Identifying Fatigue through Keystroke Dynamics

James D. Hayes

University of New South Wales at the Australian Defence Force Academy

This project aims to examine if mental fatigue induced by cognitive loading causes a change to a person's keystroke dynamics. Participants performed an experiment that involved the completion of a series of listening working memory span tests where the sentences were typed into a computer. The results of the experiment showed that there was an increase in the subjective fatigue rating of the participants between the beginning and end of the experimentation period which was not correlated by objective measures of fatigue such as memory span test performance. It was found that there was also no significant change to the participant's keystroke dynamics over the course of the experiment. It is suggested that these results were due to a lack of time on task in the experimental design and possible solutions to this are recommended.

Contents

I. Introduction .................................................. 2
II. Background Information .................................. 2
   A. Mental Fatigue ........................................... 2
   B. Keystroke Dynamics .................................... 3
   C. Use of Vibrations to Identify Keys .................. 3
III. Methodology ................................................. 4
   A. Subjects .................................................. 4
   B. Experimental Design .................................... 4
   C. Fatigue Inducing Mental Tasks ......................... 4
   D. Keystroke Dynamic Control Task ..................... 4
   E. Data Recording ........................................... 5
IV. Results Analysis ........................................... 5
V. Results ...................................................... 5
   A. Keystroke Results ....................................... 5
   B. Fatigue Indicator Results ............................. 6
VI. Discussion .................................................. 7
VII. Conclusions ............................................... 8
VIII. Recommendations ........................................ 8
    A. Vibration Data .......................................... 8
    B. Experimental Design .................................. 8

Acknowledgements .......................................... 9
References ..................................................... 9

APPENDICES
Appendix A. Sentence Sets .................................. A1

Abbreviations

LWMST  =  Listening Working Memory Span Test
PT     =  Primary Task
AT     =  Auxiliary Task
ms     =  Milliseconds

1 PLTOFF, School of Engineering & Information Technology. ZEIT4501

Final Project Report 2016, UNSW Canberra at ADFA
I. Introduction

Fatigue has long been shown to be a major contributor to human error and poor decision making which leads to reduced task performance and a decrease in alertness in complex human control tasks such as driving, the piloting of aircraft, or air traffic control [1-3]. Due to the potential safety implications of ignoring fatigue it is important to be able to effectively monitor fatigue levels to allow for actions to be taken to reduce its impact on task performance.

The aim of this thesis is to examine if mental fatigue induced through cognitive loading results in a change to a person’s keystroke dynamics. If this is shown to occur, it would potentially allow for desirable fatigue monitoring systems to be developed for computer users that would unobtrusively gather data, without the requirement for extra equipment, over a long period of time.

II. Background Information

A. Mental Fatigue

1. Defining Fatigue

In order to monitor fatigue, there must be a working definition describing what fatigue is. It has been shown that fatigue can be broken down into physical and mental components, which despite occasionally being interdependent, can be examined independently of each other [4, 5]. Mental fatigue is defined as a psychobiological state experienced by people during or after extended periods of cognitive load [6]. This state is characterized by subjective feelings of tiredness and exhaustion [6-8] along with reductions in task performance where “alertness and retrieval of information stored in memory” are required [9].

2. Methods for Inducing Mental Fatigue

Mental Fatigue and its associated feelings of tiredness can be influenced by multiple factors such as the period of time one has been awake for, the amount and quality of recent sleep, challenging physical exertion and the amount of time spent completing or working on a cognitively demanding task [10]. As the definition of mental fatigue used in this paper is caused by periods of cognitive load, the only factor from the list above that is required to be examined is the time spent on cognitively demanding tasks since it causes a change that is independent to the other factors [11]. There are three main approaches for inducing mental fatigue discussed by DeLuca [12] that relate to cognitively demanding tasks. The first of these is “Cognitive Fatigue over an Extended Period of Time” where tasks are performed over a prolonged period of time but without a constant sustained effort. This method was used by Cardillo [13] to induce mental fatigue in a driving simulator task. In this task, participants were required to complete a 5 hour block of testing that was repeated three times with a break in between each repetition. The test blocks were made up of a variety of tasks including surveys and the driving simulation task itself, and as such it is not considered to be sustained effort.

