Score Normalization for Keystroke Dynamics Biometrics

Aythami Morales, Elena Luna-Garcia, Julian Fierrez, Javier Ortega-Garcia

Departamento de Tecnología Electrónica y de las Comunicaciones, EPS, Universidad Autonoma de Madrid, C\ Francisco Tomás y Valiente, 11, 28049 Madrid, Spain

aythami.morales@uam.es, elenalunagarcia@gmail.com, julian.fierrez@uam.es, javier.ortega@uam.es

Abstract— **This paper analyzes score normalization for keystroke dynamics authentication systems. Previous studies have shown that the performance of behavioral biometric recognition systems (e.g. voice and signature) can be largely improved with score normalization and target-dependent techniques. The main objective of this work is twofold: i) to analyze the effects of different thresholding techniques in 4 different keystroke dynamics recognition systems for real operational scenarios; and ii) to improve the performance of keystroke dynamics on the basis of target-dependent score normalization techniques. The experiments included in this work are worked out over the keystroke pattern of 114 users from two different publicly available databases. The experiments show that there is large room for improvements in keystroke dynamic systems. The results suggest that score normalization techniques can be used to improve the performance of keystroke dynamics systems in more than 20%. These results encourage researchers to explore this research line to further improve the performance of these systems in real operational environments.**

Keywords—ink identification, pen verifier, hyperspectral analysis, handwritten document analysis, forensics.

I. INTRODUCTION

Biometric recognition technologies allow to authenticate users based on "something that we are" instead of traditional authentication based on "something that we know" such as passwords or PINs. The biometric technologies have become popular during last years (e.g. Apple Iphone Fingerprint sensor) and nowadays they can be considered an important player in consumer technology market. The biometric recognition technologies can be divided into physical traits (face, fingerprint, iris, DNA ...) and behavioral traits (signature, voice, gait, keystroking,...). There is no a technology which overcomes the rest and depending of the application we can opt for several solutions.

In this context, the keystroke dynamics authentication systems have attracted the interest of both researchers and industry [1-3]. The keystroke dynamics are proposed to improve the security of traditional authentication services based on passwords or PIN numbers. In the case of keystroke dynamics, the typical approaches based on fixed password authentication combine complex passwords and our keystroke dynamic biometrics. The password acts as a primary security level and the user access is not allowed until the correct password is inserted. The role of the biometric system is a secondary security level which try to detect intruders who are spoofing the identity of the legitimate user.

Among all the biometric technologies, keystroke dynamic recognition is especially interesting for Cyber Security because of: i) no need of extra sensors as the recognition of users is done according to their typing patterns using a keyboard or keypad; ii) it is possible to realize a continuous authentication based on the monitoring of the user behavior [4]; iii) keystroke dynamic technologies can be easily integrated into existing web-platforms or web-services.

The flowchart of typical keystroke dynamic recognition systems includes a classification phase in which query samples are compared with a stored template, see Fig.1. The identity of the user will be authenticated if the distance between the template and the query sample is lower than a pre-defined threshold. How to define this threshold is a challenge that has to be addressed before the deployment of a biometric system in real operational environments and score normalization can help to simplify it.

Score normalization has proved its usefulness for improving the performance of behavioral biometric traits such as signature [5-8] or voice [9-12]. The normalization of score mitigates the effects of misalignment between scores distribution from different users (this misalignment is common in behavioral traits). For the best of our knowledge, this topic has been scarcely analyzed by the research community for keystroke dynamics biometric systems.

This work studies score normalization techniques to improve the performance of keystroke authentication systems for real environments. Our experiments include four keystroke classifiers, three different normalization techniques and two publicly available databases. The results suggest that score normalization can be used to improve the performance of keystroke authentication systems. This work encourages to further explore in score normalization techniques for a better deployment of keystroke authentication in real-world services.

The rest of the paper is organized as follows: Section 2 describes the main modules of a keystroke authentication system and introduces the score normalization techniques analyzed; Section 3 presents the experimental framework and results; Finally Section 4 draws some conclusions.

