Keyboard user verification: toward an accurate, efficient, and ecologically valid algorithm

Renee Napier, William Laverty, Doug Mahar, Ron Henderson, Michael Hiron and Michael Wagner

The TRUST† Project, Research School of Information Sciences and Engineering, Australian National University, ACT 0200, Australia. email: renee @ trust. anu. edu. au

(Received 30 Sept 1994 and accepted in revised form 23 June 1995)

This paper proposes new measures of individual differences in typing behaviour which provide a means of accurately verifying the identity of the typist. A first study examined the efficacy of a multivariate measure of inter-key latencies and a probabilistic discriminator statistic in conjunction with an individual filtering system which eliminates occasional disfluent keystrokes. The results indicate that, under optimum conditions but with a very small test sample, these measures lead to better typist verification than measures suggested earlier by Umphress and Williams and then Leggett and Williams. A second study validated the improved algorithm under more ecologically valid conditions and showed that when training and test sessions were separated by one week, typist verification using the new algorithm achieved combined false-acceptance and false-rejection rates of 0.9% and 3.8% for test samples of 300 and 50 digraphs respectively. © 1995 Academic Press Limited

1. Introduction

The verification of the identity of computer users is essential in many systems which require that access to sensitive data, or specific data transactions, are restricted to authorized users. The password paradigm for such systems is notoriously insecure as it is virtually impossible to enforce the secure management of passwords by the users. Moreover, in systems which require user identity verification whenever a restricted operation is attempted, the typing of passwords may have to occur so frequently as to become entirely unacceptable to the users of the system.

Recently, biometric systems have been proposed either as an addition or as an alternative to password-based security systems. Biometric systems can be broadly delineated into those which use physiological measures, eg. digital, manual, facial or retinal patterns, and those which also use behavioural measures, eg. signature, voice and keystroke dynamics (Sherman, 1992; Miller, 1994). While both types or biometric systems potentially offer greater security than password-based systems, behavioural systems offer the advantage of unobtrusively monitoring “normal” user behaviour like typed, spoken or written input to the computer. It has also been advanced that behaviourally based verification systems tend to be relatively low in cost as well as computationally inexpensive (Sherman, 1992; Miller, 1994).

A particularly useful class of behavioural biometric techniques utilizes inter-keystroke latencies from a user's keyboard input to verify the identity of the user.

† This research has been carried out on behalf of the Harry Triguboff AM Research Syndicate.
Two consecutive characters are considered a "digraph" and the time between them the digraph latency. Umphress and Williams (1985) and later Leggett and Williams (1988) developed a technique that utilized digraph latency as an identifying characteristic. In their system a reference profile for each user is constructed consisting of the mean latency between the onsets (key-down) of successive keystrokes for each pair of keys (digraphs). The latency of digraphs typed during the test phase is then compared with the reference data for that digraph by measuring the proportion of test digraphs lying within a specified distance (expressed in terms of the S.D. of all reference profile digraph latencies) of the relevant reference profile mean. Umphress & Williams, (1985) reported that verification accuracy was optimal when the maximum allowable separation between test and reference profile digraph latencies was set at 0.5 overall reference profile standard deviations, and the proportion of test digraphs required to pass this test was set at 0.6. Throughout both reference profile construction and test phases, all digraphs with latencies above 750 ms were excluded to control for cognitive breaks and other disfluencies. The original paper reported false acceptance rates (FARs) of 6% and false rejection rates (FRRs) of 12% using a test sample of 300 characters and reference profile comprised of 1400 characters. Leggett and Williams (1988) achieved FARs and FRRs of 5 and 5.5% respectively in a study using 36 subjects, 500 ms low pass filter, a reference profile derived from 1075 keystrokes, and a test sample of 537 keystrokes (538 digraphs). The authors comment that the difference in error rates between the original result and the later one probably only reflects sample size. It must be noted that the original study (Umphress & Williams, 1985) reported a dual test system which included an addition test of mean overall latency. In a subsequent study (Leggett & Williams, 1988) it was decided that the additional test provided no more discriminating power than the digraph test alone, and so was not used.

