Regular Paper

Typing Tutor: Individualized Tutoring in Text Entry for Older Adults Based on Statistical Input Stumble Detection

Toshiyuki Hagiya[1,2,a)] Toshiharu Horiuchi[1,b)] Tomonori Yazaki[1,c)] Tatsuya Kawahara[2,d)]

Received: July 18, 2017, Accepted: January 15, 2018

Abstract: Many older adults are interested in smartphones. However, most of them encounter difficulties in self-instruction and need support. Text entry, which is essential for various applications, is one of the most difficult operations to master. In this paper, we propose Typing Tutor, an individualized tutoring system for text entry that detects input stumbles using a statistical approach and provides instructions. By conducting two user studies, we clarify the common difficulties that novice older adults experience and how skill level is related to input stumbles with a 12-key layout for Japanese. Based on the study, we develop Typing Tutor to support learning how to enter text on a smartphone. A two-week evaluation experiment with novice older adults (65+) showed that Typing Tutor was effective in improving their text entry proficiency, especially in the initial stage of use. In addition, we demonstrate the applicability of Typing Tutor to other keyboards and languages with the QWERTY layout for Japanese and English.

Keywords: text entry, mobile devices, older adults, tutoring system

1. Introduction

Smartphones offer new opportunities to improve the lives of older adults [1]. Although these individuals would like to learn about smartphones [2], those who have never used one may face difficulties because of their lack of experience. Some older adults give up using a smartphone and go back to using their old feature phone. Therefore, providing support in the initial stage is very important. To make full use of the functions of a smartphone, it is essential to master text entry on a touch screen. However, text entry is one of the operations that novice older adults find most difficult because considerable background is needed, such as knowing how each key corresponds to a character and how to select a suggestion. In addition, touch interfaces lack both the mechanical stability and tactile feedback of a keyboard, making it harder to accurately select targets [3].

Many researchers have tackled the issues associated with making text entry easier using a number of approaches, e.g. changing the layout, adjusting the key target areas, and presenting appropriate suggestions [4], [5], [6]. These aids are effective for users who are accustomed to smartphone operation. However, in addition to the problems mentioned above, smartphone novices, especially older adults, tend to have cognitive problems, such as forgetting and losing how/what to type next [7]. They therefore need patient support during the initial stages of use. According to a study [1], older adults tend to prefer an instruction manual to trial-and-error. An ideal approach is for a system to provide appropriate instructions that can be adapted to the individual user when the user has trouble during the learning process.

Accordingly, we have designed Typing Tutor, a tutoring system that automatically detects input stumbles and provides instructions that help users resolve input stumbles independently [8], [9]. We define the word "input stumble" as an occasion when a user makes a mistake or does not know how/what to type next. Because most older adults in developed countries own a mobile phone, the targeted user group is older adults who have never used a smartphone but have owned a feature phone.

In previous studies [8], we constructed an assistive typing application, which detects input stumbles and provides instructions on the typing of presented sentences. In this study, we clarified the common difficulties that novice older adults encounter in free text entry, and analyzed the relationship between the input stumbles and the user's skill. In addition, we developed an individualized tutoring system which detected input stumbles using a statistical approach with a 12-key layout keyboard for Japanese. Moreover, we demonstrate the applicability of the input stumble detection method to other keyboards and languages with a QWERTY layout for Japanese and English.

In this paper, we present the five steps to construct Typing Tutor. First, we describe a user study to clarify the problems older adults encountered in text entry. Second, we design the structure of Typing Tutor on the basis of the previous user study. Then, we improve Typing Tutor through two trials. Subsequently, we evaluate the performance of Typing Tutor in a two-week experiment. Finally, we describe the evaluation experiment of a statistical input detection method in order to demonstrate the applicability of the method to other keyboards and languages with a QWERTY layout for Japanese and English.

[1] KDDI Research, Inc., Fujimino, Saitama 356–8502, Japan
[2] Kyoto University, Kyoto 606–8501, Japan
[a)] to-hagiya@kddi-research.jp
[b)] to-horiuchi@kddi-research.jp
[c)] yazaki@kddi-research.jp
[d)] kawahara@i.kyoto-u.ac.jp

2. Related Work

2.1 Interface Design for Older Adults

Many studies have attributed older adults' difficulties in learning to use technology to a number of user characteristics, such as a decline in perceptual performance and a lack of experience in the relevant technology. Rama et al. [10] noted that older adults have difficulty learning current user interfaces because they generally have less experience with current devices than younger adults, and need to learn different types of user interfaces from former technologies. In addition, user interfaces often have buttons that perform different context-dependent tasks. To resolve such problems, Fisk et al. [11] provided a guideline for older adults. The guideline emphasizes interface changes that respond to age-related changes in perception and cognition: larger displays, fonts, and buttons, and accessibility improvements in specific contexts. With respect to the touch interface, Jin et al. [12] have investigated optimal target size, spacing, and position to derive recommendations and general guidelines for older adults. While these guidelines are very useful, needs can differ among individuals [13]. To respond to the various needs of older adults, researchers have considered adaptive solutions based on user profiles such as behaviors [14], knowledge [15], and skills [16].

2.2 Text Entry for Older Adults

Some studies proposed methods for transforming detection areas based on language models and users' touch distribution [5], [6]. Rodrigues et al. [6] analyzed the influence of the typing behavior of older adults by varying the keyboard, including the color and the width of keys. Gunawardana et al. [5] proposed methods for varying the detection areas based on the typing history in the language model. As a different approach, Bi et al. [4] optimized an algorithm for presenting suitable suggestions to make correction and completion easier. However, Kurniawan [17] reported that older adults usually dislike text-prediction features. Therefore, Komninos et al. [18] proposed a keyboard that makes users aware of any errors through highlighting text.

On the other hand, various typing apps have been proposed in the smartphone market. In most typing apps, users are prompted with a text showing what they are required to type. When users mistype a character, they are notified through auditory and visual feedback, such as beeps and underlines. This is effective for highly motivated users with a certain degree of knowledge and skill. However, a report [1] indicates that older adults tend to prefer an instruction manual to trial-and-error. Nicolau et al. [7] reported detailed analyses on how older adults learn text entry, and found their most common errors were due to cognitive problems in the initial stage. We focused on supports for cognitive problems rather than physical aspects such as mistyping.

2.3 Tutoring System for Older Adults

A wealth of research has largely focused on designing better instructional resources to assist older adults when learning how to use desktop computers. For example, Hickman et al. [19] have studied the type of guidance most suitable for older adults. Morrell et al. [20] have studied what the optimal amount of guidance is. Rogers et al. [21] investigated the kind of resources most useful in the learning process, and found that step-by-step interactive tutorials were the most effective approach in the learning process for older adults. With respect to using smartphones, Leung et al. [1] surveyed and investigated how older adults learned. According to their report, older adults tend to prefer an instruction manual to trial-and-error. Kelleher et al. [22] proposed stencil-based tutorials that overlay step-by-step instructions on the screen. Kristensson et al. [23] hypothesized that a tutoring system may reduce the time and effort when learning a new type of text entry method and proposed five design dimensions: automaticity, error correction, coverage, feedback, and engagement. However, practical tutoring systems for text entry have not been developed and the effectiveness of the tutoring system has not been verified.

Therefore, we considered that providing instructions for the next operation is an effective way for novice older adults to learn text entry on a smartphone.

3. User Study to Clarify How Novice Older Adults Make Input Stumbles

This user study has two purposes. One is to clarify how novice older adults make input stumbles in a smartphone text entry, and analyze the relationship between the input stumbles and the user's skill. Second is to discover and evaluate the necessity of support functions which should be implemented in our proposed system.

