Vivek Dhakal

Identification of typing behaviors from large keystroke dataset

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo, September 19, 2017

Thesis supervisor:

Prof. Antti Oulasvirta

Thesis advisor:

M. Sc. Anna Feit

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Vivek Dhakal

Title: Identification of typing behaviors from large keystroke dataset

Date: September 19, 2017 Language: English Number of pages: 6+49

Computer Science

Professorship: Prof. Petri Vuorimaa Code: SCI3042

Supervisor: Prof. Antti Oulasvirta

Advisor: M. Sc. Anna Feit

In this thesis work, keystroke-level typing data of over 168000 participants are analyzed to understand determinants of transcription typing behaviors. Keystroke patterns are analyzed in detail and linked to typing performance. Inter-Key Intervals of letter pairs and other statistical indicators of typing performance are calculated and their distributions and statistical relations are studied. These analyses show, among other findings, that Inter-Key Intervals in typing distant letter pairs in the keyboard are more predictive than other letter pairs, e.g. letter repetitions. Rollover typing, where the next key is pressed before the previous key is released, is prevalent widely, linked to faster typing with high correlation. Finally, medoids-based (PAM) unsupervised clustering of participants is performed to identify groups of typists with similar typing characteristics, and the findings from the clusters are interpreted in terms of performance, accuracy, hand movements and rollover behaviors.

Keywords: data analysis, keyboard, keystroke, text entry, transcription typing, typing performance

Acknowledgement

This thesis work has been carried out in the User Interfaces group in the ComNet department, School of Electrical Engineering, Aalto University in collaboration with TypingMaster Oy, Finland and Cambridge University, UK.

I would like to express sincere gratitude to Prof. Antti Oulasvirta of the User Interfaces group for his conducive supervision throughout the work. I am very much indebted to Anna, my advisor, for the excellent guidance she provided both during the work and beyond. I also extend a share of thanks to Prof. Per Ola Kristensson, Cambridge University, UK for his valuable time and effort to discuss and suggest on the work.

I deeply appreciate and thank Samuli without whose untiring help with the technicalities the work would not be possible. I am also thankful to Gregorio Palmas, former visiting postdoctoral researcher, Aalto University and Prof. Arto Klami, Asst. Professor, Dept. of Computer Science, University of Helsinki for their expert ideas and advice on the topic.

Janin, Jussi, Olli, Sunjun and the fine people at User Interfaces group have always participated in providing constructive suggestions and comments. Finally, last but not the least, I thank Shishir Bhattarai, Janaki P. Koirala, Irfan Khan and Sabin Karki among many others who have helped me with discussions, suggestions and help during the work.

Otaniemi, September 19, 2017

Vivek Dhakal

Contents

Definitions and Measures

WPM is calculated as the average number of characters words (word here meaning a 5-character transcribed string) typed every minute, calculated from the averages of each sentence in the test.

ERROR_RATE is calculated based on the Levenshtein edit distance [1]. It is the ratio between the edit distance between entered and presented sentence-strings and the length of the string, averaged over all sentences in a test.

PRESS_TIME is the timestamp when a key is pressed. **RELEASE_TIME** is the timestamp when the key is released. If a key is pressed and released during when another key is kept pressed (Example SHIFT+A), the PRESS_TIME of the later key is greater than that of the earlier key, while the RELEASE_TIME of the later key is smaller than that of the earlier. Also, **Keypress Duration** is given by the difference RELEASE_TIME - PRESS_TIME.

Error Correction refers to number of all occurrences of backspace (BKSP) and delete (DEL) keys used during typing.

Performance measures include WPM, Error Rate, IKI and Keystroke Per Character (KSPC) measures.

KSPC is the number of keystrokes (scribed as well as non-scribed key presses) recorded per correctly scribed character for the presented sentence-string in the typing test.

IKI (inter-key interval) is the time between two keypresses, computed as the difference in PRESS_TIME timestamps between two keys.

ECPC is the number of error correction keys (BKSP/DEL) pressed per correct character presented in the tests.

Rollover typing is discussed in later sections, where this refers to the techniques of typing consecutive keys without releasing the previous key [2]. The ratio or percentage of rollover keys to total typed keys (meant to show the prevalence of this technique) is used as a measure of this throughout this work.

1 Introduction

1.1 Motivations

Transcription typing has been extensively researched since the typewriters emerged. Modern keyboards are technologically different than the typewriters in spite of the semblance. This has possibly resulted in changes in the way we type, even as the basic metrics of typing performance, key arrangements and typing behaviors remain similar. As keyboard text entry is an essential means of interaction between a user and a computer, it is useful to perform both theoretical and empirical research to understand factors affecting typing performance for increased efficiency and productivity. This becomes more relevant considering the modern keyboard and newer typing techniques.

Much of the established previous works on typing performance and human factors modeling originate in the typewriter era. However, not only the modern keyboard is evolving in technological aspects but also the keyboard is used by people with different skill-sets, demographics and objectives. Among the empirical studies done on typing with the modern keyboard, most of the works are related to keystroke dynamics analysis as a method of biometric security or to mobile (soft) keyboard where data collection is easier and commercially more viable. On the other hand, large scale empirical analyses on general purpose keyboard typing, which is the mainstream typing interface for word processing, programming and communication tasks, are either rare in the literature or probably undertaken by commercial organizations, such as typing test companies, for proprietary uses. Apart from the usefulness, the Internet has also made it possible to collect large-scale data of everyday keyboard typing unlike that in typewriters.

This work tries to bridge the gaps by studying performance and strategies in keyboard typing through analyses of large datasets of transcription typing by participants with global demographics. By employing a speed-accuracy based performance criteria in the form of online typing tests, the study aims to look at the inherent trade-offs and strategies present among the participants.

1.2 Overview of the topic

Transcription typing is the act of typing previously composed text, i.e. typing sequences of characters (often meaningful words and sentences) by looking at an existing written record. Thus, the process includes hand and finger movements on the keyboards, keypresses and key-finding strategies if any [3]. Transcription process does not take into account creative activities like thinking, text modifications or proofreading. A study of transcription typing separate from cognitive factors is important when we want to study motor movement strategies discarding the cognitive factors which are complex to model and record. Although visual attention, memory and decision making in error corrections can be aspects affecting transcription per-

formance, they are not recorded in the datasets analyzed in this work, thus will be beyond the scope of analyses presented in this paper. The scope of typing strategies will be limited to study of patterns of recorded timestamps, i.e. when a key was pressed, and measures derived from it, in association with a smaller ground-truth dataset also including finger motion data collected from controlled experiments.

As for typing performance, the most common measure used for speed and accuracy are Words per minute (WPM) values and Error rates (percentage) respectively. Typing tests mostly measure the WPM speed along with the error rate. Other derived measures include average number of keypresses required per correct character, average number of Error keys (Delete/Backspaces) pressed per correct character, etc [4]. These and other metrics are defined in detail in the respective sections.

Touch typing is a strategy that has existed in widespread use for fast typing. In touch typing, people associate each finger with a resting key (home position) and the keys are pressed by a finger which needs to travel short distances from its home position, all without looking at the screen and supposedly reducing the time interval between keypresses (IKI) and therefore increasing the overall speed. However, scientific explanation that this is an optimal strategy does not exist. In fact, there are studies suggesting that even people not using all their fingers and not trained for touch typing can reach speeds comparable to or greater than touch typists [5]. These show that there is more to typing strategy than just touch typing. The typing test used to collect data for our analyses captures everyday typing behavior which can vary a lot between participants, providing us an opportunity to study the variations and identify typing patterns and strategies, not just the performances.

1.3 Problem definition

Although it is easy to measure typing performance from a typing test, it is more useful to get insights about the typist's behaviors. Identifying typing behaviors can help explain typing performance, and suggest ways to improve it, for example using a better strategy. The objective of the work presented in this thesis is to present a number of sentences to a typist for which the transcriptions are recorded as keystrokes with their timestamps, and in return to be able to identify the factors affecting hand/finger movement strategy used in typing. More specifically, the work studies aspects of hand or finger use that might be reflected in bigram level typing performance. These bigrams are then used as features in order to describe differences in typing patterns between participants. In terms of application or implementation, the participants are expected to be clustered into distinct performance and behavior groups.

1.4 Thesis structure

This thesis is structured in the order of the analyses performed. The next section will review and summarize related previous works. Some of those works highlight

the gaps this work intends to partly cover, while other works present the foundations on top of which this study explores, builds and extends. The literature review is followed by detail descriptions of the typing test and the data collected, statistical methods and analyses done on the data and the important findings.

Next, the study of typing strategies and its determinants is reported, followed by results and findings of clustering the participants and then interpreting and visualizing the differences between clusters. Discussions of results and findings are included in the following section.

Finally, the limitations of this work, directions for future works and application areas are summarized in the conclusive section.

2 Backgrounds and Related Works

Typing performance was historically useful as a measure of productivity in professional requirements. Typing speed, error rates, error types and various other metrics have been proposed to quantify typing performance. [4] introduces various measures for the same, including the sources of errors, aggregate and adjusted measures. However, most previous studies on typing performances have either focused on typewriter typing from a few decades ago [6], or more recently on keyboard layouts [7], keystroke dynamics [8] [9] and mobile (soft) keyboards [10]. These references cite works done on keyboard typing but within more specific areas or with different set of objectives.

Studies on performances of professionally trained typists have been mostly done on typewriters, and comparisons and references are made to the touch-typing strategy. However, there is no consensus that touch typing is optimal, or that any other strategy is or is not just as good. Small scale studies such as [5] show that people without typing training and not using the touch-typing strategy can type just as fast and accurate if not better. This work attempts to understand typing performance better by empirically studying the aspects of everyday typing.

Another example of a specific area of research on modern typing is keystroke dynamics. Keystroke patterns are considered consistent for a person to a high degree, and they are proposed as and commonly used as a biometric security measure [11]. Bigram IKIs, and less popularly trigram IKIs, are used to identify typing patterns in keystroke dynamics [12]. Machine Learning based predictions in this regard have been successful with high degrees to identify people based on their typing patterns. There have been studies to cluster keystroke patterns with the objective of authenticating users [13], but they do not generally outline the behaviors involved or techniques that produce those keystrokes. Specifically, these studies do not explain hand and finger movements that produce the keystroke patterns, outlining a need to translate the keystroke and timestamp data into meaningful determinants of typing behaviors.

Perceptual and cognitive aspects of typing [3] include the processes before the start of hand movement on keys, while motor aspects include the strategies for using and moving the hands and fingers over the keys. Ideas from the paper were studied and used to limit the scopes within which a typing strategy can be defined in this work. Perceptual and cognitive aspects are not expected to differ considerably between the typewriter and the modern keyboard. However, differences occurring in modern keyboard typing are rarely documented in the literature, let alone large empirical studies of modern and general purpose transcription typing. For example, [14] studies text entry strategies on miniature soft keyboards using parameters like KSPC, MERD (mean error recovery distance), where the idea is to study performance statistics which are part of this work, but for specific types of keyboards. There are other works on keystroke dynamics as biometrics and security mechanisms

or keyboard layouts such as in different languauges, however we know little about transcription typing in general.

Key-finger association is another aspect of typing. A key finger association [5] is a mapping that defines the fingers used to type keys in the keyboard. For example, a touch-typing strategy follows a distinct key-finger mapping where, although variations are observed, the keys A, Q, Z and the ones to their left are typed by the left little finger, the keys P, semicolon (;), fullstop (.) and beyond to the right are typed by the right little finger, and so on with the fingers in between. Key finger mapping can be seen as a division of keyboard regions for reaching a key quickly without looking at or searching on the keyboard, based on the hand's home (resting) position. For touch typists, this mapping is well-defined and consistent between and within people. However, day-to-day and especially untrained typing behaviors emerge to employ various finger-to-key mappings other than that in touch-typing, and among the most contrasting ones would be the preliminary 'hunt-and-peck' greedy strategy.

Combining key-finger mappings with dynamics of hand movements and keypresses can explain some observations in typing behavior. Logan et al. [15] study the effect of Fitts's law and Hick's law about the optimal mapping from fingers to keys. In addition, biomechanical studies such as [16] suggest that the dynamics of pressing a key limit the typing speed. Such lower-bounds, together with error modeling, cognitive processing time, hand and finger movement and key-search times can be combined to understand some aspects of the overall typing behavior of people. However, as it was found during the course of this work and mentioned later, these are insufficient in cases were the modern keyboard (or any new typing device) has differences from the typewriter in a way that affects the mechanics and dynamics of the device's operation. Thus, we do not know modern typing behaviors that are dependent on such dynamics. Understanding typing strategies in device-agnostic manner is another implicit attempt in this paper.

Investigating typing behaviors and the findings thereof can have implications to how a typist is tested for typing skills. One area pertaining to this is the typing test sentences. There have been various sources of sentences for testing typing performance, such as the Enron dataset [17], novels [18], news [19], Mackenzie's proposed standard test set [20] etc. Randomized sentences are used, and it is sometimes intended to include the full range of alphabets. A comparison of five public datasets has been studied by Kristensson et al. [21] which shows differences in text entry style and performance. Yi et al. [22] discuss word clarity as a metric for sampling keyboard test sets. Analyzing the suitability of test sentences becomes more evident when the suitability of keystrokes or bigram level indicators of typing performance are considered. Thus, the bigram level study of typing strategies can also aid in understanding how different test sentences work in providing appropriate typing performance measures.

3 Typing Test Data

3.1 The typing test and Data collection

3.1.1 Participants

The typing test was available through an online website which is available throughout the world. Therefore, the participant can be of any international nationality, gender and profession. The data spans over 218 countries. However, most of the participants are from English speaking countries as the language of the test is English. On the right, figure 1 tabulates the ten highest countries the participants are from. Figure 2 shows the number of participants from various countries as colours on the world map.

	COUNTRY	NUMBER
1	US	114323
2	IN	13103
3	PH	10682
4	CA	8171
5	GB	6336
6	AU	3924
7	PK	1859
8	MY	1261
9	HK	1184
10	NZ	1081

Figure 1: Most participants come from US and other english speaking countries.

3.1.2 Sentences

The 1525 Sentences from which the typing test sentences are drawn are taken randomly from newspaper headlines and the Enron Corpus, after validation checks about the length, characters used, numerals present etc. The test is in English language, and all sentences are in English. Therefore, the typing characteristics analyzed pertain to English language typing skills. Table 1 shows the examples of sentences used and their sources.