The second of these methods is “Cognitive Fatigue during Sustained Mental Effort”. This approach involves having participants sustain a constant cognitive load. This approach was implemented by Langner, Steinborn, Chatterjee, Sturm and Willmes [14] for research on the effects of mental fatigue and temporal preparation on reaction time performance. During the experimentation, the participants spent 51 minutes completing a reaction time task. This task did not involve challenging mental exertion but did show that time spent on sustained mental effort will induce fatigue. The sustained mental effort approach was also used by Mizuno et al. to induce mental fatigue. Test participants in this experiment completed 8 hours of cognitively demanding tasks that were broken into two hour blocks of sustained effort. These blocks were composed of a trail making test, a kana pick out test and a mirror drawing test and were shown to effectively fatigue the subject.

The third approach of inducing fatigue that is discussed by DeLuca [12] is “Cognitive Load after Challenging Mental Exertion”. This approach was effectively used to induce mental fatigue by Möckel, Beste and Wascher [15] with a Simon task. In their experiment they had the participants complete the task over approximately three hours. This total time period was broken down into nine even sub-blocks of testing where challenging mental exertion was experienced. Between each of the blocks there was a break time for the participants. Challenging mental exertion was also implemented by Gevins [16] who had the participants conduct an n-back spatial working memory task. In the experiment, participants completed 16 blocks of the n-back task which were composed of 8 blocks of a low load version and 8 blocks of a high load version. Each of the blocks took 200 seconds to complete with a break of 2 minutes occurring between each block.

There have also been experiments that involve the implementation of both the second and third approach discussed above. One such example of this implemented by Krupp and Elkins [17] used a four hour testing period that involved sustained effort that was broken up with the A-A test. The A-A test is cognitively complex and allows the test to fall under the third approach while the ~2hr neuropsychological tests conducted either side of this caused a sustained mental effort. Their results showed that there was a reduction in verbal memory...
performance in the neuropsychological test conducted after the A-A test which was attributed to the cognitive demands of the test.

3. Measures of Mental Fatigue

When it comes to measuring fatigue, two types of measures are typically used; subjective measures such as questionnaires and objective measures which collect data that can then be analysed to indicate fatigue [12]. Two commonly used examples of this are the Multidimensional Fatigue Inventory which examines mental, physical and general fatigue and the Chadler’s Fatigue Scale which looks at mental and physical fatigue [18]. There has also been subjective fatigue measures developed for specific occupations such as the Driver Behaviour Questionnaire developed for truck drivers [19]. A common issue with all of these subjective fatigue measures is that they all require the user to divert their attention from the task at hand making it difficult or impossible to achieve real time monitoring of fatigue levels.

Several objective fatigue monitoring methods have been developed. The difference between these and the subjective measures is that these use data to look for performance decrement [12]. One such measure of performance that can be used is reaction time which has been shown in various studies to deteriorate over time with the onset of mental fatigue [14, 20]. Similar to the subjective surveys, this method again requires the user to divert their attention away from what they are doing to be able to complete a reaction time test.

A real time method of monitoring fatigue that can be used is the use of a person’s blink rate [9, 21]. This method involves the use of computer image processing and is relatively unobtrusive. This technology has been used both individually and in combination with other behavioural monitors to effectively monitor driver fatigue in a real time and unobtrusive manner [22]. Another way that fatigue can be monitored in real time is through the use of EEG monitoring [7, 22, 23]. This has been shown to have a high level of accuracy; however the equipment for it is both expensive and obtrusive.

An unobtrusive method of fatigue monitoring was examined in [24]. In this research the use of keyboard and mouse interaction patterns as a method of observing fatigue was used. They used a total average key hold time, backspace key rate and a total average key latency time in conjunction with various mouse features and found that they were able to produce a model for detecting fatigue. To induce the fatigue for the study, a cognitive load over an extended period of time approach was used.

