A study on user-specific threshold configuration for keystroke dynamics in the context of adaptive biometric systems

Paulo Henrique Pisani[1]

[1]Centro de Matemática, Computação e Cognição (CMCC)
Universidade Federal do ABC (UFABC) – Santo André – SP – Brasil

paulo.pisani@ufabc.edu.br

Abstract. *Keystroke dynamics is a biometric modality which can be applied as an additional authentication factor. Some studies have shown that keystroke data can change over time and, consequently, a biometric system which does not update the biometric reference acquired at enrolment time may face performance degradation. Adaptive biometric systems can be applied in this case and automatically adapt the biometric reference. An important aspect of these systems is how the thresholds are defined. Thresholds can be used for classification and to define which samples will be used for adaptation. A few studies have worked on how to adapt the thresholds. This paper studies user-specific thresholds for keystroke dynamics in adaptive biometric systems.*

1. Introduction

Several services are available online nowadays. In this scenario, user authentication is a key aspect. A simple password may not provide enough security since it is prone to different types of attacks such as guessing, shoulder surfing, etc [Sae-Bae and Memon 2022]. An alternative to deal with these problems can be the adoption of additional authentication factors, such as *two-factor authentication* (2FA) and *multi-factor authentication* (MFA) [AlQahtani et al. 2021]. In addition to the password, keystroke dynamics has been proposed as a second authentication factor [Sae-Bae and Memon 2022]. Keystroke dynamics recognizes users by their typing rhythm [Killourhy and Maxion 2009].

A biometric system which implements keystroke dynamics captures keystroke samples from the genuine user at enrolment time to obtain a biometric reference. At recognition time, the biometric system receives query samples and compares them with the biometric reference. This comparison can output a score, which is then compared to a threshold to define whether the query sample will be classified as genuine or impostor. The biometric system described here considers the verification operating mode, which will be the focus of the current paper. In the literature, the threshold can be defined globally, by adopting a common fixed threshold for all users in the biometric system, or it can be defined individually (user-specific), when each user can adopt a different threshold [Giot et al. 2011b, Mhenni et al. 2019].

Keystroke dynamics is subject to changes over time and, consequently, the predictive performance of a biometric system which implements this biometric modality can decrease over time if the biometric reference is not updated [Giot et al. 2011a]. It means that the biometric reference obtained at enrolment time may become outdated and not properly represent current keystroke data of the user. These changes on the biometric data over time are also known as template ageing [Jain et al. 2016].

Adaptive biometric systems which can automatically update the biometric reference over time can handle template ageing [Ryu et al. 2023]. Some studies have studied alternatives to update the biometric reference in keystroke dynamics [Kang et al. 2007, Giot et al. 2011a, Giot et al. 2012]. These studies work by managing a gallery of biometric samples which are used to recompute the biometric reference over time.

Although the proposed adaptation strategies can increase the predictive performance, the study on how to also adapt the threshold has not been extensively explored. A method to adapt the threshold has been proposed in [Hosseinzadeh and Krishnan 2008]. The proposed method uses leave one out method (LOOM) to obtain the threshold and updates it every time the user is authenticated. Previous work from [Mhenni et al. 2016, Mhenni et al. 2019] has also investigated threshold adaptation, showing that it can increase the performance of an adaptive biometric system. Overall, both papers [Mhenni et al. 2016, Mhenni et al. 2019] proposed to make the threshold more stringent over time depending on some user parameters.

Papers which deal with keystroke dynamics frequently report results in terms of EER (equal error rate) [Roy et al. 2022]. In order to obtain the EER, the threshold value can be adjusted until false rejection and false acceptance rates are equal. Nevertheless, this method requires access to true labels of testing data and these labels may not be available in a practical application scenario. Moreover, a previous paper mentioned that reporting EER may not be a suitable choice for biometric systems with template update [Giot et al. 2012]. For these reasons, the current study does not report results in terms of EER and do not use this method to obtain the thresholds.

The current paper presents a new investigation on user-specific threshold configuration for keystroke dynamics in adaptive biometric systems, in which the biometric reference can be automatically updated over time. Two adaptation strategies were considered during the experiments: Growing and Moving/Sliding window [Kang et al. 2007, Giot et al. 2011a, Giot et al. 2012]. Along with the adaptation process, an study on user-specific threshold configuration is performed. Two methods to obtain the threshold from the user gallery are evaluated. There are two key positive aspects of these methods. Firstly, they only require the biometric samples from the user gallery to be computed. Secondly, they are not as costly as using LOOM. The experiments were performed on two public datasets, including one recently published. The remaining of the paper is organized as follows: Section 2 describes the evaluation methodology, including the evaluated threshold configuration; Section 3 reports and discusses the obtained results; and, Section 4 presents the conclusion and future work.