Sentence	**Source**
We've already eliminated his speech therapy.	mobile
Last week Ballesteros had 16 pars and two birdies in his final round.	news
I plan to be in the office tomorrow.	memorable
This is a major setback for the rookie.	email

Table 1: Examples of Sentences used and their sources

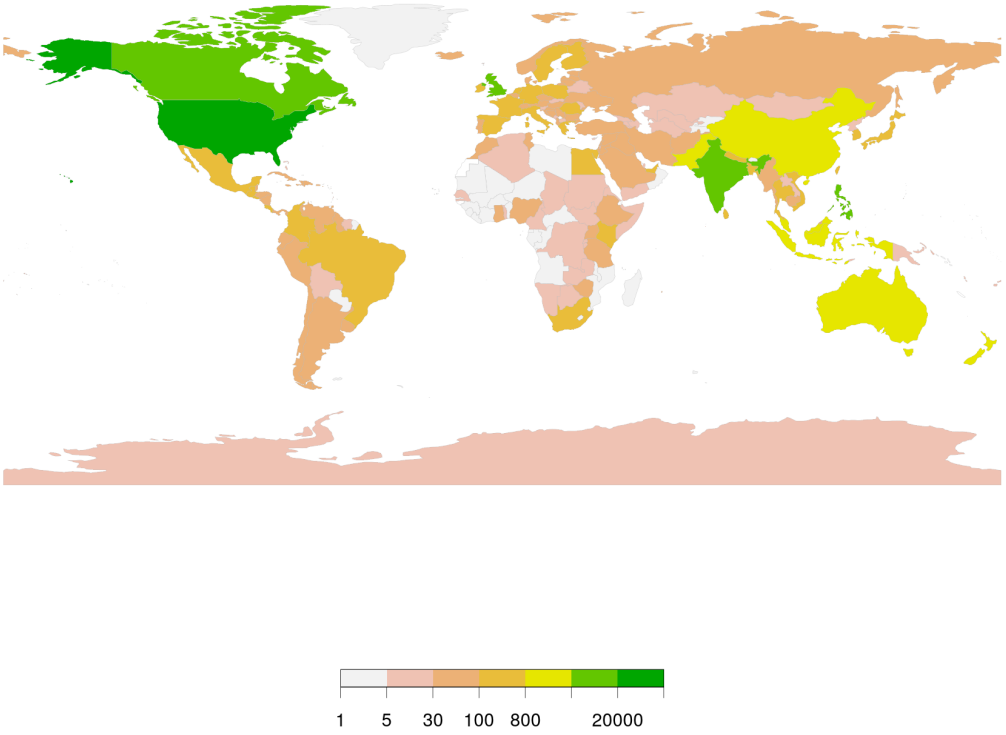


Figure 2: Participants by countries shown in the world map

3.2 Database structure

The PARTICIPANTS table is a list of all participants by their ID and IP addresses along with the demographic information as well as some basic measures of their performance and error calculations. Most of the column names above are self-explanatory. HAS_TAKEN_TYPING_COURSE is either 1 (Yes, taken a typing course) or 0 (No, Not taken such a course). LAYOUT is the keyboard layout used to take the typing test, viz. QWERTY, AZERTY or QWERTZ. FINGERS is the number of fingers the participants think they use for typing (subjective). TIME_SPENT_TYPING is the number of approximate hours the participant spends each day typing

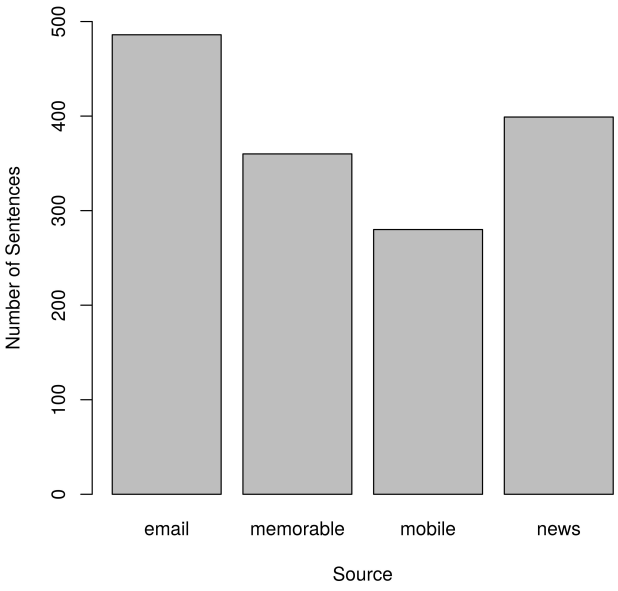


Figure 3: Distribution of sentences used in the typing test by sources

PARTICIPANTS	SENTENCES	TEST_SECTIONS	KEYSTROKES
PARTICIPANT_ID			
IP_ADDRESS			
AGE			
GENDER			
HAS_TAKEN_TYPING_COURSE			
OS		TEST_SECTION_ID	
BROWSER		SENTENCE_ID	
BROWSER_VERSION		PARTICIPANT_ID	
BROWSER_LANGUAGE		USER_INPUT	KEYSTROKE_ID
COUNTRY	SENTENCE_ID	INPUT_TIME	PRESS_TIME
LAYOUT	SENTENCE	EDIT_DISTANCE	RELEASE_TIME
WPM	SOURCE	ERROR_RATE	LETTER
OS_VERSION		WPM	TEST_SECTION_ID
EDIT_DISTANCE		INPUT_LENGTH	KEYCODE
NATIVE_LANGUAGE		ERROR_LEN	
FINGERS		POTENTIAL_WPM	
TIME_SPENT_TYPING		POTENTIAL_LENGTH	
KEYBOARD_TYPE			
REGION			
ERROR_RATE			
DETECTED_COUNTRY			
DETECTED_REGION			

Table 2: Demographic, test performance data and keystroke level data was collected from the participants of the test

(subjective).

WPM is calculated as the average number of characters words (word here meaning a 5-character transcribed string) typed every minute, calculated from the averages of each sentence in the test. ERROR_RATE is calculated based on the Levenshtein edit distance. It is the ratio between the Âăedit distance between entered and presented sentence-strings and the length of the string, averaged over all sentences in a test.

While the participant is typing each sentence presented, the web-app logs the timestamp and key-press information about each input key from the keyboard. Thus one part of the database is a table that logs the following information into the database: Keystroke_ID is the Primary key of the table, and unique for each keystroke saved into the database throughout the duration it is active.

PRESS_TIME is the timestamp when a key is pressed. RELEASE_TIME is the timestamp when the key is released. If a key is pressed and released during when another key is kept pressed (Example SHIFT+A), the PRESS_TIME of the inner key is greater than that of the outer key, while the RELEASE_TIME of the inner key is smaller than that of the outer key.

LETTER is the corresponding key character for any keystroke. KEYCODE is the JavaScript keydown/keyup event key-code [23] for the key pressed/released. TEST_SECTION_ID is the unique ID for each sentence in a test taken by any

participant. Another table SENTENCES maintains the list of all the sentences from which 15 are chosen in any order and provided to each participant for the typing test.

The last table TEST_SECTIONS is the one which stores each sentence (as SENTENCE_ID) from all the test sessions for all participants, along with the participating user, the user's input and other calculated values about the error and performance. The sentence presented at any time is randomly chosen from the unused sentences in the db, using Scala's rand function.

PARTICIPANT_ID is the participant to whom this sentence was presented, and USER_INPUT is the string transcribed by the participant. EDIT_DISTANCE, ERROR_RATE and WPM values as calculated as before. INPUT_LENGTH is the length of the transcribed string starting from the first printable character.

3.3 Smaller dataset from controlled experiment

In addition to the larger dataset from the online typing test, a smaller dataset (50 participants) collected from controlled experiments was also referred to. The dataset contains participants, both touch-typists and everyday typists, who typed various sentences for speed and accuracy, and also used in the analysis as ground truth data. In addition to the parameters collected from the online typing test, this dataset includes information about fingers used to type different letters and their movement pattern in high speed videos. Details about the controlled study on this dataset can be referred from [5].

3.4 Data Cleaning

The data is cleaned for consitency, removing redundancy, checking correctness of information (such as timestamps) and to follow criteria to make the data and the following analysis more useful and error-free. The steps involved in data cleaning are explained as follows:

1. Only data that have complete demographic information are Considered. These are users who have also completed all the sentences in the typing test. From these participants, those with error rate > 25 % were removed. This amounts to 192169 out of 517961 participants.

2. Some test sections (typed sentence based records) had wrong timestamps because of bad clock time values from browsers or because of wrongly being written to the database. Among these, the former type errors are removed by removing entries where timestamp differences between two keys was +/- 50s. (This value was selected because it was observed the jumps were generally greater than this. However, whether it removed all of it was manually checked, and no inconsistency was found in sample checks.) For the rest of the test sections, the WPM values were recalculated.

3. Then, Participants with at least 15 sentences without any timestamp errors are included. Steps 2 and 3 resulted in 168960 Participants. WPM values are recalculated based on updated input time (after correcting misplaced timestamps) for (i) first 15 test sections and (ii) all test sections. WPM values recalculated from the first 15 test sentences typed are used for further analysis, and from this point all mentions of WPM speeds denote this value.

4. The filtered and complete data along with calculated statistics (e.g. Keystrokes Per character (KSPC) measure and Number of error corrections) are stored.

5. For IKIs and key-press durations, keystrokes with timestamp differences larger than 5000ms are removed.

6. Specific cleaning methods are mentioned in the respective analysis sections later.

4 Analysis and Results

This section presents summary statistics, derived measures and statistical relations from the dataset. The large scale of data makes this analysis both novel and useful. With data from large number of participants distributed across demographics, looking at trends, regressions, correlations and statistical significance of important measures becomes more realistic.

4.1 Demographics and Performance

In this section, the distribution of the participants' demographics and their performances across the test are studied. The participants vary in age, gender, geography and their typing skill level. The distributions of demographic variables are explained below.

4.1.1 Distribution of Demographics data

Gender Among the participants who specify their gender, females outnumber males (11.2% more in number), ie. females make up 52.65% of the participants. To compare it in context with actual demographic distribution, the US (where over half of the total participants belong to) has a sex ratio of 0.9524 females/males, while the global average is 0.9346 females/males [World Bank, 2016].

Age Most of the participants are teenagers and young adults, making up about 70% of all. The overall mean age of participants is 24.5.

Use of fingers (self-reported) Touch typing, specifically with 9-10 fingers, has been considered as a technique to learn fast typing. In our data, less than half of the participants (44%) reportedly use 9-10 fingers while typing.

Hours of typing daily The number of hours of typing each day shows the experience of participants in typing. As the data shows, most participants (64%) type less than 2 hours a day, however there is significant number of people (14%) who type more than 6 hours a day.

Typing training 72% of all participants report not having taken a typing course. 64% of those who report using 9-10 fingers for typing say they have taken a typing course. This self-report is used in the analysis to define trained and untrained typists.
Table 3 summarizes the demographics statistics for the participants.

4.1.2 Performance distributions

Performance measures have been studied in detail in many previous works. Definitions for various performance measures are used as discussed in [4]. The definitions

Demographics	**Result**	**Remark**
No. of participants included in the analysis	168960	Participants with completed testsets and after filtering
Females (males)	52.65% (41.45%)	Rest preferred not to specify
Mean age (SD)	24.5 (11.24) years	75% of participants were teenagers and young adults (11-30yrs)
Number of Countries	218	68.05% (114323 participants) from US
Native language English	85%	
Took a typing course	72%	
Avg hours typing per day (SD)	3.17 (3.23) hours	64% of participants reported to type 2hrs or fewer per day, 14% reported to type more than 6 hrs per day
Qwerty layout	98.1%	Rest used local alternatives (Qwertz or Azerty.) or others (e.g. Dvorak)
Physical keyboard, Laptop keyboard	43.77%, 54.15%	Rest used on-screen (touch) or other small keyboards.

Table 3: Background statistics of participants

	All		**WPM corr**
Measure	$\overline{X}$	σ	r
WPM	51.56	20.2	–
IKI (ms)	238.66	111.6	-0.84
Keypr. duration	116.25	23.88	-0.29
Unc. Error (%)	1.167	1.43	-0.21
Error Correct. (%)	6.31	4.48	-0.36
KSPC	1.17	0.09	-0.40
Left IKI (ms)	215.23	96.8	-0.7
Right IKI (ms)	203.6	99.13	-0.68
Altern. IKI (ms)	198.26	103.95	-0.72
Repet. IKI (ms)	176.36	70.26	-0.32
Numb. fingers	6.95	2.95	0.34
Rollover ratio (%)	25.0	17.0	0.73

$\overline{X}$: Mean value σ : Standard Deviation

Table 4: Overview of results showing the mean and SD for each measure and correlation of each measure with WPM. Statistical significance of tabulated results have been tested at the 1% level using the Mann-Whitney signed rank test.

are also included in the Definitions section at the beginning of this report. In this section the performance measures for the data are reported along with the graphical representations of the distributions. Table 4 summarizes the statistics for the performance measures.

Typing Speed The average words per minute of participants is 51.56 WPM ($SD = 20.2$). Fastest typists reach words per minute score of 120 WPM or higher. Although the participants are self-selected, the standard deviation is considerable. As is commonly the case with human performance metrics, the distribution is not normally distributed and a slight positive skewness is observed. Skewness of the distribution is 0.513 and the kurtosis measure is -0.11. In addition, slightly higher typing speeds are observed in trained typists than untrained ones.

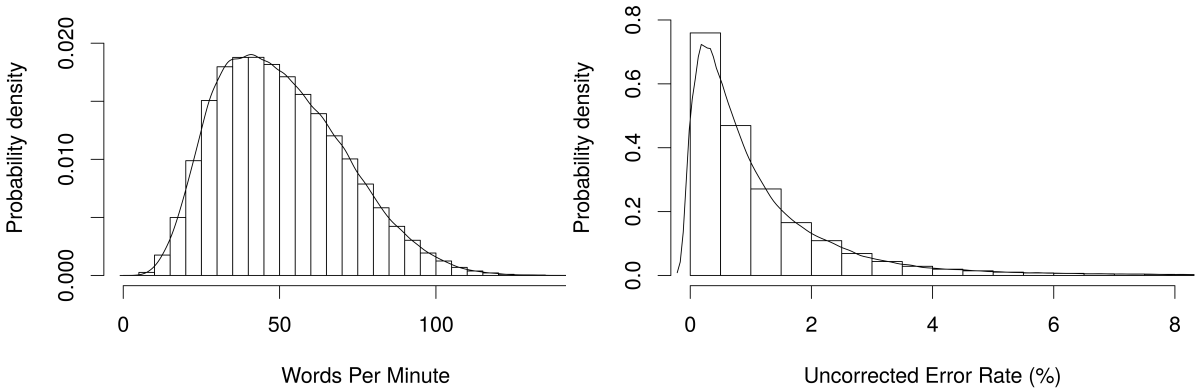


Figure 4: Distributions of WPM speed and Uncorrected Error rate among participants. The WPM distribution is skewed slightly to the right.

Error Rate and Error Correction The average uncorrected error rate of participants is 1.167% ($SD = 1.43\%$). The majority of participants left only some errors uncorrected even though they made errors while transcribing. 90% of participants had an uncorrected error rate of less than 2.66% in the transcribed text. Slow typists have significantly more uncorrected errors which could mean a reduced ability for error detection. Trained typists also have lower uncorrected errors ($\overline{x} = 1.02\%$) in comparison with untrained typists ($\overline{x} = 1.23\%$). In average, there are 2.29 *error corrections* per sentence with users at 99^{th} percentile pressing error correction keys (Backspace or Delete) up to 8.5 times per sentence on average. Also, the average KSPC rate is 1.173 ($SD = 0.094$).

Inter-Key Interval (IKI) and Keypress Duration Average inter-key interval (IKI) is 238.656 ms ($SD = 111.6$). It is observed for IKIs and keypresses that a lower bound of about 60 ms is present in the data. The IKI distribution shown in Figure 5 has a skewness of 1.98 and kurtosis measure of 7.1. As IKIs and WPM are strongly related by definition, similar comparisons are observed. Fast typists have average IKI of ~ 120 ms, while slow typists show IKIs of over 480 ms which can be as large as 900 ms and a large standard deviation of over 120 ms. The average IKI of trained typists is only slightly less than that of untrained typists, in line with the difference in typing speed.

