Analysis of the Acquisition Process for Keystroke Dynamics

Romain Giot and Alexandre Ninassi and Mohamad El-Abed and Christophe Rosenberger
Université de Caen, UMR 6072 GREYC
ENSICAEN, UMR 6072 GREYC
CNRS, UMR 6072 GREYC
{*romain.giot,alexandre.ninassi,mohamad.elabed,christophe.rosenberger*}*@ensicaen.fr*

Abstract: In order to evaluate authentication methods for keystroke dynamics, it is necessary to create new datasets. We present in this paper an analysis study of the factors, involved during the acquisition process, which affect the performance of keystroke-based authentication systems. More generally speaking, we are looking for the optimal keystroke data acquisition scenario which may deeply affect the performance of such systems.

Results show that (1) it is better to choose passwords based on short, simple and known words; and (2) during the acquisition process of a new keystroke dynamics dataset, there is no significant preference for displaying or telling to a user the text to type.

1 Introduction

Keystroke-based authentication systems are considered as a promising solution to be used for web services logical access control such as e-commerce or betting applications [Bou09]. In comparison to morphological-based solutions (*e.g.*, face or fingerprint), keystroke-based authentication systems have two main advantages: 1) keystroke dynamics is a low cost solution: no additional sensor is required; and 2) keystroke dynamics is a non intrusive solution: as shown in previous works [GEAR12], keystroke-based authentication solution is acceptable by users. However, the performance of such systems is not as efficient as other modalities such as fingerprint or facial authentication systems. Therefore, enhancing the performance of keystroke-based authentication systems is still an important challenge in order to use this modality in real life applications. In order to evaluate keystroke dynamics methods, it is necessary to create new keystroke dynamics datasets. These datasets are used in an offline way, to compute the biometric reference of each individual and the comparison scores by testing test samples to these biometric references. Thanks to these test samples, we compute the performance of the implemented method on the dataset. However, the recognition performance of one method varies depending on the datasets [KM11]. The reasons of the variations can be (i) because of the difference in the population of the various datasets, (ii) or because of the way we ask to the individuals to type the required password in the acquisition tool of the dataset. We are interested by the second case.

Towards this goal, we present in this paper an analysis study of the factors involved during the acquisition process which affect the performance of keystroke-based authentication systems. More generally speaking, we are looking the optimal keystroke data acquisition scenario which may deeply affect the performance of such systems. This paper is original because, so far to our knowledge, there is no work in the literature which analyses the relationship between the recognition performance and the way the biometric sample is acquired. The results obtained in this work will help the future creation of keystroke dynamics dataset, or the proposal of new authentication systems based on the typing of spontaneously generated passwords by the authentication system.

Section 2 presents the various datasets used in keystroke dynamics in the literature and their variations. Section 3 presents the protocol used in this study. Section 4 presents the results of the study, while section 5 concludes this paper.

2 Keystroke Dynamics Datasets

Various dataset exist in the literature. Here is a short description of the public ones. Montalvão *et al.* has used the same keystroke databases in several papers [FF06]. The maximum number of users in a database is 15, and the number of provided samples per user is 10. Each database contains the raw data. The database is composed of couples of ASCII codes of the pressed key and the elapsed time since the last key down event. Release of a key is not tracked. Four different databases have been created. Most of them were built under two different sessions spaced of one week or one month (depending on the database). Each database is stored in raw text files. The databases are available at:

`http://itabi.infonet.com.br/biochaves/en/download.htm`

Killhourhy and Maxion propose a database of 51 users providing four hundred samples captured in eight sessions (there are fifty inputs per session) [KM09]. The delay between each session is one one day at minimum, but the mean value is not stated. This is the dataset having the most number of samples per user, but, a lot of them are typed on a short period (50 at the same time). Each biometric data has been captured when typing the following password: ".tie5Roanl". The database contains some extracted features: hold time, interval between two pressures, interval between the release of a key, and the pressure of the next one. It is stored in raw text, csv or Excel files. The database is available at:

