Application of Keystroke Analysis to Mobile Text Messaging

Nathan Clarke, Steven Furnell & Benn Lines
University of Plymouth, UK
info@network-research-group.org

Paul Reynolds
Orange Personal Communication Services Ltd, UK
paul.reynolds@orange.co.uk

Abstract

The ability to verify the identity of a mobile subscriber will become increasingly important as the services and information that can be accessed by a cellular handset becomes more wide-ranging and sensitive. Current user authentication for mobile handsets is provided by the Personal Identification Number (PIN), which has a number of inherent weaknesses, such as users writing them down, telling other people and the option of not using it in the first place. As such non-intrusive, continuous and stronger authentication mechanisms are required. This paper presents the second phase of an investigation into the feasibility of one such technique, the use of keystroke analysis. Specifically, this paper evaluates the potential to authenticate subscribers by the way in which they type text messages. The study trials a number of classification algorithms based upon feed-forward multiple layer perceptron neural networks, evaluating their ability to differentiate between authorised and impostor users. Promising results have been observed, with average classification results of 18% Equal Error Rate (EER) and individual users achieving an EER as low as 3.2%.

Keywords: *Authentication, Biometric, Keystroke Analysis, Mobile Handset.*

Introduction

Text messaging has become the most successful mobile data service to date, with some 366 billion messages sent globally in 2002 (Cellular Online, 2003). This number reflects the wide proliferation of cellular networks and mobile handsets throughout society, with the current number of subscribers standing at just over a billion (GSM World.Com, 2003). However, in parallel with this rise in ownership there has been a rise in mobile related abuse, with over 700,000 handsets stolen from subscribers in 2001, in the UK (BBC, 2002). It can be conjectured that the more advanced capabilities of third generation handsets with their ability to pay for products using micro-payments and digital money, surf the internet, buy and sell stocks, transfer money and manage bank accounts will make the handsets even more desirable targets.

Current handset-based authentication is achieved through a PIN (Personal Identification Number) approach, which relies heavily on the user to ensure its validity. For example, the subscriber should not use the default setting, divulge their PIN to other people, or write it down. Apart from the technological arguments, a survey into attitudes and opinions of mobile phone subscribers found that 45% of

respondents thought the PIN to be inconvenient and, as a consequence, they did not use the facility (Clarke et al., 2002). Nonetheless, the findings also demonstrated the users' awareness of the security implications, with 81% of respondents in support of more security.

The verification of an identity can be achieved in one of three ways: using something the user *knows, has* or *is* (Wood, 1978). The first approach is a secret-knowledge technique such as the PIN and will therefore be just as inconvenient. The second is based upon the user having to carry a token. However, this would be very likely to remain within the handset permanently and thus diminish any security gains, for example, a subscriber's use of the SIM (Subscriber Identity Module) where the SIM can be described as a token, but is rarely removed from the handset. The last approach, commonly termed biometrics, is based upon some distinguishing feature of a person and includes physiological characteristics, such as fingerprints and hand geometry and behavioural traits, such as a person's voice and signature. Another behavioural biometric is keystroke analysis, which measures the typing characteristic of a user. In this context keystroke analysis has a number of advantages, including a keypad that already resides on the device and the possible non-intrusive application of the technique where the user is unaware authentication is taking place, thereby reducing user inconvenience. This paper presents the finding of an investigation into the feasibility of using text messages as a means of authenticating subscribers on a non-intrusive basis, where the user is unaware authentication is taking place. The paper begins with an introduction into keystroke analysis and the metrics used in the classification process and proceeds on to the experimental procedure and results. The paper concludes with a discussion of the results and applicability of using keystroke analysis in practice.

Keystroke Analysis

The concept underlying keystroke analysis (also known as keystroke dynamics) is the ability of a system to recognise patterns, such as characteristic rhythms, during keyboard interactions. Classification of a user is achieved by comparing an input sample against a reference template for the claimed identity and given sufficient similarity the input sample is deemed to have come from the authorised user. The reference template is securely acquired from the user when they enrolled on the system initially. However this template matching process gives rise to a characteristic performance plot between the two main error rates governing biometrics, the False Acceptance Rate (FAR), or the rate at which an impostor is accepted by the system, and the False Rejection Rate (FRR), or the rate at which the authorised user is rejected by the system. A third measure known as the Equal Error Rate (EER), is used as a comparative measure between the biometrics techniques and equates to the point at which the FAR and FRR are equal (Ashbourn, 2000).