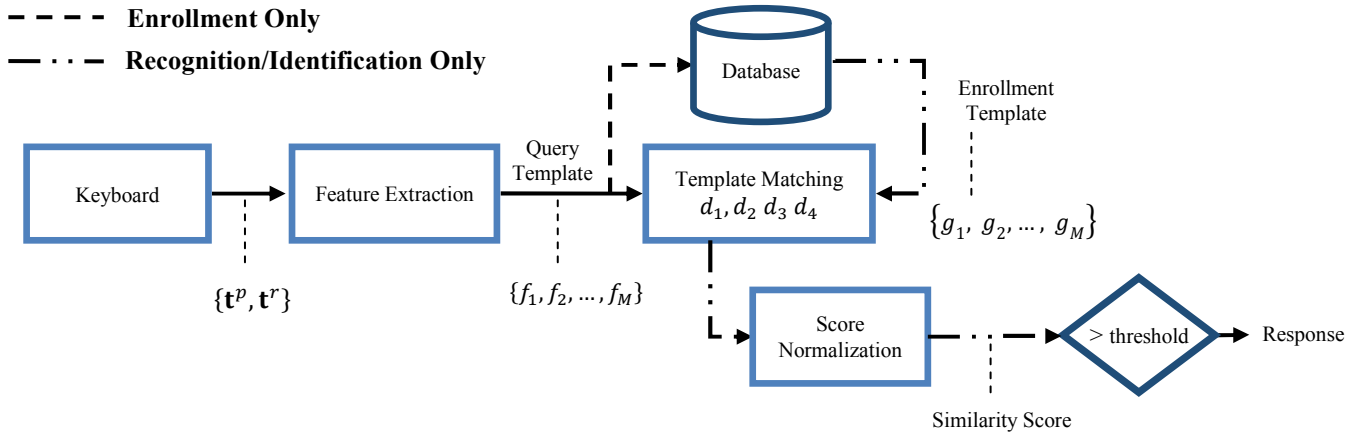

Figure. 1: Block diagram of a traditional biometric recognition system

II. KEYSTROKE DYNAMIC AUTHENTICATION

The keystroke authentication systems analyzed in this work include the traditional biometric recognition modules, see Fig. 1. The main modules are described below:

A. Feature Extraction

The keystroke dynamics extracted from a sequence of N keys consist of a vector $\mathbf{t}$ which contains the time stamp of every key-press (t^p) and key-release (t^r) event. These time stamps can be used to model the way a subject types but it is necessary to process the data to normalize the features with respect to a reference. This normalization on time can be achieved considering intervals between consecutive key events instead of absolute time stamps, see Fig. 2. The recognition systems evaluated in this paper are based on three of the most popular keystroke dynamic features:

- **Hold Time:** it is the difference between the time of pressure and release of the *i*th key:

$$H_i = t_i^r - t_i^p \qquad i = 1, \dots, N$$

- **Release-Press latency (RP-latency):** is the difference between the time of pressure of the (*i*+*1*)th key and the release of the *i*th key:

$$L_i^{rp} = t_{i+1}^p - t_i^r \qquad i = 1, \dots, N-1$$

- **Press-Press latency (PP-latency):** is the difference between the time of pressure of (*i*+1)th key and the pressure of the *i*th key:

$$L_i^{pp} = t_{i+1}^p - t_i^p \qquad i = 1, \dots, N-1$$

B. Classifiers (Template Matching)

The benchmark proposed in this work includes a baseline obtained with four keystroke dynamic classification algorithms. The algorithms could be used as baseline to further research with this dataset.

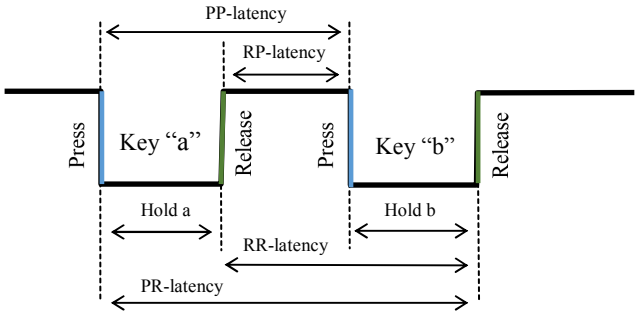

Figure 2: Keystroke dynamics features from a digraphs sequence.