`http://www.cs.cmu.edu/~keystroke/`

Giot *et al.* propose the most important public dataset from the literature if considering the number of users. It contains 133 users and, 100 of them provided samples of, at least, five distinct sessions [GEAR09]. Each user typed the password "greyc laboratory" twelve times, on two distinct keyboards, during each session (which give 60 samples for the 100 users having participated to each session). Both extracted features (hold time and latencies) and raw data are available (which allow to build other extracted features). It is stored in an sqlite database file. The database is available at:

http://www.ecole.ensicaen.fr/~rosenber/keystroke.html

Allen has created a public keystroke dynamics database using a pressure sensitive keyboard [All10]. It embeds the following raw data: key code, time when pressed, time when release, pressure force. 104 users are present on the database, but, only 7 of them provided a significant amount of data (between 89 and 504), whereas the 97 other have only provided between 3 and 15 samples. Three different passwords have been typed: ”pr7q1z”, ”jeffrey allen” and ”drizzle”. The database is available in a csv or sql file at:

http://jdadesign.net/2010/04/
pressure-sensitive-keystroke-dynamics-dataset/

We have seen that several databases for static password authentication with keystroke dynamics are publicly available. Although, it would be the best kind of dataset, no public dataset has been built with one different couple login/password for each user. Table 1 presents a summary of these public datasets.

Dataset	Type	Information	Users	Samples /users	Sessions
[FF06]	Various	Press events	< 15	< 10	2
[KM09]	1 fixed string	Duration and 2 latencies	51	400	8
[GEAR09]	1 fixed String	Press and release events. Duration and 3 latencies	> 100	60	5
[All10]	3 fixed strings	Press and release events and pressure	7/97	(89-504) /(3-15)	few months

Table 1: Summary of keystroke dynamics datasets

Although, several databases have been publicly proposed, there is no explanations on the way how the text has been proposed to the volunteer. We can suppose that 100% of the acquisition softwares write the text on the screen not far away of the input field. We think this is important information to provide. In our investigations (on keystroke dynamics authentication on tactile smartphones), we encountered bad performances because of the way we present the text to type for a user. The next section presents the protocol we used to analyse the impact of the way of presenting the text on the recognition performances.

3 Experimental Protocol

The objective of this protocol is to acquire a dataset for keystroke dynamics which follows different acquisition scenarios. The aim of this dataset is to compare the different scenarios as well as their impact on the performance of keystroke-based authentication systems. Thanks to this study, two advances are expected:

1. The creation of less noisy datasets (abnormal timing delays or too much stable patterns in comparison to real world data) by avoiding unrealistic acquisition conditions.
2. The selection of the most appropriate acquisition method to maximize the performance of keystroke-based authentication systems.

3.1 Variable data

Several parameters can vary in an authentication system based on keystroke dynamics.

3.1.1 Different kinds of password

We can identify different types of passwords:

- The passwords with words or sentences which can be separated in different subcategories:
  - *Known words*: It is assumed that the volunteer knows this word and has probably typed it before. Such passwords must be quite easy to remember.
  - *Unknown words*: These words are assumed to be unknown to the volunteer, or even randomly generated ; they can also use symbols or numbers instead of some letters. Such passwords must be quite hard to remember.
- The passwords based on numbers for which we can also find two different subcategories:
  - *Structured numbers*: They represent information not necessary known, but structured according to a known format (phone number, credit card numbers, etc.). These numbers are assumed to have a structure that makes them easier to memorize.
  - *Unstructured numbers*: These numbers can represent anything and have any size. They can be difficult to memorize.

These different kinds of passwords may be thought to lead to different typing difficulty.

