Proceedings of the IASTED International Conference
HUMAN-COMPUTER INTERACTION
November 14-16, 2005, Phoenix, AZ, USA

ACCURACY METRICS IN MOBILE TEXT ENTRY

Matti Koivisto
Principal Lecturer
Mikkeli Polytechnic
P.O. Box 181, 50101 Mikkeli
Finland
matti.koivisto@mikkeliamk.fi

Andrew Urbaczewski
Assistant Professor of Information Systems
University of Michigan – Dearborn
19000 Hubbard Dr, FCS 164
Dearborn, MI 48126
aurbacze@umd.umich.edu

ABSTRACT
There are several metrics utilized to ascertain the relative merits of human computer interaction. In the field of text entry, the Minimum String Distance (MSD) and the Keystroke Classification (KC) metrics are both used to measure text entry usability. This paper examines text entry in mobile devices to see which metric is a better measure of text entry performance. Eighty-seven subjects performed three text entry tasks, each one utilizing a different text input method. Data were collected to calculate both KC and MSD metrics. Discriminant analysis was then used with each metric to classify the 257 cases to their input method. In measuring uncorrected errors, both MSD and the non-corrected error rate (NCER) component of KC were equally weak in classifying the case to its group. In all error and speed conditions, MSD was a much better classifier of the device utilized. The data show that this may be due to KSPC being a better measure of efficiency than accuracy. For accuracy metrics the time of the measurement was an important factor. Metrics measuring the errors in original string was superior to metrics measuring the errors after error correction. Further research is required to more fully understand these phenomena.

KEY WORDS
Text entry, mobile Internet, and quality of metrics

1. Introduction

One of the key challenges of mobile device and system manufacturers is to identify an optimal input method for their devices. The small size of mobile devices prevents the usage of the standard QWERTY keyboards, handheld mice, or other traditional computing input devices. For this reason, a large variety of different input methods, including multiple virtual keyboard designs, multi-tap, and T9 have been introduced.

The importance of text input for mobile devices has made it also as a flourishing research area. Multiple metrics have been developed for analysing the performance of an input method. Text entry speed, often measured in words per minute (WPM), is a widely accepted metric for efficiency, but for effectiveness or accuracy metrics multiple alternatives exist. Two recently introduced methods to measure accuracy in text entry evaluations are: the minimum string distance (MSD) [1] and the keystroke classification (or total error rate (TER)) methods [2] [3]. Both of these methods and others have been used in multiple text entry studies (e.g. [4], [5]), but at least according to our knowledge they are not compared against each other or the quality of them is not systematically analysed.

In this paper we analyse different accuracy metrics used in text entry studies and compare the quality of them against each other. We also study the relative importance of accuracy data against the efficiency related information. Our statistical analysis is based on discriminant analysis and we use accuracy metrics of both MSD and KC methods as predictor variables while creating our classification rules.

In Section 2 we go through different performance metrics used in text entry studies. Section 3 describes the test design used for collecting text entry data and the findings from the survey. In Section 4 we discuss and explain the results in more detail and finally, in Section 5 we provide a few concluding thoughts and describe future research directions.

2. Metrics Used in Text Entry Studies

Many scholars [6], [7] agree that there are two types of usability measurements: performance and preference measurements. In performance measurements we try to collect objective metrics of the system performance. In preference measurements we are interested in user subjective preferences and opinion data. In this paper we will discuss performance measurements only.

2.1 Introduction to Performance Measurements

System performance can be measured in many ways, but if we follow the definition of the International

Standardization Organization (ISO) [8], usability related performance can be further divided into two concepts: efficiency and effectiveness. Measures of efficiency relate the effectiveness achieved to the expenditure of resources. From a user's point of view the time and effort used for the task are resources he or she consumes. Measures of effectiveness relate instead the goals or sub-goals of using the system to the accuracy and completeness with which these goals can be achieved [9].

In text entry evaluations efficiency is usually measured as input speed or throughput. Speed is usually calculated in characters per second (CPS) or even more often as WPM. These metrics are actually identical because the definition of a word for this purpose is five characters, including spaces or any other characters in the inputted text.

The effectiveness of an input method is normally analysed from accuracy point of view. If calculations of entry speed are straight forward, accuracy is another matter. Even, the intuitively simple measure "percent errors" is problematic and more or less different methods are used.

