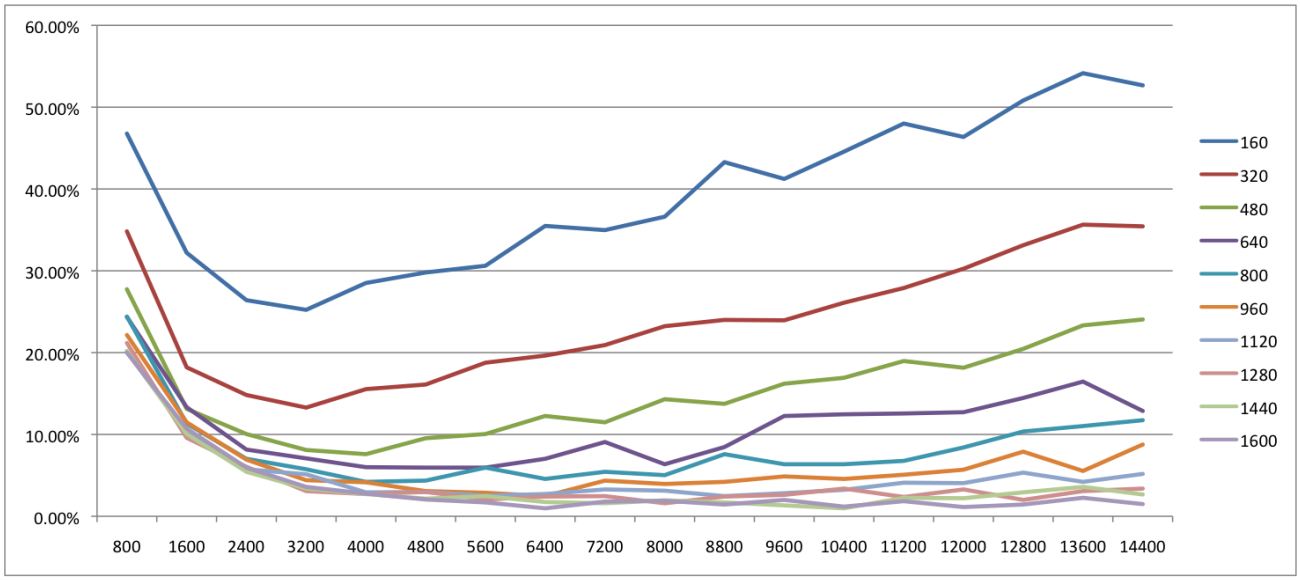

Figure 3. Effect of Reference Profile Size on False Alarm Rate (FAR). X Axis Represents Reference Profile Sizes. Y Axis Represents FAR Values. Each Line Represents the Results of Testing Test Samples of the Same Size, Ranging from 160 to 1,600 Keystrokes, against Profiles of Different Sizes, Ranging from 800 to 14,400 Keystrokes.

tance threshold values as follows, using a dataset from the authors of a previous study [2]. The dataset contains 31 users (slightly smaller than what was originally used in [2]), with the amount of data per user varying between 13,000 to 19,000 keystrokes. We fix the test sample size to 1,000 keystrokes and use the remaining data to make a user's reference profile. To calibrate the optimal d and acceptance threshold, we vary the threshold from 0.5 to 0.9, in incre-

ments of 0.01. For each threshold, the d for standard deviations is varied and tested, in increments of 0.1, between 0.8 times the threshold and 2 times the threshold. For example, when the threshold has the value of 0.5, d is varied between 0.4 (0.8*0.5) and 1 (2*0.5). This allows for a total of 346 combinations to test with each user. The combination of threshold and standard deviation that produces the best tradeoff between FAR and IPR for the user is recorded. In order to select which combination provides this best tradeoff, we plotted the FAR against the IPR yielded by each combination. We identified the point of lowest sum for FAR and IPR and the point of smallest difference between FAR and IPR among the data points. Based on inspecting the results generated by our data, we find that the point of lowest sum yielded superior FAR and IPR to the point of smallest difference.