B. Keystroke Dynamics

Keystroke dynamics are a form of behaviometrics that use the typing attributes of an individual, such as keystroke duration, keystroke latency, typing error and keystroke force that are commonly used to determine the typist’s identity [25, 26]. It was shown that it is possible to use a persons keystroke dynamics to actively monitor their identity throughout a login session [27, 28]. In these models both sets of researchers looked at using digraph pairs of letters in their analysis as they were found to be more accurate than looking at overall times. They used the average hold time of these pairs, the latency between the keys in these pairs, and the latency between the end of the pair and the next key to determine identity.

Whether or not there was a change in keystroke dynamics caused by physically fatiguing submaximal isometric finger exercises was examined by Chang, Johnson, Katz, Eisen and Dennerlin [29]. They found through the measuring of keystroke duration after the exercises that there was an observable change to the average hold time that lined up with the physiological state changes of the finger flexor and extensor muscles.

There has been some research into using keystroke dynamics for purposes other than identity management. In 2015, it was shown that it is possible to use keystroke dynamic data to determine the level of frustration in novice programmers with adequate accuracy [30]. In this research a keystroke log was used to monitor various statistics such as median keystroke latency, duration of key pairs and number of backspace keys in addition with contextual features like completion of exercises and switching between exercises.

In addition to their use for frustration monitoring, it was shown in 2009 that it is possible to use keystroke dynamic features in conjunction with language features to detect the presence of cognitive stress [31]. In this research they used features such as the average time per keystroke and the number of backspace keys per total keystrokes to provide their keystroke data.

C. Use of Vibrations to Identify Keys

Another way of tracking what keys have been pressed has been through the use of accelerometers to measure keystroke vibrations. It was shown by Marquardt, Verma, Carter and Traynor [32] that it was possible to use the three axis accelerometer in an iPhone to monitor keystrokes when placed next to a keyboard Through the data obtained from the accelerometer, the position and distance of the vibrations were matched to abstracted words in a candidate dictionary with relative success.

In 2013, de Souza Faria and Kim also looked into the use of vibrations to detect the pressing of keys [33]. In their research they used 3 MMA7260QT triaxial accelerometers, deployed around an ATM keypad to gather vibration data from which they could put into algorithms to determine which key had been pressed. They found that they were able to obtain key recognition rates of 98.4% accuracy.
III. Methodology

A. Subjects

20 volunteers aged between 18 and 25 (17 male, 3 female) were enrolled in a double crossover trial. The participants were recruited by an advertisement. Individuals who were pregnant, non-proficient in the English language, had a known intellectual or mental impairment, or dependent on medical care were excluded from the experiments. The protocol was approved by the Australian Defence Human Research and Ethics Committee (Protocol 830-16) and all subjects gave their written informed consent for the study.

B. Experimental Design

All subjects conducted 8 cycles of 2 tasks throughout the experiment. These two tasks were a robot swarm control task (PT) using keyboard and gesture control and a fatigue-inducing mental task (AT). Prior to the beginning of the task section of the experiment, participants completed a subjective fatigue Likert scale from ‘very alert’ (1) to ‘very sleepy and great effort to keep awake’ (9). In addition, participants also conducted training tasks for both methods of swarm control and completed a keystroke dynamic control task. During the task section of the experiment, participants completed four cycles where they controlled the swarm through a course while reaching specific objectives by either keyboard or gesture control. The participants had a maximum of five minutes to complete the course with the objective of finishing in the fastest possible time. This was then followed by four cycles controlling the swarm with the other control method. They would then complete a NASA TLX survey to analyse their workload and give the current rating for their mental fatigue on the Likert scale. Following this, participants would then complete a modified version of a listening working memory span task (LWMST) for a total of 3 minutes after which they would again subjectively rate their fatigue. The total time spent on the fatigue inducing mental task was 24 minutes and the time spent on the swarm control task was up to 40 minutes.

C. Fatigue Inducing Mental Task

The LWMST that was used by Daneman and Carter [35] involved reading sets of sentences to each participant where they were required to remember the last word of each sentence. At the conclusion of each set of sentences, the participants would have to recall the last words of each sentence in that set. The sentences used were regarded to be of moderate difficulty and were all 9 to 16 words in length (included at Appendix A). In order to prevent participants from solely focusing on the last word of the sentence, Daneman and Carter [35] drew their sentences from general knowledge books and the participant was asked to judge if the statement was true or false. To modify this test so that keystroke dynamic data could be obtained, participants were required to type the sentence that was played to them from a computer program and to hit enter when they had finished typing. At this point the sentence would be cleared from the screen and the next sentence played to the participant. At the conclusion of a set of 5 sentences, participants were to enter the last words of each of the sentences and then press enter. The true or false component from the original test was removed as the requirement to accurately type the sentences forced the participants to focus on each word and not just the last one. The LWMST was chosen as it allowed for sufficient amounts of keystroke data to be obtained while also remaining cognitively complex due to its need for simultaneous processing and storage [36].