2.2 Minimum String Distance Method

Efforts are underway to streamline and standardize text entry experiments. In particular, Soukoreff and MacKenzie have made an important contribution in this field. They first introduced a method based on the application of the Levenshtein String Distance Statistic [1]. The algorithm yields the minimum distance between two strings defined in terms of editing primitives. The primitives are insertion, deletion, and substitution. The idea is to find the smallest set of primitives that applied to one string (transcribed text) produces the other (presented text). The number of primitives in the set is the minimum string distance (MSD).

Using the MSD statistic they propose the following definition of text entry error rate, given a presented text string (A) and a transcribed text string (B):

$$\text{MSD Error Rate} = \frac{\text{MSD}(A,B)}{\text{Max}(|A|,|B|)} * 100\ \% \quad (1)$$

In addition to MSD based error rate they recommend the usage of another metric, keystrokes per character (KSPC) as a measure of corrected error.

2.3 Keystroke Classification Method

The latest Soukoreff-MacKenzie [2] [3] accuracy metrics are based on delineating participants´ keystrokes into four classes:

- Correct (C) keystrokes – alphanumeric keystrokes that are not errors,
- Incorrect and Not Fixed (INF) keystrokes – errors that go unnoticed and appear in the transcribed text
- Incorrect but Fixed (IF) keystrokes – erroneous keystrokes in the input stream that are later corrected, and,
- Fixes (F) – the keystrokes that perform the corrections (i.e. delete, backspace, cursor movement)

Based on this classification several statistics can be easily calculated, for example

$$\text{Total Error Rate (TER)} = \frac{(\text{INF} + \text{IF})}{(\text{C}+\text{INF}+\text{IF})} * 100\ \% \quad (2)$$

$$\text{Not Corrected Error Rate (NCER)} = \frac{\text{INF}}{(\text{C}+\text{INF}+\text{IF})} * 100\ \% \quad (3)$$

$$\text{Corrected Error Rate (CER)} = \frac{\text{IF}}{(\text{C}+\text{INF}+\text{IF})} * 100\ \% \quad (4)$$

3. Study Design And Results

3.1 Study Design

The aim of our study is to compare accuracy metrics introduced in Section 2 statistically against each other in order to analyse the quality or goodness of them. In our experiment we collected metrics used in both MSD and Keystroke Classification methods while a group of test users wrote email messages with three different input methods. After that we used discriminant analysis to provide classification rules that classifies cases back to three different groups.

The basic concept underlying discriminant analysis is fairly simple. Linear combinations of the independent variables are formed and they are used for classifying cases into one of the predefined groups. Often discriminant analysis is used to predict the outcome of the new case by comparing the characteristics of the case to those cases whose success or failure is already known (e.g. creditworthiness). The basic strategy in our case was a little bit different. We of course knew the used input method in all cases, but we wanted to study the clustering capabilities of different accuracy metrics. Separate classification rules were created for all accuracy metrics introduced in Section 2. In our analysis the classification rule that minimised the probability of misclassification was considered superior.

In our experiment the input methods were stylus pen, multi tap, and reduced QWERTY keyboard and the device used was a PDA (a Compaq iPaq 3870 PDA with

IEEE 802.11(b) WLAN connections). We also used three different message lengths (21, 63 and 197 characters) to study possible effects related to the number of characters.

To be able to collect the required information for performance metrics calculations (like presses of backspace etc.) we bypassed the operating system's standard input methods and wrote the user interface totally with Macromedia Flash. For example the pressing a letter ´a´ in a keyboard did not directly enter a letter to the text field in the user interface. Instead an Action script connected to the on release event of invisible button was called and the code added a letter to the display. For the same reason we were not using the operating system's soft keyboard but created our own soft keyboard (see Figure 1).

Figure 1. Our soft keyboard layout.

Even though multi-tap is a widely used input method in mobile phones it is not a standard feature in PDA devices. We implemented the multi-tap input method for a PDA with reduced QWERTY keyboard by re-labeling the used keys and covering the unused ones. Figure 2 shows devices used in different input methods.

The messages we used are shown in Table 1. It should be noticed that in two messages (called a standard and a start of dialogue message) users filled three fields (receiver's address, topic and message) and in reply message only one field (message).

Figure 2: Devices used in the experiment.

Table 1. Messages used in the experiment.