Assume $\mathbf{f} = [f_1, f_2, \dots, f_M]$ as the feature vector (with M features) of a given test sample and $\mathbf{g^k} = [g_1^k, g_2^k, \dots, g_M^k]\ \forall\ k \in 1, \dots T$ as an enrollment set with T samples. The four keystroke dynamics classifiers included in the benchmark are:

- **Scaled Manhattan Distance** [13]: based on the one proposed by Araujo et al. [14]. The distance between a feature vector $\mathbf{f}$ of the test sample and the enrollment set $\mathbf{g}$ is calculated as:

$$d_1 = \sum_{i=1}^{M} |f_i - \bar{g}_i| / a_i \qquad (1)$$

where $\bar{\mathbf{g}}$ is the average of the enrollment set $\bar{\mathbf{g}} = \frac{1}{T}\sum_{k=1}^{T} \mathbf{g}^k$ and $\mathbf{a}$ is the average absolute deviation of the enrollment features, $\mathbf{a} = \frac{1}{T}\sum_{k=1}^{T} \left| g_i^k - \bar{g}_i \right|\ \forall\ i \in 1, \dots M$.

- **Mahalannobis + Nearest Neighbor:** this classifier was proposed by Cho et al. [15]. The distance between a test sample $\mathbf{f}$ and each of the enrollment samples $\mathbf{g^k}$ is calculated as:

$$d_2{}^k = \left(\mathbf{f} - \mathbf{g}^{\boldsymbol{k}}\right)\mathbf{S}^{-1}\left(\mathbf{f} - \mathbf{g}^{\boldsymbol{k}}\right)^{\mathrm{T}} \quad (2)$$

where the covariance matrix of the gallery set, $\mathbf{S}$, is introduced to increase the impact of those features with a smaller variance and $(\cdot)^{\mathrm{T}}$ is the transpose. The final distance d_2 is obtained as the minimum in k.

- **Combined Manhattan-Mahalannobis distance:** this distance metric was proposed in [16]. The test samples $\mathbf{f}$ and the enrollment set $\mathbf{g}$ are first normalized as $\hat{\mathbf{f}} = \mathbf{S}^{-1/2}\mathbf{f}^{\mathrm{T}}$ and $\hat{\mathbf{g}} = \mathbf{S}^{-1/2}\mathbf{g}^{\mathrm{T}}$, where $\mathbf{S}$ is the covariance matrix of the enrollment set. The distance d_3 is then calculated applying the L_1 distance between the normalized test sample and the average normalized enrollment set:

$$d_3 = \left\|\hat{\mathbf{f}} - \hat{\bar{\mathbf{g}}}\right\|_1 \quad (3)$$

- **Modified Scaled Manhattan distance** [14]**:** the distance between a feature vector of the test sample $\mathbf{f}$ and the enrollment set $\mathbf{g}$ is calculated as:

$$d_4 = \sum_{i=1}^{M} |f_i - \bar{g}_i|/\sigma_i' \quad (4)$$

where $\boldsymbol{\sigma}'$ is a modification of the standard deviation:

$$\sigma_i' = \begin{cases} \frac{0.2}{M}\sum_{j=1}^{M}\sigma_j & if\ \ \sigma_i < \frac{0.2}{M}\sum_{j=1}^{M}\sigma_j \\ \sigma_i & rest \end{cases} \quad (5)$$

This simple modification tries to mitigate the effects of samples with very low variance during the normalization (low variance means high weight).

The matching score between the test sample and the enrollment set is worked out as the inverse of each of the distances: $s_i = 1/d_i$.

C. Score Normalization

It is well-accepted [5-7] that user-dependent thresholds (one threshold per user) outperform global thresholds (same threshold for all users). The rationale behind this statement is that score normalization reduces the misalignment between score distributions from different users. There are several works in the literature [5-12] analyzing the effectiveness of score normalization techniques in behavioral biometrics (e.g. voice and signature). The impact of the user is high in behavioral biometrics such as keystroke and therefore, it is possible that user-dependent score normalization helps to improve the verification performance.