3.1.2 Different ways to ask a password to type

Each kind of password can be presented with different ways of presentation. These ways of asking to type the text can depend on the kind of text to type. We can list the following ones:

- Common ways of presentation, whatever is the text to type:

  - Display it in the graphical user interface (can be read when typing).
  - Display it in a modal window[1] before typing (it requires memorization).
  - Display it with different graphical presentations (font sizes, font colors, etc.).
- Specific ways of presentation of textual data:
  - Listening to the pronunciation of the words.
  - Listening to the spelling of the words.
- Specific ways of presentation for numerical data:
  - Display the number by sets of L digits (*i.e ddd ddd ddd* with d representing a number when number displayed by sets of 3).
  - Listening to the pronunciation of the numbers.
  - Listening to the pronunciation of the numbers by packet of L digits.

3.2 Acquisition constraints

Data are acquired by means of a dedicated application developed for this study (figure 1). The basic principles of the application is to present a password (word, number) and to record keystroke dynamic information of volunteers. The way to present the password is the varying factor of this study. To avoid dependencies between results and passwords, several passwords are tested for each way of presentation. Each password must be entered several times to reduce the risk to have noisy data. Data acquisition should be performed on a keyboard with a numeric keypad.

Let M be the set of passwords selected for the experiment, M_c the set of passwords supposed beings known, and M_i the set of passwords supposed to be unknown.

$$M = M_i \cup M_c \tag{1}$$

Let N denotes the set of numbers selected for the experiment, N_s the set of structured numbers, and N_n the set of unstructured numbers.

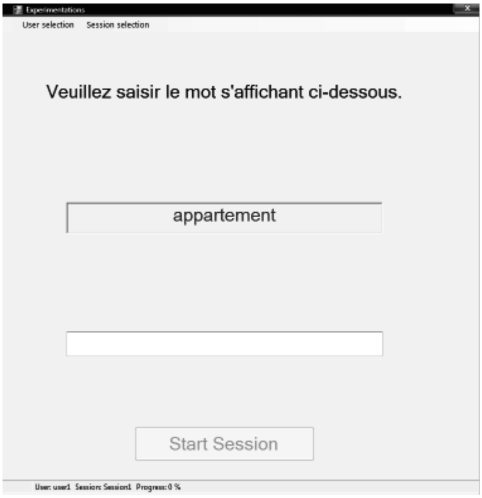
$$N = N_s \cup N_n \tag{2}$$

The data acquisition must be made in several sessions. The organization of the sessions should be as follows:

- Each session at time i must be identical for each user ($S_i^{u_a} = S_i^{u_b}, \forall u_a, u_b, i$), so they can be represented as S_i, with i the session number:

$$S_i = \{E_i^1, E_i^2, \ldots, E_i^{\#}\} \tag{3}$$

[1] a modal window is a child window that requires users to interact with it before they can return to operating the parent application



Veuillez saisir le mot s'affichant ci-dessous.

appartement

User: user1 Session: Session1 Progress: 0 %

Figure 1: Screenshot of the data acquisition application.

where E_i^k is the kth event to be recorded in the ith session.

An event is characterized by a pair: (i) the information to type, and (ii) the way to present this information.

$$E_i^k = \{I_i^k, P_i^k\}, \left\{ \begin{array}{c} I_i^k \in M \cup N \\ P_i^k \in \mathcal{P}(I_i^k) \end{array} \right. \tag{4}$$

where $\mathcal{P}(password)$ the set of ways to present a password for a specific type.

Each event in a session must be presented at least F times.

$$\forall k, \sum_l \mathbb{1}\{E_i^k = E_i^l\} \geq F \tag{5}$$

- Each session S_i must be unique (they are different from the others), to limit the habituation factor ($S_{i_a} \neq S_{i_b}, \forall i_a, \forall i_b, i_a \neq i_b$).

$$\forall_{i,j,j \neq i} \exists_k \quad E_i^k \neq E_j^k \tag{6}$$

- As it is not possible to be exhaustive in an unique session (because of the time it would take), each possible pair must appear in at least two sessions of all the sessions.