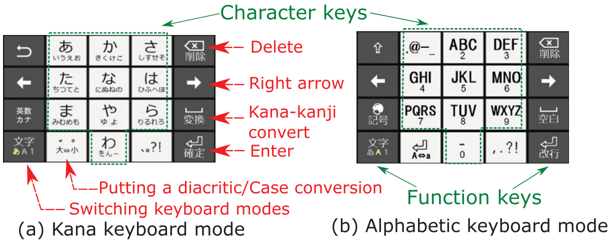
3.1 Participants

Thirty-five participants took part in this experiment. Five were colleagues who had advanced IT skills, four males and one female, ranging in age from 29 to 48 years with a mean age of 36.6 (sd = 6.6). The other thirty were older adults who were recruited from a local social institution. There were fifteen males and fifteen females, ranging in age from 60 to 83 years with a mean age of 72.1 (sd = 8.2). None of them had any previous experience of using a smartphone, but had owned a feature phone for more than one year, i.e. one with a physical keyboard with a 12-key layout. They were familiar with this kind of keyboard because the same layout has been used in feature phones for a long time. Twenty-six of them had often communicated via e-mail, while the others had seldom done so, using the phone only for making phone calls. Twenty-eight had previously entered text with PCs. Ten had used their own PC routinely. None of them had a tremor disorder, eye problems or other relevant health problems.

3.2 Apparatus

The software keyboard of the smartphone had a 12-key multi-tap layout as shown in **Fig. 1**. The smartphone recorded all touch events in an input log using the standard Android API, plus all linguistic information, e.g., typed keys, displayed characters, and suggestions. All participant operations were recorded by an overhead video camera and were displayed on a monitor that a human tutor observed.

All colleagues and the fifteen older adult participants used a Samsung Galaxy S3 running Android 4.1.2 with a 4.8-inch screen having a resolution of $1{,}280 \times 720$ pixels (306 ppi). The other



(a) Kana keyboard mode (b) Alphabetic keyboard mode

Fig. 1 Key layout of 12-key.

Table 1 Japanese syllabic alphabet represented by Roman alphabet.

	A	I	U	E	O
	a (あ)	i (い)	u (う)	e (え)	o (お)
K	ka (か)	ki (き)	ku (く)	ke (け)	ko (こ)
S	sa (さ)	si (し)	su (す)	se (せ)	so (そ)
⋮			⋮		
W	wa (わ)	-	-	-	wo (を)

fifteen used an Asus ZenFone5 whose size and weight are similar to that of a Galaxy S3, running Android 4.4 with a 5.0-inch screen having a resolution of 1,280 × 720 pixels (294 ppi). The difference in the two conditions was due to a delay in the study. However, we assumed there would be little influence on the result because we did not focus on physical factors such as mistypes but on cognitive factors such as the input stumble in a text entry.

Next, we describe the characteristics of the keyboard with a 12-key layout. The keyboard has several points in common with the QWERTY layout, such as function keys (arrow, switching keyboard, and case conversion) and a suggestion list. A character string being edited is not finalized until either a suggestion is selected or an enter key is pressed. A point of difference is that the characters are input by the combination of a key and a number being pressed. For example, the letter "c" is input by typing the "a" key three times. Therefore, when typing two consecutive characters allocated to the same key, the first character must be fixed by pressing the right arrow key before typing the second character.

Next, we describe the characteristics of the Japanese input. Kana characters, which comprise Japanese phonetics, are combinations of a consonant and a vowel as shown in **Table 1**. In this keyboard, a key corresponds to a consonant, and the number of presses corresponds to a vowel. For some characters, the diacritic key needs to be used after inputting kana. For example, the kana of "gu (ぐ)" is input by pressing a diacritic key after the kana for "ku (く)." Although the suggestion list usually shows the auto-complete list predicted by a typed string, the kanji list is shown after typing the kana-kanji convert key.

3.3 Procedure

First, the participants were given explanations and examples of how to operate a smartphone, including touch and swipe operations, by a human tutor. Next, they were instructed on how to use the software keyboard, including the correspondence between each key and a character, and how to select suggestions.

Then, they typed ten sentences in Japanese using an e-mail application. Participants created the sentences as a response to given sentences, such as "You expect to be ten minutes late for a meeting with a friend. Please create a sentence to let him know." They operated the smartphone while holding it in their hand and sitting on a chair. Participants were instructed to type by themselves if at all possible, with no help from the human tutor who sat next to them. However, they were permitted to ask the human tutor only when they lacked the confidence on what to do next. The experimental period was limited to 60 minutes. After the experiment, they took part in an interview.

3.4 Annotation

First, three annotators independently extracted the pattern of input stumbles from the logs and the videos of the study, and based on the discussion classified the input stumbles into 30 classes. The 30 classes of input stumbles were grouped into 10 categories based on similarities. **Table 2** shows the list of classified input stumbles along with the categories. Next, the three annotators labeled input stumbles and possible stumble opportunities according to the logs of each input. The possible stumble opportunity was labeled when an annotator judged that an input stumble is likely to occur based on the time series information before an input, such as the unfixed string and the elapsed time from the previous input. We adopted the labels applied by at least two annotators. The concordance rate using Fleiss' kappa [24] was 0.44. The reason for the low kappa value was that occasionally there were conflicts in determining when an input stumble started even if annotators labeled the same input stumble as belonging to the same operation. The kappa value of seven classes (4-1), (4-2), (5-1), (6-1), (6-2), (6-3), and (10-4) was under 0.20. The kappa value without the seven classes was 0.68.

At the same time, the logs for each sentence were also labeled according to five skill levels. To determine the criteria, the annotators watched a video in which two sentences were being typed and defined the skill level as being from one to five. A sentence of skill level one was typed at an average rate of 4.1 characters per minute (CPM) [25]. On the other hand, that of skill level five was typed at an average rate of around 20.3 CPM, which is comparable to the rate for an average smartphone user with this kind of keyboard. A sentence of skill level two-four was defined between one and five subjectively. The average rate of typing speed was 7.9 CPM for skill level two, 11.8 CPM for skill level three, and 16.4 CPM for skill level four. After watching the video, the annotators placed the sentences into one of five skill levels subjectively from the standpoints of the typing speed, the typing errors and the number of input stumbles. We adopted the labels applied by at least two annotators. The concordance rate by Fleiss' kappa was 0.72.

3.5 Results of the User Study

All colleagues and the twenty-eight older adults completed the task. The average completion time for colleagues and the older adults was 12.4 minutes (sd = 6.0) and 32.4 minutes (sd = 10.9), respectively. The other two participants did not complete the task within 60 minutes. These were the two who had had no previous experience in entering text with their own feature phones.

Table 3 shows the distribution of labeled skill levels for the initial and the final sentences of the participants. Most of the sentences are below level three, whereas the final sentences of

Table 2 Input stumble categories; the asterisk mark (*) shows the particular input stumble in Japanese.