There have been some developments made on the traditional system. Leggett, Williams, Usnick and Longnecker (1991) utilized a sequential statistics theory to develop a dynamic verifier system. They utilized the eight most common digraph cells from a test sample of 501 digraphs that achieved a combined FAR plus FRR of 24%. Later, in an examination of the traditional approach, Mahar, Henderson, Laverty, Lawrie, Hiron, Gough and Wagner (1994) showed that the use of individual S.D.s for each profile cell instead of a pooled S.D. of all reference profile digraph latencies produced superior verification rates. Maher et al. (1994) also closely examined the maximum separation and passed proportion values for the algorithm. They argue that the optimum values will vary according to particular data sets because the algorithm is greatly affected by differences in digraph distribution variability, which may vary from data set to data set. For their particular data set, they established that the optimum values were 0.5 for maximum separation and 0.7 for passed proportion of digraphs. Further, Laverty (1994) suggested that a filter method based on Tukey's (1977) interquartile method for identifying outliers would, logically, be a more appropriate filtering system than the blanket 500 ms low pass filter system adopted by Leggett & Williams (1988).

There remains a number of areas in the original Leggett and Williams (1988) approach yet to be examined. Their traditional approach used the key down-to-down time as the base unit of measure. However, down-to-down time may be further delineated into two orthogonal components: total time the first key is
depressed and, the time between first key release and second key depression. Brown and Rogers (1993) have used these two components in their type signature neural network verification system, however, they did not provide adequate data to ascertain if the use of separate orthogonal digraph components added significant predictive power to the more traditional down-to-down digraph latency approach.

Another area for examination is the use of a dichotomous, pass/fail criterion on individual digraph information. If individual differences between subjects occurs in only a small number of digraphs then Umphress and Williams' (1988) pass/fail discrimination procedure will ensure that these differences are given very little weight, especially if those digraphs have an infrequent occurrence in the test sample. To address this likelihood, a discrimination procedure that provides a distance measure between test score and reference profile mean is necessary.

2. Study 1

The aim of the first study was to compare the performance of proposed improvements to the original typist verification algorithm with the performance of the unadulterated original version proposed by Leggett and Williams (1988). To this end, two components of the algorithm have been considered. Firstly, a bivariate measure of latency was compared with the single measure and, secondly, a chi-square based distance between test digraphs and reference profiles was compared with their pass/fail \( z \)-test. Both new elements should improve the performance of the algorithm and the inclusion of both in the algorithm should improve its performance even further.

2.1. Method

2.1.1. Subjects

Sixty-seven subjects participated in the study. Fifty-eight were undergraduate psychology students while nine were TRUST project staff. Subjects were chosen regardless of typing skill, age or gender.

2.1.2. Design

This study employed a two-factor, repeated-measure design. The first factor was the latency measure used. The measure was either univariate which consisted of only key down to key down time (latency), or multivariate which consisted of key down to key up time (period) as well as key up to key down time (gap).

The second factor was the statistic used to test for maximum separation between test and profile latencies. Two statistics were compared. The first was the traditional Leggett and Williams (1988) \( z \)-test set to 0.5 S.D.s which produces a pass/fail for each digraph. The valid proportion required for a "pass" was set to 0.7. The second statistic was the chi-square which measured the distance between test digraph and reference profile. This statistic calculated a Mahalonobis distance (Dempster, 1969) score for both gap and period as well as the correlation between the two. The
dependent measures were the FARs and FRRs observed for each subject in each condition.

2.1.3. Procedure
The procedure used generally followed that of Leggett and Williams (1988). A transcription paradigm was used in which the subjects typed approximately 1000 characters presented as a series of either single words or phrases displayed on the VDU. Words and phrases were chosen to provide a fair representation of digraphs as they are used in typed English. All subjects typed the same text, although the order of stimuli was randomized across subjects. Both period and latency were recorded for each digraph and data was analysed either as period and latency or as the sum of the two. A reference profile for each subject was then constructed. These profiles included only those digraphs consisting of a combination of lower case alpha keys and/or the space bar, but excluded outlier digraphs as defined by a subject–digraph specific filtering algorithm (Laverty, 1994). This system excludes digraphs whose gap time exceeds limits of 1.5 S.D.s from the 75th and 25th quartiles. Finally, to ensure the robustness of the digraph latency means and S.D.s, digraph cells for which less than five observations were available were also excluded.