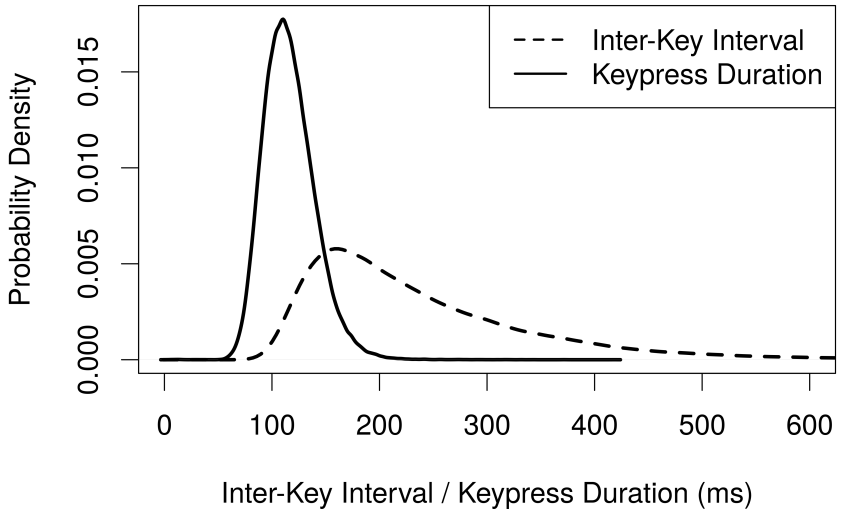


Figure 5: Distributions of average IKI and Keypress durations. The IKIs vary widely between fast and slow typists, however, Keypress duration does not. A biomechanical lower bound of $\tilde{6}$0ms is present.

In contrast, the average *keypress duration* is 116.24 ms ($SD = 23.88$) and is not found to vary largely. The distribution has a skewness of 0.8 and a kurtosis of 2.36, much smaller than that of the IKI distribution. Keypress durations do not vary much between trained and untrained typists. Prior works report similar keypress durations ([24]).

4.2 Bivariate Analyses

In this section, findings about relationship between various performance and demographic measures are reported. Although measures such as WPM and IKI are quite correlated by their nature, bivariate analysis can help learn relationship between more non-obvious variables.

4.2.1 Reported finger use

A widespread belief about typing is that using many fingers enables achieving higher number of words per minute. The 'touch typing' technique is based on this assumption along with the idea of assigning keys to be typed by each finger without looking. It was observed that in average and over all participants the larger the number of fingers people report using, the faster they are, both in terms of high WPM score and low IKI.

In addition, self-reported number of fingers used for typing and typing speed show a positive correlation. Participants reported using any number of fingers from 1 and 2 to 9 and 10. The average number of fingers used is 6.95 ($SD = 2.95$). 9-10 fingers

was reportedly used by 47.6% of participants . Figure 6 shows that the larger the self-reported number of fingers, the higher the typing speed ($r = 0.38$, $p < 0.001$). In addition, Trained typists report using more fingers than untrained typists (on average 8 versus 6.5).

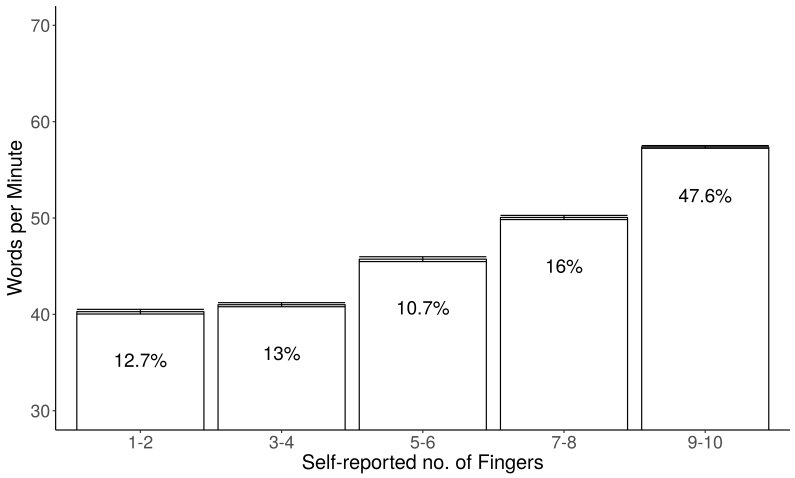


Figure 6: Average typing speed is higher for people reporting more fingers used in typing. Error bars at the top represent 95% confidence intervals.

4.2.2 Speed-Error distribution

Speed-Error trade-off is an important phenomenon studied for a long time as an important aspect of learning [25]. Keyboard typing as a motor skill also inherits this trade-off: the faster one tries typing the more errors one makes, and there are studies that attempt to investigate reasons and models for this behavior [26]. However, for a large pool of participants, a study of their speed-error distribution has a different meaning than for an individual. Specifically, such a distribution shows the prevalence of typists who may either show fast speeds at the expense of reduced accuracy, or those who maintain both speed and accuracy, or even those who fair poorly at both measures.

Figure 7 shows the typing speed in WPM plotted against ERROR_RATE, along with a fitted curve and shaded confidence regions. Figure 8 is a smooth regression of the two variables within the range of WPM values 20 to 80, showing that the participants who type more accurately tend to have better typing speeds.

It is observed that faster users generally make fewer mistakes. A negative correlation was observed between uncorrected error rate and WPM ($r = -0.21$) as well as for ECPC ($r = -0.36$) and KSPC ($r = -0.4$). Furthermore, it is observed that the correlations of substitution and omission errors with WPM are higher than for insertion errors ($r = -0.45$ and $r = -0.33$ versus r=−015). This could indicate that the number of insertion errors changes less with higher performance.

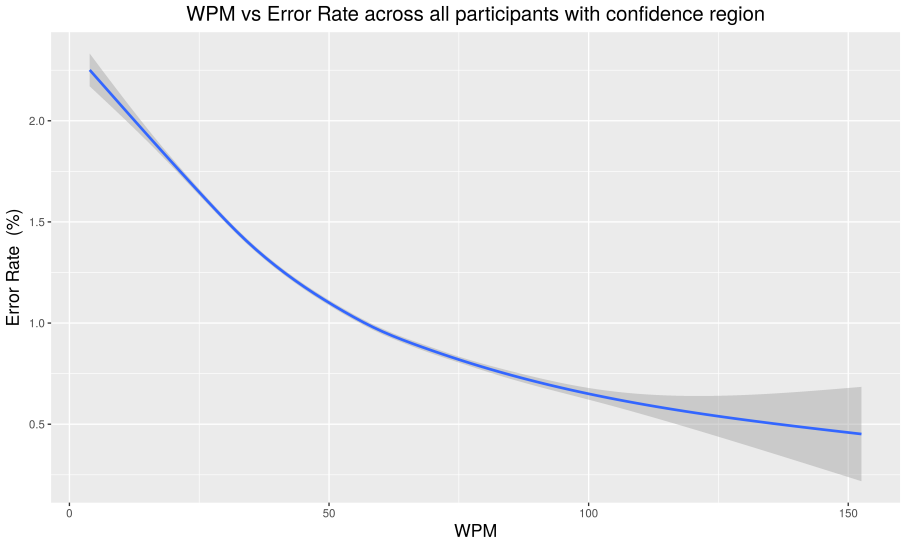


Figure 7: Typists with lower uncorrected error rates are generally faster.

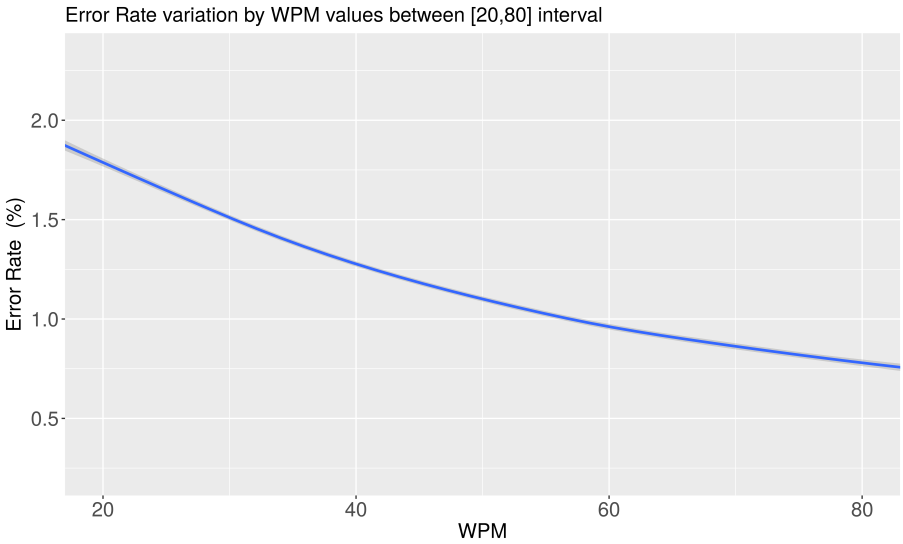


Figure 8: Smoother comparison of WPM-Error rates in the 20-80 WPM interval

4.2.3 Typing speed vs Error corrections

Unlike uncorrected error rate, error corrections are associated with corrected errors. ECPC (number of error corrections per character) denotes how often a participant makes and corrects errors. ECPC is related to KSPC, another measure that describes how many keys are input for every character in the string to be transcribed. Figure 9 shows the relation between WPM and KSPC across participants. Very slow typists ($< 25WPM$) appear to make and correct many mistakes, resulting in a lower speed. At about 25-30 WPM, a slight flat region in the curve suggests that a group of typists is present who differ in how fast they are able to press keys, while making and correcting errors. A linear correlation is observed between WPM and KSPC for speeds higher than about 30WPM: faster typists type less keystrokes per character as they make and leave less mistakes. Errors does not only take more

time in typing additional keys but also in locating errors. Therefore, as the Figure shows, typing speed can be largely improved even by reducing errors by a small margin.

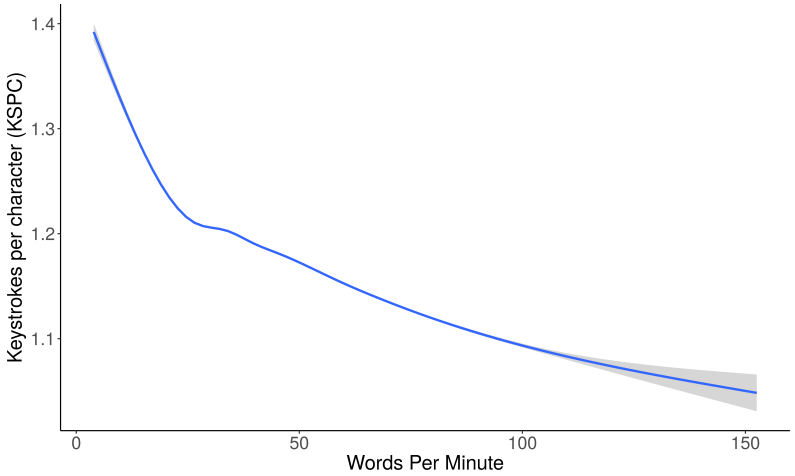


Figure 9: Faster typists also make and leave fewer mistakes, bringing the KSPC measure down too.

Similar to KSPC, similar flat region is observed in the ECPC-WPM curve as well.

4.3 Typist groups

In this section, characteristics of groups of typists defined by their performance and behavior criteria are reported. From the questionnaire, self-reported information about whether a participant has taken a training before and how many fingers he/she uses for regular typing are used. For the purpose of the analysis, the participants are grouped as touch or non-touch typists, using their self-reported information. The typists are also studied as either fast or slow groups based on their typing performance recorded during the test.

4.3.1 Typing Speeds

Because it is not possible to otherwise know and validate whether a typist is actually a touch typist or not, the self reported information in the questionnaire are used to define the terms as follows:

Touch typists (blue): Participants using 9-10 fingers to type and who have taken a typing course. **Non-touch typists (red)**: Participants using 1-6 fingers to type and who have not taken a typing course. Particpants reportedly using 7-8 fingers or with trainings were excluded from either group because:

1. It is expected to improve the distinction between touch typist and non-touch typist groups, by excluding any possible effects due to wrongly reporting a touch typing strategy as using 7-8 fingers.

2. Most typing trainings somehow take touch typing as a standard method, so taking a training could mean one may perform touch typing even if 9-10 fingers are not reportedly used.

Figure 10 shows the similarity in typing performance distribution of trained and untrained typists among the overall fast typists. As we narrow down participants by their WPM speeds, we see that untrained typists show similar distribution of IKI and ECPC measures significantly as well as the trained typists. This shows that untrained typists can reach the same level of performance as trained typists.

Furthermore, similar distribution is also observed for the groups of typists who are reportedly touch typists vs those who report they are not. However, after the filtering mentioned in section 4.3.1, the sample contains about twice as many touch typists as non-touch typists. The similar ranges of IKI and ECPC measures in both groups, especially for those with WPM>80, show that non-touch typists can have similar performance as touch typists, and are in significant numbers. If we exclude the effect of training in touch typing as a standard technique, it might be more interesting to see comparable sizes of touch or non-touch typists with high performance.

4.3.2 IKI and ECPC distribution among participants

Figure 11 is a scatterplot of all participants coloured by their WPM speeds and plotted with their mean IKI and ECPC (error corrections per character). It shows clear distinction between participants who type faster and those who type slower, mostly in the IKI intervals. Also, faster typists seem to have a narrow range of low IKI values, and also a smaller range of error corrections.

Also, it can be seen that there are very few typists who are fast while being careless (here meaning more error corrections). In addition, while there are significantly more participants who are slower because of high IKI values even with low error corrections, there are also considerable group of careless typists who are slow primarily because of high errors made.