Input stumble category	Input stumble class
(1) Typing characters allocated to the same key consecutively	(1-1) Typing characters allocated to the same key consecutively
(2) Selecting a suggestion	(2-1) Not selecting despite display the desired word (2-2) Not displaying the desired word due to a long input (2-3) Not displaying the desired word because there are too many suggestions (2-4) Not knowing how to manipulate the suggestion list to find the desired word (2-5) Not knowing how to select a word in the suggestion list in kana-kanji convert mode* (2-6) Not displaying the desired word due to a long input in kana-kanji convert mode* (2-7) Not displaying the desired word because there were too many suggestions* (2-8) Not knowing how to switch the modes*
(3) Enter Key	(3-1) Fixing the sentence (3-2) Linefeed
(4) Deleting character	(4-1) Deleting the last character (4-2) Deleting the character before several characters
(5) Switching keyboard modes	(5-1) Switching the keyboard from kana mode to alphabetic or numeric mode (5-2) Switching the keyboard from alphabetic or numeric mode to kana mode
(6) Moving cursor	(6-1) Moving cursor to the previous position (6-2) Moving cursor to the end of sentence (6-3) Moving cursor in order to change target of kana-kanji conversion*
(7) Entering a diacritic	(7-1) Not knowing how to select the diacritic (7-2) Selecting the symbol key by mistake (7-3) Selecting the case conversion by mistake
(8) Case conversion	(8-1) Case conversion of alphabetic character (8-2) Case conversion of kana character*
(9) Entering a symbol	(9-1) Not knowing how to select a period or exclamation mark (9-2) Not knowing how to select a question mark (9-3) Selecting the diacritic key by mistake
(10) Other	(10-1) Touching multiple keys simultaneously (10-2) Failure of key response due to short-duration touch (10-3) Over toggling (10-4) Not knowing how to type emoji (pictorial symbols)

Table 3 The distribution of labeled skill levels for the initial and final sentences of the participants.

Skill level	1	2	3	4	5
Initial sentence	8	11	9	3	4
Final sentence	7	8	11	6	6
All sentence	56	96	93	43	47

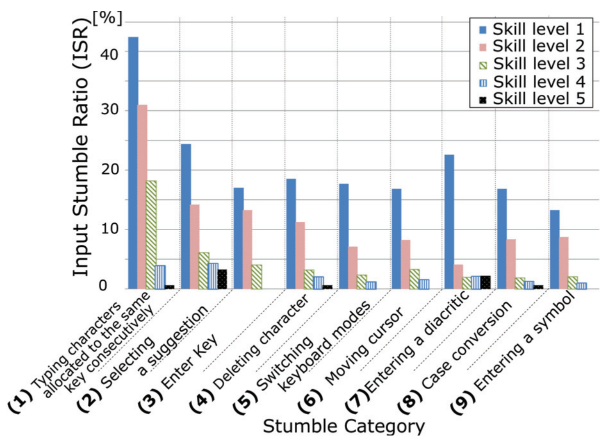


Fig. 2 The relationship between the input stumble ratio (ISR) and the input stumble category for each skill level.

all colleagues are at level five. Comparing the skill level for the initial sentences with that for the final sentences, the skill level generally improved. **Figure 2** shows the relationship between the input stumble ratio (ISR) and the input stumble category for each skill level. The ISR was defined as the percentage of the number of input stumbles in the total of possible stumble opportunities for each category. The category "(10) Other" was excluded because it was difficult to define the possible stumble opportunity uniquely. The overall average of ISR, which is the percentage of the number of input stumbles in the total of possible stumble opportunities of all categories, was 8.3%. A Mann-Whitney U-test ($\alpha = 0.05$) did not show a significant difference between the two different smartphones.

We confirmed that higher-skills had lower ISRs. The category with the highest ISR was "(1) Typing characters allocated to the same key consecutively." This means that when typing two characters allocated to the same key consecutively, the first character must be fixed by pressing the right arrow key before typing the second character. This operation is confusing because no character or string changes when this key is pressed. And it wasn't until the human tutor had repeated instructions several times that many participants realized that they needed to press this key. The category with the second highest ISR was "(2) Selecting a suggestion." This category includes eight classes. These operations need deep knowledge, especially in Japanese because the suggestion list contains the results of auto-complete and kana-kanji conversion. Therefore, this category had the highest ratio among sentences whose skill level was four or five. Many participants found it difficult to judge the timing for selecting the desired suggestion. The ISR of "(3) Enter key" was higher than that of the overall average. The enter key has two functions allocated to it, (3-1) fixing the sentence and (3-2) linefeed. Some participants did not understand the difference so they became confused. On the other hand, the ISR of other operations were lower than that of the overall average because participants had only to check the last character entered and the method of the next operation. Once they had received instructions, most of them were able to carry out these operations.

These results were also supported by comments of the interview after the tasks were completed.

"The tutor's instructions were helpful. But I sometimes couldn't understand why I should do something. If I'm not given the reasons, I will encounter the same problems again."

"We [older adults] can use a smartphone if we get used to it. But it will be tough to keep using it without any support if we only receive instructions from the tutor on one occasion."

In the experiment, we extracted two things that need to be incorporated into the instructions provided through the tutoring system: One is that older adults need instructions to be repeated for more complex operations. The other is that they not only need instructions on how/what to type next but also the reasons and tips on how to perform these operations.

4. System Design of Typing Tutor

In this section, we describe the system design of Typing Tutor based on the previous user study.

4.1 System Design

In the previous user study, we confirmed that the tendencies of an input stumble were quite different according to the skill level. In addition, we confirmed that older adults needed instructions to be repeated for more complex operations, and not only needed instructions on how/what to type next but also the reasons and tips on the operations. Therefore, we designed Typing Tutor based on the following principles: The instructions should be provided based on the skill level and whether the user has previously succeeded or not. Typing Tutor should provide not only instructions on how/what to type next but also provide the reasons and tips on how to perform difficult operations.

Following these principles, we designed Typing Tutor having four functions, as shown in **Fig. 3**: (i) skill classification, (ii) input stumble detection, (iii) success detection and (iv) instruction control. First, Typing Tutor monitors the input data, such as the keys typed and suggestions. Then, the function of skill classification identifies the skill level. At the same time, the functions of input stumble detection and success detection work in concert. These functions detect whether the operation was a stumble or successful. Lastly, the function of instruction control determines whether the instruction should be provided or not, based on information collected through the other three functions. Specifically, Typing Tutor provides instructions even for easy operations when the skill level is low. In addition, no instructions are provided for easy operations that the user has already succeeded in doing without any instruction. This makes it possible to provide only those instructions that are needed by the individual user. We describe the details of each function in this section.

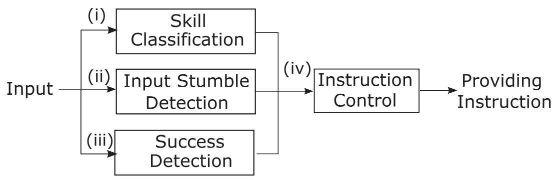


Fig. 3 Block diagram of Typing Tutor.

4.2 Skill Classification

The skill classification model was constructed based on the data identifying the five skill levels. We trained a linear support vector machine (SVM) classifier based on the skill level data. The models are constructed for the number of touches of 10, 20, 50 and 100 from the beginning of typing. The features for the skill classification are nine dimensions as follows: the average and standard deviation of pressing time, those of the time required to type the same/different key, the number of input characters, the number of suggestions used, and the number of times the delete key was used. The cross-validation (CV) for evaluation was conducted 35-fold for each user. As shown by the accuracy rate in **Table 4**, we confirmed that the greater the number of touches, the higher the accuracy rate. The accuracy rate of the 10-touch model was 88.7%. Accuracy within ±1 skill level is greater than 95%, even for the 10-touch model.

