Investigating, Implementing and Evaluating Client-Side Keystroke Analysis User Authentication for Web Sites

C.G.Hocking and P.S.Dowland

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

Abstract

In today's electronic information society security is certainly the challenge of the 21st century. Driven by the need to counteract ever more determined cyber criminals, focus is turning upon biometric security as a means of ensuring protection and privacy of the most sensitive information. Keystroke dynamics and the analysis of the way people type their password is one method drawing significant attention because of its non-invasive nature and lack of requirements for expensive additional hardware. The majority of research has been executed in a local area network environment but this paper examines the possibility of implementing a solution for web sites and whether refinement of comparison data over time would lead to increasing improvement. Although the web site solution is not conclusive further investigation into profile refinement indicates that this may not have a significant impact on authentication rates. The observed typing characteristics of subjects at the beginning, throughout and at the end of the testing period offer little evidence for implementing a profile refinement strategy.

Keywords

Keystroke, authentication, biometric

1 Introduction

Telephones, computers and other electronic gadgetry have evolved from office-based machinery into household necessities, fashion accessories and even symbols of status. Reliance upon them has proportionately grown and as everyday financial and business activities move evermore into cyberspace, the temptation for others to misuse the trust we place in such devices has correspondingly increased.

Protection of software systems and the information they hold has generally relied upon the ubiquitous user identification and password security concept. Although this is not necessarily a bad idea human beings are inherently forgetful creatures and have tendencies to trust others. Passwords are often written down, placed in desk drawers, left attached to computer monitors on post-it notes or even or divulged for a bar of chocolate (BBC, 2004), and so the search to either replace or bolster this method of authentication is underway.

Electronic communication stretches back as far as the 1830s when Samuel Morse invented the electric telegraph and enabled messages to be transmitted from one side of a country to the other. Since this time expert users have always anecdotally been reported as being able to identify other operators via the style and rhythm with which they sent their communications. Extrapolating the principle this work investigates the possibility of using keystroke dynamics and the way in which people type as a secondary tier of user authentication on web sites. Biometric authentication using a keyboard is an ideal candidate because it requires no expensive additional equipment; it can be implemented in nearly any situation and does not rely upon specific knowledge but rather an individual's natural characteristics.

Biometric systems such as this are generally measured by three factors; the False Rejection Rate (FRR), the proportion of times an authentic user is rejected as being an impostor; the False Acceptance Rate (FAR), the rate at which an impostor successfully passes the authentication procedure; and the Equal Error Rate (EER), the point at which the FRR and FAR are identical. As figure 1 indicates the tolerance or threshold setting directly influences both the FRR and FAR; any solution aims to achieve an EER as near zero as possible with the ultimate system yielding a zero rate.

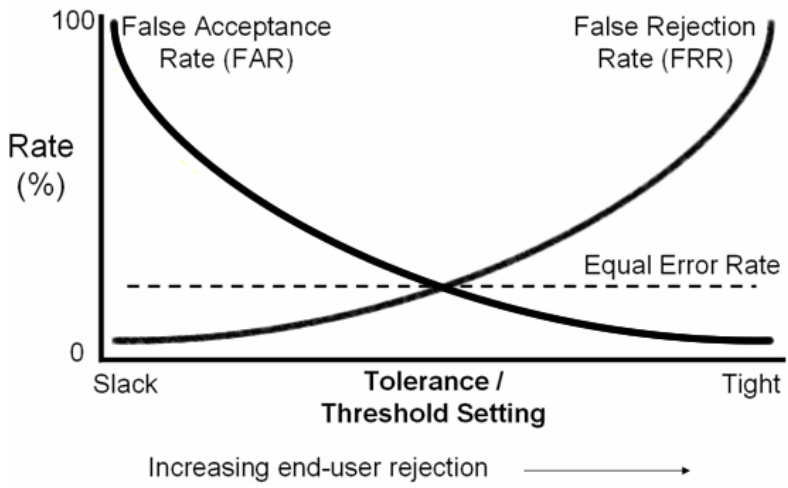

Figure 1: Biometric system result rates

A web site will be developed that allows individuals to register themselves and create a personal timing pattern that can then be tested against other subjects typing the same pass phrase. Once a period of testing has been completed the captured data will be retrieved and analysed off-line to establish if this is indeed a viable method of identity confirmation and whether further refinement over time would improve the observed success rate.

2 Background

Classification of typing characteristics or keyboard dynamics is generally based upon the analysis of consecutive keystroke (digraph) timings; the interval between key depressions, the gap between the release of the first key and the pressing of the second, and the total digraph length. In some instances the inter-key latency can be negative as a fast typist will depress the second key before releasing the first. Figure 2 outlines the significant events when typing the word 'the' where $K2_D$ represents the precise time of the depression of key two and $K3_U$ the release (key up) of key three.