Type	Field	Content
Reply message	Address Topic Message	- - tuesday is ok see you
Standard message	Address Topic Message	sara@rock.net message the quick brown fox jumps over the lazy dog
Start of dialogue message	Address Topic Message	joe@mail.com hi joe how are you want to meet tonight want to go to the movie with sue and me what show do you want to see we are meeting in front of the theatre at eight let me know if we should wait

3.2 Data Collection and Results

As mentioned earlier data collection took place in a laboratory study in which undergraduate students of a large polytechnic school in Finland wrote email messages with three different input methods. The number of subjects in our study was 87 (64 male, 23 female). Because each participant used three different input methods the total number of the cases was 261. Four cases were removed from the analysis because the test failed for some reason (e.g. the mobile phone of the test user rang during the test).

We used the Latin square technique in organizing the order of the input methods and message length to avoid a learning effect tainting the subjects and the results. Key statistics of the collected data is shown in Table 2.

Table 2. Item Statistics for Three Input Methods.

	STYLUS (N=85)		KEYBOARD (N=87)		MULTI TAP (N=85)	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
WPM	5.1	1.8	13.8	5.2	5.1	2.1
MSD	1.4%	0.028	0.7%	0.018	0,6%	0.014
KSPC	1.3	0.59	1.15	0.53	2.79	1.35
NCER	1.4%	0.028	1.0%	0.026	0.7%	0.023
CER	12.1%	0.11	5.1%	0.092	15.2%	0.16

Based on the collected data we first created multiple classification rules for the MSD and the keystroke classification. The classification results of these discriminant functions are shown in Table 3.

Table 3. The Percentages of Original Grouped Cases

	MSD method		Keystroke classification	
	Value	Metrics	Value	Metrics
Not corrected errors metric only	36.6%	MSD	37.4 %	NCER
Corrected errors metric only	70.8%	KSPC	47.9%	CER
Accuracy metrics together	61.5%	MSD+ KSPC	50.2%	NCER+ CER
Speed metric only	61.9 %	WPM	61.9%	WPM
Speed metric added	86.4%	WPM+ MSD+ KSPC	63.4%	WPM+ NCER+ CER

The results in Table 3 show that metrics measuring uncorrected errors (MSD and NCER) are equally weak in correctly classifying the cases to their correct groups. Their values (36,6% and 37,4%) are so close to a random classification result (33,3 % with three groups) that their can be considered almost worthless in identifying the input method.

For corrected error metrics the situation is the opposite. Both CER and KSPC give a higher success rate, but KSPC has a much higher grouping capability compared to CER. Its strength makes the MSD method in general a better classifier, giving an 86,4 % result when used together with the speed metric WPM.

4. Discussion

The results of our test suggest that from the two accuracy measurement methods studied here, the metrics used in the MSD method are stronger and better. However, we are not ready to make that conclusion for two reasons. First of all, the two suggested metrics for uncorrected errors (MSD error rate and NCER) have equally weak clustering capability and the difference between two methods is totally based on the difference between KSPC and CER. Secondly, we do not agree with Soukoreff and MacKenzie when they suggest KSPC as a metric for corrected errors. It is without any doubt a strong metric but the evidence indicates it is an efficiency metric, not an accuracy metric. KSPC is measuring the effort made by a user (number of keystrokes) needed to create a character, which is the task he or she is doing in text entry.

More importantly our test indicates that it is very important to pay attention when the accuracy is measured. In our unconstrained text entry study the users had freedom to correct errors or to leave them uncorrected. This divided the text entry task to two overlapping processes: initial entry and correction process as shown in Figure 3. From accuracy metrics analysed in this study only CER is measuring error rate during entry (entry process metric) and MSD and NCER are measuring error rate of the transcribed or corrected string.

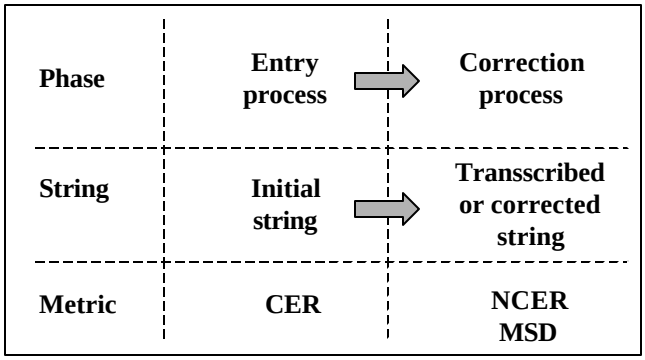

Figure 3. Sub-processes of text entry task

Our results suggest that if accuracy is measured after the correction process the accuracy measure is no longer correlating with the used input method. Instead if the accuracy is measured before correction processes the correlation between accuracy and input method still exists.