4.3 Input Stumble Detection

We implemented statistical models for input stumble detection in Fig. 3 (ii). To select an appropriate learning machine, we compared the performances of four methods, two kinds of SVM [26] (linear, RBF), a logistic regression and a C4.5 [27]. The number of covered stumble classes was 23 after excluding 7 classes for which the kappa value was under 0.2 in the annotation as described in the previous section. Each model was trained by using the data labeled into 24 classes: the 23 stumble classes and no input stumble. Eighty-two features were obtained on the basis of the input sequence which includes unfixed string, fixed string, typing histories, and suggestions. Features for adopting two SVMs and a logistic regression were selected by L1 regularization [26]. Those for a C4.5 were selected by stepwise backward selection of each participant 35-fold cross validation (CV). Each method was evaluated by the F-measure from the 35-fold CV. The F-measure was 0.67 with linear-SVM, 0.70 with RBF-SVM, 0.68 with logistic regression and 0.66 with C4.5. Anova (α = 0.05) showed no significant differences among these models, so we adopted RBF-SVM hereinafter, with the highest accuracy rate. The effective features adopted by RBF-SVM are shown in **Table 5**.

Table 4 Accuracy of skill classification.

Number of touches	Accuracy rate	Accuracy within ±1
10	81.7	96.3
20	82.1	97.5
50	84.4	98.4
100	88.4	98.4

Table 5 Ten effective features to classify input stumbles with RBF-SVM.

ID	The feature
1	The time interval between touches
2	The difference from the average in the time interval
3	The touch holding time
4	The type of current keyboard
5	The number of simultaneous touches
6	The type of typed key
7	The number of consecutive touches of the same key
8	Whether the final character can take diacritic
9	The type of morpheme of the last word
10	The number of words in an unfixed string

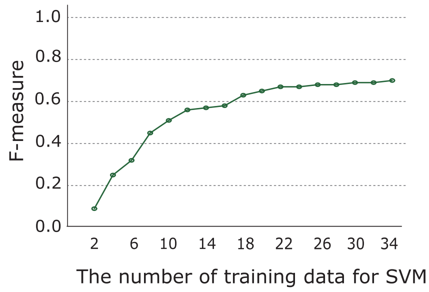


Fig. 4 Relationship between F-measure of RBF-SVM and the number of training data.

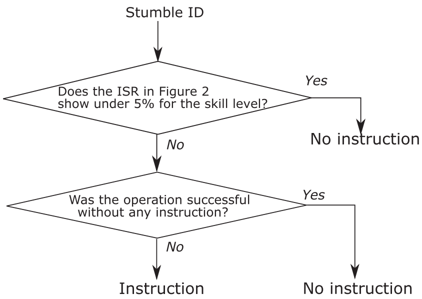


Fig. 5 Flowchart for instruction control.

Figure 4 shows the relationship between the training data per participant and the F-measure. The performance of the model trained from the data of N persons was evaluated by the following procedure. First, the model was trained using the data of N persons, then the F-measure was calculated using the data of the remaining 35-N persons. Next, N variable selection was randomized and calculations were repeated a fewer number of times: all possible combinations of N or 5,000 times. The F-measure of the model trained from the data of N persons was defined as the average value. In Fig. 4, the F-measure increases as N increases. The F-measure exceeds 50% with the training data of 10 persons and becomes saturated with the training data of 20 persons.

4.4 Success Detection

This function detected successful operations after their completion, unlike input stumble detection. This after-the-fact characteristic made detection easier. However, a premature evaluation can lead to errors, so successful operations were registered only if the operation was not canceled before the input of the next two characters.

4.5 Instruction Control

This function determines whether an instruction on how to correct a detected input stumble is provided or not. The flowchart is shown in **Fig. 5**. Even if an input stumble is detected, no instruction is provided when the relevant ISR in Fig. 2 is under 5% for the user's skill level or when the success of the operation without any instruction was detected in the past. Otherwise, the instruction is provided. The instruction corresponding to each type of input stumble was prepared manually, based on the effective advice observed by human tutors in previous studies. Specifically, Typing Tutor provides only operating instructions for categories (4)-(10), but provides both instructions on the reason behind them and the tips for categories (1)-(3). Both text instruction and voice

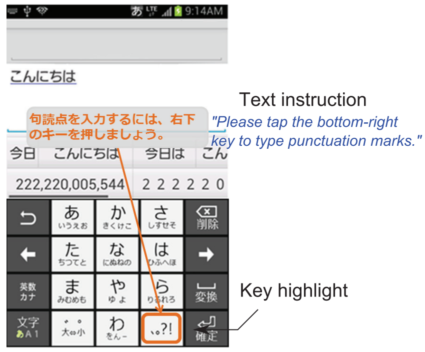


Fig. 6 Screen shot showing instruction provision.

instruction are provided, plus overlaid key highlighting, the visibility of which was improved as shown in **Fig. 6**.

5. Experiment to Improve Typing Tutor

In this section, we describe the preliminary experiment of two trials designed to improve Typing Tutor.

5.1 Description of the Two Trials

5.1.1 Participants

In the first trial, six participants, three males and three females between the ages of 65 and 72 years with a mean age of 68.2 (sd = 2.9), took part. The second trial involved the same number of males and females between the ages of 66 and 74 years with a mean age of 69.4 (sd = 3.2). They were not participants in the other preliminary experiments, and were recruited from a local social institution. None of them had previous experience with a smartphone, but had owned a feature phone for more than one year. They had had previous experience in entering text using a feature phone. Four participants in each trial had used a PC to create documents when employed. Two in the first trial and three in the second trial had routinely used their own PCs. None had a tremor disorder, eye problems or other relevant health problems.

5.1.2 Procedure and Apparatus

The procedure and apparatus were the same in both trials. The tutorial in which the participants were shown how to use a smartphone and a keyboard was the same as that in the previous experiment. After the tutorial, they typed ten sentences as responses to given sentences, using an e-mail application. Participants in the first trial used Typing Tutor, which was described in the previous section. Participants in the second trial used Typing Tutor which was improved after the first trial. All of them made ten sentences as responses to given sentences related to personal conversations, just as in the previous user studies. The experiment period was limited to 60 minutes, with a one-minute rest after each sentence. After the experiment, they filled out a Likert-scale questionnaire and took part in an interview. An Asus ZenFone5 was used in both trials.

5.2 Results of the First Trial

All participants completed the task. The solid bars in **Fig. 7** show the results of the questionnaires administered in the first trial. There was a positive response to the item "Instruction is useful to improve my skill." The instruction method was also

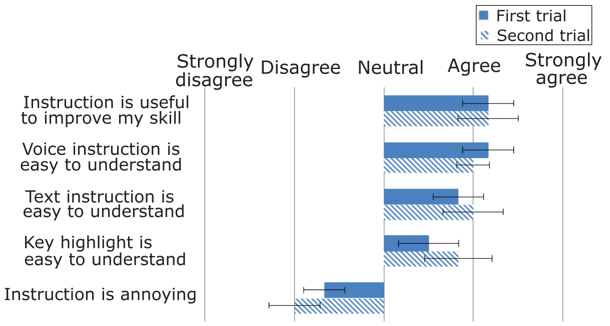


Fig. 7 Questionnaire responses regarding the two trials.

rated positively, with voice instruction the most favored. These results show that Typing Tutor was well accepted. The interviews yielded some important feedback:

"I am not confident that I can repeat an operation that I did today. I would like to receive instructions again."

"I felt some instructions were annoying because I could understand where to type next, needing only the key highlight."

These interviews indicated a need for two improvements: One is to repeat instructions sometimes even though participants have already succeeded in performing an operation. Another is to improve the handling of instructions for easy operations, which can be understood simply from the position and the label.

5.3 Improvement of Typing Tutor

The two points above, which were deemed to require improvement based on the first trial, were dealt with as follows.