In order to minimise the amount of ethics clearances required and due to time constraints for testing, the test was designed in conjunction with two other students completing separate theses, Adam Rogers and Sam Clark. As a result, a time period of 3 minutes was selected for the fatigue inducing mental task in order to minimize the total experiment time. This decision had the effect of reducing the amount of time spent on the specific mentally fatiguing task with a total of 24 minutes being spent conducting the fatigue inducing mental task. A similar amount of time on task has been used previously by Gevins [16] where a total time on task of 27 minutes was used to effectively cognitively load the working memory. This lack of constant time on task is also somewhat alleviated by the primary task which can be considered as the second part of a task switching arrangement similar to that seen in [23]. This is because the primary task is also considered to be cognitively complex due to its requirement for path control, monitoring and planning of multiple unmanned systems [37]. Consequently, there is a time on task of up to 64 minutes for each participant.

D. Keystroke Dynamic Control Task

The keystroke dynamic control task involved the participants completing the modified LWMST for 10 sets of 20 sentences. These sentences fulfilled the same conditions as those in the fatigue inducing task of being of moderate difficulty and containing 9 to 16 words. In addition to its data gathering purposes, this task allowed for the participants to practice how the test worked while also not being fatiguing due to the limited amount of recall required [35].
E. Data Recording
To measure the vibrations of each key press, 4 MMA7361 accelerometers were used in conjunction with Arduino Mega 2560 microcontrollers. The accelerometers were arranged so that three formed a triangle around the keyboard and one was placed farther away on the desk to measure the ambient vibrations (Fig.1). The data from these accelerometers was passed through the microcontroller’s serial connection to the computer where the time received of the data from the start of the test in milliseconds (ms) and the value in the X, Y and Z axes were recorded. This combination of accelerometer and microcontroller was chosen due to its similar characteristics to those used by de Souza Faria and Kim [33].

To record the required keystroke information, a custom C# software program was used. This program recorded which key had been pressed, the time the key was held for in ms, the time from the start of the test that the key down event occurred in ms and the time from the start of the test that the key up event occurred in ms.

IV. Results Analysis
The data gathered by the C# recording program was used to examine the following keystroke metrics:
- Single Key Hold Time – Amount of time a specific key was held for
- Digraph Pair Hold Time – Amount of time from the key down event for the first key to the key up event of the second key of the 12 most common digraph pairs across the sentence sets.
- Digraph Pair Latency – Amount of time from the key up event of the first key to the key down event of the second key of the 12 most common digraph pairs across the sentence sets.

These metrics are used to examine the change throughout the experiment and the overall change. The change through the experiment will be examined by looking at each of the metrics for each block over the course of time while the overall change will be examined by comparing the control test to the results obtained in Set 8. The program was also used to determine various metrics that would be used as indicators of the onset of fatigue:
- Backspace Count – Amount of times the Backspace key was pressed for each sentence set.
- Backspace Rate – The number of Backspace key presses divided by the number of key presses for each sentence set.
- LWMST Recall – Amount of last words correctly recalled for each sentence set.
- LWMST Recall Rate – The time spent in the recall section of each test for each word correctly recalled.

In addition to these metrics, the subjective fatigue ratings that were obtained will also be used as an indicator to fatigue.

It should be noted that there were issues with the data recording for Participant 1 and as such the limited data that was obtained from this test has been excluded from the results analysis. Due to time constraints, the data recorded by the accelerometers was unable to be analysed.

V. Results
A. Keystroke Results
When the average hold time for single keys was examined across the course of the experiment (Fig. 2), it was seen that there was a very slight upward trend in the hold time. However despite this slight upward trend there is no significant difference that can be observed between the average hold time in the control test and the average hold time at the conclusion of AT8.

