# Keystroke analysis as an authentication method for thumb-based keyboards on mobile handsets

S.Karatzouni and N.L.Clarke

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail:info@network-research-group.org

## Abstract

The evolution of mobile networking has opened the door to a range of possibilities for mobile devices, increasing at the same time the sensitivity of the information stored and access through them. Current PIN-based authentication has proved to be an insufficient and an inconvenient approach. Biometrics that have proved to be a reliable approach to identity verification can provide a more robust mean of security as they rely on personal identifiers. Amongst various biometric techniques keystroke analysis combines features that can offer a cost effective, non-intrusive and continuous authentication solution for mobile devices. This research is being undertaken in order to investigate the performance of keystroke analysis on thumb-based keyboards that are being widely used in PDA's and Smartphone devices. The investigation was based on the scenario of authenticating users while typing text messages, using two keystroke characteristics, inter-key latency and hold-time. The results showed to be promising achieving an EER=12.2% with the inter-key latency, whereas unusually hold-time did not prove to be a feasible feature to utilise in such tactile environment.

## Keywords

Keystroke analysis, Biometrics, Authentication, Mobile

## 1. Introduction

The proliferation of mobile devices and mobile networking has introduced new challenges for the protection of the subscribers' assets. The security risks are no longer associated only with safeguarding the subscribers account. With the introduction of 3rd generation mobile networks the services and information accessible through mobile handsets have increased in sensitivity, as micro-payments, mobile banking, location-based services are all reality for the mobile world and more potential is arriving in the future. But moreover the attraction that high-tech devices can result places a further concern for enhanced security, as underlined by looking at the statistics for mobile theft, which in the UK accounts the 45% of overall theft (British Transport Police, 2006).

Current authentication, mainly achieved by PIN's, is not enough to substantially safeguard today's mobile handsets and the data access enabled. As a secret knowledge technique it has several drawbacks, as it can be shared or written down, but also being a 4-6 digit number is not difficult for a potential impostor to acquire (Lemos, 2002). Furthermore as survey results show, subscribers consider it as an inconvenient method and as such they do not use them in the first place leaving their device unprotected (Clarke *et al*, 2002). Even a secondary measure, SIM cards due to

their functionality it is unlikely to be removed from the device, thus provide no protection in case a device is stolen or lost.

Alternative authentication based on biometrics could provide an enhancement on the security currently provided. Biometrics rely on the personal identifiers and therefore they can provide authentication based on something a person is, a fact that introduces a unique level of security that other approaches do not meet as it relates the process to a person and not to a possession of knowledge or token. A biometric method that can provide a cost-effective and a non-intrusive solution for mobile handset authentication is keystroke analysis, which is based on the typing dynamics of a user. The purpose of this research is to investigate keystroke analysis in thumb-based keyboards based on text messaging input, looking at the feasibility of applying this technique as an authentication method for mobile handsets that offer that tactile interface.

## 2. Keystroke analysis

Keystroke analysis is a behavioural biometric that attempts to verify identity based on the typing pattern of a user looking at certain characteristics of his interaction with a keyboard. A lot of research has been undertaken on the method since first introduced in 1980's, identifying two main characteristics to provide valuable discriminative information:

- Inter-key latency, which is the interval between two successive keystrokes, and
- Hold-time, which is the interval between the pressing and release of a key

The majority of the studies have looked at the feasibility of keystroke analysis on full QWERTY keyboards (Umpress & Williams, 1985; Joyce & Gupta, 1990; Brown & Rogers, 1993; Obaidat & Sadoun, 1997), showing satisfactory results for both of the characteristics mentioned. In general inter-key latency has showed to provide better information for the classification in comparison to hold-time.

As in all biometrics the way to access the performance of keystroke analysis, two measures are used. The False Acceptance Rate (FAR) that indicates the percentage of an impostor falsely granted access to the system, and the False Rejection Rate (FRR), which represents the percentage of a legitimate user getting rejected. There is a trade-off between increasing security (and therefore decreasing the FAR) and increasing user convenience (and thus decreasing the FRR). As of the different security requirements for each system, the point that those two rates cross - the Equal Error Rate (%), is used as a more objective mean for the comparison of different biometrics.