As an example, Figure 1 depicts the results for a single user (user "4989044"). The threshold and standard deviation that yielded the lowest sum point for this user was a threshold of 0.60 and a standard deviation of 0.60. (Both happen to have the same value of 0.6.) This threshold and standard deviation is recorded and deemed to be most suitable to accurately authenticate this user.

2.3. Description of Dataset and Data Processing

For both experiments, we use a keystroke dynamics dataset that we share publicly in the 2014 International Joint Conference on Biometrics [10]. This dataset contains free-text data for 39 users collected in a laboratory. Instead of directly using the raw data, we preprocess it by removing blanks, backspaces, keystrokes deleted by backspaces, which may have been used to correct misspellings or simply edit the text, and several other keys that we do not use. After the preprocessing, we are able to obtain on average about 16,000 keystrokes for each of the 39 users.

Experiment 1: We first record all occurrences of digraphs and trigraphs in each user's data. We then randomly pick digraph and trigraph instances, based on the given size information, to create a test sample and a reference profile, respectively. To apply the Gunetti-Picardi algorithm described above, we divide each reference profile's data into five samples. We call matching a test sample to the user's own reference profile data a "genuine access," and matching the test sample to another user's profile an "attack." All the tests are repeated 50 times to reduce random errors. There are 39*50 genuine accesses and 39*38*50 attacks. Based on these tests, we calculate and report the average False Alarm Rate (FAR) and Imposter Pass Rate (IPR) for all 39 users.

Experiment 2: We split each user's data into groups of 1,000 digraph instances. Based on the outcome of Experiment 1, which suggests that test samples should contain about 1,000 keystrokes, we keep the test sample size at 1,000 instances. Reference profile size was varied from 6,000 digraph instances to 17,000 digraph instances to test the effect of reference profile sizes on authentication performance. These tests involved $n - 1$ genuine tests, where n thousand is the number of keystrokes provided by the user. For instance, if a user provided 19,000 keystrokes, the algorithm would run through 18 genuine tests. The test sample size was then raised to 2,000 digraph instances, and the reference profile was varied from 6,000 digraph instances to 17,000 digraph instances. These tests involved $\binom{n}{2}$ genuine tests. In the end, only 25 users, out of the 39, had sufficient data for all of our tests.

Regardless of test sample size, the number of impostor tests was kept at 500, each test consisting of data randomly selected from one of the remaining users. The genuine test sample size was kept consistent with the impostor test sample size during all tests. The average FAR and IPR of 25 tested users was taken for each test sample and reference profile size.

3. Results for Experiment 1

Based on the results from Experiment 1, we discuss the effect of data size on FAR and IPR in Sections 3.1 and 3.2, respectively. We summarize our findings in Section 3.3.

3.1. Effect of Data Size on False Alarm Rate (FAR)

Based on Figure 2 and Figure 3, we observe the following points for FAR:

1. From the 18 lines in Figure 2, we can observe that *for reference profiles of the same size, as the test sample gets larger, from left to right, the FAR value improves and becomes smaller.* This can also be seen from the 10 lines in Figure 3, which exhibit the overall trend across all the lines that FAR can improve as the size of a test sample increases from top to bottom, from 160 to 1,600 keystrokes.

2. The top seven lines in Figure 3 have higher average FAR values than the remaining ones. Moreover, their FAR values exhibit a U shape trend as reference profiles become larger. These seven lines are results for test sample sizes ranging from 160 to 1,120 keystrokes. This indicates that *holding the test sample size constant, increasing reference profile size does not always improve FAR. Test samples must be large enough for this to happen*, about 1,000 keystrokes for this case (between the 960 line and 1,120 lines in Figure 3).