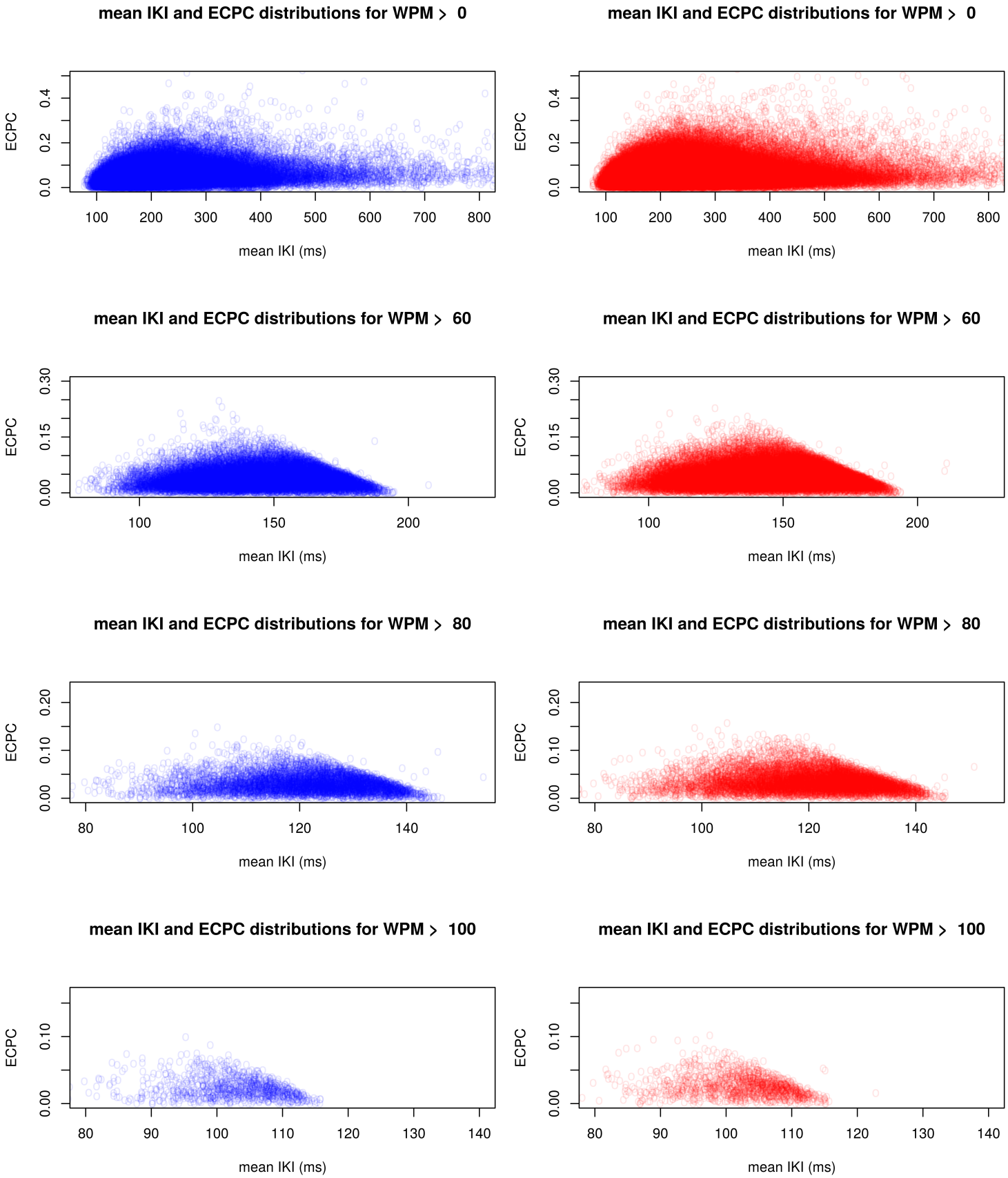


Figure 10: IKI and ECPC distributions by speed groups. blue=touch typists, red=non-touch typists.

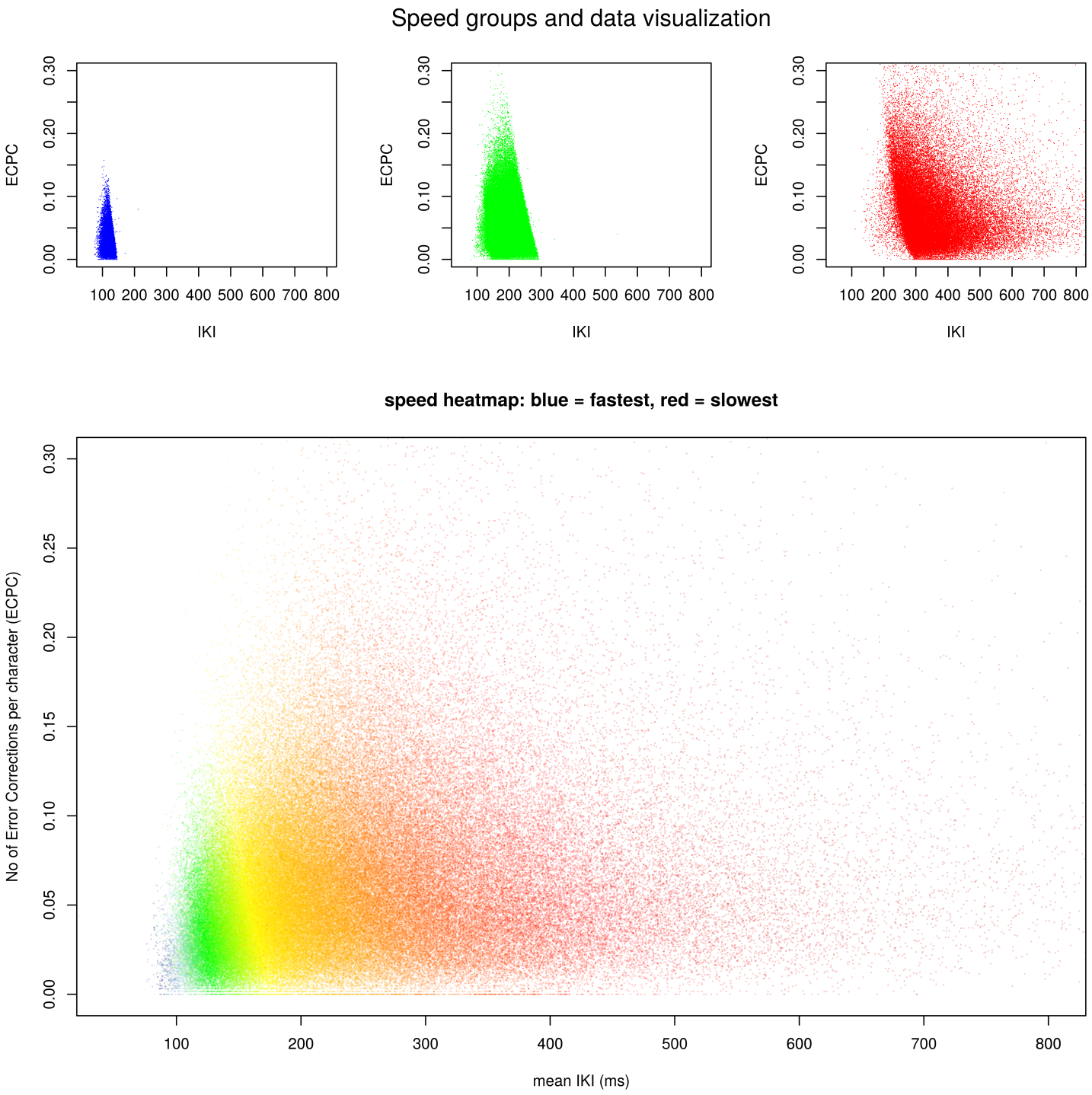


Figure 11: Visualizing IKI and Error Correction rates in continuous speed differences

4.4 Rollover typing

It is observed that for most participants the timestamps of keypress duration of one key overlapped with that of the following key by varying degrees. In physical terms, the fingers are reaching out to press a new key before the previous key being pressed by another finger is released effectively. This is assumed to denote a preparatory behavior in typing. Modern operating systems can detect multiple keys of the keyboard pressed at once in sequence, hence it is possible to type in the rollover fashion reducing the effective time for which a key is pressed down. For this 'rollover ratio' is define as the proportion of keys pressed with rollover (number of overlapped keypairs/total keypairs). Rollover ratio is found to be as high as 0.7 in people who type faster, i.e. about 70% of the total keypresses being pressed with preparation.

Figure 12 shows a user typing two keys, where the next key is typed before the former is released.

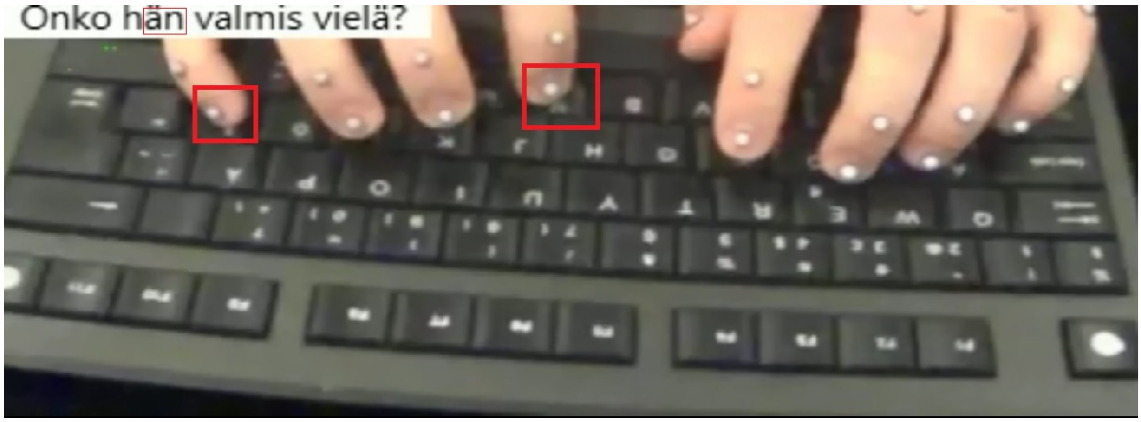


Figure 12: Rollover behavior in action, from recorded typing session in controlled experiment in our lab. The poor quality is due to zoomed image from a high-speed footage.

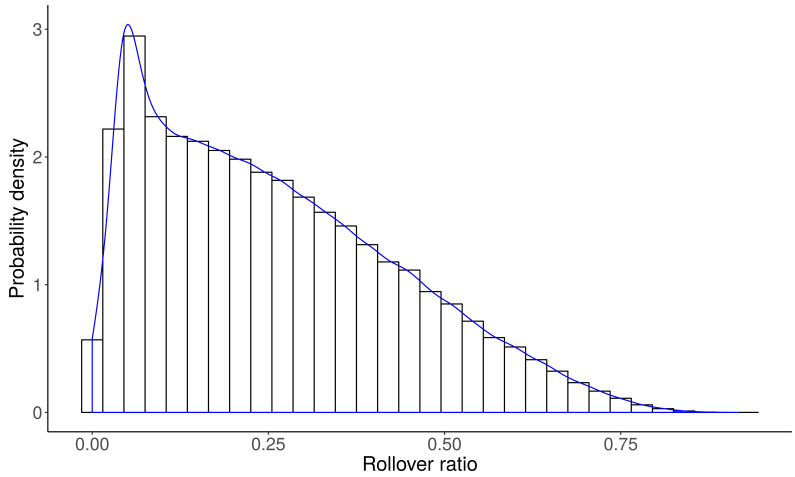
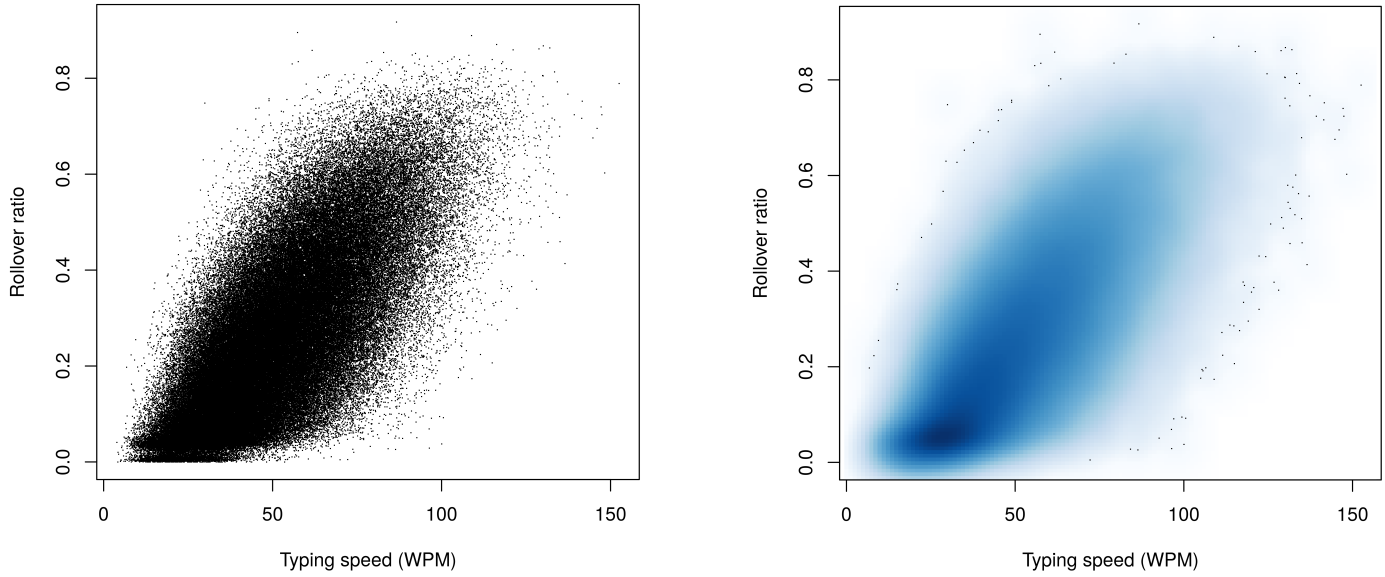


Figure 13: Histogram distribution (probability density) of rollover ratio (R). Rollover is quite prevalent, especially in fast typists (0.5), but also in average (0.25).

On studying how the Rollover ratio is distributed among all participants, it was found that most of the people had about 10-20% rollover (ie. 0.2 as a ratio). Figure 13 shows the density distribution of the rollover ratio measure among the participants.

4.4.1 Rollover ratio vs WPM

It was observed that this behavior has high correlation with typing speed (correlation coefficient = 0.73). The assumption is that rollover typing is an indicator of preparation and thus improves typing performance. Further, the scatterplot in figure 14 shows the elliptical scatterplot of R vs WPM, showing high correlation.



(a) Scatterplot (b) Smooth Scatterplot

Figure 14: Scatterplot visualization of R vs WPM showing high correlation between the two

Figure 15 shows the effect of rollover typing by participants' fingers usage. As seen in the graph, the more rollover typing is used, the higher the speed. This is measured by ratio of rollover keypairs which here refers to the proportion of consecutive keypairs typed in this fashion to the total number of keys typed. Interestingly, the graph shows that at very low rollover ratios, using 9-10 fingers for typing has no benefits over using less fingers. Similarly, at high proportions, even participants that state they use very few fingers can achieve as high WPM scores as those using 9-10 fingers. In intermediate proportions, using more fingers seems to have benefit in achieving increased typing speeds.

It shows that typing with overlaps and preparation is very useful in determining and hence improving typing speeds for all kinds of typists (based on number of fingers).