The false acceptance rate was calculated for each subject by drawing a random sample of 10 individual digraphs from each of the other 66 subject’s data and assessing the proportion of these samples which passed each of the maximum separation statistics with α set to 0.05 for each statistic. Similarly, the FRR was calculated for each subject by drawing 66 random samples of 10 individual digraphs from the subject’s reference profile, then assessing the proportion of these samples which passed each statistic.

It must be noted that the “bootstrapping” sampling method used in the current study produces an optimum outcome for both conditions. The method samples test data from the common pool of data used to construct the reference profile, and filters test samples against their own reference profile distributions. Caution must be advised when interpreting the absolute results of the current study as this sampling method will clearly produce optimistic error rates. However, as the purpose of the current study was to refine the algorithm by comparing components of the system, the sampling method was kept constant across conditions thus providing a consistent platform for comparison.

2.2. RESULTS AND DISCUSSION
FARs and FRRs were added together to produce an indice of overall verification accuracy. FARs and FRRs ranged from 24% to 2%, with the poorest cell appearing in the univariate z-test condition, and the best in the multivariate, chi-square condition. Examination of mean combined FARs and FRRs (see Figure 1) shows that the multivariate measure produced significantly lower error rates (M = 4%, and 2%) than respective univariate measures (24% and 11%) in both statistic conditions, $F(1,66) = 237.82, P < 0.001$. These results show that the measurement of latency, in terms of its two orthogonal components of gap and period, significantly increases the sensitivity of the approach.

Similarly, there was a significant effect of statistic type on combined FAR and
FRRs. Using both multivariate and univariate measures the chi-square rule (M = 11% and 2%) produced significantly fewer errors than the simple z-test (24% and 4%), $F(1,66) = 206.49, P < 0.001$. The existence of a significant interaction between statistic type and inter keystroke latency measure, $F(1,66) = 166.21, P < 0.001$, however, suggests that this effect of statistic is not constant over types of measurement. An examination of Figure 1 reveals that while the multivariate measure improves error rates in both statistic conditions, this improvement is larger in the z-test condition than the chi-square condition.

An examination of means suggest that the multivariate chi-square condition produced the best combined error rates (2%). Tukey’s HSD test demonstrates that this condition produced significantly lower error rates than all other conditions (all significant at the 0.001 level) including the multivariate z-test condition which produces the next best error rates (4%). These results suggest that the two variations on the traditional Leggett and Williams (1988) algorithm are additive; individually, the use of the multivariate measure of latency and the use of the chi-square statistic improves verification rates, while the combined use of both improves the performance of the algorithm even further.

While the absolute verification rates achieved here may have been influenced by the particular sampling procedure used, the relative differences between conditions remain valid. The present procedure used a “bootstrapping” sampling method, that is, it sampled test data from the data pool which forms the reference profile. However, when applied to a real security system, this would be impossible, and indeed illogical. Test data, by definition, comes from an unknown distribution. For the algorithm to be considered ecologically valid it must sample test and reference profile data from separate data sets. Indeed, the use of separate data sets, recorded at different times, would also enable testing of temporal stability. Further, impostor test data, in the current study, was filtered using information from the impostors reference profile. This information would, obviously, not be available to the
algorithm in an applied setting where filtering of test data would only be based on the distribution of test data itself. As a result, there is still some work to be done to ensure that the current algorithm is transferable to a more realistic or "ecologically valid" setting.

The results of the current study clearly demonstrate that under optimal experimental conditions Leggett and Williams' (1988) keyboard verification paradigm benefits substantially from systematic refinement. Verification rates were significantly improved by the use of a multivariate measure of keystroke latency, and by the use of a distance statistic to test for maximum separation between test and reference profile digraph latency. Further, the effects of these two variations appear to be additive: the best performing algorithm is that which utilises both a multivariate latency measure and a chi-square statistic.