The U shapes for the top seven lines in Figure 3, which are resulted from testing with smaller test samples, can be explained as follows. When the test sample and the reference profile are both too small, their feature values (duration for digraphs and trigraphs) occupy only a subset of the

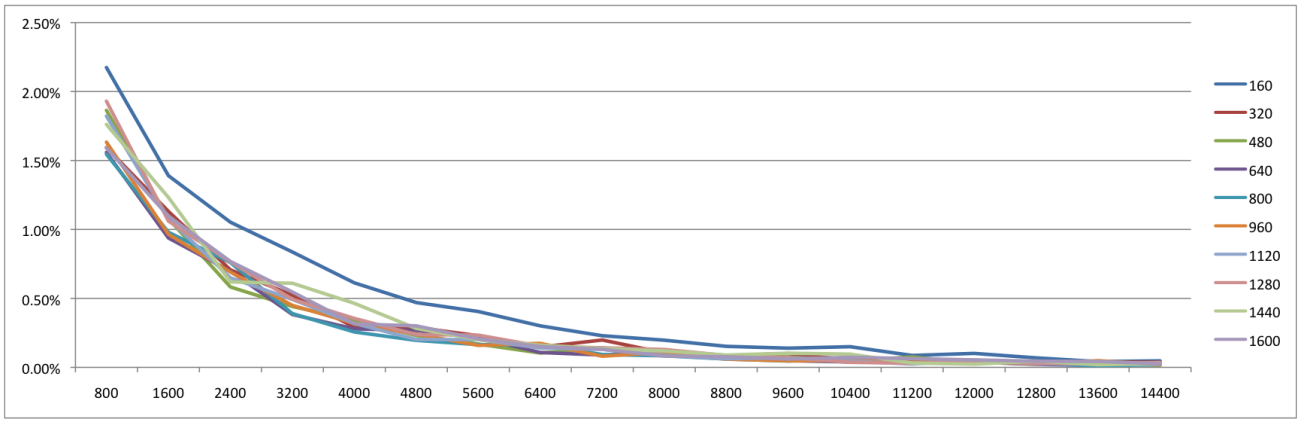

Figure 4. Effect of Reference Profile Size on Imposter Pass Rate (IPR). X Axis Represents Reference Profile Sizes. Y Axis Represents IPR Values. Each Line Represents the Results of Testing Test Samples of the Same Size, Ranging from 160 to 1,600 Keystrokes, against Profiles of Different Sizes, Ranging from 800 to 14,400 Keystrokes.