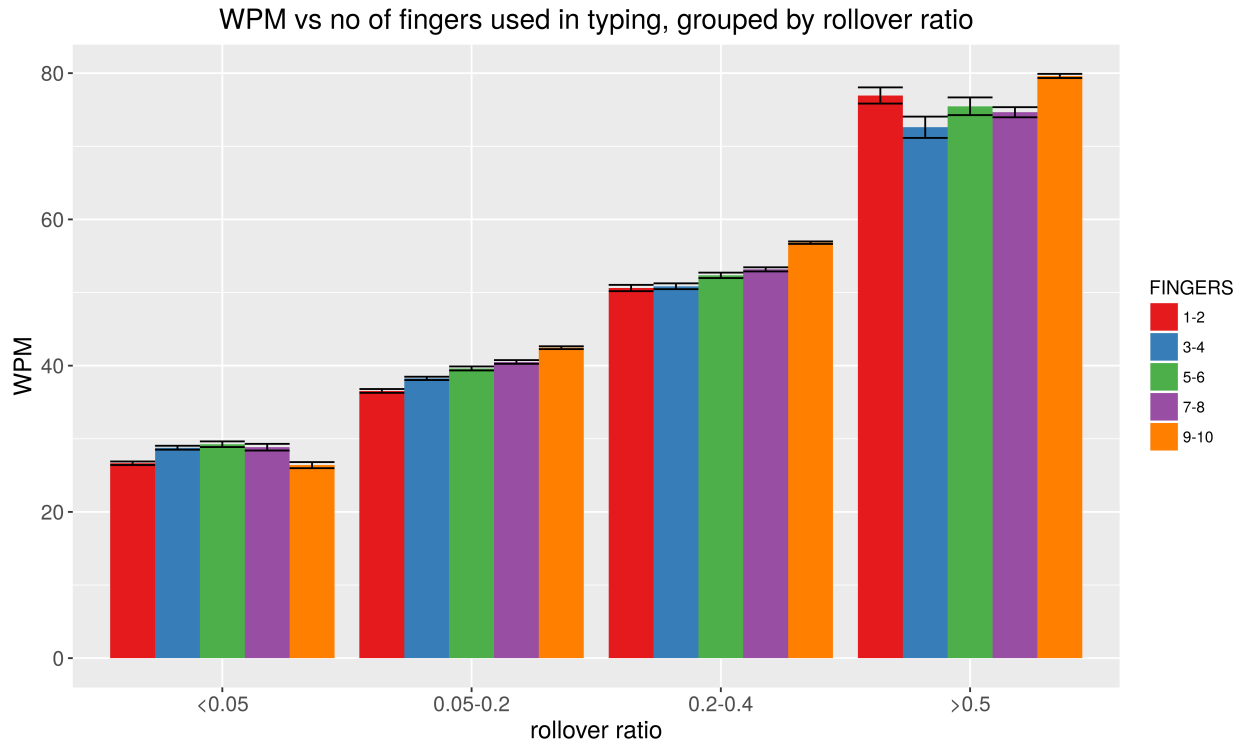


Figure 15: With high rollover, people type faster irrespective of the reported number of fingers used.

4.4.2 Rollover Typing behavior: Comparisons between Speed groups and Touch/Non Touch Groups

In this section, the participants are grouped into two groups based on their typing speeds: those with WPM>=80 belonging to the 'fast typists' group and those with WPM<=40 belong to the 'slow typists' group. Following this, the distribution of Rollover ratio in both groups is studied, and the results correspond to the earlier findings about the correlation of rollover ratio with typing speed. The following figure shows the distribution of rollover ratio for both the groups.

In addition, a comparative study of the distribution of rollover ratio was studied for the 'touch typist' and 'non-touch typists' groups, which are defined in the same way as mentioned earlier. The differences are not as prominent as that in explicit speed groups. This indicates that the preparatory behavior is present in touch typists in similar extent compared to non-touch typist groups.

4.5 Bigram IKI analysis

In this section, we study the Inter-Key intervals of different bigrams and their relation to other measures such as the relative frequencies of the bigrams in English alphabet and the performance measures of participants.

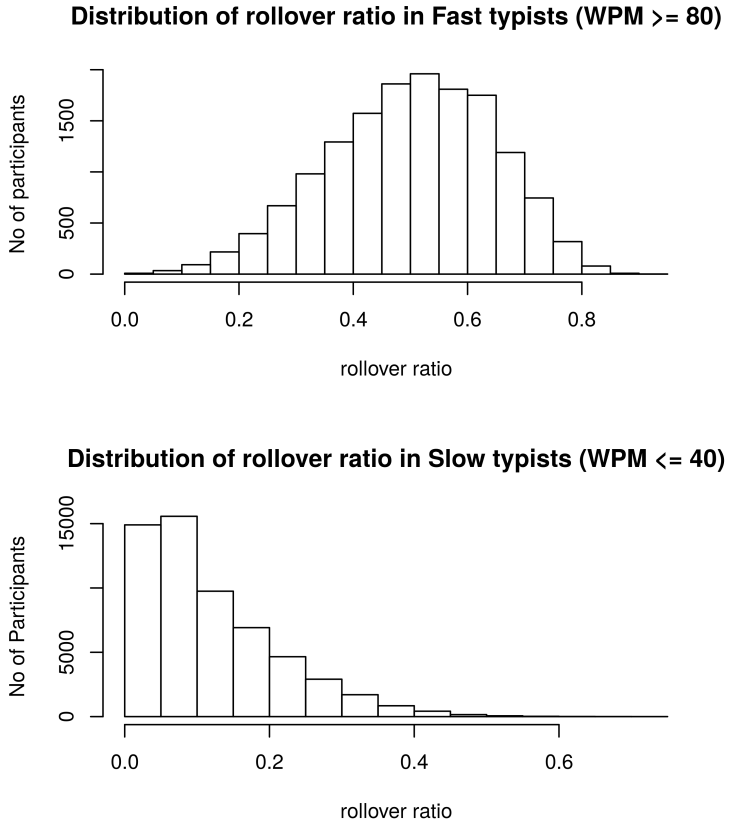


Figure 16: Faster typists leverage rollover typing more than slower typists.

4.5.1 Bigram frequencies

The English language has an inherent distribution of frequencies with which bigrams occur in written text. This has been studied, calculated and documented in various previous works with variety of sources of texts. Google books and Project Gutenberg books are two large text sources from which the studies have been drawn.

The following figure shows the IKI in ms for some frequent bigrams, using extended Google Books corpus's analysis [27] as the reference for bigram frequencies.

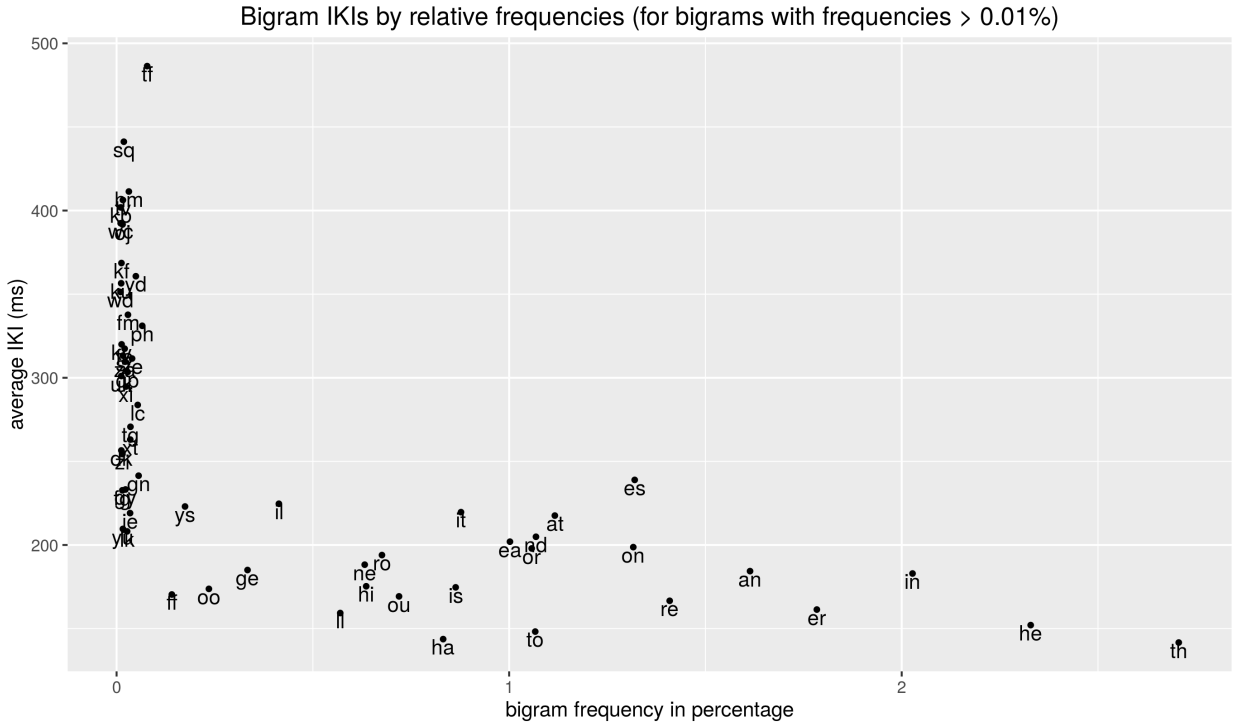


Figure 17: More frequent bigrams are generally typed with lower IKIs than less frequent ones.

4.5.2 IKI and keypress patterns

IKI distribution of selected keypairs: Graphs 18 through 20 below show how IKI is distributed among participants belonging to certain groups of typists based on speed or typing behaviors (fingers/typing course). The keys are selected based on their position on the keyboard and categorized into following types:

1. Repeated character, e.g. aa, cc, ll These character pairs show the distribution of IKI when participants type keys with the same finger.

2. Distant character-pairs, e.g. al, sp, an These character pairs are expected to be typed better with separate hands.

3. Middle-position key pairs, e.g. gh, rt

4. Close keypairs towards one end, e.g. as, lk

5. Left/(or Right) and Middle keypairs, e.g. at, st, lt

6. Keys at different positions with spacebar, e.g. a_, t_, p_ (_ stands for a spacebar)

4.5.3 Fast and slow touch typists

Observations from figures 18 through 20 indicate some important points:

1. Certain keypairs are more determining when it comes to typing performance, based on the position on the keyboard. This may directly be in relation to participants' strategy.

2. For fast typists, typing the same key twice (with the same finger) is slower than typing different keys (with different fingers or hands). For slower, typists, variations are noticed.

3. Additionally, obviously fast typists have IKIs in the narrower range of variance and towards lower levels.

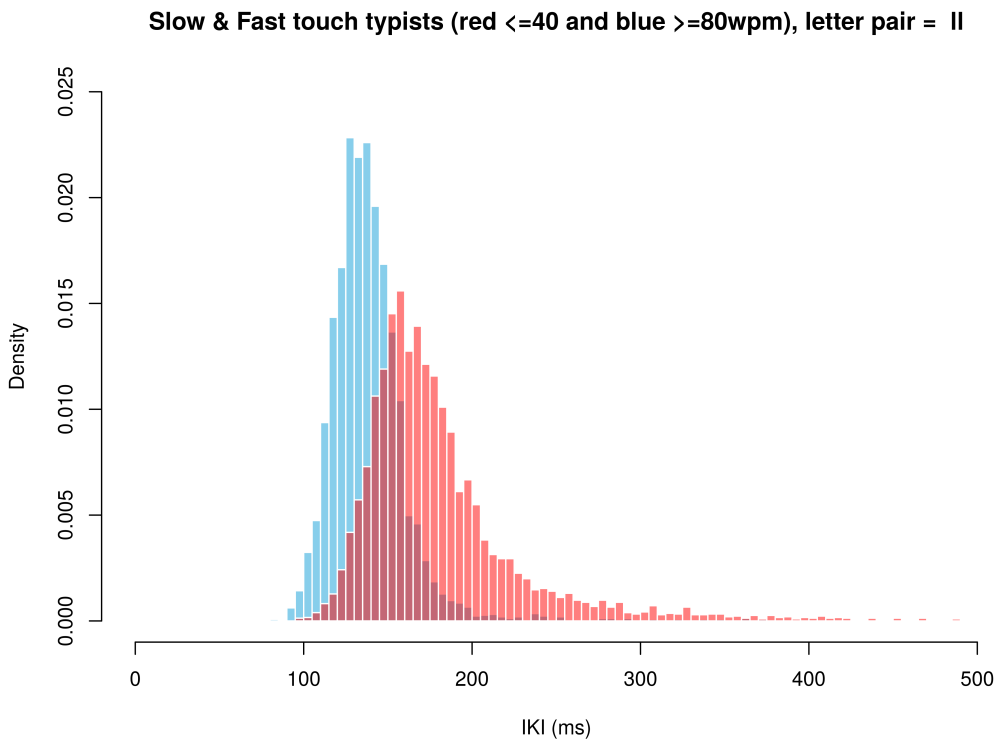


Figure 18: Repetitive bigrams such as *ll* have similar distribution among fast and slow typists.

4.5.4 Touch and Non-Touch Typists

Similar distributions of IKI for touch typists (those who report using 9-10 fingers and have taken a typing course) and non-touch typists (those who report using less than 9-10 fingers and have not taken a typing course) are presented here, irrespective of the speed groups. In comparison, these graphs show less distinction between the groups than the previous ones, suggesting that both of these (touch and non-touch) groups include mixed strategies.

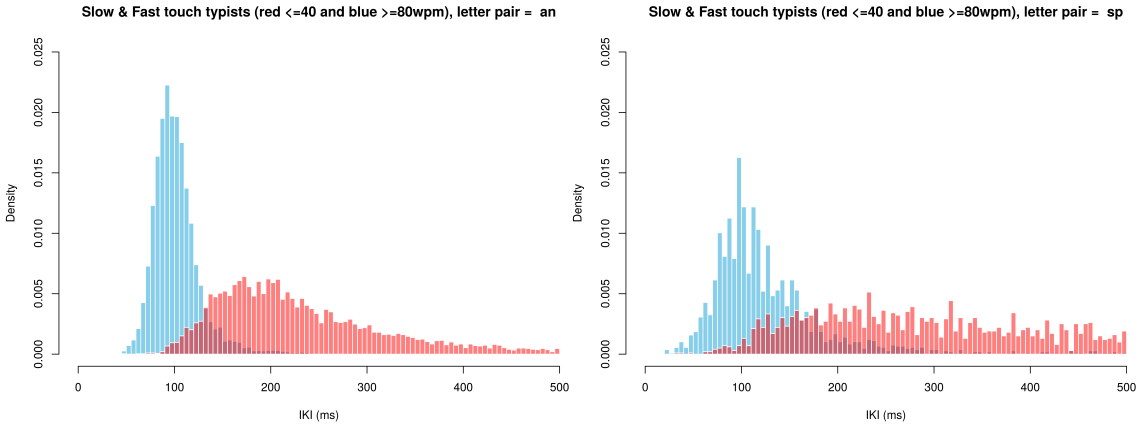


Figure 19: Distant bigrams such as *an* and textitsp have distinct distributions for fast and slow typists and are more indicative of typing speed.

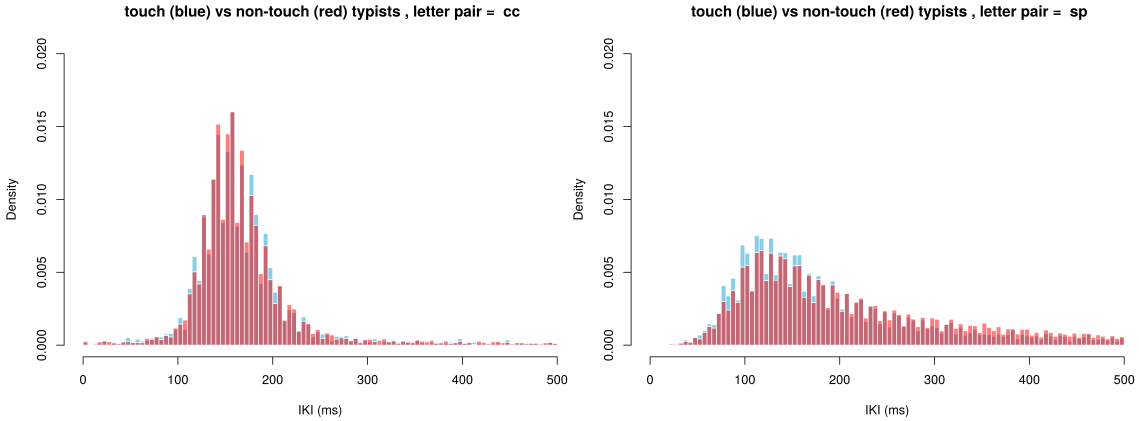


Figure 20: Differences in distribution of distant bigrams are not clear in case of touch and non-touch typists.