3. Study 2

The first study has shown that typist verification using keystroke digraph statistics can be substantially improved by the use of a bivariate measure of latency and an appropriate distance measure. However, there is also a need to examine the verification algorithm under more ecologically valid conditions and, more specifically, to test for temporal stability. In an applied setting, data forming the reference profile and data tested against it are gathered at different times. Furthermore, disfluency filtering of test data must not be based on any knowledge about a specific impostor.

The aim of the second study was therefore to validate the proposed algorithm under more "difficult" but more "realistic" and thus ecologically more valid conditions. To this end, the new algorithm was tested with independent test data collected 1 week after the training data, and results were again compared with those from the algorithm proposed by Leggett and Williams (1988). Under these conditions both the original and improved algorithms will perform more poorly than when tested using the bootstrapping sampling method, but verification rates in the improved condition should perform significantly better than those in the original condition.

3.1. METHOD

3.1.1. Subjects
Twenty-four subjects participated in the study. Ten were undergraduate psychology students while 14 were TRUST project staff. Subjects were chosen regardless of age or gender, although age ranged from 17 through to 45 and gender was fairly evenly split. The majority (22) of the subjects described themselves as either nine or 10 finger typists.

3.1.2. Design
This study employed a simple comparison between two methods of analysis: the traditional Leggett and Williams (1988) approach vs. the "improved" version which uses an individual S.D., multivariate measure chi-square statistic and individual subject/digraph filtering. The dependent measures were the FARs and FRRs
observed for each subject in each condition. The data was first analysed using 300
digraph test samples then 50 digraph test samples.

3.1.3. Procedure
The procedure used generally followed that of Leggett and Williams (1988). A
transcription paradigm was used in which the subjects typed a series of either single
words or phrases displayed on the VDU. All subjects typed the same text, although
the order of stimuli was randomized across subjects. Both period and latency were
recorded for each digraph and data was analysed either as period and latency or as
the sum of the two.

In the traditional condition the analysis of data was modelled strictly on Leggett
and Williams (1988). That is, from each subject’s first data set, a reference profile
was constructed from combined gap and period latencies using approximately 1300
keystrokes. These profiles included only those digraphs consisting of a combination
of lower case alpha keys and/or the space bar and excluded digraphs whose latency
was greater than 500 ms. Finally, to ensure the robustness of the digraph latency
mean and S.D., digraph cells for which less than five observations were available
were also excluded. After filtering, reference profiles contained 79 to 86 digraph
examples, with a mean of 82.

The mean latency for each digraph, and the S.D. of all digraphs in the reference
profile were then calculated. The FAR was calculated for each subject by drawing a
random sample of 300 digraphs, low pass filtered at 500 ms, from each of the other
24 subject’s second data set and assessing the proportion of these samples which
passed the S.D. test at 0.5 S.D.s of all digraphs, and valid proportion criterion at 0.7.
Similarly, the FRR was calculated for each subject by drawing 24 random samples of
300 filtered digraphs from each subjects’ own second data set, then assessing the
proportion of these samples which passed the S.D. test at 0.5 S.D.s of all digraphs,
and valid proportion criterion at 0.7. The procedure was then repeated using test
samples of 50 digraphs.

In the improved condition data was analysed as both period and latency and
digraphs were excluded from the reference profile as defined by a subject–digraph
individual filtering system. Three hundred digraph samples from client and impostor
second data sets were then compared against the reference profile and a chi square
distance measure for each test digraph calculated. After the individual filtering
system was applied to test data, a chi-square statistic was calculated for each test
digraph and the sample was deemed to have passed if it produced an average
distance measure of less than 4.61 ($P < 0.05$). From these passed or failed test
samples FARs and FRRs were calculated. This procedure was then repeated using
test samples of 50 digraphs.

While great care was taken to ensure that the procedure outlined by Leggett and
Williams (1988) was followed for the “original” condition of the current study, one
change was logically unavoidable. In the original work FARs were calculated by
comparing each subject against all others while false rejection rates were calculated
by comparing a single sample of each subject’s data against themselves. Following
this procedure FRRs in the current study would comprise of only 24 comparisons
while FARs would comprise of $23 \times 24$ comparisons. In the current study this inequality was addressed by ensuring that 24 samples of each client's own data was used to produce FRRs.