3.3 Selected Acquisition Protocol

The following protocol has been applied from the constraints previously specified. For obvious reasons of feasibility, it cannot be exhaustive, and not all the combinations of the points presented before have been used. 28 volunteers participated for the experiment during 2 sessions. The minimum delay between the acquisition of the sessions is of 1 week. The targeted machine is a laptop. The Text is typed on an AZERTY keyboard and we asked participant to type number with the numerical keyboard. We have selected 3 passwords by type of complexity of password and there are 10 presentations of each combination of password. The words selected are:

- Known words:
  MC1 "voiture"
  MC2 "poisson"
  MC3 "appartement"
- Unknown words (they are based on anagrams of the known words):
  MI1 "vertuio" (simplest anagram of "voiture")
  MI2 "ospsoni" (most complex anagram of "poisson")
  Mi3 "entappremat" (anagram of "appartement" with the same complexity)
- Numbers:
  MF1 "118218"
  MF2 "982491840"
  MF3 "234567"

The kinds of presentation selected are:

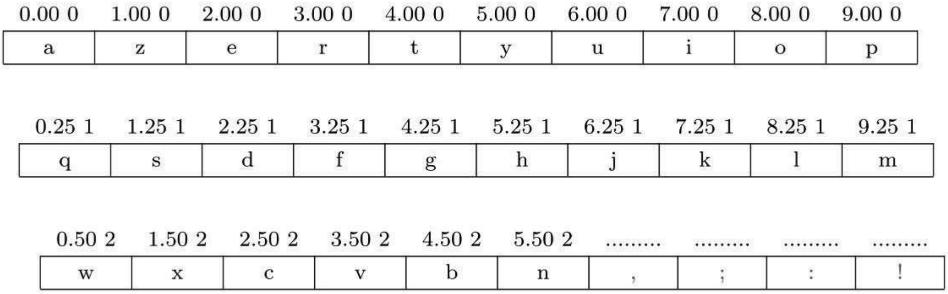
- Known words (in French as typists are French)
  PC1 Display in the graphical user interface.
  PC2 Listening to the pronunciation of the words.
- Unknown words (based on permutations of the known words):
  PI1 Display in the graphical user interface.
  PI2 Listening to the spelling of the words.
- Numbers:
  PF1 Display in the graphical user interface, all digits together.
  PF2 Display in the graphical user interface, by packet of 3 digits.
  PF3 Listening to the pronunciation of the numbers by packet of 3 digits.

There is then $3 * 2 + 3 * 2 + 3 * 3 = 21$ scenarios per session in average which gives $21 * 5 = 105$ inputs per sessions. For the choice of the unknown words, a complexity measurement in terms of typing difficulty has been used. With a password composed of L characters denoted $C_i,\ i = 1 : L$, its complexity of typing is related to the distance between keys of characters composing it[2]. The complexity is computed with the equation (7).

$$Complexity(password) = \sum_{i=1}^{L-1} \sqrt{(X_{C_{i+1}} - X_{C_i})^2 + (Y_{C_{i+1}} - Y_{C_i})^2} \qquad (7)$$

where X_{C_i} and Y_{C_i} are the locations of the character C_i on the keyboard. Figure 2 shows a possible quantification of the locations of each character in an AZERTY keyboard.

[2] However, the method does not take into account the use of the two hands at the same time, more investigations must be done on this point



0.00 0	1.00 0	2.00 0	3.00 0	4.00 0	5.00 0	6.00 0	7.00 0	8.00 0	9.00 0
a	z	e	r	t	y	u	i	o	p

0.25 1	1.25 1	2.25 1	3.25 1	4.25 1	5.25 1	6.25 1	7.25 1	8.25 1	9.25 1
q	s	d	f	g	h	j	k	l	m

0.50 2	1.50 2	2.50 2	3.50 2	4.50 2	5.50 2	....	....	....	....
w	x	c	v	b	n	,	;	:	!

Figure 2: Locations of characters in an AZERTY keybord.