4.6 Correlation Annalysis

Figure 21 shows graphically the correlation between different variables studied, along with the correlation coefficient values. The list of variables included in the figure are age, wpm, error rate, keypress duration, iki, KS (total keystrokes count), err_corr (no of error correction keys ie BKSP and DEL), KSPC, ECPC), IKIs of common bigrams (including space as a character, represented by '_'). Kp, IKI and R denote average Keypress durations, Inter-key Intervals and the rollover ratio respectively. The coloured boxes at any cell in the lower triangle represent the correlation between the intersecting variables, while the upper triangle symmetrically states the corresponding coefficients numerically.

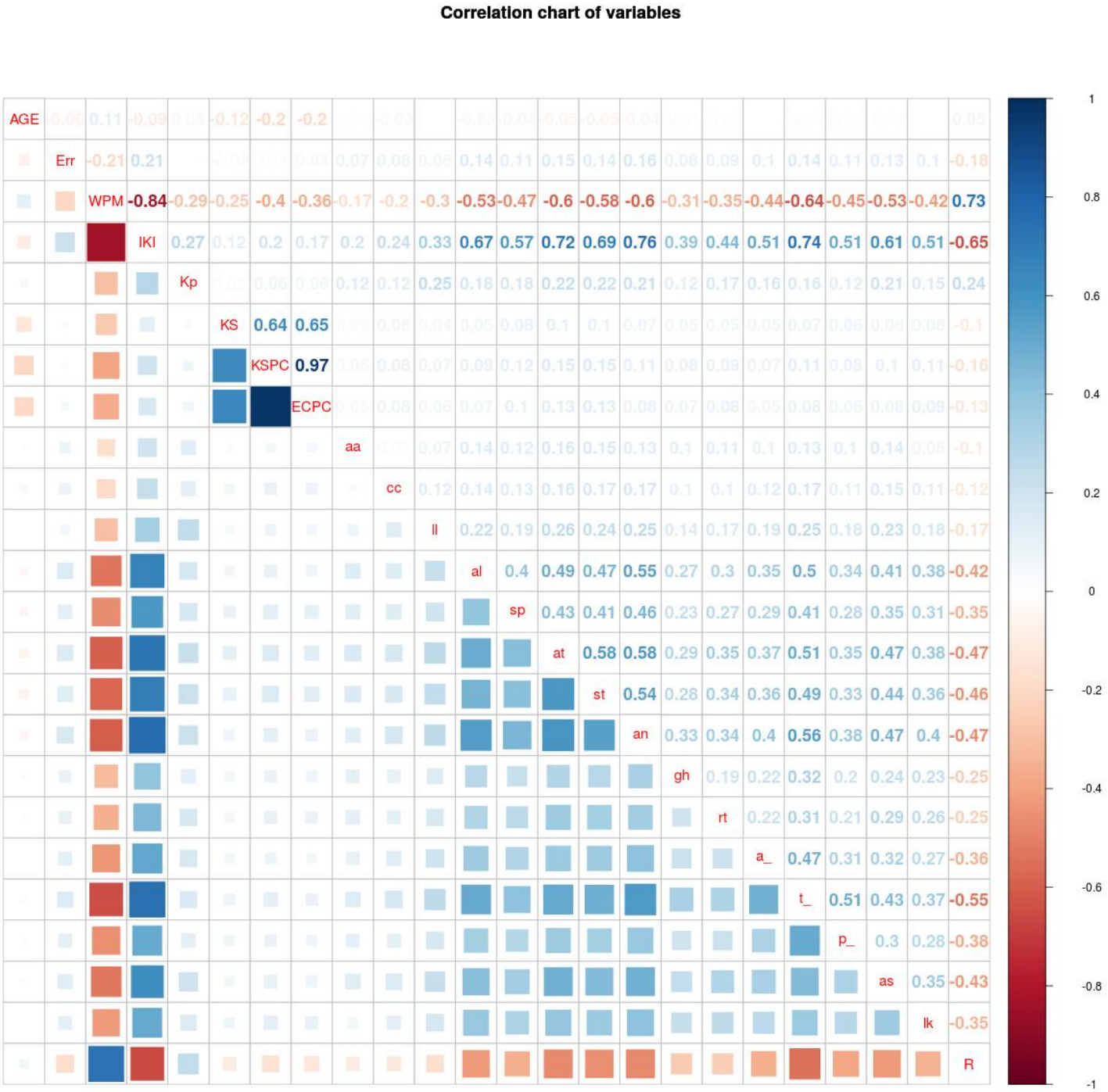


Figure 21: Corelation plot. Blue colour (in the row and columns) at the intersection of two variables (along the diagonal) shows high positive correlation and Red shows high negative correlation. On the upper part, numbers show the exact coefficients. Err: Error Rate, KS: Keystrokes Count, Kp: Keypress Duration, R: Rollover Ratio

5 Keystroke-Level Unsupervised Clustering

This section describes the features studied to perform clustering of the typists, and the measures used to interpret the clustering. Interpretation of the clusters is very important in an unsupervised study because the relative differences between typist groups can be better explained using those measures. In this work, along with performance measures, detailed error analysis (various types of errors and their relative differences) was performed and is described in one of the sections.

5.1 Motivations

Typing strategies can tell why certain (high or low) levels of performances are achieved. More specifically, besides identifying various factors affecting typist's performances, it can shed light on what the best technique to impart typing skill could be. What are the strategies used by fast typists, so we can identify and train new or poor typists using those strategies and techniques. Individual strategies can say about what is missing and how one individual can improve further, for example if certain typing performance is temporarily due to situational factor or in intrinsic to the typing behavior. Personalized remedial and constructive suggestions on learning or improving typing skill of an individual can be generated.

5.2 Bigram IKIs as features for classification of typists

The typing behavior of people can be studied based on the pattern with which bigrams or letter pairs are typed [12]. Keystroke dynamics analysis employs, among other features, bigram IKI patterns to detect the individual; however, we would be more interested in generalizing it so as to describe the typist group's strategy in relation to the group's observable typing characteristics (speed, accuracy, behavior etc). It was observed earlier in this work that certain bigrams are more indicative of the overall performance of a person than others.

Another aspect about bigrams that can be leveraged is the position of the keys on the keyboard. For example, for the bigram 'as', the keys A and S on the keyboard are adjacent, so it can be assumed with good probability that they are typed with the same hand irrespective of the person's strategy. This could always be violated depended on the typist, however, for the sake of simplicity, it is assumed that typing strategies that employ using hand alternation for adjacent keys on the left most (or equivalently on the right-most) region of the keyboard do not sustain longer owing to the inefficiency of hand movements. Following this, we can divide the keyboard into three categories:

1. the left hand bigrams where both letters are located at the left end of the keyboard and assumed to be typed with the left hand,

2. the rightmost bigrams where both letters are located at the right end of the keyboard and assumed to be typed with the right hand, and

3. the bigrams that contain two keys at distant positions in the keyboard, assumed to be typed by alternate hands.

The last assumption above, although a simplification, is based on the idea that otherwise a single finger would be employed to type the two successive keys most of the times, producing either an unsustainable typing behavior in long run or a determinate behavior that can be separately analyzed by a pointing model.

One-hand bigrams		**Hand-alt. bigram**	**Letter repet.**
Left hand	**Right hand**		
as, sa, er, re, sd, ds, ec, ce, ew, we, wa, aw, cr, sc, cs	lk, lo, ol, op, po, io, oi, no, on, in, ni	al, la, ak, ka, am, ma, an, na, ai, ia, so, os, sp, ps, en, ne, em, me, el, le, ep, pe	ll, cc, aa, nn, tt, ss, pp

Table 5: Categorization of bigrams depending on which hand is used to type the corresponding letters at least 90% of the time.

For the purpose of implementation, certain bigrams are enumerated and classified as one of the three categories as mentioned above. Also, various derived measures are used to compare typing behavior between participants. This categorization was done using the ground truth data from our lab [5], by taking bigrams that are typed at least 90% of the time by the left or right or alternately by both hands. The respective bigrams are tabulated in Table 5.

5.2.1 Bigram IKI Distribution

Although different bigrams are typed with different IKIs, with variances between each occurrence of even the same bigram, it becomes interesting to learn about how these IKIs are distributed as a probability distribution. In other words, studying how the IKIs of bigram(s) typed by one person are probabilistically distributed is a part of this work. For this, different probability distributions of IKIs of bigrams typed by a person were fitted to the actual data, and the closest fit in terms of shape of the probability distribution curve, skewness and kurtosis [1]. The distribution analysis is pictorially shown in figure 22 .

[1] R fitdistr package

The Q-Q (quartile againts quartile) plot maps the actual quartile values against theoretical values based on the fitted distribution (in this case a general two-parameter gamma distribution), whereas the P-P (cumulative probability against cumulative probability) plot compares the cumulative probabilities of the data and the fitted distributions. Figure 22 shows the fit is close and IKIs can be described as a accurately with a parametrized gamma distribution.

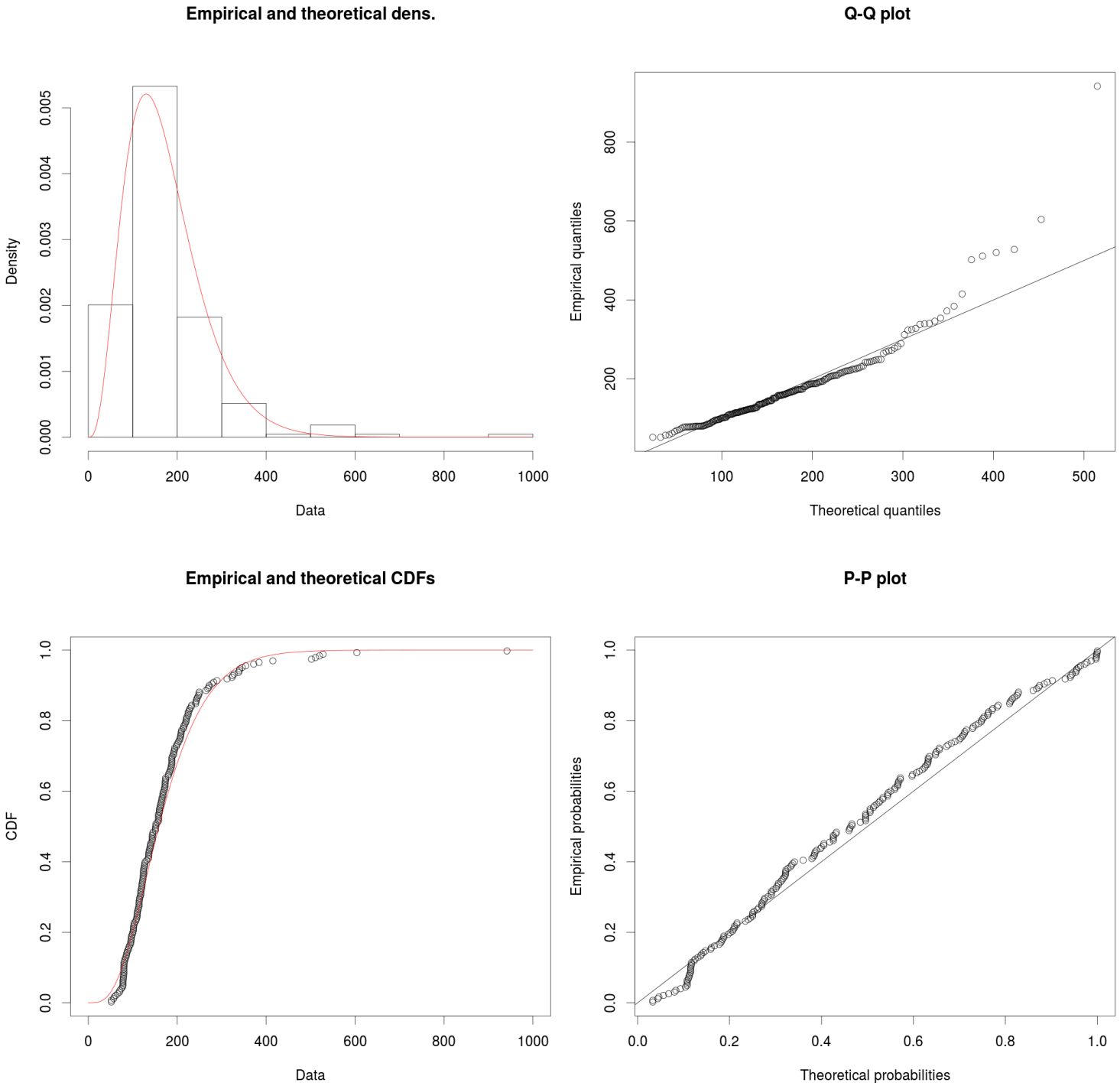


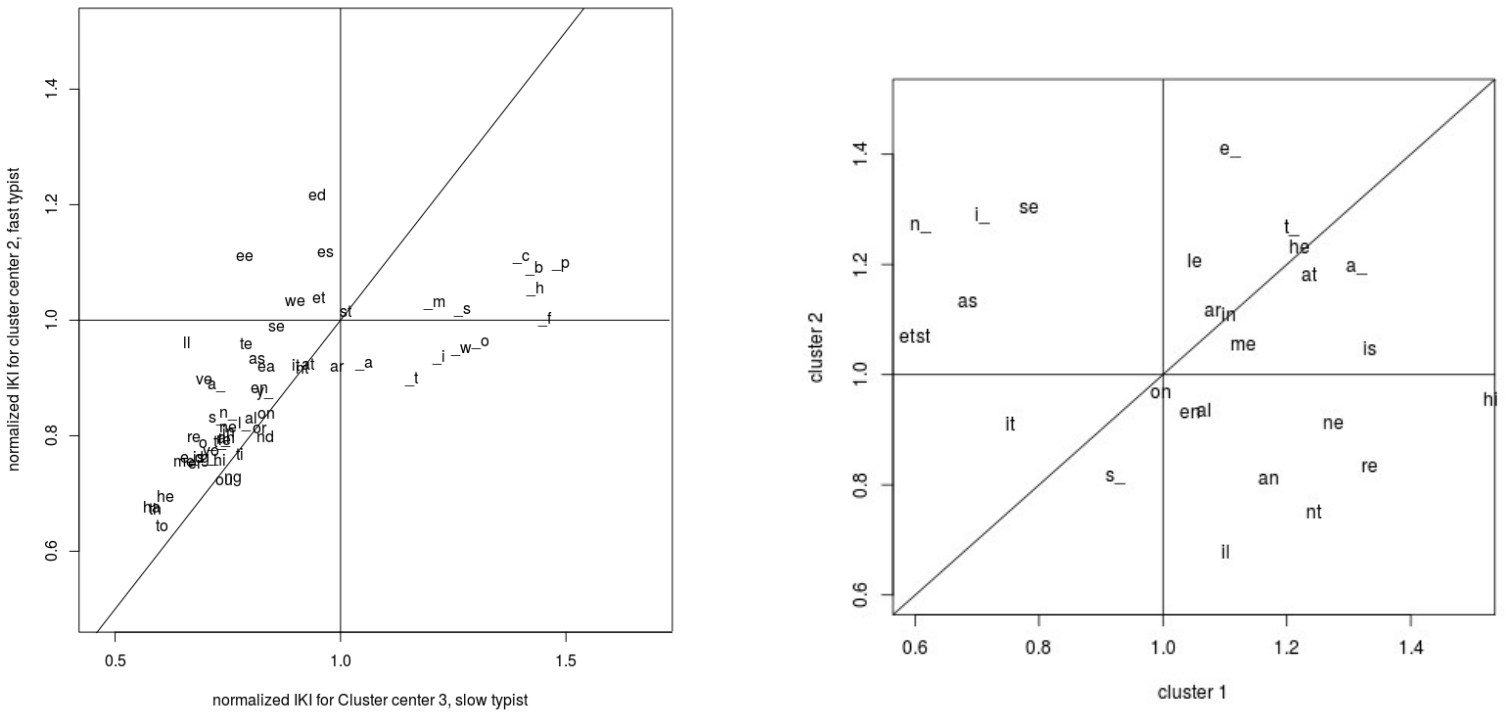
Figure 22: Fitting average IKIs of various bigrams as a gamma distribution.