5.3.1 Consideration of Forgetting

Typing Tutor in the first trial did not provide further instructions after the participant had succeeded once. To allow the system to forget some successes, we use forgetting curves. This memory retention problem has been examined in many studies [28], [29], [30]. According to the review in Ref. [28], the major mathematical models can be classified into three categories: log-linear model, exponential model, and hyperbolic model. We used the exponential model in Ref. [30] as follows:

$$f = \exp\left(-b_i \sqrt{t}\right) \tag{1}$$

Here f is the retention rate, t is the elapsed time since the last operation, i is the number of times the operation is executed and b_i is the coefficient difference based on i. In the system, the retention rates of each operation are computed, and when the retention rate falls below 50%, the instructions are provided again.

5.3.2 Improvement of Instructions for Easy Operations

Typing Tutor in the first trial provided instructions in a similar way regardless of the degree of difficulty of the operation. Therefore, the operations were divided into two difficulty levels by three annotators. The criterion was whether the operation could be understood simply from an indication of the key position and a word. The input stumbles made when executing the easy operations are the stumble classes (4-1,2), (5-1,2), (7-1,2), (8-1,2) and (9-1,2,3) as shown in Table 2. The instruction for the operations was improved by providing the key highlight and text above the key highlight, as shown in **Fig. 8**.

Fig. 8 Screenshot of instruction for easy operations after improvement.

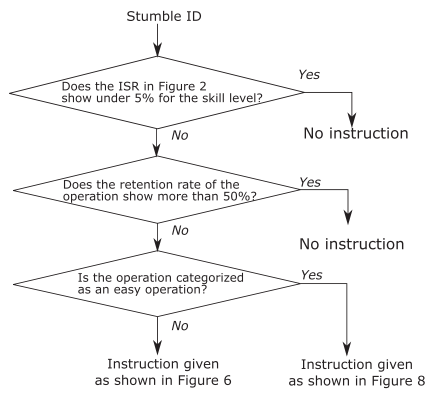


Fig. 9 Flowchart for instruction control improved after the first trial.

5.3.3 Modification of Instruction Control

To take the two improvements into consideration, the function of the instruction control was modified. **Figure 9** shows the flowchart of the instruction control. Even if an input stumble is detected, the instruction is provided only when the relevant ISR in Fig. 2 is not under 5% for the user's skill level and when the retention rate of the operation falls below 50%. Moreover, the way of instruction varies depending on the difficulty of the operation.

5.4 Results of the Second Trial

All participants completed the task. The hatched bars in Fig. 7 show the results of the subjective questionnaires administered in the second trial. Compared to the results of the first trial, there appeared to be an improvement for the items "Key highlight is easy to understand" and "Instruction is annoying," although a Mann-Whitney U test ($\alpha = 0.05$) did not show a significant difference. The half of the participants responded negatively to the voice instruction:

"The voice instruction was easy to understand. But I wouldn't like to use it outside because it embarrasses me."

Therefore, Typing Tutor was altered to allow the choice of using voice instruction or not.

6. Two-week Evaluation Experiment

We developed Typing Tutor, a tutoring system that detects 23 stumbles and provides instructions in two ways according to the difficulty of the operation. Finally, we assessed the performance by means of a two-week evaluation experiment. In this experiment, the participants would chat for one hour with an operator

Table 6 The distributions of classified skill levels on the first day and the final day for each group.

Skill level		1	2	3	4	5	Average skill level
Group A	First day	1	6	2	1	0	2.3
	Final day	0	1	3	4	2	3.7
Group B	First day	1	4	3	2	0	2.6
	Final day	0	1	4	2	3	3.7

using the smartphone's chat application.

6.1 Descriptions of the Evaluation Experiment

6.1.1 Participants

Twenty participants, eleven males and nine females between the ages of 65 and 74 years with a mean age of 68.4 (sd = 4.6) who did not participate in the other experiments took part in this experiment. They were recruited from a local social institution. Ten of the participants were provided instructions by Typing Tutor, the input detection model of which was trained by RBF-SVM with the data of 35 participants (group A). The other ten participants were not provided instructions. However, the participants of group B were permitted to ask for advice on how to operate the smartphones through a chat application, but only if necessary. None of them had previous experience with a smartphone, but all had owned a feature phone for more than one year. All of them had entered text using a feature phone before and had often communicated via e-mail. Seven in group A and seven in group B had routinely used their own PCs. None of them had a tremor disorder, eye problems or other relevant health problems.

6.1.2 Procedure and Apparatus

First, the participants were given explanations and examples of how to operate a smartphone and how to use the software keyboard, just as in previous experiments. After that, they were taught how to use the chat application, "Hangout," which is the default chat application for Android. They kept the smartphone for two weeks (twelve days from Monday to the second Friday) and had a chat with an operator for one hour every day except Saturday and Sunday. The chat topics were not limited and included such subjects as hobbies, their work, and daily events. After two weeks, we held interviews. An Asus ZenFone5 was used in the experiment.

6.2 Experimental Results

All participants sent more than six messages each day. The daily average number of messages and the average length of a message are 8.5 (sd = 1.7) and 43.2 characters (sd = 12.3) for group A, and 8.4 (sd = 1.5) and 42.2 characters (sd = 11.0) for group B. We evaluated the performance by four metrics: the skill level, the number of input stumbles, the typing speed and the typing accuracy. Each metric was tested by a Mann-Whitney U-test ($\alpha = 0.05$) each day.

6.2.1 Classified Skill Level

Table 6 shows the distributions of the classified skill level on the first day and the final day for each group. On the first day, the skill levels were almost the same. On the final day, the skill level of group B was higher than that of group A. However, a U-test did not show a significant difference between the two groups.

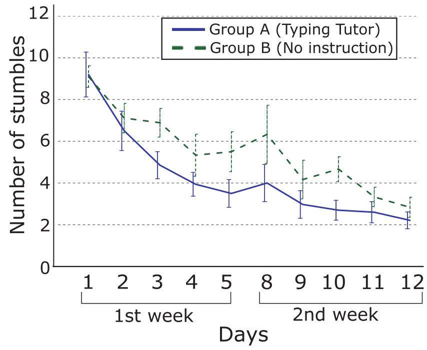


Fig. 10 User progress in the number of input stumbles for each group; vertical bars show standard errors.

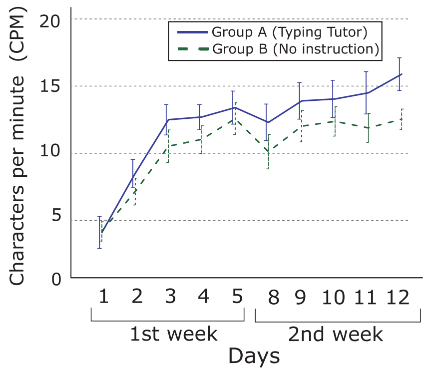


Fig. 11 User progress in the typing speed for each group; vertical bars show standard errors.

6.2.2 Number of Input Stumbles

Figure 10 shows the number of input stumbles in both groups. At the beginning of use, the number of input stumbles was almost the same in all groups. After the second day, there were fewer stumbles in group A than in group B. A U-test showed a significant difference from the third day to the tenth day. The number of input stumbles on the eighth day was higher than on the fifth day because the participants had not used their phones on the previous two days.

6.2.3 Typing Speed

We compared the typing speed in characters per minute (CPM) [25] excluding the time used for instructions. **Figure 11** shows the CPM of both groups over two weeks. The CPMs of both groups were almost the same on the first day. After that, the CPM of group A increased dramatically. In contrast, that of group B did not increase monotonically. This difference was maintained up until the final day. A U-test showed a significant difference except for the first day, the second day, and the fifth day.