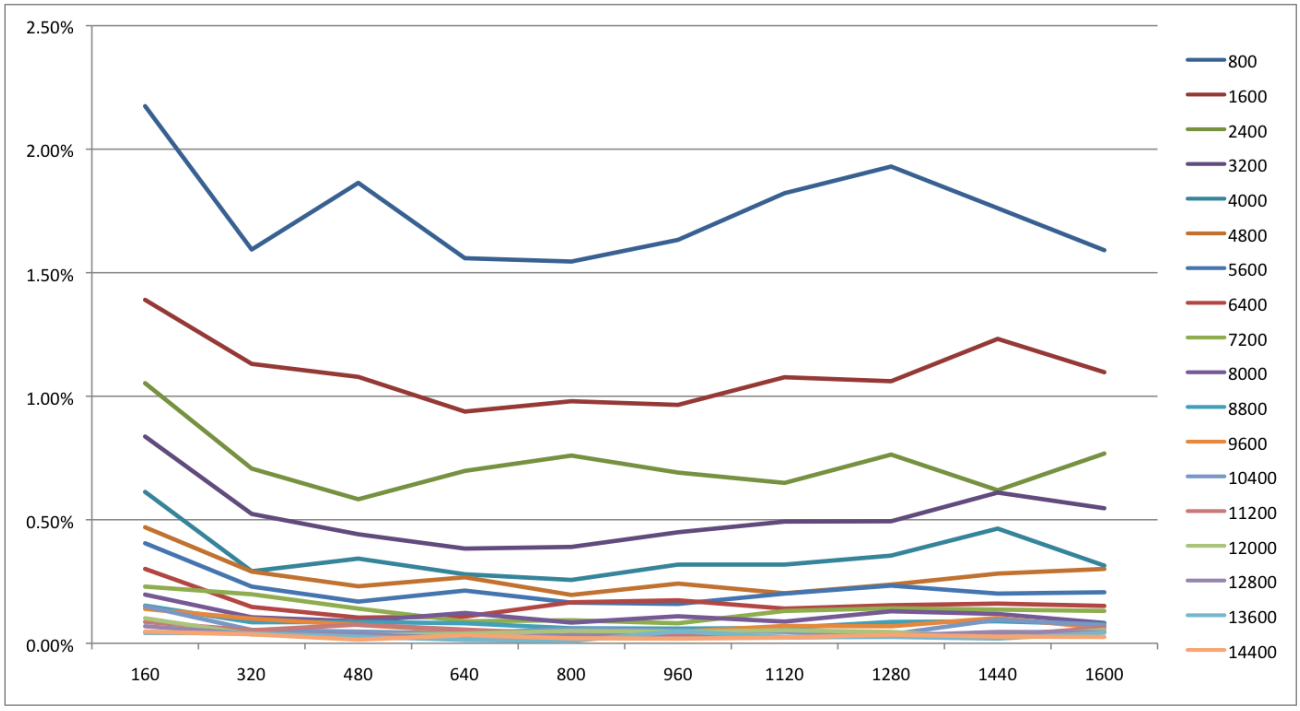

Figure 5. Effect of Test Sample Size on Imposter Pass Rate (IPR). X Axis Represents Test Sample Sizes. Y Axis Represents IPR Values. Each Line Represents the Results of Testing Reference Profiles of the Same Size, Ranging from 800 to 14,400 Keystrokes, against Test Samples of Different Sizes, Ranging from 160 to 1,600 Keystrokes.

user's complete feature space. As a result, it is more likely for them to come from different value ranges and be different. That is why FAR is high when both the test sample and the reference profile are too small, as shown by the blue line on the left in each line in Figure 3. When the size of the reference profile increases, it will share more feature values with the test samples, and thus, FAR gets better/smaller. But increasing the profile size further, beyond a certain point, will start to decrease its similarity with the test sample, resulting in an increase in FAR values.

3.2. Effect of Data Size on Imposter Pass Rate (IPR)

Based on Figure 4 and Figure 5, we observe the following points for IPR:

1. Figure 4 shows clearly that *the reference profile size can have a major effect on IPR. Larger reference pro-*

file will lead to smaller IPR values. This can also be seen from the 18 lines in Figure 5, which from top to bottom, exhibit the overall trend across all the lines that IPR can get better as the size of a reference profile increases, from 800 to 14,400 keystrokes.

2. However, the 18 lines in Figure 5 reveal no clear trend that increasing the test sample size will improve IPR. Indeed, when reference profiles are smaller than 6,400 keystrokes, we can see from Figure 5 that the lines exhibit a U shape or otherwise irregular shapes for their IPR values. A separate plot (not shown) created for those lines whose reference profiles larger than 6,400 reveals similar trends. This indicates that *increasing test sample size may not improve IPR*, unlike what the reference profile size does, as shown in Figure 4.

3.3. Selection of Data Sizes Based on Required FAR/IPR

Integrating the discussions in Sections 3.1 and 3.2, we conclude that

1. Larger reference profiles will drive down both IPR and FAR values, provided that the test samples are large enough.

2. Larger test samples have no obvious effect on IPR. This is true regardless the size of the reference profiles.

Furthermore, we recommend the following process for determining the appropriate sizes of a reference profile and a test sample, based on the required FAR and IPR values:

1. Figure 5 is convenient for selecting a minimal size for a reference profile, based on a given IPR value. For example, if it is necessary to have an IPR lower than 0.25%, then a reference profile of 6,400 or more keystrokes, is needed. On the other hand, we can see that a reference profile of 11,200 keystrokes can probably further drive down the IPR to below 0.05%.