For the assessment of keystroke analysis traditionally statistical approached were used, though more recently the use of neural network pattern recognition proved to provide better performance. A summary of the literature results underlying keystroke analysis on PC keyboards is provided in Table 1.

| Study | Users | Input | Inter-Key | Hold-time | Approach | FAR | FRR |
|---|---|---|---|---|---|---|---|
| Umpress & Williams | 17 | Alphabetic | ● | | Statistical | 11.7 | 5.8 |
| Joyce & Gupta | 23 | Alphabetic | ● | | Statistical | 0.3 | 16.4 |
| Brown & Rogers | 25 | Alphabetic | ● | ● | Neural N. | 0 | 12 |
| Obaidat & Sadoun | 15 | Alphabetic | ● | ● | Statistical | 0.7 | 1.9 |
| | | | | | Neural N. | 0 | 0 |
| Ord & Furnell | 14 | Numerical | ● | | Neural N. | 9.9 | 30 |

**Table 1: Literature summary results on keystroke analysis on PC keyboards**

Although the extensive research on keystroke analysis, it was not till recently that the method was assessed on interfaces provided on mobile phones where the tactile environment differs. A series of studies (Clarke & Furnell, 2006) accessed the method on regular mobile phone keypads with promising outcomes, achieving an EER= 8% based on numerical input. Nevertheless, the performance of keystroke analysis for thumb-based keyboards was undocumented. Thumb-based keyboards constitute an interesting gap in research as they provide the extensive interface of a PC keyboards and the thumb-based keystrokes of a mobile phone.

## 3. Methodology

This study looked into the feasibility of authenticating a user while typing text messages. Two different types of analysis were used in the context of this research-static and pseudo-dynamic accessing inter-key latency and hold-time respectively. A number of thirty messages, comprised the input of the experiment, which were designed to fulfil certain requirements.

| Keyword | # Inter-key latencies | #Samples after outliers' removal | Training Set | Testing Set |
|---|---|---|---|---|
| everything | 10 | 27 | 18 | 9 |
| difficult | 9 | 26 | 18 | 8 |
| better | 6 | 27 | 18 | 9 |
| night | 5 | 27 | 18 | 9 |
| the | 3 | 26 | 18 | 8 |
| and | 3 | 27 | 18 | 9 |

**Table 2: Keywords used for inter-key latency**

For the static analysis six varying sized keywords were included in the text messages providing a static component to use. The keywords were selected based on the criteria that it should be likely to appear often in a text message, while no abbreviations could be used as substitutes. Thirty repetitions of each keyword were included, a number of which though were removed as outliers. The words selected are listed in Table 2, along with the number of inter-key latencies that they involve and the number of samples used for training and testing.

The pseudo-dynamic analysis was based on the hold-time of the six most recurrent letters in the English language – e t a o n i, adequate number of repetitions of which were included. Literature has showed that attempts to perform dynamic analysis on keystroke dynamics (Leggett, Napier) did not yield satisfactory results. As such an attempt was made to utilize a static component – the recurrent letters, in a dynamic form of analysis.

Fifty participants were recruited to type the series of the text messages, using an XDA II handset that deploys a representative example of today's thumb-based keyboard, as illustrated in Figure 1. In order to capture the keystroke data, appropriate software was implemented using Microsoft's Visual Basic .NET, and deployed on the handset. A screenshot of the software is provided in Figure 2. As usual in keystroke analysis studies, corrections were not permitted in case the user misspelled a word as this would undesirably interfere with the data of the inter-key latency (Umpress & Williams, 1985). Instead the whole word should be retyped in the correct form. The data collection was performed in a single session, although it would be preferred to collect the data during multiple sessions, as thus a more indicative typing profile of the users could be captured.

**Figure 1: An XDA IIs thumb-based keyboard**


**Figure 2: Screenshot from experiment software**

## 4. Results

### 4.1 Inter-key latency

An initial analysis of the input data showed a fairly large spread of values on the inter-key latencies, even for the smaller keywords that were expected to be more concise because of the commodity and length. Additionally to that the difference of the values attributed to each user was not large, so that many of the users overlapped. This puts a burden on the classification algorithm, as those two factors make the definition of limits to differentiate between users very difficult as the values are interfering. Figure 3, illustrates the mean and standard deviation for the larger keyword across all users as an example of the problem.