2. Evaluation methodology

This section describes the evaluation methodology adopted in the current study.

2.1. Datasets, enrolment and biometric data stream

This paper used two keystroke dynamics datasets: CMU [Killourhy and Maxion 2009] and KeyRecs (only the fixed text part) [Dias et al. 2023]. The CMU[1] dataset contains data from 51 users, while KeyRecs[2] contains data from 99 users.

[1] `https://www.cs.cmu.edu/~keystroke/`

[2] `https://zenodo.org/records/7886743`

For each user, the first 50 samples were used for enrolment (training). It corresponds to the first session in the CMU dataset, and the first 50 samples of the first session in the KeyRecs dataset. The remaining samples were used to generate a test biometric data stream, which were presented in the same order they were stored in the dataset. Among the genuine biometric samples, some impostor samples were randomly introduced. Impostor samples were chosen randomly from the other users in the dataset. The generation of the test sequence followed a protocol similar to the one used to generate a pool in [Giot et al. 2013]. In the current study, the test biometric data stream contains 70% of genuine samples and 30% of impostor samples.

Regarding adaptation, three cases were considered in the experiments: no adaptation, *Growing* window and *Sliding/Moving* window [Kang et al. 2007, Giot et al. 2011a, Giot et al. 2012]. *Growing* adds every sample recognized as genuine to the user gallery and recomputes the biometric reference afterwards. *Sliding* works in a similar way, but it also removes the older sample from the gallery.

2.2. Score computation and threshold configuration

The algorithm from [Magalhães et al. 2005] was used to obtain the biometric reference and compute similarity scores. The biometric reference of the user j will be defined as ref_j in this paper. The experiments performed in this work adopted an approach in which the biometric system only has access to the data from the genuine user to define the threshold. It means that the system only has access to G_j, which is the genuine gallery of the user j. At enrolment time, the gallery contains only enrolment samples from the genuine user. In this scenario, the biometric reference ref_j is computed based on G_j. In an adaptive biometric system, this gallery can be updated later, depending on the adaptation strategy [Kang et al. 2007, Giot et al. 2011a, Giot et al. 2012].

A possible method to obtain the threshold is to compute the scores that each gallery sample in G_j obtains when using ref_j. It results in a set of scores S_j. A simple method is to define the threshold is $min(S_j)$. This method will be named as *min* throughout the paper. By adopting this method, all samples in the gallery would be correctly matched, meaning zero false non-matches for these samples. However, outliers may lead to a threshold value which is too low and, consequently, false matches may also increase. In order to deal with it, outliers can be removed, similarly to [Hosseinzadeh and Krishnan 2008]. This method, named as *min-non-outlier* here, computes the threshold as $min(S_j^n)$. S_j^n is the set of scores from S_j which are higher than $mean(S_j) - 2.5 \times std(S_j)$, where $mean$ and std are the mean and the standard deviation, respectively. The choice of 2.5 standard deviations was based on the value adopted in [Hosseinzadeh and Krishnan 2008], which presented a similar procedure.

During the experiments, the adaptive strategies were applied using a *fixed threshold* and a *dynamic threshold* configuration. In the *fixed threshold* configuration, the threshold is computed at enrolment time, using either *min* or *min-non-outlier*, and the threshold remains fixed during the test biometric data stream. Conversely, in the *dynamic threshold* configuration, the threshold does not remain fixed over time. Each time the gallery is updated, the threshold is also recomputed using the same method applied at training time (*min* or *min-non-outlier*). The dynamic configuration is similar to the LOOM threshold from [Hosseinzadeh and Krishnan 2008], although no leave-one-out is performed here, decreasing computational cost.

3. Results and discussion

This section discusses the results obtained in the experiments. Firstly, the overall results are shown and then an evaluation of genuine scores over time is presented.