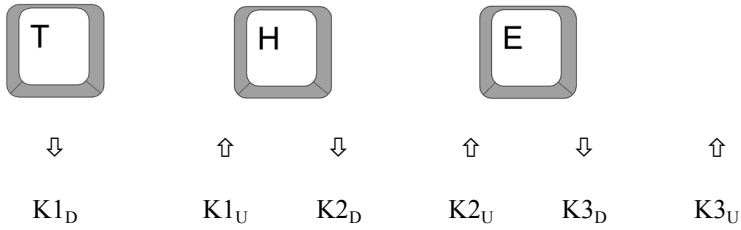

Figure 2: Sequence of events when the word 'the' is typed

This leads us to understand that for the digraph 'th', 't' was held down for $K1_U - K1_D$ time intervals, the inter-key latency was $K2_D - K1_U$ and the entire digraph took $\max(K1_U, K2_U) - K1_D$ to type.

The first study into keystroke analysis was in 1980 and involved the observation of a secretarial pool of seven typists. Each typist provided two typing samples, the digraph means of which were examined to see if they had come from the same subject using a T-test. An EER of around 5% was achieved in this preliminary study suggesting that keystroke analysis was a viable method for user identification and authentication (Gaines et al, 1980). First Leggett and Williams (1988) and then Joyce et al (1990) extended this early work by introducing the classification technique when comparing sampled timings with a mean test pattern synthesised from eight master samples. Although probability of digraph occurrence was not found to have a significant impact upon lowering the EER, the use of structured text during a training cycle was identified as being of greater benefit (Monrose et al, 1994). More recently neural networks have been used to identify typing patterns and have returned observations as low as 1% across 21 subjects (Cho et al, 2000).

Many researchers have also established the suitability of using such a technique to identify and authenticate (Bartolacci et al, 2005; Bergadano et al, 2003; Dowland and Furnell, 2004; Napier 1995) but in all of these cases the testing and data capture has been performed either on stand-alone PCs or across a local area network. Ngo et al (2006) successfully performed authentication across the Internet and it is with these and similar works that research is now turning towards refinement of a solution provision as opposed to proof of concept.

3 Research Detail

This work has involved the creation of a web site using a mixture of HTML and JavaScript with the server processing being undertaken by PHP and the data written to an Access database. The site required subjects to register a personal timing template which was synthesised from the 10 best entries of 12 attempts. Details of each digraph punched were stored in readiness for data analysis which was to occur off-line. Following registration individuals took part in the testing phase during which they were repeatedly requested to enter their own and a mystery user's password. Digraph timings were captured using HTML events such as 'onKeyDown', 'onKeyUp' and 'onBlur' to trigger appropriate JavaScript routines which used the internal PC clock to record the precise time. If at any time during typing of a password a mistake was made or inappropriate keys pressed the entry was entirely rejected and the individual requested to resubmit.

The approach to analysis was based upon proposals made by Magalhães and Santos (2005) but this work parameterised some of the settings and thresholds to investigate the impact upon the results. Additional work beyond the normal FRR, FAR and EER analysis was designed to investigate whether refinement of an individual's profile over time would indeed impact and improve the results. Upon validation of an entered password the timing of each character pair would be tested to see if it lay between an upper and lower bound, calculated from the password owner's profile. For each success 1 point was scored and for each failure zero. The average point score was then calculated and if this exceeded a specified 'acceptance' threshold the user was deemed to be authentic. The lower (b_l) and upper bounds (b_u) are calculated as shown below, using the profile mean ($\overline{p}$), profile median (p_m), a lower threshold (T_l), an upper threshold (T_u) and the profile standard deviation (σ).

$$b_l = \min(p_m, \overline{p}) \times \left(T_l - \left(\frac{\sigma}{\overline{p}} \right) \right) \quad \text{and} \quad b_u = \max(p_m, \overline{p}) \times \left(T_u + \left(\frac{\sigma}{\overline{p}} \right) \right)$$

The two thresholds are to be varied in the range (0.900-0.975) and (1.025-1.100).