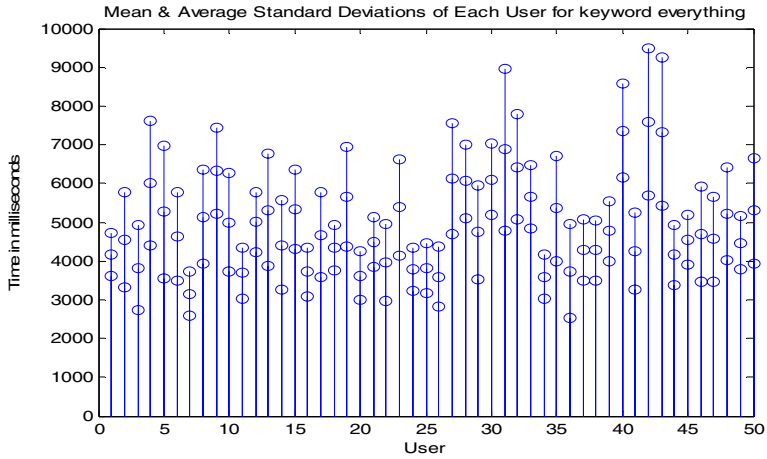


**Figure 3: Mean & Standard deviation for keyword "everything"**

A number of tests took place, using Feed Forward Multilayer Perceptron neural network as it has showed very good performance in previous research (Clarke & Furnell, 2006). Different configurations were tested, changing the network size and weights but also the training time, looking for optimum performance. The best results were outcome of the keyword 'everything' as expected because of providing a larger input vector, giving an EER=23.4% with FAR=19.3 and FRR=27.5, the last of two are indicated in Figure 4.

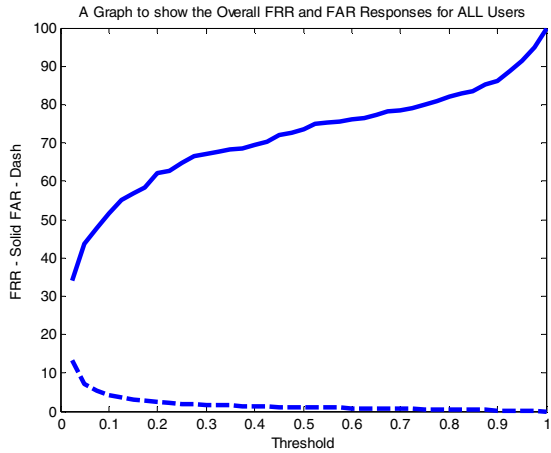


**Figure 4: Overall FAR and FRR for best case network for keyword "everything"**

As can be seen, but also for all of the tests, the results showed an FRR much higher from the FRR which can be explained by the large amount of 49 impostors extensively training the network versus the one authorised user. Furthermore the number of samples assigned to the testing of the classification was small, resulting to the FRR encountering large steps in its transitions.

The error rate is fairly high, nevertheless, there were cases of users reaching an EER below 10% with best case user 1 achieving an EER of 0.3% that shows a good ability of classification. The rest of the keywords resulted even higher error rates, as it was though expected as they provide a smaller input vector. The best results for each keyword are listed in Table 3.

| Keyword | FAR (%) | FRR (%) | EER (%) |
|---|---|---|---|
| everything | 12.8 | 34.2 | **23.5** |
| difficult | 13.2 | 43.0 | **28.1** |
| better | 18.0 | 43.1 | **30.5** |
| night | 21.3 | 45.8 | **33.5** |
| the | 23.7 | 41.5 | **32.6** |
| and | 24.3 | 43.6 | **33.9** |

**Table 3: Best results for each keyword**

The results of different networks showed minimal change in the EER's, though the FAR and FRR showed much variation. This indicates the fact that the network tries to optimise for the population of users, averaging the performance and as such each user can not train to the best suited way.