3.1. Overall performance

The results were reported in terms of false match rate (FMR) and false non-match rate (FNMR) [Precise Biometrics 2014, Mhenni et al. 2019]. FMR measures the rate in which the impostor samples are wrongly classified as genuine, and FNMR measures the rate in which the genuine samples are wrongly classified as impostor. Balanced accuracy was also reported and is defined as $1 - \frac{(FMR+FNMR)}{2}$. In general, the *min-non-outlier* method reached better balanced accuracy, which is a result of an improvement over the FMR. The higher threshold of the *min-non-outlier* compared to the *min* method contributed to this result. Due to these results, only the *min-non-outlier* method will be further studied. The overall performance is shown in Table 1 (average from 30 runs).

Table 1. Overall performance for both datasets. Standard deviation is shown between parenthesis and the best results are highlighted in bold.

	CMU			KeyRecs		
Adaptation strategy	$BAcc$	FMR	$FNMR$	$BAcc$	FMR	$FNMR$
No adaptation	0.774 (0.120)	0.238 (0.213)	0.214 (0.213)	0.671 (0.124)	0.613 (0.248)	0.045 (0.090)
Growing (fixed threshold)	0.824 (0.116)	0.312 (0.244)	0.040 (0.056)	0.626 (0.122)	0.733 (0.234)	**0.016** (0.064)
Sliding (fixed threshold)	0.881 (0.088)	0.140 (0.184)	0.097 (0.113)	0.709 (0.130)	0.547 (0.263)	0.036 (0.074)
Growing (dynamic threshold)	0.817 (0.114)	0.336 (0.232)	**0.029** (0.038)	0.634 (0.118)	0.715 (0.226)	0.018 (0.065)
Sliding (dynamic threshold)	**0.889** (0.066)	**0.110** (0.152)	0.112 (0.063)	**0.762** (0.122)	**0.415** (0.246)	0.060 (0.088)

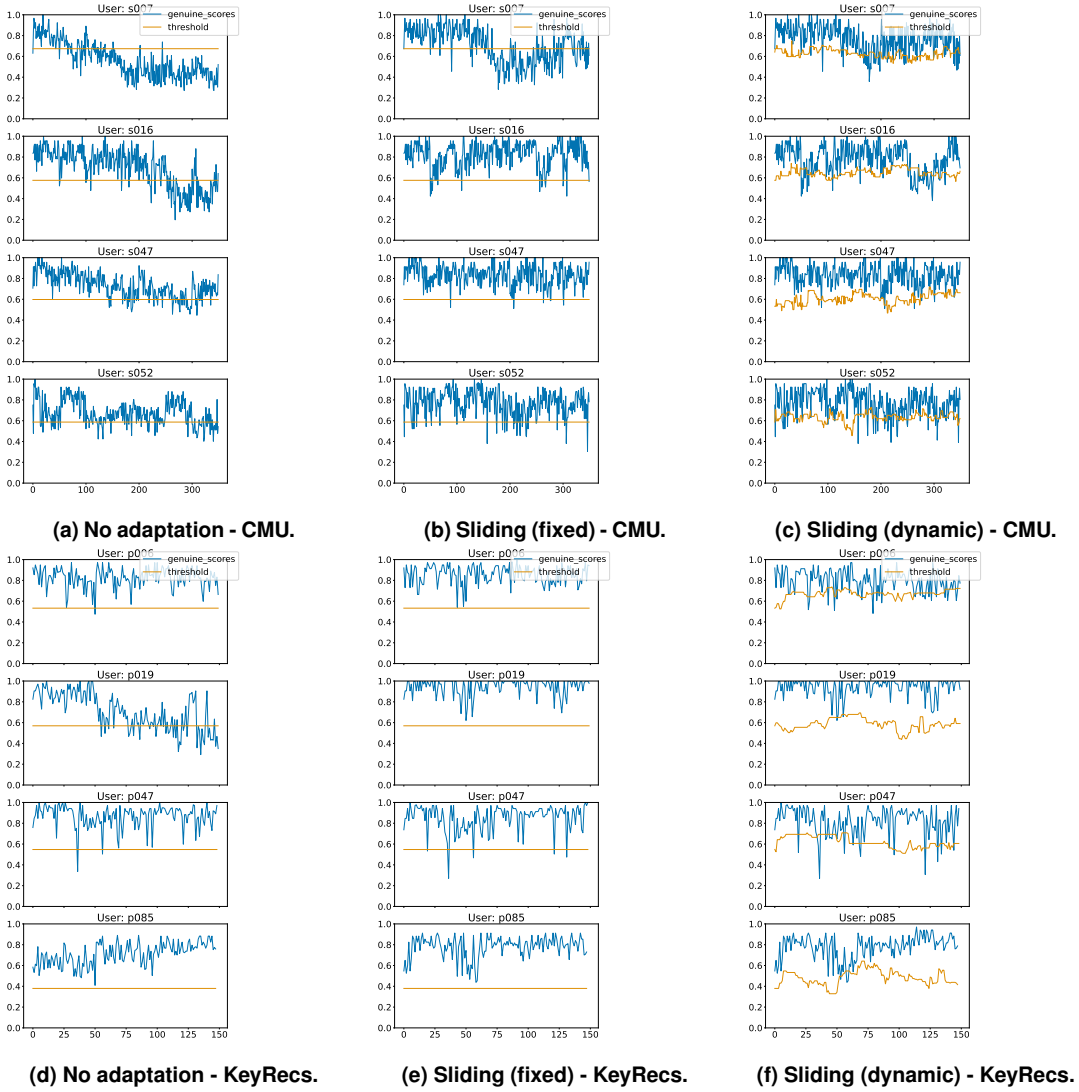
The results in CMU and KeyRecs datasets were different by adopting a dynamic threshold. Overall, the balanced accuracy was higher using a dynamic threshold. In CMU, however, Growing is an exception. It obtained higher balanced accuracy by using a fixed threshold. These results illustrates that adapting the threshold over time can be a promising approach for the Sliding adaptation strategy.