2. Once a minimal reference profile size is selected, we can use Figure 2 to select the test sample size, based on a given FAR value. For example, if it is required to have an FAR lower than 10%, from Figure 2, we can see that the reference profile of size 6,400 selected in the last step, is not good since it will result in an FAR higher than 10%, regardless the test sample size. On the other hand, if the required FAR is 5%, the reference profile of size 11,200 selected in the last step, with a test sample of size 1,000, will be able to meet the required 5% for FAR.

3. A practical limit on the test sample size is the decision time of the authentication system, which is directly impacted by a user's typing speed. Requiring a larger test sample implies that the authentication system will need to wait longer between consecutive decisions. This is undesired since it may reduce the level of security of a system. For example, average computer users type 33 words, or 165 characters, per minute [3]. With a test sample of 1,000 keystrokes required, the authentication system must wait about six minutes before conducting the next authentication.

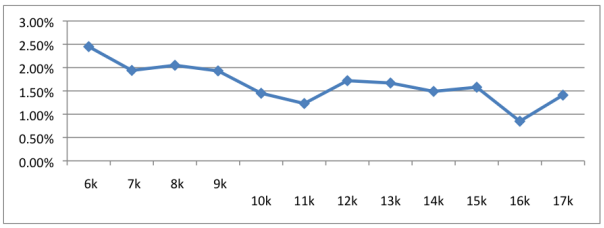

Figure 6. Effect of Reference Profile Size on FAR (False Alarm Rate). X Axis: Number of Keystrokes in Reference Profiles. Y Axis: FAR. Test Sample Size: 2,000 Keystrokes.

4. Results for Experiment 2

Our goal for Experiment 2 ("Zone-of-Acceptance" Algorithm) is to provide additional evidence to confirm some of the observations about the effect of data size that we obtain from Experiment 1. Indeed, we observe similar effect of reference profile sizes as in Experiment 1. Figure 6 and Figure 7 show the effects of reference profile sizes on FAR and IPR, respectively. Due to data limitations, we were only able to test up to a reference profile size of 17,000 digraph instances. From our results shown in Figure 6 and Figure 7, it can be noted that FAR and IPR continue to decrease as the reference profile size is increased from 6,000 to 17,000 keystrokes, with several fluctuations. Moreover, it appears that a reference profile of 10,000 or above is needed to achieve good authentication performance. These observations are consistent with the ones that we make in Experiment 1.

However, because the way our algorithm optimizes for its parameter values (Section 2.2) is very computation-intensive and time-consuming, also partly motivated by the observation from Experiment 1 about the selection of the test sample size, we test only for test sample sizes of 1,000 and 2,000 keystrokes. As a result, we will not discuss further about the effect of test sample sizes on authentication performance in Experiment 2.

5. Discussion

Gunetti-Picardi [2] and Leggett *et al.* [5] represent two different tradeoffs between FAR and IPR. By comparing

[3] `http://en.wikipedia.org/wiki/Words_per_minute`, last accessed 12/12/2014.

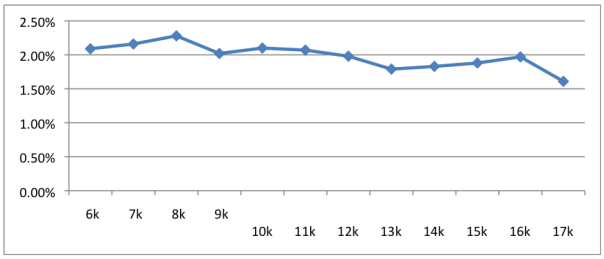

Figure 7. Effect of Reference Profile Size on IPR (Imposter Pass Rate). X Axis: Number of Keystrokes in Reference Profiles. Y Axis: IPR. Test Sample Size: 2,000 Keystrokes.

Figure 2 and Figure 6, we can see that Gunetti-Picardi [2] is generally more lenient on FAR than Leggett *et al*. [5]. On the other hand, comparing Figure 4 and Figure 7, we can see that Gunetti-Picardi [2] is more strict in terms of IPR, thus better optimized for security applications than Leggett *et al*. [5].