To overcome that problem a different approach was utilised based on the improved results it gave on previous study (Clarke & Furnell, 2006). A gradual training was performed, training the network for an extensive amount of time but periodically evaluating the performance. The results showed a noticeable decrease on the error rates with best case achieving an EER=12.2% for the larger keyword. The summary of the gradual training results are listed in Table 4.

| Keyword | FAR (%) | FRR (%) | EER (%) |
|---|---|---|---|
| everything | 15.8 | 9.1 | **12.2** |
| difficult | 16.8 | 12.0 | **14.4** |
| better | 23.5 | 14.4 | **18.9** |
| night | 24.2 | 14.4 | **19.3** |
| the | 29.3 | 19.5 | **24.4** |
| and | 28.7 | 17.6 | **23.1** |

**Table 4: Gradual training results for all keywords**

For the keyword "everything", 20 users achieved an FRR=0% with the respective FAR below 10%, which provides a very promising result, with the best user achieving an FAR=0.7% and FRR=0%. The list of best and worst case users for all keywords are listed in Table 4. The results underline the requirement of different training intensiveness for each user, but mainly that inter-key latency offers the discriminative data to classify users in the specific tactile interface.

| Keyword | Best Case | | | | Worst Case | | | |
|---|---|---|---|---|---|---|---|---|
| | User | FAR | FRR | EER | User | FAR | FRR | EER |
| everything | 2 | 0.7 | 0 | **0.4** | 6 | 42.6 | 22.2 | **32.4** |
| difficult | 11 | 2.6 | 0 | **1.3** | 46 | 18.1 | 50.0 | **34.1** |
| better | 49 | 3.2 | 0 | **1.6** | 27 | 35.1 | 33.3 | **34.2** |
| night | 34 | 4.5 | 0 | **2.3** | 25 | 25.6 | 55.5 | **40.5** |
| the | 26 | 12.8 | 0 | **6.4** | 39 | 41.6 | 50 | **45.8** |
| and | 11 | 10.9 | 0 | **5.4** | 5 | 32.2 | 66.7 | **49.4** |

**Table 5: Best & Worst Case results from gradual training**

As due to time limitations the network was not optimised it is believed that further testing will be able to provide even lower results.

### 4.2 Hold-time

In the contrary to inter-key latency, hold-time did not seem to be able to provide any data to help classify different users. A series of tests on different network configurations using all six letters (as to provide the larger possible input vector) resulted in an EER of around 50%, showing that little classification could be performed. The same error rate derived using different size subsets of the letters with smaller input vectors but with more repetitions of each letter, but also when a larger input of eight letters was used adding in the set also the letters 'r' and 's', as next on the reoccurrence list.

In order to further access the performance of hold-time, a group of only 20 users was used aiming to help the classification as the population to discriminate against would be less, though with no change in the results. Even when gradual training was tested, using the six letters set, no improvement came. Sample results from various tests are provided in Table 6. Although there were users with FRR or FAR of 0% the respective FAR or FRR was reaching over 80%. Even though there was a 10% decline on the EER using gradual training, the results are still very high to suggest that hold-time can offer valuable discriminative information.

| Set | Training | Users | FAR | FRR | EER |
|---|---|---|---|---|---|
| 6 letters | normal | 20 | 49.5 | 49.4 | **49.5** |
| 6 letters | normal | 50 | 31.3 | 69.0 | **50.2** |
| 8 letters | normal | 50 | 26.7 | 72.9 | **49.8** |
| 3 letters | normal | 50 | 22.1 | 77.6 | **49.9** |
| 6 letters | gradual | 50 | 34.2 | 36.8 | **36.8** |

**Table 6: Sample results from various tests on hold-time**

## 5. Discussion

As the results showed inter-key latency can provide a mean of differentiating between users, when based on a latency vector of 10, being able to achieve a 12.2%

EER with the gradual training approach. Using a smaller input vector, although classification was able to be performed there were increased error rates, though it must be noticed that no network optimization was researched for the smaller keywords.

In regards to the inter-key latency, the results did not have the low rates that research on regular keyboards has showed, though there are a number of factors that differentiate this study. An issue to underline is that the keyboard used provides a more restricted keystroke interface as the distance between the keys is smaller in comparison with a PC, but also the number of fingers likely to be used is two in contrast with ten in the respective case. Both of these factors limit the typing dynamics as the combinations of the fingers in conjunction with the timing of the keystrokes and movement to achieve them, are restricted. This results in a smaller value area for the keystrokes of the users, making the distinction between them more difficult. Furthermore, although the layout was familiar to all users as it shares the same layout with a PC keyboard, some of the participants experienced difficulty in identifying the placement of the keys due to the different way of typing.