4 Results

The web site and all appropriate data capture routines were developed and implemented. During a 12 day testing period 16 users fully registered their details and performed suitable amounts of testing to deem them significant candidates. The testing process was incentivised by the display of a 'Top 10 Testers' league table on the data entry webpage, with each scoring one point for the entry of a mystery password. In all 6999 tests were recorded, 4083 'own' and 2916 'mystery' passwords. Off-line analysis of this data using combinations of acceptance=0.40,0.45,0.50,0.55,0.60; lower bound threshold=0.900,0.950,0.975;

upper bound threshold=0.900,0.950,0.975; produced the following summarised results:

Acceptance	Lower Threshold	Upper Threshold	Self Failed	Other Passed	FRR	FAR
0.40	0.950	1.050	434	820	10.62944%	28.12071%
0.45	0.950	1.050	986	544	24.14891%	18.65569%
0.50	0.950	1.050	986	544	24.14891%	18.65569%
0.55	0.950	1.050	1226	290	30.02694%	9.94513%
0.60	0.950	1.050	1882	173	46.09356%	5.93278%
0.40	0.900	1.050	191	1142	4.67793%	39.16324%
0.45	0.900	1.050	517	833	12.66226%	28.56653%
0.50	0.900	1.050	517	833	12.66226%	28.56653%
0.55	0.900	1.050	668	526	16.36052%	18.03841%
0.60	0.900	1.050	1229	331	30.10042%	11.35117%
0.40	0.900	1.100	87	1351	2.13079%	46.33059%
0.45	0.900	1.100	264	1044	6.46583%	35.80247%
0.50	0.900	1.100	264	1044	6.46583%	35.80247%
0.55	0.900	1.100	378	717	9.25790%	24.58848%
0.60	0.900	1.100	863	479	21.13642%	16.42661%
0.40	0.950	1.100	279	1064	6.83321%	36.48834%
0.45	0.950	1.100	668	751	16.36052%	25.75446%
0.50	0.950	1.100	668	751	16.36052%	25.75446%
0.55	0.950	1.100	864	458	21.16091%	15.70645%
0.60	0.950	1.100	1508	275	36.93363%	9.43073%
0.40	0.975	1.025	687	595	16.82586%	20.40466%
0.45	0.975	1.025	1315	364	32.20671%	12.48285%
0.50	0.975	1.025	1315	364	32.20671%	12.48285%
0.55	0.975	1.025	1632	162	39.97061%	5.55556%
0.60	0.975	1.025	2283	92	55.91477%	3.15501%

Table 1: Summarised testing results

The combination of bounds that exhibited the lowest EER was 0.900 and 1.100 and the graphical representation of these is shown in figure 3.

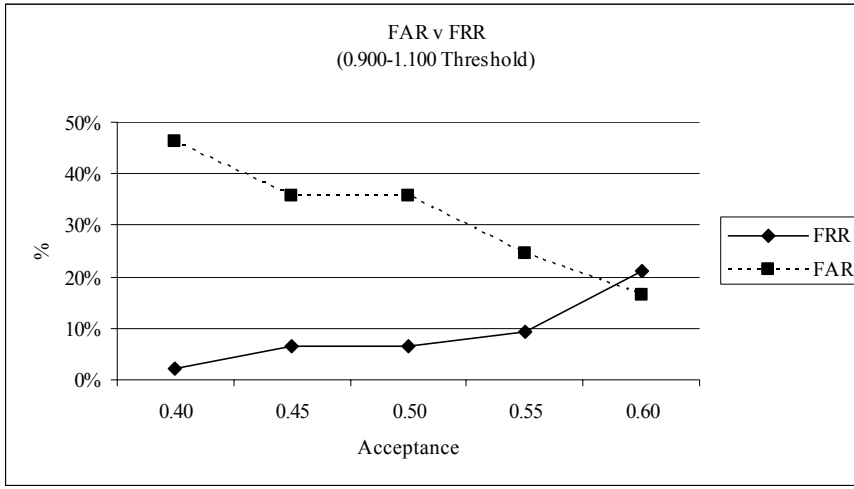

Figure 3: Graph of best results from table 1

Investigation then continued into the longevity of profiles; for this purpose the 'self test' data for the most prolific tester was isolated. This particular subject completed 2004 entries of their own password which consisted of nine single case alphanumeric characters (ilewamh04) and the data provided the following results:

	Mean				Standard deviation			
Digraph	Profile	All entries	First 10	Last 10	Profile	All entries	First 10	Last 10
i-l	238.600	254.880	251.000	264.300	7.031	56.902	29.833	20.174
l-e	223.100	266.605	221.500	333.900	20.940	88.799	61.136	32.191
e-w	244.800	246.415	249.500	245.400	15.465	73.758	18.511	30.771
w-a	262.200	254.624	260.600	247.700	33.409	114.204	30.474	15.963
a-m	197.900	198.471	208.900	200.500	31.297	70.080	25.324	28.218
m-h	284.000	264.134	271.400	284.400	11.498	42.525	17.374	14.678
h-0	426.000	428.607	480.300	555.400	81.884	120.599	140.155	173.959
0-4	276.200	225.124	255.700	259.800	39.364	67.979	29.678	76.349