3.4 Validation of the results

For each scenario, we want to analyse the performances with the following error rates:

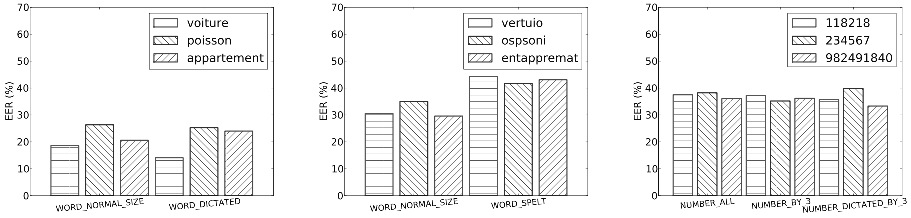
- The Equal Error Rate (EER) which gives the recognition error when the ratio of accepted impostors is the same as the ratio of rejected genuine users. It is a commonly used metric.
- The Failure To Acquire Rate (FTAR) which gives the ratio of acquisition problems. In keystroke dynamics, an acquisition problem is a typing mistake which forces the individual to type again the text from scratch. As this figure is important for this biometric modality, it also annoys a lot the user in keystroke dynamics, it is so mandatory to analyse it.

In order to determine whether there is a significant relationship between each acquisition scenario, we use the Kruskal-Wallis test (KW) [Hig]. It is a non-parametric (distribution free) statistical test, which is used to decide whether K independent samples are from the same population. In other words, it is used to test two hypothesis given by Equation 8: the null hypothesis H_0 assumes that samples originate from the same population (*i.e.*, equal population means) against the alternative hypothesis H_1 which assumes that there is a statistically significant difference between at least two of the subgroups.

$$\begin{cases} H0: \mu_1 = \mu_2 = \ldots = \mu_k \\ H1: \exists\,(i,j) \text{ with } i \neq j, \mu_i \neq \mu_j \end{cases} \tag{8}$$

The KW test statistic H is given by Equation 9, and the $p-value$ is calculated using a χ^2 distribution with $k-1$ degrees of freedom.

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{g} n_i \, \bar{r}_{i.}^2 - 3\,(N+1) \tag{9}$$



(a) Known (or easy) words (b) Unknown (or hard) words (c) Numbers

Figure 3: Equal Error Rate for each kind of text and way of presenting it.

where n_i is the number of observations in group i, r_{ij} is the rank of observation j from group i, N is the total number of observations across all groups.

$$\bar{r_{i.}^{2}} = \frac{\sum_{j=1}^{n_i} r_{ij}}{ni} \; and \, \bar{r} = \frac{1}{2}\left(N+1\right)$$

The decision criterion allowing to choose the appropriate hypothesis is given in Equation 10.

$$\begin{cases} p - value \geq 0.05 & accept\ H_0 \\ otherwise & reject\ H_0 \end{cases} \tag{10}$$

4 Experimental Results

In this section, we present the results of the proposed study[3]. 28 users participated at the two distinct sessions. The users have different background and some of them have difficulties to use a keyboard. Each session was done in approximately 10 minutes. The Equal Error Rate (EER) is computed by using the first 5 samples to compute the biometric references and the 5 last samples to compute the scores with the statistical method presented in [HRC07] ($d(q, \{\mu, \sigma\}) = 1 - \frac{1}{N} exp\left(-\frac{|q-\mu|}{\sigma}\right)$, with q the query, μ the mean sample of the user and σ its standard deviation sample). We merge the two sessions together for the statistical analysis. Figure 3 presents the Equal Error Rate (EER) for each text and way of typing it, while table 2 presents the results of the KW comparisons. By analysing these results, we make the following assertions:

- Known words give better performances than randomized ones (p-value=0.00394). We must use real common words in keystroke dynamics, instead of complicated ones as for non keystroke dynamics password authentication.
- Keystroke dynamics authentication based on numbers, instead of known words, gives worst performance (p-value=0.00146). It may be not a good idea to use

[3] The configurable acquisition software and the collected database can be given on demand

Table 2: Kruskal-Wallis comparison table. Two sources are compared together.