In this work we analyze the performance of different score normalization techniques. We can distinguish between two different approaches:

- **Normalization a posteriori:** the normalization is done using statistics from the results obtained during the test phase. This normalization allows to explore the limits (in terms of performance) of the normalization but it is a non-realistic approach. In a real application scenario, the scores of the test phase are unknown and not available during the training phase.
- **Normalization a priori:** the normalization is done using statistics obtained exclusively from the enrollment data. This is a more realistic methodology in which the normalization is tuned according to the information available during the enrollment phase.

The normalization techniques analyzed in this work are based in the z-score:

$$\hat{s}_i = \frac{s_i - \mu_i}{\sigma_i} \qquad i = 1, \ldots, 4 \quad (6)$$

where $\hat{s}_i$ is the normalized matching score, s_i are the distances obtained with the four classifiers presented in section II.B, μ_i and σ_i are the mean and standard deviation of the data used during the normalization (μ_i and σ_i are estimated for each classifier). In this work we explore three different normalization schemes based on the nature of the data used [5]:

- **Genuine-Impostor Centric (GIC):** in the GIC method it is used information from both genuine scores (intra-person variability) and impostor scores (inter-person variability). In this method μ_{GI} and σ_{GI} are obtained as the mean and standard deviation from both genuine and impostor scores calculated during the test phase. Therefore, this is an *a posteriori* normalization technique.
- **Genuine Centric (GC):** in the GC method it is used information exclusively from genuine scores of the enrollment samples. In this method μ_{GC} and σ_{GC} are obtained as the mean and standard deviation from genuine scores calculated from the available enrolment data using the Leave-One-Out methodology. Therefore, this is an *a priori* normalization technique.
- **Genuine-Modified Centric (GMC):** the previous GC method does not include the impact of the impostor scores in the score normalization. The GMC method tries to model such an impact by estimating the effects of impostor scores in the statistics. It is expected that impostor scores show a lower mean and greater standard deviation. Therefore the new mean and standard deviation are obtained as follows:

$$\mu_{GMC} = \mu_{GC}/R \quad \text{and} \quad \sigma_{GMC} = P\sigma_{GC} \quad (7)$$

where R and P are factors to introduce the impact of the impostor scores. They are calculated empirically for each classifier (one R and P for each classifier).

III. Experiments and Results

A. Databases

This work analyzes two different keystroke recognition scenarios:

Table 1: Feature performance in terms of Average EER (%) and standard deviation (in brackets). Last row includes the performance of the combination of RP and HT features.

	d_1	d_2	d_3	d_4
HT	11.0 (0.12)	15.9 (0.13)	13.8 (0.12)	8.87 (0.12)
RP	7.10 (0.10)	11.9 (0.12)	10.4 (0.12)	4.10 (0.08)
PP	6.70 (0.10)	11.9 (0.11)	12.1 (0.12)	4.4 (0.08)
HT+ RP	**4.30 (0.07)**	**9.00 (0.12)**	**9.60 (0.11)**	**2.22 (0.06)**

- **Scenario A – ATVS-Keystroke database** [17]**:** The database comprises 63 users with 12 genuine access (two sessions with time lapse greater than one day) and 12 impostor access for each user for a total number of samples equal to 1512 (63 users × 24 access). There are people from two different nationalities with 60% of males and 40% females. The design of the acquisition platform is inspired in the traditional web-platforms based on application forms (e.g. USA Electronic System for Travel Authorization). The idea was to provide a familiar environment which allows a natural user behavior. The acquisition platform includes 5 forms to provide the following personal data: given name, family name, email, nationality and national ID number. Therefore, in this scenario each user has his/her own password (in form of his/her personal data). The impostor samples are made by users who try to access the system with the information from other users. The keystroke dynamics of the users are captured for all 5 forms. These features include: Hold Time, PP-Latency, PR-Latency, RR-Latency and RP-Latency.
- **Scenario B – CMU database [13]:** this dataset comprises 51 subjects and 8 sessions with 50 repetitions per session. The time lapse between sessions is more than one day and the 400 typing samples were collected with an accuracy of ±200 microseconds. In this scenario the password is the same for all users and it consists of a ten characters typical strong password which includes uppercase, lowercase and symbols: *tie5Roanl*. The feature data for each sample includes: Hold Time; PP-Latency; RP-Latency.