Table 2: Results of the most prolific tester over time

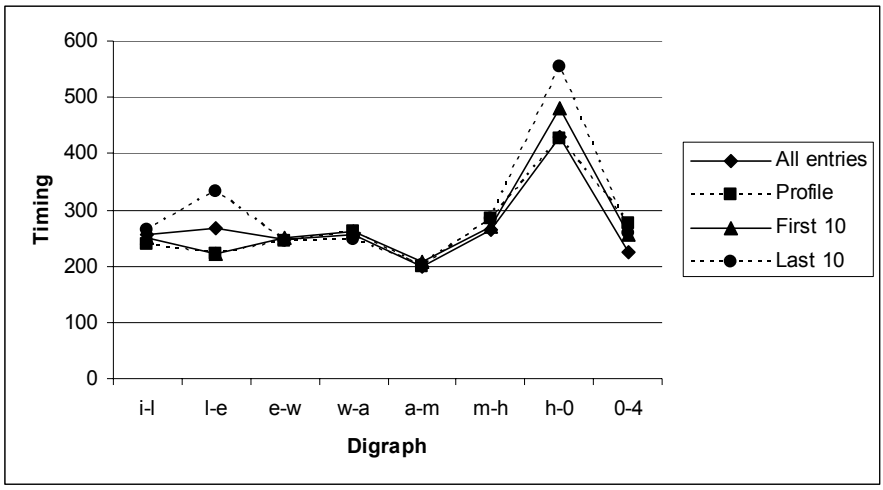

Figure 4: Mean timings of the most prolific user

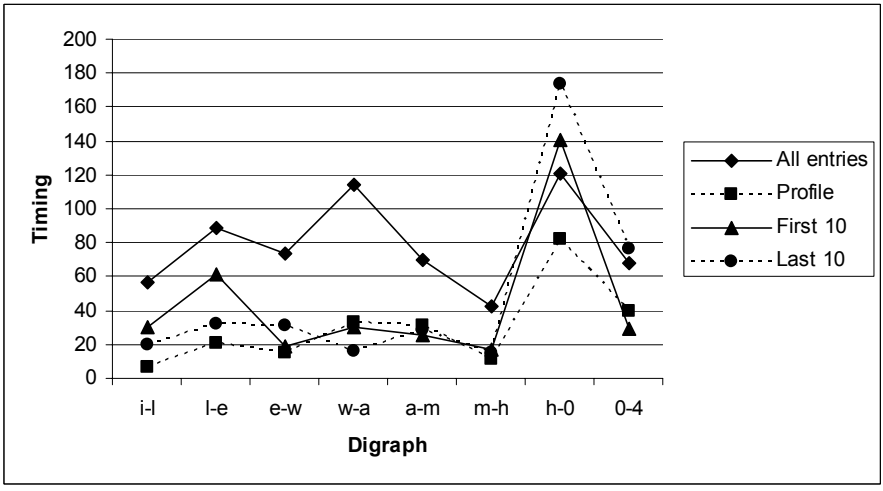

Figure 5: Timing standard deviation of the most prolific user

5 Discussion

Analysis of the keystroke timing results yields a plot (figure 3) which suggests an Equal Error Rate of approximately 18%, much higher than desired. This is certainly not low enough to conclude that keystroke analysis during this testing cycle was sufficiently accurate to be employed as a method of user authentication. It should be noted that the testing phase was compressed into a 12 day period and with the nature of the competition some subjects submitted a large number of tests. Of course the

quantity of data was welcomed to enable some meaningful calculations to be undertaken but contrary to this with so many repetitions passwords become familiar. Indeed at the outset because passwords were being openly displayed to other users, individuals were encouraged to select an unfamiliar password rather than pick one that was in regular use. Consequently at the beginning the only experience they had of using a particular pass phrase was their 12 repetitions during profile registration. Nearly everybody was as inexperienced as anybody else and one would expect a high standard deviation to be exhibited leading to higher FRR and FAR indices.

Typing expertise is a factor in exhibited results although it is not suitable to be used as a metric during the evaluation process. Anecdotally one of the subjects, a touch typist who uses all 10 fingers, reported they had trouble typing one of the test passwords, the name 'rebecca'. This for them was solely left-handed typing because of the position of the keys on the keyboard and the transition between 'e-b-e' was an "unnatural finger spread" which caused a disruption to fluency. A trait that was clearly visible in the captured data.