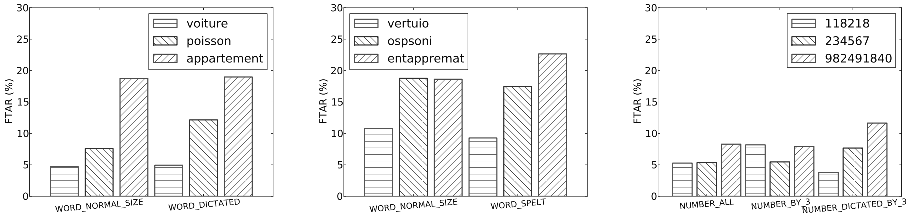
S1	S2	avg(S1)	avg(S2)	p-value	Conclusion
EER easy passwords	EER hard passwords	21.49	37.37	0.00394	$S1 < S2$
EER easy passwords	EER numbers	21.49	36.58	0.00146	$S1 < S2$
EER hard passwords	EER numbers	34.34	36.58	0.906	$S1 \simeq S2$
EER easy written	EER easy oral	21.86	21.13	0.8272	$S1 \simeq S2$
EER hard written	EER hard oral	31.69	43.04	0.049	$S1 < S2$
EER numbers written	EER numbers oral	36.73	36.27	0.6055	$S1 \simeq S2$
FTAR text	FTAR numbers	12.81	7.06	0.1023	$S1 \simeq S2$
FTAR easy passwords	FTAR hard passwords	11.89	16.26	0.2623	$S1 \simeq S2$
FTAR easy written	FTAR easy oral	10.35	12.03	0.5126	$S1 \simeq S2$
FTAR hard written	FTAR hard oral	16.06	16.46	0.8282	$S1 \simeq S2$
FTAR numbers written	FTAR numbers oral	6.74	7.70	1	$S1 \simeq S2$

keystroke dynamics in PIN based contexts. But, there is no significant difference between numbers and unknown words (p-value=0.906). We could have expected a difference.

- Oral presentation increases the EER for the randomized words (p-value=0.049). It may be explained because we are not used to hearing messages when using a computer. However, there is no difference for the known words (p-value=0.8272) or the numbers (p-value=0.6055).

Figure 4 presents the Failure To Acquire Rate (FTAR) for each text and way of typing it. We make the following assertions:

- FTAR is lower for numbers than texts. It can be explained by the fact that there are fewer keys to use in order to type it and less activities for the brain and the hands. However, the KW test does not prove it statistically (p-value=0.1023). We believe it is because there are not enough available examples. If we remove “appartement”, which behaves as an outlier, as well as its anagram, p-value=0.55, so there is still no differences. However if we compare hard word to numbers, we see there is fewer errors for numbers (p-value=0.00321).
- FTAR is lower for known words than other words. It can be explained by typing habits for known words. Once again, the KW test does not prove it statistically (p-value=0.26).
- Smaller text size holds in smaller FTAR (cf. figure 4).
- Oral presentation does not increase the FTAR, in comparison to textual comparison (p-value=0.5126 for known words, p-value=0.8272 for unknown words and p-value=0.8272 for numbers). However, we have noticed during the acquisition procedure, that after a certain amount of typing, people remembered the passwords and typed them without listening for the whole password.



(a) Known (or easy) words (b) Unknown (or hard) words (c) Numbers

Figure 4: Failure to Acquire Rate for each kind of text and way of knowing it.