Hold-time did not provide any proof that it can be utilised in the specific typing interface though there are a number of factors that may explain the inability of the keystroke feature.

Firstly the keys that the thumb-based keyboard deploys are very small related to the chunky tactile environment that a normal keyboard offers, restricting the interval length between the pressing and release of a key and thus not providing much differentiation in values. Although hold-time has performed well on regular mobile phone keypads (Clarke & Furnell, 2006), where still the keys were larger than the keyboard used in this experiment, a further factor was that, in a mobile keypad in order to access the preferred letter more than one pressings are often required, with the hold-time being calculated from the first keystroke till the last key release, increasing immediately the range of values and thus allowing an easier distinction between them.

Furthermore in a thumb-based keyboard, fingers stay almost static due to the limited area, thus keystrokes hardly differentiate, as no other factors such as hand movement appears as in PC keyboards which may affect the pressing of a key. What must be also noticed is that some participants complained about the feedback from the keyboard, as they could not at all cases be sure if they had pressed a key, which might led to a continual change of the hold-time.

## 6. Conclusion

This research was a feasibility study on the utilisation of keystroke analysis as an authentication method in devices that offer the tactile environment of a thumb-based keyboard. The results showed that from the two traditionally used keystroke characteristics- inter-key latency showed promising results, whereas hold-time gave no clues of a potential use in that kind of keystroke interface, though research must be undertaken to further access them.

Future work will search upon an optimised network configuration that was not extensively research during this study, in regard to the inter-key latency. Furthermore the use of different keywords will be investigated as also the combined use of more than one, looking also to use abbreviations as keywords as they are more likely to appear in a text message more often. In respect to hold-time, further tests are required before concluding to its ineffectiveness, exploring the use of longer input vectors and different letter subsets. A future experiment will also look to utilise thumb-based keyboards that offer a slight different tactile environment than the one used in this study, to have a mean of comparison, of the performance of the keystroke characteristics and an insight on the factors that may affect it.

Nevertheless, the study showed promising results for the use of keystroke analysis in thumb-based keyboards. Although the accuracy of the method does not compete in distinctiveness with other biometrics such as fingerprints, the nature of keystroke analysis can provide a monitoring authentication mechanism, transparent to the user that is not feasible for other techniques. In that basis it can provide continuous authentication based on the regular use of the device, and if used in conjunction with other authentication approaches that can fulfil the lack of the method in accuracy, a more enhance security can be achieved.

## 7. References

British Transport Police (2006): “Mobile phone theft”, http://www.btp.police.uk/issues/mobile.htm

Brown, M., Rogers, J. (1993): “User Identification via Keystroke Characteristics of Typed Names using Neural Networks”, International Journal of Man-Machine Studies, vol. 39, pp. 999-1014

Clarke, NL., Furnell, SM.,(2006) : “Authenticating Mobile Phone Users Using Keystroke Analysis”, International Journal of Information Security, ISSN:1615-5262, pp1-14,2006

Clarke, N., Furnell, S.M, Rodwell, P.M, Reynolds, P.L (2002): “Acceptance of subscriber authentication method for mobile telephony devices”, Computers & Security, vol. 21, no.3, pp220-228

Joyce R., Gupta, G. (1990): “Identity Authentication Based on Keystroke Latencies”, Communications of the ACM, vol. 39; pp 168-176.

Lemos, R. (2002): “Passwords: The Weakest Link? Hackers can crack most in less than a minute”, CNET.com, http://news.com.com/2009-1001-916719.html

Obaidat, M. S., Sadoun, B. (1997): “Verification of Computer User Using Keystroke Dynamics”, IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics, Vol. 27, No.2.

Ord, T. (2000): “User Authentication for Keypad-Based Devices using Keystroke Analysis”, MSc Thesis, University of Plymouth, UK.

Umphress, D., Williams, G. (1985): “Identity Verification through Keyboard Characteristics”, International Journal of Man-Machine Studies, Vol. 23, pp. 263-273