Further, two other changes were made to the original procedure to ensure accurate representation of its performance. Firstly, the valid proportion criteria was set, in the current study to 0.7, in contrast to the traditional setting of 0.6. This setting has been found to produce optimal error rates for the algorithm (Maher et al., 1994). Secondly, in the current study measures are taken to protect the robustness of the reference profiles of both conditions by including only digraph cells with at least five observations.

3.2. RESULTS AND DISCUSSION

FARs and FRRs were combined to produce an index of verification accuracy. When using a test sample of 300 digraphs the combined FAR and FRR in the traditional condition was 20.6% and in the improved condition was 0.9% (see Figure 2). These scores were analysed using a repeated measure analysis of variance which showed that the "improved" condition lowered error rates significantly, $F(1,23) = 10.41$, $P < 0.005$.

To examine the robustness of these algorithms a second analysis was conducted using a test sample of 50 digraphs. This smaller sample size which more accurately reflects a practical application of the algorithm produced a combined FAR plus FRR of 29.5% in the traditional condition and 3.8% in the improved condition (see Figure 2). An analysis of variance showed that the improved condition produced

![Figure 2](#)  
**Figure 2.** Combined FARs and FRRs generated by the Leggett & Williams (LW) algorithm and the improved algorithm using 300 and 50 test digraphs. ■: FAR; ◻: FRR.
significantly more accurate verification on 50 test digraphs, $F(1,23) = 57.2; P < 0.0001$.

4. GENERAL DISCUSSION

Results of Studies 1 and 2 clearly demonstrate that Leggett and Williams’ (1988) algorithm for typist verification benefits substantially from refinement. With the use of a multivariate measure of latency, a distance measure for maximum separation between test and reference data, individual filtering and individual cell S.D.s combined error rates improved dramatically. This significant improvement was seen when the algorithms were tested in an environment constructed purely for clear comparison. However, the dramatic improvement was also clearly evident in an ecologically valid environment.

Another important feature of the results of the current paper, especially those of Study 1, is the replication of the results reported in the original studies. Umphress & Williams (1985) reported, a combined FAR plus FRR of 18% when they used a test sample size of 300 keystrokes. The current study has produced a combined error rate of 20% using the same test sample size.

The effects of these results on computer security are important. The current study has demonstrated that user identity can be accurately verified within 50 digraphs, i.e. potentially within 51 keystrokes. The most useful aspect of this small sample size is that it may be facilitate a continual security system. At 50 digraphs (approximately one line of text) the system may verify often enough and at short enough intervals for ongoing verification. In addition, these small test samples can utilise the advantages of behaviourally based verification systems, i.e. high user acceptability, low cost, and low computation demands while still producing high levels of accuracy.

In addition, the improved algorithm offers the facility to reduce work place intrusion even further by trading the probability of FARs against that of FRRs. By reducing the stringency of the maximal distance statistic the algorithm can be weighted to reduce the FRR and increase the FAR. Conversely, it is possible to reduce the chance of illegal access further by decreasing FAR and increasing the chance of false rejection. Thus the proposed algorithm allows a fine balance between the opposing requirements of limited intrusion and tight security.

While the results of the current study are positive, there is still room for improving the verification algorithm. Ideally, verification should occur within the number of keystrokes it takes to type a password. A preferred system would, perhaps, build verification into the password system itself as well as continuous monitoring, producing a very much improved security system. In this regard aspects of the algorithm are yet to be optimized. There may remain room for improvement in the individual data filtering system, reference profile construction, or perhaps with the examination of Type II errors and the probabilities of impostor attack.

The results of the current studies clearly suggest that it is possible to use keystroke latencies to accurately verify computer users in a practical security setting. The current algorithm allows keyboard users to be monitored accurately, continuously and with limited intrusion on the work environment. An algorithm that produces 3.6% combined verification error rate on test samples of 50 digraphs goes some way toward a usable, practical system for user verification.
References


Paper accepted for publication by Associate editor Dr. I. Witten