We find that larger reference profiles will drive down both IPR and FAR values, provided that the test samples are large enough. On the other hand, larger test samples have no obvious effect on IPR, and this is true regardless of the size of the reference profiles. These are good news for practical deployment of keystroke authentication, since in practical settings, it should be easy to create a reference profile with several tens of thousands of keystrokes, if not more. Furthermore, for test samples, we have mainly looked at 1,000 keystrokes, which on average still requires six minutes of typing to create such as a test sample. So more work is needed to investigate whether smaller test samples would still work well enough.

In Experiment 1, we create a reference profile and a test sample by randomly picking $2,3$-graphs instances, without preserving their temporal order. Intuitively, timing is likely an important factor, i.e., times are not "independent." In other words, if you at one time are typing to your boss and another your spouse, your patterns may be different. But since the algorithm currently does not utilize such temporal information anyway, we believe that our approach is justified for the purpose of this study. In a subsequent experiment, we preserve the timing information in reference profiles and test samples, but we observe no obvious difference in the results.

6. Conclusion

To investigate the effect of data size on the performance of free-text keystroke authentication, we conduct two experiments, using two existing algorithms and a public dataset. By systematically varying the sizes of the reference profile and the test sample, we find that increasing reference profiles will drive down both IPR and FAR values, provided that the test samples are large enough; On the other hand, larger test samples have no obvious effect on IPR, and this is true regardless of the size of the reference profiles. These are good news for practical deployment of keystroke authentication.

Future work can be devoted to replicating our study on larger datasets, and with additional algorithms, to ensure the external validity of our findings. Recently, the biometrics community has started to systematically design datasets [6] and apply statistical methods to more rigorously evaluate performance [7] [9] [4]. Statistical methods can rigorously quantify and control the parameters involved in such studies and results (e.g., ensuring an adequate number of subjects and samples are included in a dataset, and considering conditions such as the types of keyboard used and a subject's mood). In addition to the existing public dataset [10], we are currently establishing a new one, which can benefit from incorporating these methods to guide our data collection.

Acknowledgments

This work is partially supported by United States National Science Foundation under Award No. CNS-1314792.

References

[1] S. D. Connell and A. K. Jain. Template-based online character recognition. *Pattern Recognition*, 34(1):1–14, 2001.

[2] D. Gunetti and C. Picardi. Keystroke analysis of free text. *ACM Trans. Inf. Syst. Secur.*, 8(3):312–347, Aug. 2005.

[3] A. Jain, R. Bolle, and S. Pankanti, editors. *Biometrics: Personal Identification in Networked Society*. Kluwer Academic Publishers, 1999.

[4] K. Killourhy and R. Maxion. Why did my detector do that?!: Predicting keystroke-dynamics error rates. In *Proceedings of the 13th International Conference on Recent Advances in Intrusion Detection*, pages 256–276. Springer-Verlag, 2010.

[5] J. Leggett, G. Williams, M. Usnick, and M. Longnecker. Dynamic identity verification via keystroke characteristics. *Int. J. Man-Mach. Stud.*, 35(6):859–870, Nov. 1991.

[6] S. Sarkar and Z. Liu. Evaluation of gait recognition. In S. Z. Li and A. Jain, editors, *Encyclopedia of Biometrics*, pages 281–289. Springer US, 2009.

[7] M. Schuckers. *Computational Methods in Biometric Authentication: Statistical Methods for Performance Evaluation*. Springer, 2010.

[8] U. Uludaga, A. Ross, and A. Jain. Biometric template selection and update: a case study in fingerprints. *Pattern Recognition*, 37:1533–1542, 2004.

[9] G. Veres, M. Nixon, and J. Carter. Is enough enough? what is sufficiency in biometric data? In A. Campilho and M. Kamel, editors, *Image Analysis and Recognition*, volume 4142 of *Lecture Notes in Computer Science*, pages 262–273. Springer Berlin Heidelberg, 2006.

[10] E. Vural, J. Huang, D. Hou, and S. Schuckers. Shared research dataset to support development of keystroke authentication. In *IJCB*, pages 1–8, Sept 2014.