Figure 1. Layout of the vibration sensors on the keyboard

Figure 2. Average single letter hold time across the experiment.
The other keystroke metrics that were examined are the average hold time and average latency for each of the 12 most common digraph pairs present in the sentence sets. It can be seen in Fig. 3 that there were no noticeable trends in the average hold time data for the majority of the digraph pairs. The trends only seem to appear for the less frequently used digraph pairs as can be seen in the results for ‘AR’ which appears to be trending upward and the results for ‘OR’ which appears to be trending downwards. The four most commonly used pairs all appear to be relatively stable with no significant change.

![Figure 3. Digraph pair hold time for the 12 most common digraph pairs over the course of the experiment.](image1)

Similar to the results of the digraph pair hold time, there is no consistent trend to the latency between each of the keys in the digraph pair (Fig. 4). There appears to be a slight upward trend for the latency in the ‘TH’ pair, however this trend is not visible for any of the other digraph pairs. It can also be seen in Fig. 4 that there is are no latencies apart from the ‘HE’ that maintain a relatively consistent time period across the experiment.

![Figure 4. Digraph pair latency time for the 12 most common digraph pairs over the course of the experiment.](image2)

### B. Fatigue Indicators

Over the course of the testing it was found that the subjective fatigue ratings from the participants increased as the testing went on as shown in Fig. 5. The average fatigue rating that was provided prior to the commencement of the main tasks was 4.58 with a standard deviation of 1.61 while the average fatigue rating at the conclusion of the eighth fatigue inducing mental task was 6.11 with a standard deviation of 1.76. When a paired t-test was conducted on these results it was found that the change in subjective fatigue rating was statistically significant (p<0.01).
Throughout the course of the experiment there was no observable consistent change to performance on the LWMST. It can be seen in Fig. 6a that there is no consistent change to the amount of words correctly recalled as the experiment progressed. This lack of change is reflected in the average amount of words recalled for AT1 (4.95 with a standard deviation of 1.32) and for AT8 (4.74 with a standard deviation of 1.77) along with the result of a paired t-test (p>0.05).

Another way the performance on the LWMST was examined was through the recall rate. It can be seen in Fig. 6b that similar to the amount of correctly recalled words, there was no consistent change over the course of the experiment. When comparing the start of the experiment to the end, there was no statistically significant difference between AT1 (5522.52 with a standard deviation of 2207.95) and AT8 (6445.53 with a standard deviation of 5523.44).

It was expected that there would be an increase in the use of the backspace key as participants became more fatigued due to an increased number of mistakes occurring in their typing entries. When examining Fig. 7a it can be seen that the backspace key was most used in the control task and not AT8. This is attributed to the greater amount of keys pressed during the control task where you would expect backspace to be used more due to sample size. However when you examine Fig. 7b it can be seen that the rate of use of the backspace key appears to fluctuate over the course of the experiment before finishing at its highest rate in AT8. This difference between the control and AT8 was not found to be significant (p>0.05).

VI. Discussion

The results from the various keystroke metrics proved inconclusive as to whether or not there was any change over the course of the experiment. In Fig. 2 it was shown that there was a slight increasing trend in the average single letter hold time across the experiment, but this was found to not be statistically significant. This lack of significant change was also observable in the average key hold time for the digraph pairs. The four most commonly occurring pairs can be seen to fairly stable in their hold times (Fig. 3) while there is some instability present in the later pairs, some of which appear either a slight upward or downward trend. It was seen that there were also no
consistent trends present in the digraph latency times (Fig. 4) apart from ‘HE’ which appeared to have a slight upward trend. These lack of significant changes points towards one of two causes which will be discussed later; either not enough keystroke data being collected or the fatiguing effect of the LWMST not occurring.

The results seen in the subjective fatigue test show that the first indicator from our definition, a subjective feeling of tiredness, has been produced. However, this in itself does not indicate the onset of fatigue. The reason for this is that subjective fatigue tests are prone to mood, emotional and motivational bias [12]. It was noted throughout the course of the experiment that participants frequently commented on how they were not motivated to complete the LWMST to the point where some participants did not provide a reasonable attempt to either accurately type the sentence (by frequently skipping words) or to recall the words at the end of each sentence set.