A significant amount of prior research has been conducted in this domain, dating back to the 1980s (Leggett & Williams 1988; Joyce & Gupta 1990 and Monrose & Rubin 1999). However, all of these studies have focussed upon alphanumeric inputs from a standard PC keyboard. Little work to date has considered the application of keystroke analysis to a mobile handset, which has obvious tactile and interoperability differences. This paper presents the findings in the second phase of a feasibility study conducted by the authors. The first phase investigated the feasibility of authenticating

users based upon entering numerical data, such as telephone numbers and PIN codes (Clarke et al., 2003), and measuring the inter-keystroke latency (or time between two successive keystrokes) as the unique characteristic rhythm. This study compared and contrasted a number of different pattern recognition and neural network classification algorithms and concluded positively, with a feed-forward multi-layered perceptron neural network proving most successful achieving an EER of 11.3% and 10.4% with static a telephone number and PIN code respectively.

The second phase of the study, as presented in this paper, seeks to evaluate the feasibility of using text message entry as a means of authenticating users. However, unlike typing numbers where 10 digits give 100 possible digraph pairs (different combinations of two keys), in a alphabetic input scenario where there are 26 characters, over 600 digraph pairs exist. Although it would be possible to develop classification algorithms based on a small number of these digraphs, a subscriber could not be guaranteed to enter them in any one situation. As such, this second phase investigation attempts to differentiate between subscribers using the keystroke hold-time (i.e. the time taken to press and release a single key) (Obaidat & Sadoun, 1997). This has the resultant effect of reducing the number of key combinations back down to 26.

Experimental Procedure

The objective of this investigation was to evaluate the feasibility of authenticating subscribers based upon the way in which they type text messages. To this end, software was written to permit the capture of key press (down and up) data, as illustrated in Figure 1. However, it was considered that the standard numerical keypad on a PC keyboard would not be an appropriate means of data entry, as it differs from a mobile handset in terms of both feel and layout, and users would be likely to exhibit a markedly different style when entering data. As such, the data capture was performed using a modified mobile handset, interfaced to a PC through the keyboard connection.