Finally, it can be observed that adaptation strategies can result in higher balanced accuracy compared to not using any adaptation strategy. An exception occurred in the KeyRecs dataset, in which Growing results in lower performance regardless of the threshold configuration. Between Growing and Sliding, usually, FNMR values are lower for Growing and FMR are lower for Sliding.

3.2. Genuine scores and threshold

After evaluating the overall performance, this section deepens the study by observing the genuine scores and threshold values over time. Figure 1 contains the genuine scores and the threshold value over time of the first run. The genuine scores are the scores obtained by the verification performed on the true genuine samples only. The impostor samples were not considered in the plots of this section. These figures focus on the Sliding adaptation strategy, which obtained the best results on both datasets. The *no adaptation* is also illustrated for comparison.

In the CMU dataset, among the 51 users, four of them were selected since they present different behaviours over time. In these plots, if the blue points (genuine scores) are above the orange curve (threshold value), it means that the biometric sample was correctly classified as genuine. Firstly, in general, the use of the adaptation strategy resulted



(a) No adaptation - CMU. **(b) Sliding (fixed) - CMU.** **(c) Sliding (dynamic) - CMU.**

(d) No adaptation - KeyRecs. **(e) Sliding (fixed) - KeyRecs.** **(f) Sliding (dynamic) - KeyRecs.**

Figure 1. Genuine scores and threshold over time for both datasets.

in higher scores, particularly in later moments of the biometric data stream. It illustrates that adaptation can improve predictive performance by decreasing $FNMR$.

Regarding the threshold, the plots show that it changes in different ways over time. For example, for users s016, s047 and s052, the dynamic threshold value were higher than the fixed value in several cases. The threshold increase for user s047 was still below most genuine scores. However, for user s016, the higher threshold results in many false non-matches in the last part of the biometric data stream. Some false non-matches also occurred for user s052. Conversely, the threshold decreased for user s007 compared to the initial fixed threshold. The lower threshold improved the adaptation performance, reducing false non-matches. It can be observed by comparing the results of the dynamic threshold and the other approaches with fixed threshold. By reducing false non-matches, a higher number of true genuine samples can be used to update the gallery, which can result in a better biometric reference over time.

In the KeyRecs dataset, among the 99 users, four users were selected. Three of

these users (p006, p047 and p085) showed a tendency to obtain increased threshold values over time, while one (p019) presented a different tendency. In general, the higher threshold values were still below the genuine scores. It can potentially result in lower FMR while not substantially increasing the $FNMR$. However, some users may have been negatively impacted by a higher threshold. For example, the higher threshold value of user p006 when using dynamic threshold for Sliding, compared to the fixed configuration, resulted in some false non-matches.

The average threshold value considering all biometric samples (and all runs) were higher for the dynamic threshold approach in Sliding compared to the fixed threshold (on both datasets). It illustrates that the dynamic threshold can potentially decrease the FMR by adopting a higher threshold for Sliding.