Figure 12 shows the time taken by both groups to input character keys and non-character keys. The non-character keys refer to the function keys as shown in Fig. 1 and the suggestion list, which require a longer decision time to select than the character keys. The time taken by the two groups to input character keys decreases gradually from about 1,500 ms to 1,000 ms in a similar manner. However, the time taken by group A to input the non-character keys decreased markedly compared with group B, although times on the first day and the final day were almost identical in both groups. A U-test showed that a significant difference existed for each day from the third day to the eighth day.

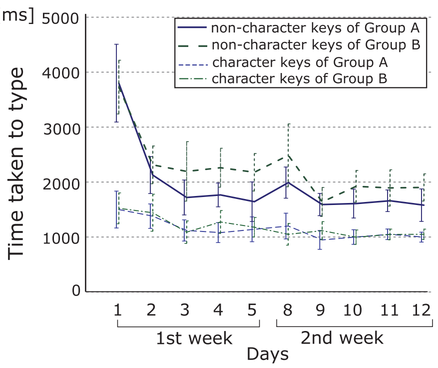


Fig. 12 Time taken to type character keys and non character keys for each group.

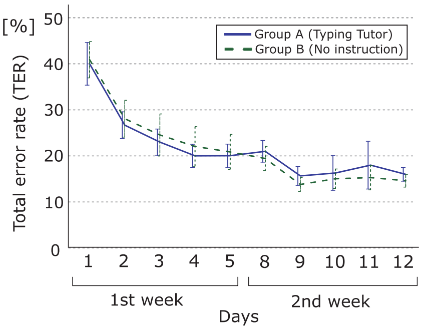


Fig. 13 User progress in total error rate for each group; vertical bars show standard errors.

6.2.4 Typing Accuracy

We compared typing accuracy in terms of total error rate (TER) [25]. Although TER is the metric for typing presented sentences including error correction, sent sentences were used as presented sentences. The difference between the two groups was negligible as shown in **Fig. 13**. A U-test did not show a significant difference between the two groups. The metric is affected by a number of physical factors such as mistyping rather than by cognitive factors. Therefore, this metric is outside the scope of the evaluation of Typing Tutor.

7. Applicability to Other Keyboards and Languages

In the previous section, we confirmed the effectiveness of Typing Tutor. However, the applicability to other keyboards and languages was not taken into consideration. The skill classification is not heavily dependent on keyboards and languages because the features are not language dependent. On the other hand, input stumble detection is influenced by keyboards and languages because some features are language dependent. Therefore, we examine the applicability of the input stumble detection focusing on the QWERTY keyboard layout for Japanese and English.

7.1 Participants

Twenty-four participants took part in this experiment. Twelve of them, six males and six females between the ages of 61 and 77 (mean 68.5, sd = 4.9), typed Japanese with a QWERTY. The other twelve, six males and six females between the ages of 60 and 73 (mean 65.7, sd = 3.8), typed English with a QWERTY. They had used English for work such as composing documents when they were still employed. All of them were recruited from a local social institution. None of them had any previous experience of using a smartphone, but had owned a feature phone for more than one year. All of them had often communicated via e-mail with a feature phone and had previously entered text with PCs. Eight had used their own PC routinely. None of them had a tremor disorder, eye problems or other relevant health problems.

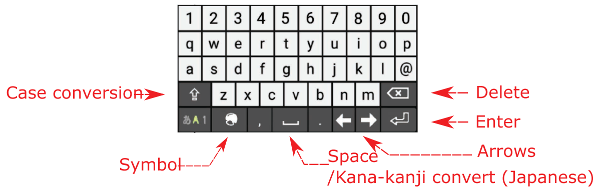


Fig. 14 Key layout of QWERTY.

7.2 Procedure and Apparatus

The experiment was conducted in a similar way to the experiment in Section 3. In brief, first, the participants were given explanations and examples of how to operate a smartphone and how to use the software keyboard. Using a smartphone and while sitting on a chair, they created ten sentences in the form of a response to given sentences. They could rest for one minute after writing each sentence. They were instructed to type by themselves if at all possible. An Asus ZenFone5 was used in the experiment. The software keyboard of the smartphone had a QWERTY layout as shown in **Fig. 14**.

Japanese characters are typed by combinations of a couple of Roman alphabet letters as shown in Table 1. English letters are typed by tap input, which means one touch corresponds to one letter. The smartphone recorded all touch events using the standard Android API and all linguistic information. All participant operations were recorded by an overhead video camera.

7.3 Annotation

All of the older adult participants completed the task. The average completion time was 28.5 minutes (sd = 8.5) with a QWERTY Japanese keyboard and 40.5 minutes (sd = 15.0) with a QWERTY English keyboard.

Similar to the experiment in Section 3, three annotators independently extracted the pattern of input stumbles from the logs and the videos of the study. The input stumbles were classified into 22 classes under 7 categories for the QWERTY Japanese keyboard and 12 classes under 8 categories for the QWERTY English keyboard. **Tables 7** and **8** show the list of classified input stumbles along with the categories. Then, the three annotators labeled input stumbles according to the logs of each input. We adopted the labels that were applied by at least two annotators. The concordance rate using Fleiss' kappa [24] was 0.54 in QWERTY Japanese and 0.42 in QWERTY English. The kappa value of input stumbles (3) and (4) for both keyboards was under 0.20. The kappa value without the input stumbles was 0.70 for QWERTY Japanese and 0.63 for QWERTY English.

Focusing on the differences in the keyboards by comparing the input stumbles for the 12-key layout in Table 2 with the QWERTY in Table 7, there were a lot of common points between the

Table 7 Input stumble category by QWERTY Japanese; the asterisk mark (*) shows the particular input stumble in Japanese.

Input stumble category	Input stumble class
(1) Selecting a suggestion	(1-1) Not selecting despite display the desired word (1-2) Not displaying the desired word due to a long input (1-3) Not displaying the desired word because there are too many suggestions (1-4) Not knowing how to manipulate the suggestion list to find the desired word (1-5) Not knowing how to select a word in the suggestion list in kana-kanji convert mode* (1-6) Not displaying the desired word due to a long input in kana-kanji convert mode* (1-7) Not displaying the desired word because there were too many suggestions* (1-8) Not knowing how to switch the modes*
(2) Enter Key	(2-1) Fixing the sentence (2-2) Linefeed
(3) Deleting character	(3-1) Deleting the last character (3-2) Deleting the character before several characters
(4) Moving cursor	(4-1) Moving cursor to the previous position (4-2) Moving cursor to the end of sentence (4-3) Moving cursor in order to change target of kana-kanji conversion*
(5) Case conversion	(5-1) Case conversion of alphabetic character (5-2) Case conversion of kana character*
(6) Entering a symbol	(6-1) Not knowing how to select a period or exclamation mark (6-2) Not knowing how to select a question mark (6-3) Selecting the diacritic key by mistake
(7) Other	(7-1) Touching multiple keys simultaneously (7-2) Failure of key response due to short-duration touch

Table 8 Input stumble category by QWERTY English.

Input stumble category	Input stumble class
(1) Selecting a suggestion	(1-1) Not selecting despite display the desired word (1-2) Not displaying the desired word due to a long input
(2) Enter Key	(2-1) Linefeed
(3) Deleting character	(3-1) Deleting the last character (3-2) Deleting the character before several characters
(4) Moving cursor	(4-1) Moving cursor to the previous position (4-2) Moving cursor to the end of sentence
(5) Case conversion	(5-1) Case conversion of alphabetic character
(6) Entering a symbol	(6-1) Not knowing how to select a question mark
(7) Inserting a space	(7-1) Not inserting space between words
(8) Other	(8-1) Touching multiple keys simultaneously (8-2) Failure of key response due to short-duration touch

two keyboards. This was because the basic concept of operation was similar due to the same language. The number of stumble classes for the QWERTY layout was less than that of the 12-key layout. This was attributed to the fact that it is simpler to execute operations using the QWERTY layout than it is using the 12-key layout. Next, we focus on the differences in language by comparing input stumbles in Japanese (Table 7) and English (Table 8). Japanese was associated with a number of input stumbles in terms of selecting suggestions because Japanese needs kana-kanji conversion. English had the unique stumble of not inserting a space between words.