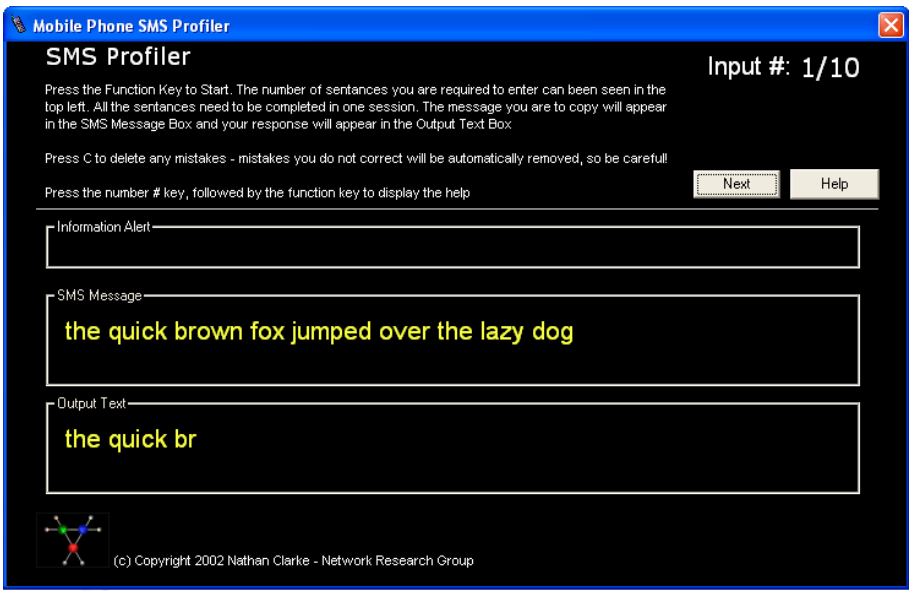


Figure 1 Text Message Input Interface

A total of thirty subjects participated in the study, where each participant entered 30 text messages, split over three sessions, in order to achieve a better representation of subscriber input characteristics. The messages themselves were a mixture of quotes, lines from movies and typical text messages, where the only proviso was to ensure that enough of the characters were repeated to enable classification. This particular study utilised up to 6 of the most common recurring characters in the dataset, as illustrated in Table 1 as inputs to the classification algorithm.

Character	# of Repetition
E	202
T	151
A	140
O	129
N	115
I	107

Table 1 Character Repetitions

The classification techniques utilised in this study were based on the results from the earlier numeric input study, which identified one particular neural network configuration, a feed-forward multi-layered perceptron (FF MLP), as being the most effective. FF MLP networks have particularly good pattern associative properties and provide the ability to solve complex non-linear problems, through the use of multiple perceptron layers (Bishop, 1995). The network topology of a FF MLP is illustrated in Figure 2. For more information regarding keystroke analysis and neural networks, and neural networks in general refer to Brown & Rogers (1993), Cho et al. (2000), Hagan et al. (1996) and Haykin (1999).

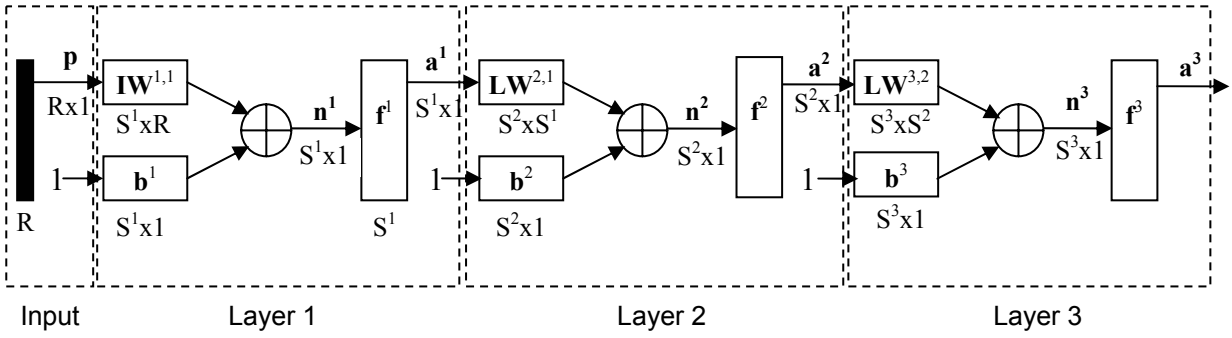


Figure 2 Feed-Forward Multi-Layer Perceptron Neural Network

Using the network, three different investigations were performed. The first was to vary the number of the network parameters, such as the number of neurons in the first and second layers (S^1 and S^2) and to vary the number of training epochs. The second and third investigations are very similar as they both utilise algorithms where the network is fixed but the number of epochs varies in order to optimise the overall performance. The second investigation employs a gradual training schema, where a network is trained for an unusually large number of epochs, but evaluated in terms of performance every 200 epochs, at which stage the network parameters are also stored. The third investigation employs early stopping (Bishop, 1995). Using this schema a user's input data is split into three datasets (training, validation & test datasets), instead of the usual 2 (training and test datasets) and the network uses both the

training and validation datasets in the training algorithm, evaluating the network after each epoch. If the error associated with the network begins to increase then the network stops training and resets the network settings back corresponding to the lowest error. Both of the algorithms have their advantages; the gradual training schema does not stop training if the error increases and the network performance is measured on half of the training data, whereas the early stopping schema has a epoch resolution of 1 and takes considerably less time to train.