More meaningful results were exhibited by the investigation into user profile improvement. The two graphs produced by the results (figures 4 and 5) provided some very interesting results. Figure 4 exhibits the diagraph mean timings for the profile, all tests, the first 10 tests and the last 10 tests of a user who submitted 2004 individual entries of their own password. There is very little variation between the four profiles but the biggest change is noted in the 'l-e' diagraph combination which worsens by approximately 50% over time. This is perhaps related to the number of executed cycles and although the subject stated they are a 'good' typist the distance between the letters 'l' and 'e' on a keyboard is relatively large indicating that tiredness and perhaps the use of few fingers impacted on the timings. With a touch typist the digraph combination would be completed using both hands and so the distance to travel eliminated as a factor. The tiredness theory is further supported by figure 5 which shows an increased standard deviation across 'all tests' from the lowest set, the profile.

These findings suggest that to refine the profile as time goes on would not significantly tighten variation and is therefore of little benefit. It may be argued that with a high repetition rate apathy increases and concentration wanes leading to the observed discrepancies.

6 Conclusion

This research was carried out under time constraints and the development of the web site with the complexities within impinged upon the testing phase and the quantity of data gathered. Sixteen significant users are less than would be ideally chosen and so the results are a little unclear. Reflection upon self testing and the improvement exhibited over time appears to suggest that profile refinement would not impact upon the investigated user. Further work would ideally be targeted at this stronger vein of enquiry and compare the effects of time and repetition on other subjects.

The quantity of data would provide reasonable scope to refine the algorithms used in an attempt to identify typing patterns. An Equal Error Rate of 18% is too large to conclude this is a viable form of user identification and authentication with the techniques employed but eminent research and anecdotal reports compound to dispute this and further investigation is certainly required. Some of the fuzziness maybe due to JavaScript being employed to capture the precise timings and comparisons should be executed using alternative methods of timing to clarify this dilemma.

7 References

Bartolacci, G., Curtin, M., Katzenberg, M., Nwana, N., Cha, S. and Tappert, C. (2005), 'Long-Text Keystroke Biometric Applications over the Internet' *Proceedings of Student/Faculty Research Day, CSIS, Pace University*

BBC (2004), 'Passwords Revealed by Sweet Deal' available: http://news.bbc.co.uk/1/hi/technology/3639679.stm [accessed 14 Mar 2007]

Bergadano, F., Gunetti, D. and Picardi, C. (2003), 'Identity Verification Through Dynamic Keystroke Analysis' *Intelligent Data Analysis,* vol. 7 pp. 469–496

Cho, S., Han, C., Han, D. H. and Kim, H. I. (2000), 'Web-Based Keystroke Dynamics Identity Verification Using Neural Network' *Journal of Organizational Computing and Electronic Commerce,* vol. 10 no. 4 pp. 295–307

Dowland, P. S. and Furnell, S. M. (2004) 'A Long-term Trial of Keystroke Profiling using Digraph, Trigraph and Keyword Latencies' *Proceedings of IFIP/SEC 2004 - 19th International Conference on Information Security*, pp. 275-289

Gaines, R., Lisowski, W., Press, S. and Shapiro, N. (1980), 'Authentication by Keystroke Timing: some preliminary results' *Rand Report R-256-NSF. Rand Corporation*

Joyce, R. and Gupta, G. (1990) 'User authorization based on keystroke latencies' *Communications of the ACM,* vol. 33 no. 2 pp. 168–176

Leggett, J. and Williams, G. (1988) 'Verifying identity via keystroke characteristics' *International Journal of Man-Machine Studies,* vol. 28 no. 1 pp. 67-76

Magalhães, S. T. and Santos, H. D (2005) 'An Improved Statistical Keystroke Dynamics Algorithm', Proceedings of the IADIS Virtual Multi Conference on Computer Science and Information Systems, 2005

Monrose, F., Rubin, A. D. (2000) 'Keystroke dynamics as a biometric for authentication' *Future Generation Computer Systems*, vol. 16 pp. 351–359

Napier, R., Laverty, W., Mahar, D., Henderson, R., Hiron, M. and Wagner, M. (1995) 'Keyboard user verification: toward an accurate, efficient, and ecologically valid algorithm' *International Journal of Human Computer Studies*, vol. 43 pp. 213-222

Ngo, G., Simone, J. and St. Fort, H. (2006) 'Developing a Java-Based Keystroke Biometric System for Long-Text Input' *Proceedings of Student/Faculty Research Day, CSIS, Pace University*