Despite this increase in subjective fatigue, there was no observable reduction to either the recall count or the recall rate to indicate that the definition’s second indicator, a reduction in performance and memory retrieval tasks, had been satisfied. A lack of motivation for the task or boredom are thought to have played some role in the results of that were observed in Fig. 6a and Fig. 6b, particularly AT6 and AT7 where the lack of interest in putting maximum effort into completing the task was most commonly observed. It is also thought that a marked increase in motivation occurred for all participants when they realized they were completing the last LWMST which potentially explains the upturn of results in AT8 for both the recall count and recall rate. A response like this was seen by Boksem, Meijman and Lorist when they reached the conclusion of their research into mental fatigue and motivation [38].

The other indicator of task performance that was monitored throughout this experiment was the usage rate of the backspace key. It was seen in Fig. 7a that there was a slight but statistically insignificant increase in the amount of use of the backspace key during the experiment with what appears to be an upward trend beginning in AT6. This upward trend in usage amount lines up with the increase in usage rate seen from AT6 onwards in Fig. 7b. Despite these apparent trends late on in the experiment, no conclusive observations were able to be drawn from these metrics with relation to fatigue. The late upwards trend suggests that more test sets were required for a statistical increase in usage to be observed.

One experimental design flaw that was noted throughout the conduct of the experiment was that some of the participants with slower typing rates only had enough time to complete one recall task in the three minute time period. This is a potential cause of the desired fatiguing effects of the LWMST not being observed during the conduct of the experiment. Two possible solutions to this are to extend the amount of time the AT is conducted for or to have each participant complete the same amount of sentence sets. Both of these options were considered during the design of the joint experiment, however due to the time constraints that were placed on it, unable to be implemented.

Along with the slow typists not completing more than one recall task per cycle, they also had a much smaller data set that could be examined. This meant that any outliers from their participation had an effect on the results that were seen. This could be rectified through having more volunteers participate in the experiment or by increasing the amount of typing data available through the two options discussed above for fatigue.

VII. Conclusions

The purpose of this project was to identify if there was a noticeable change to keystroke dynamics with the onset of mental fatigue induced by cognitive loading. This testing was conducted as part of a joint experiment with two other thesis students as it would reduce the ethics clearance work required and the competition for volunteers. The effect of fatigue on keystroke dynamics was unable to be determined in this report due to issues in the joint experimental design. Over the course of the experiment there was a significant increase in the subjective fatigue of the participants, but this was not reflected in the other fatigue metrics meaning that there was not enough observable change to state that there had been an onset in fatigue. There were no significant observed changes to the keystroke metrics examined in this project.

VIII. Recommendations

A. Vibration Data

Over the course of the experiment keystroke vibration data was gathered for each of the 8 auxiliary tasks. The vibration data that was captured during this experiment was unable to be effectively analysed for any potential changes over the course of the experiment. Continuous wavelet transforms were produced from the data that was collected, however further signal processing still needs to be conducted before these can be used for results analysis.

B. Experimental Design

The experimental design that was implemented in this experiment was found to be insufficient to induce the fatigue required to examine the stated hypothesis. The recommended experimental design proposed is to maintain the two task nature of the experiment. However it is recommended that the PT be replaced by cognitively loading
with a task similar to those used by Mizuno et al [3]. This should be followed by the LWMST used in this experiment. To determine the time for each task and the number of cycles in the experiment, a pilot study needs to be conducted to determine the right settings for the experiment. In this pilot study it is recommended that the variable typing rates of the participants be taken into account. These changes to the design would allow for the mitigation of the lack of time on task problems experienced here while potentially providing a larger set of keystroke data to be analysed.

Acknowledgements

I would like to thank my supervisor, Prof. Hussein Abbass, for his valuable advice, insight and guidance throughout the course of this project. I would also like to thank my fellow students, Adam Rogers and Sam Clark, for their work in the design, testing and ethics clearance of the joint experiment and the questioning of ideas throughout the project. Finally I would like to thank my friends and family, who have helped keep me motivated and on task throughout the project.

References