5.2.2 Comparative bigram IKIs

Comparative bigram IKIs are important tools to study typing strategies and behavior. Some people are more adept with typing certain bigrams better than others, depending on the position of the keys and their demographics or skill in the use of hands and fingers [28]. A comparative bigram IKI graph as the one shown below compares two typists based on their normalized IKIs for various bigrams: for example if the bigram 'an' is under consideration one user may type it faster than his/her overall average IKI, while another may type it slower than his/her overall average. The normalization is done by dividing the absolute IKI values of feature bigrams by their average, ie.

normalised IKI of a bigram $IKI_{norm} = \frac{IKI_{bigram}}{Averageoffeaturebigrams}$

The following graph shows the bigram behavior comparison between two participants, where each axis represents the centre of a cluster.



(a) effect of spacebar-letter bigrams in comparative IKIs

(b) Comparative IKIs of typists belonging to two clusters

Figure 23: comparative IKI plot can be used to compare relative bigram typing behavior of typists

The clustering used as features bigrams that included, among others, spacebar-letter pairs (eg a_, s_ etc). However, the results (the above figure) show that most of the differentiation between clusters is based on these space-letter bigrams, which denote more than hand strategy: since space-letter bigrams are at the start of a new word, [28] mentions the word-initiation effect where the first keystroke is typed

generally with a longer delay and with higher probability of error; As this could be influenced by factors pertaining to parsing and execution, this is excluded in order to ignore such factors in this work.

5.3 Clustering and results

In this section, we discuss the approach, methods and results of another objective of the thesis, ie. to be able to cluster the participants into groups based on typing strategies.

Various methods were used in order to prepare the data for clustering, including feature selection, normalization, Principal Component Analysis tests etc. Essentially, building on the idea that relative bigram speeds capture elements of hand and finger movements, IKIs of bigrams are used as features. The bigrams are selected on the basis of their position, following the observations in previous sections, in order to capture hand and finger movement strategies such as use of one hand, hand alternation,etc. In addition, the IKIs were normalized by dividing by the participants' average overall IKIs, as it preserves the relative IKI differences of various bigrams without over-emphasizing the overall typing speed reflected in the average IKI. This ensures we do not merely get clusters based on typing speed instead of actual typing strategies that we tend to observe.

5.3.1 Methods and approaches

k-medoids clustering The dataset was organized with performance indicators such as WPM, IKI, error measures etc along with different bigram IKIs as fields. However, only Normalized IKIs of bigrams were used as features to avoid the clusters to be overly influenced by typing speed. Finally, the Medoid based partitioning, specifically PAM (Partitioning around Medoids), was chosen as the clustering method. The reasons for choosing this algorithm are:

1. Each user is represented by a data point defined by the normalized IKI in a multi-dimensional space.
2. It is easy to implement considering the large dataset.
3. As a validation model is not in place, it makes more sense to carry out the clustering and then interpret it so that we can devise baselines and validation models for example for further analyses and approaches.

Using normalized bigram IKIs as fields, clara PAM clustering approach was used with Euclidean distance as a metric between data points. The bigrams in the fields were selected based on their proportion in the collected data to avoid missing values. Bigrams which occurred for at least 90% of participants were selected as features, and participants with at least 20 out of 38 such bigram data were

clustered, resulting in $\sim$ 165300 participants to be clustered. For better interpretation, R's clara implementation, which is a median partitioning approach, was used.

To select the number of clusters, the clustering algorithm was run with different cluster numbers and the clustering which resulted in maximum isolation of clusters was selected. Here, isolation refers to the measure of how compact a cluster is (i.e. the average distance of any cluster member from the cluster center) and how well-separated different clusters are (i.e. the minimum distance between a cluster center and any member of a different cluster). Mathematically,

$$isolation_i = \frac{diameter\ of\ cluster\ i}{\arg\min(distance\ from\ x_i\ to\ y_j, \forall j \neq i)} \tag{1}$$

Then, a weighted average of the ratio (weighed by resulting cluster sizes) was compared for various clusterings. Using this criteria in the results of PAM clustering with different numbers of clusters, n=8 was found to result in the minimum isolation value.

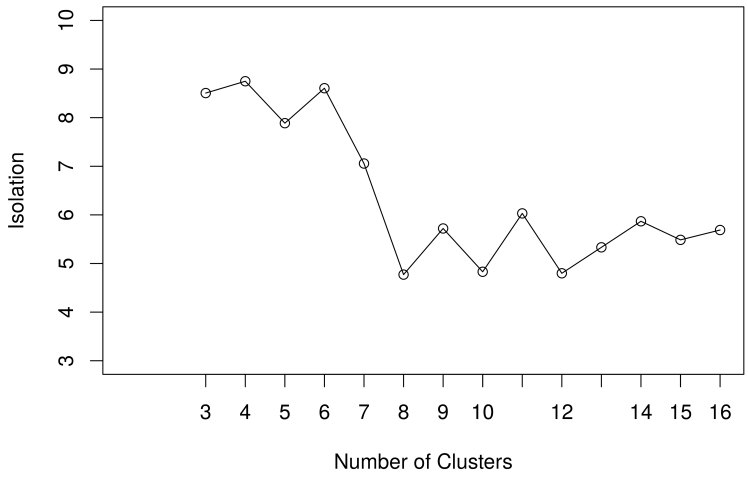


Figure 24: Isolation values obtained by clara PAM clustering for different number of clusters. N=8 clusters yield the best value while giving a low number of meaningful clusters.

Figure 24 compares the weighted average of the isolation values of each cluster (weighted by resulting cluster sizes) for various numbers of clusters. Eight clusters was found to result in the minimum isolation value while yielding a low number of meaningful clusters. The weighted average of isolation values was used to select a clustering with balanced clusters and not very small or very large clusters.

5.3.2 Performance and Error Measures

The following measures were used to interpret the clusters: average left hand bigram IKIs, average right hand bigram IKIs, their ratios, average hand alternation

bigram IKIs etc besides WPM, IKI, ECPC and others to interpret the clusters.

A detailed error analysis aims to look at different types of errors prevalent in the typing of a participant. How errors are made and corrected can also show differences in typing behavior. It can also answer questions about whether a fast participants necessarily makes fewer errors, or there can further be room for improvement as much time is spent in correcting the errors.

Since Wobbrock's TextTest software [29] cannot handle large datasets for detailed error analysis of all participants, it was performed for a subset of the participants (783 participants) closest to the cluster center (within a distance of 60% of the cluster diameter). This was done by first selecting a smaller set of participants whose typing patterns are closest to the cluster centers, and then for each cluster the distributions of various error measures were observed. The error analysis includes both corrected as well as uncorrected error rates. For this, the following types of errors which as used as measures for interpreting the clusters are explained below.

Substitution Errors Substitution error is when a participant wrongly types in a character instead of another. For example, if the word 'rose' is typed as 'rode', then the letter 's' is substituted by the letter 'd'.

Omission Error An omission error is when a letter is omitted from a typed sequence completely. For example, typing 'occuring' for 'occurring' has an omission error as an 'r' is missing.

Insertion Error An insertion error is one where an extra letter not present in the correct text is wrongly inserted in the transcription text. For example, typing 'flight' for 'fight' has an extra (inserted) 'l' and therefore is an insertion error.

In order to study which typist group makes what kind of error the most, these error rates are computed for each cluster. In our data, 783 participants belonging to various clusters, who were closest to their cluster centres, were studied for error analysis. Wobbrock [29]'s TextTest software was used to calculate these measures. The error measures are based on the Input Stream, include both corrected and uncorrected errors, and only take into account errors in typing alphanumeric characters and spacebar (punctuation errors are excluded in the analysis due to issues with the software.)

5.3.3 Clusters

Short descriptions of the clusters are given below (numbers denote clusters).

1 Among various clusters, this group of typist are comparatively slower typists ($\sim$ 48 WPM). However, they leverage hand alternation in typing distant bigrams

as they type such bigrams faster than one-handed bigrams. The typists have a smaller average rollover ratio (0.193) compared to other groups. The typists have the highest error rates.

2 These typists are average typists ($\sim$ 56 WPM), with similar error rates and slightly better rollover compared to cluster 1. They are relatively faster with their hands in typing, with IKI levels close to the faster typists, however do not leverage hand alternation very well compared with other clusters. These typists can improve more with hand alternation practice and through low error rates.

3 This group of typists are average typists ($\sim$ 53 WPM) with only slight improvements in error rates compared to cluster 1, and with similar use of rollover. However, they leverage hand alternation better than other clusters while typing with lower overlaps between keys (rollover). The higher average typing speed compared to cluster 1 can be attributed to the increased hand alternation leverage. They can improve by improving their rollover typing behavior.

4 This group of typists are slow ($\sim$ 46 WPM) with low average rollover ratio of 0.2, and they do not leverage hand alternation as well as other clusters.

5 These are faster than average typists ($\sim$ 65 WPM), with high rollover behavior (avg rollover ratio 0.36). They leverage hand alternation better than typing with single hand. However, they have only slight improvements in error rates compared to the other clusters, so improving accuracy can make them faster.

6 These are average typists ($\sim$ 53) with average rollover ratio (avg. 0.26). They leverage hand alternation as much as the faster ones, however their overall IKI with their hands make then slower. They can improve by bettering their hand movements with practice and by making lesser errors.

7 These are average typists ($\sim$ 52 WPM), with the average overall IKI larger than average IKIs with either hand or with hand alternation, which could mean certain bigrams have more pronounced effect in reducing their overall typing speed. They also have a higher ECPC (error correction rate) than the other groups. This group can improve by careful typing such that errors are minimized and less corrections are required, and by practising certain problematic bigrams more.

8 This is the fastest group of typists ($\sim$ 68 WPM), with high average rollover behavior (38%) and the lowest error rates. They leverage hand alternation well, however the average ratio of left hand to right hand IKI is at 1.09 which is among the highest differences. The difference shows the dominance of the right hand by a slight margin (yet higher than any other cluster) which further hints towards the possibility of improving the typing speed by practicing further with the non-dominant hand. The error rates, especialy substitution error rate, are also

Clusters Summary: Various measures

Cl #	WPM	R	IKI	Left Hand bigrams IKI	Right Hand bigrams IKI	Hand alternation bigrams IKI
1	48.12	0.1929	235.3	217	216.2	185.3
2	56.5	0.272	197.8	180.5	179.8	161.2
3	53.87	0.2117	205.3	205.9	199.9	159.3
4	46.5	0.1998	245.8	221.9	218.5	202.1
5	64.59	0.3575	181.9	173.1	163.2	153.9
6	53.12	0.2623	212.3	204.6	192.7	174.5
7	52.36	0.2444	214.9	205.3	203.8	175.4
8	68.35	0.3776	161.9	159.5	150.1	138.2

	Error Rate				Left vs Right	One hand vs	
Cl #	Uncorrected	Omission	Insertion	Substitution	IKI ratio	alternation ratio	ECPC
1	1.260	0.009	0.0077	0.02	1.037	1.212	0.061
2	1.313	0.0081	0.0067	0.017	1.031	1.145	0.059
3	1.263	0.009	0.0064	0.018	1.064	1.325	0.064
4	1.187	0.007	0.0059	0.017	1.049	1.129	0.06
5	1.220	0.007	0.0063	0.016	1.091	1.116	0.051
6	1.147	0.008	0.0076	0.016	1.102	1.168	0.058
7	1.094	0.0092	0.0081	0.017	1.038	1.208	0.064
8	0.969	0.0061	0.0064	0.011	1.09	1.139	0.051

Table 6: Summary of clusters across averages of different measures

significantly the lowest of all.

Detailed statistics of all clusters is reported in Table 3.

6 Discussions

Study of the participants' key-finger mapping and bigram behaviors helps us learn similarities and differences between their strategies. The results from this work show that it is possible to identify strategies about hand and finger use from the empirical data of the keystroke patterns. This generalized information can tell us more about:

which hand is dominant (used more),

which hand is more leveraged for performance,

weak areas, especially considering errors and hand movement, leveraging of hand alternation,

whether a factor other than hand movement is causing the behavior (such as cognitive factors), etc.

Besides, the work produced large data and analysis useful for reporting typing performances/behavior among global participants. Such data and reports can be used for further analysis as well, using other various techniques and objectives.

6.1 Results and Findings

Different statistics obtained in this work reveal several characteristics about the phenomena and measures of keyboard typing in general. The work started with a large-scale data collection and demographic analysis of the online typing test. A mix of trained and untrained participants, typing with different number of fingers, and at different performance levels, was observed. Studies about speed-error distribution, relation between various measures their implications were reported.

Comparative studies of trained vs untrained typists and reportedly touch vs non-touch typists were done. Untrained typists were found to reach similar levels of performances in speed and accuracy as trained ones. Although in average the reported touch typists were found to be faster than non-touch typists using fewer fingers, it was also observed that many non-touch typists also reach performances similar to their counterparts in terms of both speed and accuracy. This fact stands out even more when we consider the fact that currently touch-typing is the most widespread technique in the training of typing skill.

Rollover behavior was observed as a preparatory behavior with high correlation with typing speed. Fast typists used rollover keys as high as 50-60% of the times or even more. Rollover behavior would not be possible with traditional typewriters as it would jam the keys, and this particular behavior is important for the keyboard and modern devices that support multiple keypress detection. In addition, rollover behavior was observed in fast typists irrespective of where they were touch typists or not, and whether they reported using all the finger while typing or only a few.