7.4 Performance of Input Stumble Detection

A model of input stumble detection for each language was trained by RBF-SVM. The number of stumbles was 17 for QWERTY Japanese and 8 for QWERTY English, after excluding the stumbles for which the kappa value was under 0.2 in the annotation. The number of features obtained from input was 79 for Japanese and 70 for English. The features included unfixed string, fixed string, typing histories, and suggestions. Next, the features were reduced to 28 for Japanese and 26 for English by L1 regularization.

Table 9 shows the performance of each model evaluated by 12-CV. The F-measures were 0.63 for QWERTY Japanese and 0.53 for QWERTY English. Comparing the F-measure of these two keyboards with that of 12-key Japanese trained by 11 persons in Fig. 4, the F-measure of QWERTY Japanese was significantly larger than that of 12-key Japanese by an ANOVA with post hoc Tukey's test ($\alpha = 0.05$).

Table 9 Classification performance of input stumble detection.

Keyboard	F-measure (Precision)
QWERTY Japanese	0.63 (0.88)
QWERTY English	0.53 (0.80)
(12-key Japanese trained by 11 person in Fig. 4)	0.56 (0.87)

8. Discussion

8.1 Two-week Evaluation

The number of input stumbles in group A decreased more rapidly than in group B for the first three days as shown in Fig. 10. In a similar way, the CPM of groups A increased more rapidly as shown in Fig. 11. However, there were no significant differences between the two groups in the first two days. This can be attributed to the fact that the participants of both groups were in the process of learning by trial and error during this time. In contrast, significant differences were seen from the third day to the final day. It is considered that the trial-and-error period continued in group B.

Regarding the time taken to type characters, we observed that a difference existed between the time taken to type character keys and non-character keys, as shown in Fig. 12. The time taken to type non-character keys in group A decreased more rapidly than that in group B as the number of input stumbles decreased. Typing Tutor focused on support for these operations which have a cognitive dimension. These results show that Typing Tutor is an effective way to improve the skills of participants from a cognitive viewpoint. However, in terms of the time taken to type character keys, the differences between the two groups were negligible. This metric is affected by the ability to construct sentences. These results indicate that Typing Tutor did not affect the performance.

In the interview, participants of group A made similar comments:

"Initially, I thought that the smartphone operations were markedly different from those of my feature phone. But as I came to realize the differences and common points between my feature phone and the smartphone through the instructions, I gradually became able to operate the smartphone."

These comments indicate that the Typing Tutor works effectively for novice older adults.

Finally, we focus on the participants for whom not much effect was observed. One participant had a classified skill level of two on the first day and his level remained the same until the final day. His CPM was 3.33 on the first day and 8.15 on the final day, which was the worst result among all participants. During the interview, he made the following comment;

"The touch operation doesn't suit me because a key that I didn't intend respond. I prefer a feature phone with a physical keyboard."

His TER exceeded 32%, which was also the worst rate among all participants. This was caused by physical rather than cognitive factors. Hence, it can be concluded that Typing Tutor was not effective in this case.

8.2 Applicability to Other Keyboards and Languages

First, we discuss the results regarding the different keyboards. Table 9 shows that the accuracy of input stumble detection of the QWERTY for Japanese was higher than that of the 12-key layout for Japanese. This would appear to be because the number of classes for input stumble using the QWERTY is fewer than that for the 12-key layout and the operation of the QWERTY layout is simpler than that of the 12-key layout. For example, in order to input the kana "gu (ぐ)," with a 12-key layout, it is necessary to press the diacritic key after the kana "ku (く)." In contrast, with QWERTY, only "gu" needs to be typed using alphabetical letters.

Next, we discuss why the accuracy of input stumble detection of QWERTY for English was lower than that of the other two keyboards in Table 9. One reason is the difference in the sentence structure between Japanese and English. Some input stumbles concentrate around the end of a sentence. The order of Japanese is subject-object-verb and Japanese does not have that many expressions that appear at the end of a sentence. This makes it easy to predict the end of a sentence. By contrast, the order of English is subject-verb-object and the object can be post-modified. Therefore it can be difficult to predict the end of a sentence without information on the syntactic dependency or grammar, e.g., whether the morpheme at the end of a sentence is grammatically correct or not, which was not incorporated as an SVM feature in this study. The other reason is that the lower concordance rate of Fleiss' kappa was responsible for the lower accuracy of stumble detection. Although the annotation was controlled based on the discussion in advance, inconsistencies arose because the three annotators were non-native English speaker. On the other hand, as shown in Fig. 4, the F-measure improved as the training data increased in the 12-key layout for Japanese. This means that the F-measure of other keyboards and languages may have the same tendency. These indicate that Typing Tutor for other keyboards and languages could be also effective.

9. Limitations and Future Work

In terms of the detection of input stumbles, the F-measure was around 0.7 in this study. It should be possible to improve the precision, but it is difficult to detect input stumbles with a perfect accuracy. Therefore, providing instructions in a way that does not interrupt the user's actions is essential. In addition, there is a possibility that the acceptability of the way instructions are given depends on the cultural background of the user. Thus, we need to research the acceptability of Typing Tutor for people in other cultures.

Finally, we would like to refer to support for physical factors, e.g., the mistyping problem. Typing Tutor does not focus on this problem, even though this is also a serious problem for older adults [7]. In the field of text entry, many studies have focused on making it easier to type text by adjusting key target areas or presenting suitable suggestions [4], [5], [6]. It should be possible to support both cognitive and physical factors by combining Typing Tutor with these techniques. This is the issue that we will address in the future.

10. Conclusion

In this paper, we have presented Typing Tutor, a tutoring system for text entry that automatically detects input stumbles and provides individualized instructions for older adults. Through a user study, we clarified the common difficulties that older adults experience and how a user's skill level is related to input stumbles using a 12-key layout for Japanese. Based on the study, we developed Typing Tutor. A two-week evaluation experiment with novice older adults (65+) showed that Typing Tutor was an effective way to improve their proficiency in the area of text entry, especially in the initial stage of use compared with instructions through a chat application online. It was particularly beneficial for users who had previously used feature phones but had trouble with self-instruction. In addition, we evaluated the applicability of statistical input stumble detection to other keyboards and languages. Based on the result, we were able to confirm that the performance of the input stumble detection of QWERTY layout for Japanese and English was almost same as that of a 12-key layout for Japanese.

References

[1] Leung, R., Tang, C., Haddad, S., Mcgrenere, J., Graf, P. and Ingriany, V.: How older adults learn to use mobile devices: Survey and field investigations, *ACM Trans. Accessible Computing*, Vol.4, No.3, pp.11:1–11:33 (2012).

[2] Beisge, B. and Kraitchman, M.: *Senior Centers*: *Opportunities For Successful Aging*, Springer Publishing Company (2003).

[3] Hoggan, E., Brewster, S.A. and Johnston, J.: Investigating the Effectiveness of Tactile Feedback for Mobile Touchscreens, *Proc. SIGCHI Conference on Human Factors in Computing Systems*, *CHI '08*, pp.1573–1582, ACM (2008).