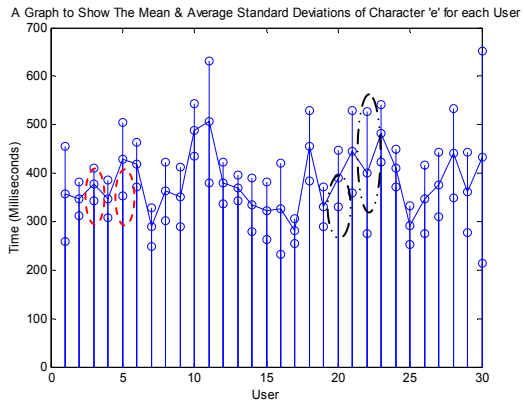
Results

The analysis of the input data allows an insight into the complexities of successfully authenticating a person from a single input vector of hold-times. The problem is that hold-time vectors observed from a single user may incorporate a fairly large spread of values. This spread, otherwise known as variance, is likely to encompass input vectors that closely match other users. Because users' input vectors do not exist on clearly definable classification regions, the problem is that much more complex for the authentication algorithm.

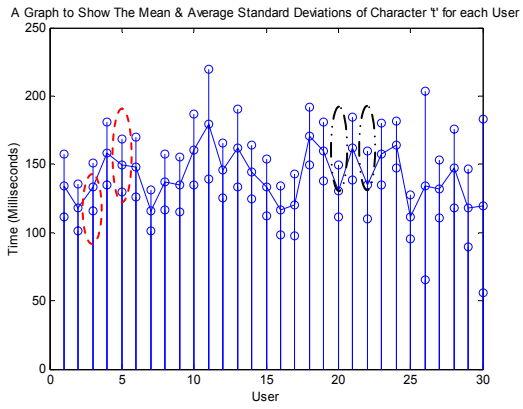
Two types of variance have been identified within the input data:

- inter-sample variance, which ideally would be zero, so that every sample a user inputs would be identical and therefore easier to classify,
- inter-user variance, a measure of the spread of the input samples between users, which would be ideally as large as possible in order to widen the boundaries between classification regions.

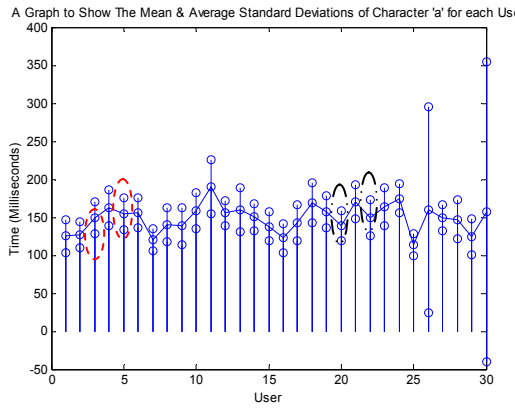
An initial analysis of the inter-sample and inter-user variances indicates that they are by no means ideal. Figure 3 illustrates these relationships with the four most recurrent characters; 'e', 't', 'a', 'o'. It can be seen that many of the subscribers share a similar mean and standard deviation plot, e.g. participants 2 and 4 with character 'e' and participants 19 and 21 with character 't', thereby giving rise to possible poor network performance due to the inability to discriminate between users. However, this relationship does not continue throughout all the characters, as shown in the figure, and since the input vector to the neural network is constructed from a number of characters this should provide the disparities in input data for the network to discriminate against.



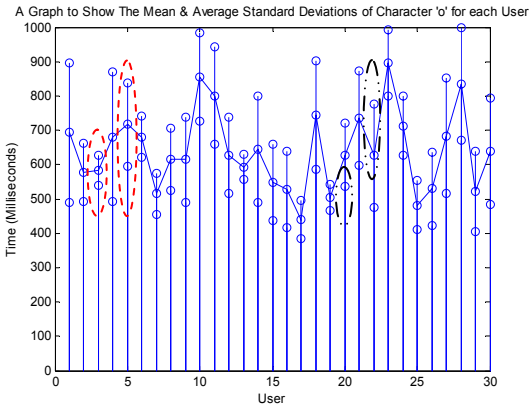
(a) Character 'e'



(b) Character 't'



(c) Character 'a'



(d) Character 'o'

Figure 3 Subscriber Mean & Standard Deviation

The performance results for the three classification algorithms are illustrated in Figure 4, with the gradual training schema proving most successful with an input vector constructed of 5 characters achieving an EER of 17.9%. On average, the early stopping algorithm proved least successful, with the worst EER of 27% - although further fine tuning of the algorithm parameters should decrease the error rate to a similar level as the other algorithms.