B. Experimental Protocol

The main aim of the experimental protocol is to establish the framework to evaluate the different features (*section II.A*), classifiers (*section II.B*) and normalization techniques (*section II.C*). Due to the different characteristic of both databases, the protocols used present some differences:

- Experimental protocol for the scenario A: the six feature vectors from the first session are used as enrollment set. The remaining six feature vectors from the second session and the twelve impostor vectors are used as test set to calculate the FRR and FAR respectively. Therefore we will obtain $6 \times 63 = 378$ genuine scores and $12 \times 63 = 756$ impostor scores.
- Experimental protocol for the scenario B: the methodology used for the CMU database is the same proposed in the literature [13] for this dataset. The first four sessions are used as enrollment set while the remaining four session are used as test set. As all the users share the same password, the first five samples from the first session of each user are used as impostor set. Therefore we will obtain $400 \times 51 = 20{,}400$ genuine scores and $5 \times 51 \times 50 = 12{,}750$ impostor scores.

Table 2: Score Normalization performance (HT,RP feature combination) in terms of EER (%) for ATVS-Keystroke database..

Ceiling	d_1	d_2	d_3	d_4
From table 1	4.30	9.00	9.60	2.22
Normalization	d_1	d_2	d_3	d_4
none	12.0	14.3	16.4	7.3
GIC (*a posteriori*)	5.0	10.5	10.0	3.1
GC (*a priori*)	10.6	24.2	17.2	7.5
GMC (*a priori*)	9.6	15.6	15.6	5.5

C. Results – Scenario A

The first experiment is aimed to establish a baseline with the performance obtained using the best conditions. These conditions include the calculation of the Average Equal Error Rate obtained according an optimal threshold per user (threshold obtained a posteriori from genuine and impostor scores and average EER obtained as mean of the EERs obtained for all users). This can be considered as an over optimistic performance evaluation but it will be useful to establish the performance ceiling. Table 1 shows the performance obtained for all four classifiers and different features.

The results suggest that RP-Latency is more discriminant than Hold Time but combination (by concatenating the features) outperforms individual performances. The rest of experiments will be carried out using the combined feature scheme. The differences among classifiers is clear and it is evident the superior performance of $\{d_1, d_4\}$ against $\{d_2, d_3\}$. The best performance is obtained for the Modified Scaled Manhattan Distance with an EER equal to 2.22%.

As was mentioned, using optimized thresholds produces over optimistic performances. A common threshold for all users is a more suitable approach for real applications. Table 2 shows the results obtained when a common threshold is used for all subjects (all the genuine and impostor scores are calculated and the EER is obtained based on a unique threshold). The table shows the performance ceiling (from Table 1), the performance obtained when the normalization is not applied and the performance obtained using the different normalization techniques proposed in Section II.C.