4. Conclusion and future work

Adaptive biometrics systems can automatically update the biometric reference, avoiding or decreasing predictive performance loss over time. In this context, several adaptation strategies involve managing a gallery, such as *Growing* window and *Sliding* window. Apart from adapting the gallery, the threshold is also a component which can be updated over time. Some previous papers have discussed threshold adaptation for keystroke dynamics [Hosseinzadeh and Krishnan 2008, Mhenni et al. 2016, Mhenni et al. 2019].

The current study contributes by studying some methods to update the threshold value of each user and how they impact the performance over time. This study involved the evaluation of different adaptation strategies (*Growing window* and *Sliding window*) and two threshold approaches: fixed and dynamic. According to the reported results, each adaptation strategy has a different behaviour while using each threshold approach. *Sliding window* attained the best results on both datasets. This adaptation strategy also was the one which most benefited from the dynamic threshold approach using *min-non-outlier*. The possibility of assigning a lower threshold in some cases seems to have improved the performance for some users, while the threshold increased over time for other users.

Some limitations of the current study can be explored in future work. For instance, this study considered two methods to define the initial threshold: *min* and *min-non-outlier*. However, other methods could be investigated in future studies. Moreover, additional algorithms to compute scores may also be studied and how different threshold configuration approaches impact the predictive performance. Other datasets or biometric modalities could also be considered, along with different methods to generate the biometric data stream. Keystroke dynamics in the forensic context may also be part of additional studies. Future work can also investigate the challenges of deploying the studied methods in real-world scenarios, including how adaptation strategies contribute to FMR and FNMR over time.

References

AlQahtani, A. A. S., El-Awadi, Z., and Min, M. (2021). A survey on user authentication factors. In *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0323–0328.

Dias, T., ao Vitorino, J., Maia, E., Sousa, O., and Praça, I. (2023). KeyRecs: A keystroke dynamics and typing pattern recognition dataset. *Data in Brief*, 50:109509.

Giot, R., Dorizzi, B., and Rosenberger, C. (2011a). Analysis of template update strategies for keystroke dynamics. In *2011 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, pages 21–28.

Giot, R., El-Abed, M., Hemery, B., and Rosenberger, C. (2011b). Unconstrained keystroke dynamics authentication with shared secret. *Computers & Security*, 30(6):427 – 445.

Giot, R., Rosenberger, C., and Dorizzi, B. (2012). Hybrid template update system for unimodal biometric systems. In *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE.

Giot, R., Rosenberger, C., and Dorizzi, B. (2013). A new protocol to evaluate the resistance of template update systems against zero-effort attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Hosseinzadeh, D. and Krishnan, S. (2008). Gaussian mixture modeling of keystroke patterns for biometric applications. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 38(6):816–826.

Jain, A. K., Nandakumar, K., and Ross, A. (2016). 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recog. Letters*, 79:80 – 105.

Kang, P., Hwang, S.-s., and Cho, S. (2007). Continual retraining of keystroke dynamics based authenticator. In Lee, S.-W. and Li, S. Z., editors, *Advances in Biometrics*, pages 1203–1211, Berlin, Heidelberg. Springer Berlin Heidelberg.

Killourhy, K. S. and Maxion, R. A. (2009). Comparing anomaly-detection algorithms for keystroke dynamics. *Proceedings of the International Conference on Dependable Systems and Networks*, pages 125–134.

Magalhães, S. T., Revett, K., and Santos, H. M. D. (2005). Password secured sites - stepping forward with keystroke dynamics. In *International Conference on Next Generation Web Services Practices (NWeSP'05)*, page 293–298.

Mhenni, A., Cherrier, E., Rosenberger, C., and Essoukri Ben Amara, N. (2019). Double serial adaptation mechanism for keystroke dynamics authentication based on a single password. *Computers & Security*, 83:151–166.

Mhenni, A., Rosenberger, C., Cherrier, E., and Ben Amara, N. E. (2016). Keystroke template update with adapted thresholds. In *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 483–488.

Precise Biometrics (2014). Understanding biometric performance evaluation.

Roy, S., Pradhan, J., Kumar, A., Adhikary, D. R. D., Roy, U., Sinha, D., and Pal, R. K. (2022). A systematic literature review on latest keystroke dynamics based models. *IEEE Access*, 10:92192–92236.

Ryu, R., Yeom, S., Herbert, D., and Dermoudy, J. (2023). The design and evaluation of adaptive biometric authentication systems: Current status, challenges and future direction. *ICT Express*, 9(6):1183–1197.

Sae-Bae, N. and Memon, N. (2022). Distinguishability of keystroke dynamic template. *PLOS ONE*, 17(1):1–17.