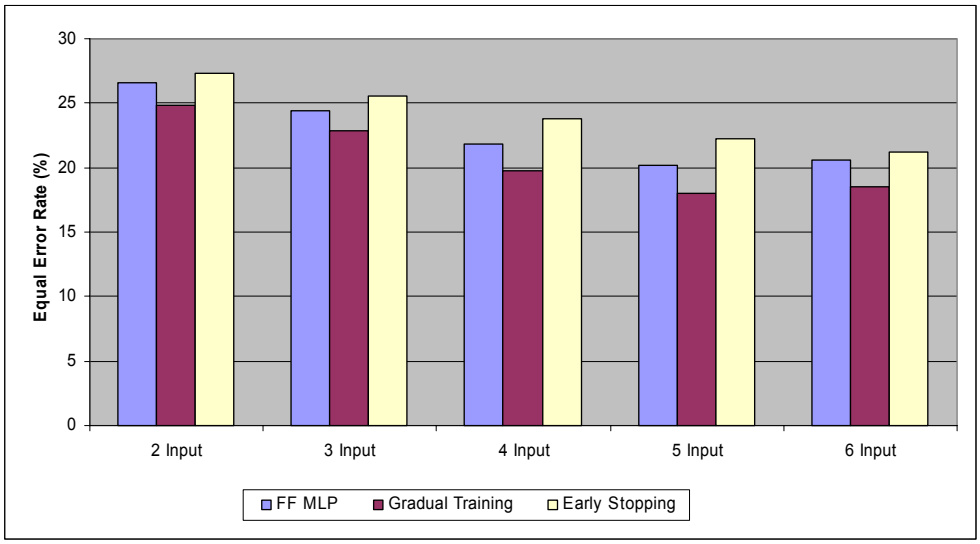


Figure 4 Classification Results for Character Input

An analysis of the results on an individual basis raises some interesting results, as illustrated in Table 2. The most discriminative participant was number 17, achieving an EER of 7.2% with the gradual training algorithm and an EER of 8% with the early stopping technique. Conversely, however, the classification algorithms found it difficult to discriminate between a number of participants, with subscriber 15 achieving an EER of 42.6% - representing a user where keystroke analysis would be unreliable.

Classification Algorithm	# of Inputs	Best		Worst	
		User	EER (%)	User	EER (%)
FF MLP Network Configurations	5	13	8.8	9	34.4
Gradual Training	5	17	7.2	23	30.0
Early Stopping	6	17	8.0	15	42.6

Table 2 Best & Worst Individual EER's

A comparative analysis of the three classification algorithms has shown that although network parameters such as the number of neurons a network has in its hidden layer and indeed to some extent the number of hidden layers it has is important, given a network with enough neurons to compute the problem, which can be arbitrarily set high, network performance can be vastly improved solely dependant on fine tuning the number of epochs the network is trained with – as over and under training the network will result in poor performance measures.

Discussion

The investigations have shown the potential for classification algorithms to correctly discriminate between users with a relatively good degree of accuracy for the majority of subscribers, based on the hold-time of a key. Although the Gradual Training algorithm proved to be the most successful, in a practical situation a trade-off would exist between the computational complexities of one technique next to another. The gradual training algorithm is the most intensive of the three techniques as it is required to perform a large number of training epochs. The Early Stopping algorithm is the least computationally intensive, as the number of training epochs is (usually) smaller than the standard FF MLP algorithm. Therefore the chosen technique would somewhat depend upon on the time permitted for training and the computational power of the processing device.