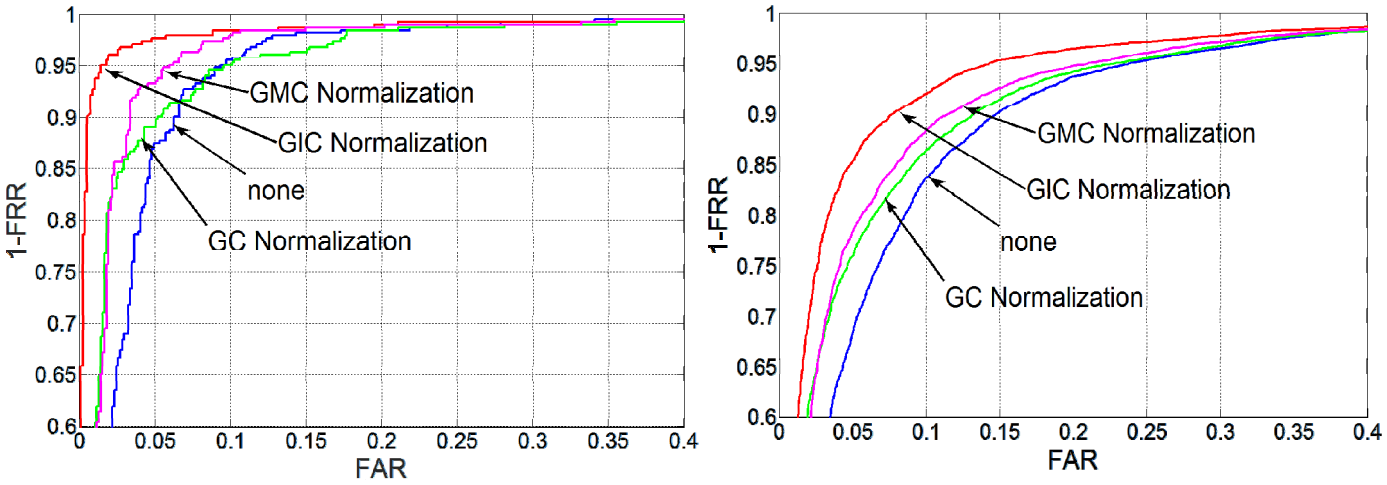

Figure 3: ROC curves for different normalization strategies using the Modified Scaled Manhattan Distance and ATVS-Keystroke database (on the left) and CMU database (on the right).

Table 3: Score Normalization performance in terms of EER (%) for CMU dataset.

Ceiling	d_1	d_2	d_3	d_4
HT+PP+RP	8.6	8.9	8.2	8.8
Normalization	d_1	d_2	d_3	d_4
none	10.3	9.6	9.9	12.6
GIC (*a posteriori*)	8.6	9.0	7.9	8.9
GC (*a priori*)	10.7	11.5	10.0	11.7
GMC (*a priori)*	10.2	10.5	9.8	10.7

The results show how the performance decreases when a unique threshold for all users is applied with EER three times greater in some cases (in comparison with the performance ceiling obtained in Table 1). These results suggest a large misalignment between score distributions from different users. However, the normalization techniques mitigate this effect and it is possible to achieve competitive performances with GIC and GMC normalization. The differences between GC (only genuine scores used to calculate the statistics) and GMC (genuine statistics modified according the impostor models) normalization suggest the great impact of impostor scores in the normalization strategies.

D. Results – Scenario B

In the case of CMU database, the performance ceiling is established according the best feature combination reported in the literature with this dataset [13]. Therefore, the feature vector is composed by Hold Time, PP Latency and RP-Latency. The second row of Table 3 reports the performance of the Average Equal Error Rate obtained using the optimal threshold per user (as in Table 1). The rows 4 to 7 shows the performance obtained according the different normalization techniques.

The results suggest that scores distribution from CMU dataset are more aligned than those from ATVS-Keystroke database. This can be observed comparing the moderate degradation of the performance when we use a unique threshold without normalization (fourth row) instead of optimized thresholds (second row). While ATVS-Keystroke database showed degradation up to 300%, the degradation on CMU database is around 40%. The reason of this alignment can be explained because all users in CMU share the same password while users in ATVS-Keystroke database have specific passwords per subject. However, for some classifiers (e.g. Modified Scaled Manhattan distance) the normalization clearly improves the performances (see Fig. 3).

IV. CONCLUSIONS

This paper has analyzed different score normalization techniques for keystroke biometric authentication. The experiments include two different databases and four state-of-the-art keystroke dynamics classifiers. The results suggest that score normalization can be used to improve the performance of keystroke authentication systems. The improvement varies depending of the dataset and the classifier but there is still room for research and the results obtained encourage to further investigate. Future work will include larger dataset, supervised machine learning algorithms (e. g. SVMs or NN) and score calibration techniques.