The classification of accuracy metrics into two categories is supported by Wobbrock [5]. In his joystick text entry study he found out that even though some text entry method had a higher error rate during entry subjects´ transcribed phrases were more accurate.

The correlation matrix in Table 4 highlights the difference between the initial and corrected string error metric. NCER and MSD error rate have significant correlation at the 0.01 level. But CER does not correlate with either NCER or MSD. This clearly indicates that even though

they are both measuring accuracy they are measuring it at different levels.

Table 4. Correlation of the accuracy metrics

	CER	NCER	MSD ERR
CER	1	-.009	-.033
NCER	-.009	1	.829**
MSD ERR	-.033	.829**	1

** Correlation is significant at the 0.01 level

5. Conclusion and Future Steps

As with all research studies our work has several limitations. First of all our study was carried out in controlled laboratory instead of real environment. Mobile devices are mostly used in dynamic context and in our experiment user movements were restrained. In addition to that our study was limited to one device and to a student population, which raise the issue of generalization.

It is too early to make final conclusions about the quality of different metrics used in text entry studies and further studies are needed to understand the relationships between different usability aspects in more detail. However, our results suggest that in text entry, task efficiency related information is more valuable than accuracy data. For this reason WPM should not be used as an only efficiency metric but it should be used together with other efficiency metrics like KSPC.

When evaluating the accuracy of the text entry it seems to be very important to know when the accuracy is measured. In this study we divided the text entry into two phases: initial entry and correction process. If accuracy is measured before the error correction process the input method used can be identified and correlation between accuracy and input method exists.

If accuracy is measured after the error correction process accuracy and the input method no longer correlate. In that case one is not measuring the accuracy of the input method alone but participants' precision level or conscientiousness is affecting to the measurement results.

Our results are indicating that accuracy is an attribute of both the user and the device. For this reason we do not recommend the usage of NCER or MSD when comparing accuracy of different input methods against each other but consider CER as a better accuracy metric in that situation. Instead NCER and MSD can be used when accuracy of the user is the main interest.

Acknowledgements

We want to thank students of Mikkeli Polytechnic taking part of our experiment. We would also like to thank the reviewers of this paper for their useful comments.

References

[1] R.W. Soukoreff, & I.S. MacKenzie, Measuring errors in text entry tasks: An application of the Levenshtein string distance statistic. Extended Abstracts, *Proc. ACM Conference on Human Factors in Computing System – CHI 2001*, New York, NY, 2001, 319-320.

[2] R.W. Soukoreff, & I.S. MacKenzie, Metrics for text entry research: An evaluation of MSD and KSPC, and a new unified error metric. *Proc. ACM Conference on Human Factors in Computing Systems – CHI 2003*, New York, NY, 2003, 113-120.

[3] R.W. Soukoreff, & I.S. MacKenzie, Recent developments in text entry error rate measurements. Extended Abstracts, *Proc. ACM Conference on Human Factors in Computing Systems*, New York, NY, 2004, 1425-1428.

[4] P.M. Commarford, An investigation of text throughput speed associated with Pocket PC input method editors. *International Journal of Human – Computer Interaction, 17*(3), 2004, 293-308.

[5] J.O. Wobbrock, B.A. Myers, H.H. Aung,, & E.F. LoPresti, Text Entry from Power Wheelchairs: EdgeWrite for Joysticks and Touchpads. *Proc. 6th ACM Conference on Computers and Accessibility*, Atlanta, GA, 2004, 110-117.

[6] N. Bevan, Measuring usability as quality of use. *Journal of Software Quality, 2*(4), 1995, 115-150.

[7] J. Larson, The what, why and how of usability testing. *Speech Technology Magazine, 7*(5), 2002.

[8] ISO, Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability, *International Organization for Standardization 9241-11*,1998.

[9] N. Bevan, & N. Macleon, Usability Measurement in Context, *Behaviour and Information Technology, 13*(1), 1994, 132-145.