The hold-time is an unusual keystroke characteristic to use on its own, but has proved useful in this investigation as it avoided the problem of sampling and profiling the large number of digraph pair combinations. Authentication performance could however be increased if the classification algorithm utilised a number of techniques to classify the subscriber, capitalising on the specific content of the message. For instance, in a worst case scenario, the hold-time classification algorithms presented in this paper could be used on messages with dynamic content utilising between 2 and 6 of the most commonly recurrent characters. However, a next stage, dependant on content would be to perform classification on commonly reoccurring static words, such as "hello", "meeting" and "c u later" where both inter-keystroke latency and hold-time characteristics can be used to better classify the subscriber (Obaidat, 1997). A further stage again would be to authenticate the user, using either or both of the previous techniques, more than once within the same text message, with the system responding on the aggregate result. For instance, a system with a FAR of 20% would see this reduce to 4% with two authentications and 0.16% with three authentications.

Conclusions

The investigation as a whole (including the first phase) has shown that it is feasible to authenticate some mobile handset users by their typing patterns. In particular, this study has demonstrated the ability of neural networks to differentiate users based upon two different keystroke characteristics – the traditional keystrokes latency, or time between two successive keystrokes and the hold-time of a key press. Undoubtedly, the overall process of discrimination would be improved if both techniques could be used in conjunction with each other; however this only performs well with static inputs, such as a PIN code or frequently occurring words.

Keystroke analysis has proven to be a promising technique, having achieved good results in both the numeric and alphabetic input investigations. Although, the ability for classification algorithms to correctly discriminate between subscribers very successfully is low, the majority of users are experiencing fair to good performance. However, as with all biometrics a number of participants remain that are unable to be correctly authenticated to a reasonable degree. Therefore, any practical implementation of keystroke analysis would require a flexible framework, ensuring that even those subscribers for whom keystroke analysis is not a viable approach are still provided with an adequate level of security. The flexible framework would form a hybrid authentication algorithm, capable of utilising a number of difference biometrics (such as facial recognition, speaker verification and keystroke analysis) in an intelligent manner in order to achieve both non-intrusive and continuous authentication of the subscriber.

References

Ashbourn, J. (2000). Biometric - Advanced Identity Verification. The Complete Guide. Springer.

BBC. (2002). Huge surge in mobile phone thefts, BBC News Report, 8th January 2002. http://news/bbc.co.uk/hi/english/uk/newsid_1748000/1748258.htm

Bishop, M. (1995). Neural Networks for Pattern Classification. Oxford University Press.

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

Cellular Online. (2003). Latest Global, Handset, Base Station & Regional Cellular Statistics. June 2003. www.cellular.co.za

Cho, S., Han, C., Han D., Kin, H. (2000). Web Based Keystroke Dynamics Identity Verification Using Neural Networks. Journal of Organisational Computing & Electronic Commerce, vol. 10, pp 295-307.

Clarke, N., Furnell, S., Rodwell, P., Reynolds, P. (2001). Acceptance of Subscriber Authentication for Mobile Telephony Devices. Computers & Security, vol. 21, no.3, pp. 220-228.

Clarke, N., Furnell, S., Lines, B., Reynolds, P. (2003). "Using Keystroke Analysis as a mechanism for Subscriber Authentication on Mobile Handsets". Proceedings of the IFIP SEC 2003, pp97-108.

GSM World.com. (2003). World Cellular Subscribers 2002. www.gsmworld.com/news/statistics/index.shtml.

Hagan, M., Demuth, H., Beale, M. (1996). Neural Network Design. PWS Publishing Company

Haykin, S. (1999). Neural Networks: A Comprehensive Foundation (2nd Edition). Prentice Hall.

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

Monrose, R., Rubin, A. (1999). Keystroke Dynamics as a Biometric for Authentication. Future Generation Computer Systems, 16(4) pp 351-359.

Obaidat M., Sadoun, B. (1997). Verification of Computer Uses Using Keystroke Dynamics. IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics, vol. 27, no. 2, pp.261-269

Wood, H. (1977). The Use of Passwords for Controlling the Access to Remote Computer Systems and Services. Computers and Security, Vol.3. C.T. Dinardo, Ed., p.137. Montvale, New Jersey: AFIPS Press.