Bigram level IKI patterns were studied for certain bigrams, which showed that how we press distant keys on the keyboard indicates better our overall typing performance. It follows that making use of hand alternation and finger alternation is a strategy that can be used to enter keys fast. Hand and finger alternation are also important sources of rollover behavior. The frequencies with which bigrams occur in the English alphabet also affect the time taken to type them: faster bigrams are observed to have lower IKIs in general. In addition, it was observed that in the case of many bigrams people consistently typed each letter using the same hand or finger. These information about bigrams were used to identify important bigrams whose IKIs can be used to classify the participants' typing strategies.

Finally, a clustering of participants was carried out revealing eight groups with specific characterisitcs in terms of performances, error and rollover criteria. Different groups were found to have different levels of the left, right or hand leverages utilized in typing. Some were faster and more accurate than others. It was possible to identify probable causes of the shortcomings in performance and suggest general ways to remedy them. The clustering was carried out with more focus on interpretability.

The next subsections discuss the limitations of the approaches and suggestions for future extensions to this work.

6.2 Limitations

The analysis and reports prepared during this work are based on the data from self-selected online participants, hence it missed out many aspects of a controlled experiment. Although it was closely observed and studied in relation to similar (but with higher scope) controlled data present in Aalto UI lab, the ground-truth dataset was small in size and there are limitations of linking it with a database of the scale the online typing data is.

Clustering was performed with a higher emphasis on interpretation and less emphasis on the clustering technique. Moreover, results from clustering show that, even after normalizing, bigram IKIs resulted in dense overlapping clusters with varying distributions of indicators.

6.3 Future works

Naturally, a more controlled and validated yet large collection of data, even though online, would be an upgrade to the data collection approach. Specific keyboard and typing environments such as mobile (small, soft) keyboards or touch typing in large keyboards can be other approaches.

As keystroke analysis is widely used in biometrics, along with Machine Learning techniques, certain approaches could be used for the purpose of personalized

remedial and constructive typing analysis. One of them could for example be an alternate formulation of the clustering approach as a time-series pattern analysis of the keystroke data. By modelling the keystroke pattern of a typed sentence as a time series, each keypress could be associated with the hand or finger as a latent state. These could be implemented using a Hidden Markov Model [30] or a sequence-to-sequence Neural Network [31] [32]. Of course these would have their own limitation with the large number of participants' models, but they nevertheless represent additional work that might produce further useful insights.

One application can be to use the cluster centers and develop a classifier to categorize new participants and tell about their behaviors. In ideal case, such a classifier would be beneficial for remedial and well as constructive purposes. The information deduced can then be used for applications about typing, such as:

Personalized training program for performance typing

Virtual (custom) Keyboard design

Further Typing test (sentences/method) design

Personalized summary of a typing pattern

Modelling typing behavior can be supplemented with applications in the area of personalized typing assistance. With enough data, learning typing behavior to assist in both typing performance as well as typing accuracy, training new typists optimally or corrective assistance to poor typists would be some objectives of such works.

7 Conclusions

Understanding general typing behavior in terms of keystroke patterns, although even for classification purpose and with high accuracy, has many non-trivial factor in play. Large amount of data is required to be collected, within a method which itself cannot be controlled well to eliminate noise (as in the case of online data collection), and from people whose general shared pattern of keystrokes is hard to predict. However, study and analysis of large dataset with parts of the objectives was well accomplished within the scope of this thesis work.

Despite the limitations, the data and approach can be useful in building on further works on typing behavior and performance.

References

[1] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[2] Rollover key.

[3] Timothy A Salthouse. Perceptual, cognitive, and motoric aspects of transcription typing. *Psychological bulletin*, 99(3):303, 1986.

[4] Jacob O Wobbrock. *Measures of text entry performance*. San Francisco: Morgan Kaufmann, 2007.

[5] Anna Maria Feit, Daryl Weir, and Antti Oulasvirta. How we type: Movement strategies and performance in everyday typing. In *Proceedings of the 2016 chi conference on human factors in computing systems*, pages 4262–4273. ACM, 2016.

[6] William E Cooper. *Cognitive aspects of skilled typewriting*. Springer Science & Business Media, 2012.

[7] Siddharth Jain and Samit Bhattacharya. Virtual keyboard layout optimization. In *2010 IEEE Students Technology Symposium (TechSym)*, pages 312–317, April 2010.

[8] Rajkumar Janakiraman and Terence Sim. Keystroke dynamics in a general setting. *Advances in Biometrics*, pages 584–593, 2007.

[9] Terence Sim and Rajkumar Janakiraman. Are digraphs good for free-text keystroke dynamics? In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–6. IEEE, 2007.

[10] Robert William Soukoreff. *Text entry for mobile systems: Models, measures, and analyses for text entry research*. National Library of Canada= Bibliothèque nationale du Canada, 2003.

[11] Rick Joyce and Gopal Gupta. Identity authentication based on keystroke latencies. *Communications of the ACM*, 33(2):168–176, 1990.

[12] Paul S. Dowland and Steven M. Furnell. *A Long-Term Trial of Keystroke Profiling Using Digraph, Trigraph and Keyword Latencies*, pages 275–289. Springer US, Boston, MA, 2004.

[13] Ngoc Tran Nguyen. Distance-based classification of keystroke dynamics, 2016.

[14] Frode Eika Sandnes. Evaluating mobile text entry strategies with finite state automata. In *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices &Amp; Services*, MobileHCI '05, pages 115–121, New York, NY, USA, 2005. ACM.

[15] Gordon D Logan, Jana E Ulrich, and Dakota RB Lindsey. Different (key) strokes for different folks: How standard and nonstandard typists balance fittsâĂŹ law and hickâĂŹs law. 2016.

[16] Francisco J Valero-Cuevas. An integrative approach to the biomechanical function and neuromuscular control of the fingers. *Journal of biomechanics*, 38(4):673–684, 2005.

[17] Bryan Klimt and Yiming Yang. Introducing the enron corpus. In *CEAS*, 2004.

[18] Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 568–575. ACM, 1999.

[19] Shumin Zhai, Alison Sue, and Johnny Accot. Movement model, hits distribution and learning in virtual keyboarding. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 17–24. ACM, 2002.

[20] I. Scott MacKenzie and R. William Soukoreff. Phrase sets for evaluating text entry techniques. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '03, pages 754–755, New York, NY, USA, 2003. ACM.

[21] Per Ola Kristensson and Keith Vertanen. Performance comparisons of phrase sets and presentation styles for text entry evaluations. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, pages 29–32, New York, NY, USA, 2012. ACM.

[22] Xin Yi, Chun Yu, Weinan Shi, Xiaojun Bi, and Yuanchun Shi. Word clarity as a metric in sampling keyboard test sets. 2017.

[23] Browser keycode.

[24] Sunjun Kim, Jeongmin Son, Geehyuk Lee, Hwan Kim, and Woohun Lee. Tapboard: making a touch screen keyboard more touchable. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 553–562. ACM, 2013.

[25] Richard P Heitz. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in neuroscience*, 8:150, 2014.

[26] Motonori Yamaguchi, Matthew JC Crump, and Gordon D Logan. Speed–accuracy trade-off in skilled typewriting: Decomposing the contributions of hierarchical control loops. *Journal of Experimental Psychology: Human Perception and Performance*, 39(3):678, 2013.

[27] Peter Norvig. English letter frequency counts: Mayzner revisited, 2013.

[28] Timothy A Salthouse. Effects of age and skill in typing. *Journal of Experimental Psychology: General*, 113(3):345, 1984.

[29] Jacob O Wobbrock and Brad A Myers. Analyzing the input stream for character-level errors in unconstrained text entry evaluations. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 13(4):458–489, 2006.

[30] Andreas D Lattner and Otthein Herzog. Unsupervised learning of sequential patterns. In *ICDM 2004 Workshop on Temporal Data Mining: Algorithms, Theory and Applications (TDMâĂŹ04)*, 2004.

[31] Alex Graves. Supervised sequence labelling. In *Supervised Sequence Labelling with Recurrent Neural Networks*, pages 5–13. Springer, 2012.

[32] Lalit Gupta, Mark McAvoy, and James Phegley. Classification of temporal sequences via prediction using the simple recurrent neural network. *Pattern Recognition*, 33(10):1759–1770, 2000.

[33] Van Long Tran. *Visualizing High-density Clusters in Multidimensional Data.* PhD thesis, Jacobs University Bremen, 2010.

[34] Davoud Moulavi, Pablo A Jaskowiak, Ricardo JGB Campello, Arthur Zimek, and Jörg Sander. Density-based clustering validation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 839–847. SIAM, 2014.

[35] Edward Clarkson, James Clawson, Kent Lyons, and Thad Starner. An empirical study of typing rates on mini-qwerty keyboards. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1288–1291. ACM, 2005.

[36] Jing Gao, Pang-Ning Tan, and Haibin Cheng. Semi-supervised clustering with partial background information. In *Proceedings of the 2006 SIAM International Conference on Data Mining*, pages 489–493. SIAM, 2006.

[37] Päivi Majaranta. *Text entry by eye gaze.* Tampereen yliopisto, 2009.

[38] Li Zheng and Tao Li. Semi-supervised hierarchical clustering. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 982–991. IEEE, 2011.

[39] Bilal Esmael, Arghad Arnaout, Rudolf Fruhwirth, and Gerhard Thonhauser. Multivariate time series classification by combining trend-based and value-based approximations. *Computational Science and Its Applications–ICCSA 2012*, pages 392–403, 2012.

[40] Jean-Charles Lamirel, Pascal Cuxac, Aneesh Sreevallabh Chivukula, and Kafil Hajlaoui. A new feature selection and feature contrasting approach based on quality metric: application to efficient classification of complex textual data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 367–378. Springer, 2013.

[41] Xiaojin Zhu. Semi-supervised learning literature survey. 2005.

[42] Miika Silfverberg. Historical overview of consumer text entry technologies. *Text entry systems: Mobility, accessibility, universality*, pages 3–25, 2007.

[43] I Scott MacKenzie and K Tanaka-Ishii. *Evaluation of text entry techniques*, volume 2007. Morgan Kaufmann San Francisco, CA, 2007.

[44] David E Rumelhart and Donald A Norman. Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive science*, 6(1):1–36, 1982.

[45] Jack Tigh Dennerlein, CD Mote Jr, and David M Rempel. Control strategies for finger movement during touch-typing the role of the extrinsic muscles during a keystroke. *Experimental Brain Research*, 121(1):1–6, 1998.

[46] Elizabeth A Bosman. Age-related differences in the motoric aspects of transcription typing skill. *Psychology and aging*, 8(1):87, 1993.

[47] Donald R Gentner. Keystroke timing in transcription typing. In *Cognitive aspects of skilled typewriting*, pages 95–120. Springer, 1983.

A Screenshots of the typing test

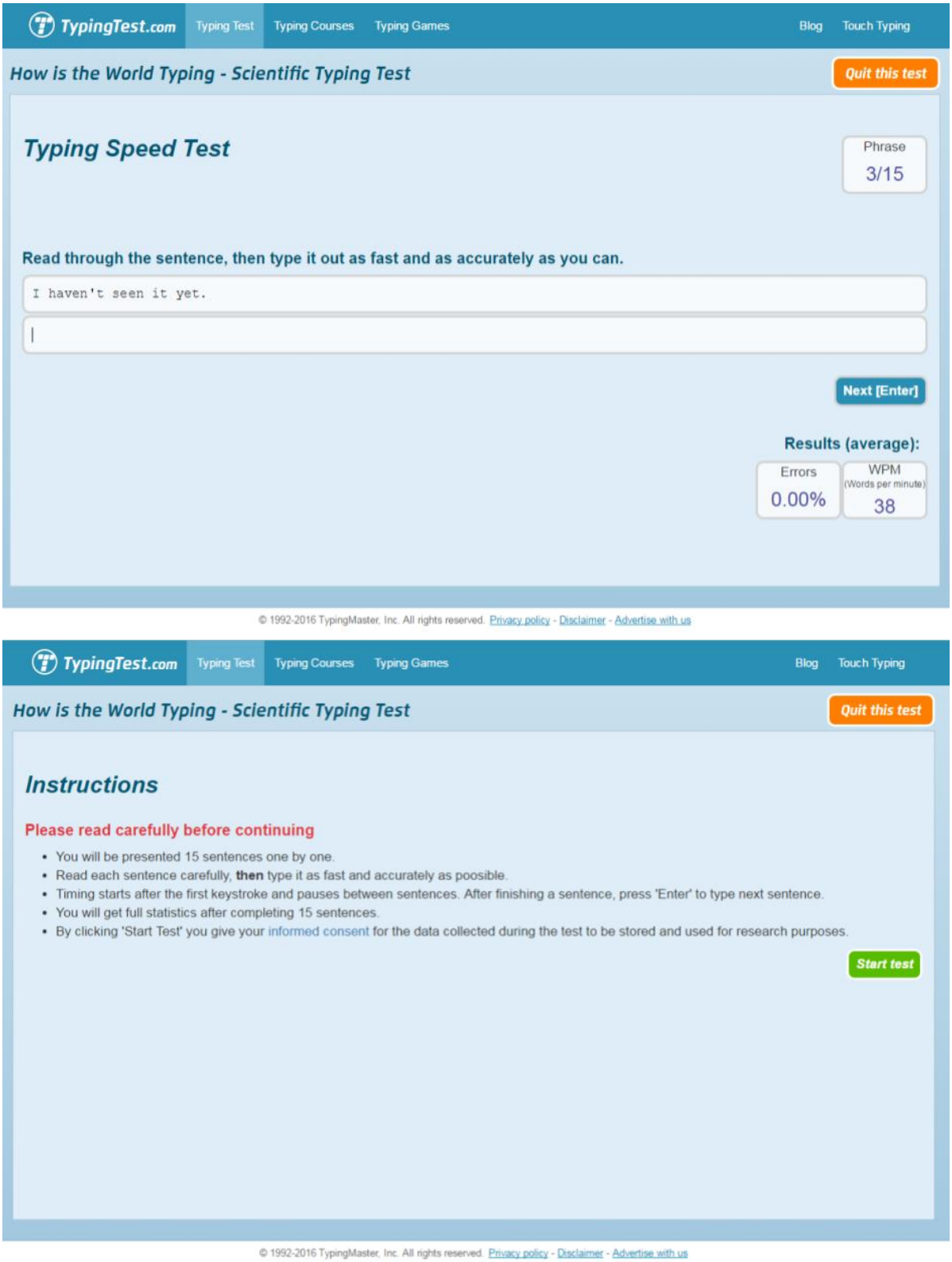


Figure 25

B Typing Speed Test - Questionnaire

1. What is your age?

2. What is your gender?

- (•) Prefer not to specify
- ( ) Male
- ( ) Female

3. Have you ever taken a typing course?

- ( ) Yes
- (•) No

4. What is the highest degree or level of school you have completed?

No schooling completed ▾

5. How many fingers do you use while typing?

1-2 ▾

6. On average, how many hours do you spend writing with your computer per day?

7. (Optional) Do you have any motor impairment or illness that could impact your ability to type?

8. Look at left side of the uppermost row of letters on your keyboard. Which are the first six letters?

QWERTY ▾

9. What kind of keyboard are you using?

Laptop

10. What is your native language?

11. Which country are you from?

Show result

Figure 26