ACKNOWLEDGMENT

A.M. is supported by a post-doctoral Juan de la Cierva contract by the Spanish MECD (JCI-2012-12357). This work has been partially supported by projects: Bio-Shield (TEC2012-34881) from Spanish MINECO, BEAT (FP7-SEC-284989) from EU, CECABANK and Cátedra UAM Telefónica.

REFERENCES

[1] A. Peacock, X. Ke, and M. Wilkerson, "Typing patterns: A key to user identification". *IEEE Security and Privacy*, 2(5), pp. 40–47, 2004.

[2] D. Gunetti and C. Picardi, "Keystroke analysis of free text". *ACM Transactions on Information and System Security*, 8(3), pp. 312–347, 2005.

[3] A. Morales, J. Fierrez and J. Ortega-Garcia, "Towards predicting good users for biometric recognition based on keystroke dynamics". In *Proc. of European Conf. on Computer Vision Workshops*, Springer LNCS-8926, Zurich, Switzerland, pp. 711-724, 2014.

[4] S. Mondal and P. Bours, "A computational approach to the continuous authentication biometric system". *Information Sciences*, 304 (20) pp. 28–53, 2015.

[5] J. Fierrez-Aguilar, J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Target dependent score normalization techniques and their application to signature verification". *IEEE Trans. on Systems, Man & Cybernetics - Part C*, 35 (3), pp. 418-425, 2005.

[6] R. Plamondon and G. Lorette, "Automatic signature verification and writer identification - The state of the art". *Pattern Recogni*tion, 22 (2), pp. 107–131, 1989.

[7] M. Paulik, N. Mohankrishnan, and M. Nikiforuk, "A time varying vector autoregressive model for signature verification". In *Proc. 37th Midwest Symp. Circuits Syst.*, 2, pp. 1395–1398, 1994.

[8] A. Jain, F. Griess, and S. Connell, "On-line signature verification". *Pattern Recognition*, vol. 35, no. 12, pp. 2963–2972, 2002.

[9] S. Furui, "Cepstral analysis technique for automatic speaker verification". *IEEE Trans. Acoust. Speech, Signal Process.*, 29 (2), pp. 254–272, 1981.

[10] T. Matsui, T. Nishitani, and S. Furui, "Robust methods of updating model and a priori threshold in speaker verification". *In Proc. IEEE Intl. Conf. Acoustics, Speech Signal Process.*, ICASSP, 1996, pp. 97–100, 1996.

[11] R. Auckenthaler, M. Carey, and H. Lloyd-Tomas, "Score normalization for text-independent speaker verification systems". *Digital Signal Process.*, 10, pp. 42–54, 2000.

[12] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D.-A. Reynolds, "A tutorial on text-independent speaker verification," *J. Appl. Signal Process.*, 2004 (4), pp. 430–451, 2004.

[13] Kevin S. Killourhy and Roy A. Maxion, "Comparing Anomaly Detectors for Keystroke Dynamics". *In Proceedings of the 39th Annual International Conference on Dependable Systems and Networks (DSN-2009)*, pp. 125-134, 2009.

[14] L. C. F. Araujo, L. H. R. Sucupira, M. G. Lizarraga, L. L. Ling, J. B. T. Yabu-uti, "User Authentication Through Typing Biometrics Features". *IEEE Trans. On Signal Processing*, 53 (2), pp. 851:855, 2005.

[15] S. Cho, C. Han, D. H. Han, H. Kim, "Web-based keystroke dynamics identity verification using neural network". Journal of Organizational Computing and Electronic Commerce, 10 (4), pp. 295–307, 2000.

[16] Y. Zhong, Y. Deng, and A. K. Jain, "Keystroke dynamics for user authentication". *In Proc. Of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 117-123, 2012.

[17] A. Morales, M. Falanga, J. Fierrez, C. Sansone and J. Ortega-Garcia, "Keystroke Dynamics Recognition based on Personal Data: A Comparative Experimental Evaluation Implementing Reproducible Research". *In Proc. the IEEE Seventh International Conference on Biometrics: Theory, Applications and Systems*, 2015.