[4] Bi, X., Ouyang, T. and Zhai, S.: Both Complete and Correct? Multi-Objective Optimization of Touchscreen Keyboard, *Proc. SIGCHI Conference on Human Factors in Computing Systems*, *CHI '14*, pp.2297–2306, ACM (2014).

[5] Gunawardana, A., Paek, T. and Meek, C.: Usability Guided Key-Target Resizing for Soft Keyboards, *Proc. 15th International Conference on Intelligent User Interfaces*, *IUI '10*, pp.111–118, ACM (2010).

[6] Rodrigues, É., Carreira, M. and Goncalves, D.: Enhancing typing performance of older adults on tablets, *Universal Access in the Information Society*, pp.1–26 (2014).

[7] Nicolau, H. and Jorge, J.: Elderly Text-entry Performance on Touchscreens, *Proc. 14th International ACM SIGACCESS Conference on Computers and Accessibility*, *ACCESS '12*, pp.127–134, ACM (2012).

[8] Hagiya, T., Horiuchi, T., Yazaki, T., Kato, T. and Kawahara, T.: Assistive Typing Application for Older Adults Based on Input Stumble Detection, *J. Inf. Process.*, Vol.25, pp.417–425 (2017).

[9] Hagiya, T., Horiuchi, T. and Yazaki, T.: Typing Tutor: Individualized Tutoring in Text Entry for Older Adults Based on Input Stumble Detection, *Proc. SIGCHI Conference on Human Factors in Computing Systems*, *CHI '16*, pp.733–744, ACM (2016).

[10] Rama, M.D., Ridder, H. and Bouma, H.: Technology generation and Age in using layered user interfaces, *Journals of Gerontechnology*, Vol.1, No.1, pp.25–40 (2001).

[11] Fisk, A.D., Rogers, W.A., Charness, N., Czaja, S.J. and Sharit, J.: Designing for Older Adults: Principles and Creative Human Factors Approaches (2009).

[12] Jin, Z.X., Plocher, T. and Kiff, L.: Touch screen user interfaces for older adults: button size and spacing, *Universal Access in HCI, Part I, HCII2007, LNCS4554*, pp.933–941 (2007).

[13] Gregor, P., Newell, A.F. and Zajicek, M.: Designing for dynamic diversity: Interfaces for older people, *Proc. 5th International ACM Conference on Assistive Technologies*, pp.151–156, ACM (2002).

[14] Sloan, D., Atkinson, M.T., Machin, C. and Li, Y.: The Potential of Adaptive Interfaces as an Accessibility Aid for Older Web Users, *Proc. 2010 International Cross Disciplinary Conference on Web Accessibility 35*, ACM (2010).

[15] Wobbrockand, J.O., Kane, S.K., Gajos, K.Z., Harada, S. and Froehlich, J.: Ability-Based Design: Concept, Principles and Examples, *ACM Trans. Accessible Computing*, Vol.3, No.3, Issue 9 (2011).

[16] Edlin-White, R., Cobb, S., D'Cruz, M., Floyde, A., Lewthwaite, S. and Riedel, J.: Accessibility for Older Users through Adaptive Interfaces: Opportunities, Challenges and Achievements, *Human-Computer Interaction, Part III, HCII2011, LNCS6763*, pp.483–489 (2011).

[17] Kurniawan, S.: Older people and mobile phones: A multimethod investigation, *International Journal of Human Computer Studies*, Vol.66, No.12, pp.889–901 (2008).

[18] Komninos, A., Nicol, E. and Dunlop, M.D.: Designed with Older Adults to Support Better Error Correction in SmartPhone Text Entry: The MaxieKeyboard, *Proc. 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, *MobileHCI '15*, pp.797–802, ACM (2015).

[19] Hickman, J.M., Rogers, W.A. and Fisk, A.D.: Training older adults to use new technology, *Journals of Gerontology: Series B*, Vol.62 (2007).

[20] Morrell, R.W., Park, D.C., Mayhorn, C.B. and Kelley, C.L.: Effects of age and instructions on teaching older adults to use eldercomm, an electronic bulletin board system, *Educational Gerontology*, Vol.26, No.3, pp.221–235 (2000).

[21] Rogers, W.A., Cabrera, E.F., Walker, N., Gilbert, D.K. and Fisk, A.D.: A survey of automatic teller machine usage across the adult lifespan, *Human Factors*, Vol.38, No.1, pp.156–166 (1996).

[22] Kelleher, C. and Pausch, R.: Stencils-based Tutorials: Design and Evaluation, *Proc. SIGCHI Conference on Human Factors in Computing Systems*, *CHI '05*, pp.541–550, ACM (2005).

[23] Kristensson, P.O.: Analysis of Text Entry Performance Metrics, *Proc. 15th International Conference on Artificial Intelligence in Education*, *AIED 2011*, *Lecture Notes in Computer Science*, Vol.6738, pp.494–496, Springer (2011).

[24] Fleiss, J.L.: Measuring nominal scale agreement among many raters, *Psychological Bulletin*, Vol.76, No.5, pp.378–382 (1971).

[25] Arif, A.S. and Stuerzlinger, W.: Analysis of Text Entry Performance Metrics, *Proc. International Conference on Science and Technology for Humanity*, *TIC-STH '09*, pp.100–105, IEEE (2009).

[26] Bishop, C.: *Pattern Recognition and Machine Learning*, Springer (2007).

[27] Quinlan, J.R.: *Programs for Machine Learning*, Morgan Kaufmann Publisher (1993).

[28] Anzanello, M.J. and Fogliatto, F.S.: Learning curve models and applications: Literature review and research directions, *International Journal of Industrial Ergonomics*, Vol.41, No.5, pp.573–583 (2011).

[29] Ebbinghaus, H.: *Memory; A Contribution to Experimental Psychology*, New York by Teachers College (1885).

[30] Wixted, J.T.: Analyzing the Empirical Course of Forgetting, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol.16, No.5, pp.927–935 (1990).

Toshiyuki Hagiya received his B.E. in 2008, M.E. in 2010, and Ph.D. in 2018 from Kyoto University. He is currently with KDDI Research, Inc. He has been engaged in research and development of spoken dialogue systems and interactive user interface. He is a member of IPSJ, ACM.

Toshiharu Horiuchi received his Bachelor's, Master's and Ph.D. degrees from Nagaoka University of Technology in 1999, 2001 and 2004. Since 2009, having joined the KDDI Corporation, he has worked as a R&D manager at KDDI Research, Inc. His research interests include acoustic signal processing and adaptive signal processing, and spatial signal processing. He received the 17th Telecommunications Advancement Foundation Award in 2002, the 20th Image Information and Media Future Award in 2012, and the Best Technical Demonstration Runner-up at ACM Multimedia 2012. He is a member of ASJ, IEEE, and IEICE.

Tomonori Yazaki is a senior manager in User Interface Laboratory at KDDI Research, Inc. He has been engaged in the research and development of Human Computer Interaction system/application. He received his B.E. and M.E. degrees from The University of Electro-Communications in 1996 and 1998, respectively. He is a member of IEICE, JSSD, and JSKE.

Tatsuya Kawahara received his B.E. in 1987, M.E. in 1989, and Ph.D. in 1995, all in information science, from Kyoto University, Kyoto, Japan. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a Professor in the School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR and NICT. He has published more than 300 technical papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been conducting several speech-related projects in Japan including speech recognition software Julius and the automatic transcription system for the Japanese Parliament (Diet). He is VP-Publications of APSIPA, a board member of ISCA, and a Fellow of IEEE.