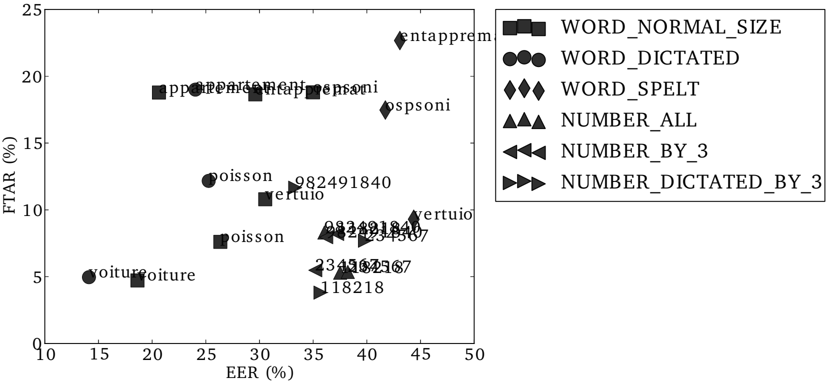


Figure 5: Failure to Acquire Rate and Equal Error Rate for each couple of text and way of giving it

The average EER among all the couples is 32%, while the mean FTAR among all the couples is 11%. These values are quite important. EER value is bad because there are not a lot of samples used for the enrollment, and there is a big difference between the text performing the better and the text performing the worst. FTAR value is quite bad too. Keystroke dynamics must be the biometric modality having the highest FTAR, as any typing mistake implies to start again the typing from scratch. The ratio of typing errors is also quite different between the various texts. There is no correlation between the EER and the FTAR among the various couples (Pearson correlation coefficient of 0.005). This can be confirmed in the figure 5 displaying the couples of EER/FTAR for each couple of presentation/text.

5 Conclusion

In this paper, we have analysed the performance of keystroke dynamics systems by varying the type of text to type and the way to present it to the user. We have shown that:

- It is better to choose passwords based on short, simple and known words.
- During an acquisition procedure of a new keystroke dynamics dataset, there is no preference to ask vocally the user to type a password.

These results are useful for people planning to acquire new datasets, or for people planning to produce systems generating spontaneous passwords in order to let them generate and present passwords allowing to have a good authentication performance.

It would be interesting to increase the size of the database with more words in order to apply statistical analysis on the results. These statistical analysis would help to prove or refute the presented facts.

References

[All10] Jeffrey D. Allen. An Analysis of Pressure-Based Keystroke Dynamics Algorithms. Master's thesis, Southern Methodist University, Dallas, TX, May 2010.

[Bou09] P. Bours. Feature Selection in Keystroke Dynamics. In *Norsk informasjonssikkerhetskonferanse (NISK)*, 2009.

[FF06] Jugurta R. Montalvo Filho and Eduardo O. Freire. On the equalization of keystroke timing histograms. *Pattern Recognition Letters*, 27:1440–1446, 2006.

[GEAR09] Romain Giot, Mohamad El-Abed, and Christophe Rosenberger. GREYC Keystroke: a Benchmark for Keystroke Dynamics Biometric Systems. In *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS 2009)*, pages 1–6, Washington, District of Columbia, USA, September 2009. IEEE Computer Society.

[GEAR12] R. Giot, M. El-Abed, and C. Rosenberger. *Keystroke dynamics*. Intech Book on Biometrics, 2012.

[Hig] J. J. Higgins. An Introduction to Modern Nonparametric Statistics. *The American Statistician*.

[HRC07] Sylvain Hocquet, Jean-Yves Ramel, and Hubert Cardot. User Classification for Keystroke Dynamics Authentication. In *The Sixth International Conference on Biometrics (ICB2007)*, pages 531–539, 2007.

[KM09] K.S. Killourhy and R.A. Maxion. Comparing Anomaly-Detection Algorithms for Keystroke Dynamics. In *IEEE/IFIP International Conference on Dependable Systems & Networks, 2009. DSN'09*, pages 125–134, 2009.

[KM11] Kevin S. Killourhy and Roy A. Maxion. Should Security Researchers Experiment More and Draw More Inferences? In *4th Workshop on Cyber Security Experimentation and Test (CSET'11)*, pages 1–8, August 2011.