A Comparative Usability Study of Two Japanese Gaze Typing Systems

Kenji Itoh[*] Hirotaka Aoki[†] John Paulin Hansen[‡]

Tokyo Institute of Technology Tokyo Institute of Technology IT-University of Copenhagen

Abstract

The complex interplay between gaze tracker accuracy and interface design is the focus of this paper. Two slightly different variants of *GazeTalk*, a hierarchical typing interface, were contrasted with a novel interface, *Dasher*, in which text entry is done by continuous navigation. All of the interfaces were tested with a good and a deliberate bad calibration of the tracker. The purpose was to investigate, if performance indices normally used for evaluation of typing systems, such as characters per minute (CPM) and error-rate, could differentiate between the conditions, and thus guide an iterative system development of both trackers and interfaces. Gaze typing with one version of the static, hierarchical menu systems was slightly faster than the others. Error measures, in terms of rate of backspacing, were also significantly different for the systems, while the deliberate bad tracker calibrations did not have any measurable effect. Learning effects were evident under all conditions. Power-law-of-practice learning models suggested that Dasher might be more efficient than GazeTalk in the long run.

CR Categories: H.5.2 [Information Interfaces and Presentation]: User Interface – Ergonomics; Interaction Styles; D.2.8 [Software Engineering]: Metrics – Performance Measures; H.1.2 [Models and Principles]: User/Machine Systems – Human Factors

Keywords: Gaze interaction, usability, Japanese text typing, assistive technology, alternative communication.

1 Introduction

Gaze interaction is a promising assistive technology that may contribute to enhancing the quality of life of severely disabled people, such as patients with Amyotrophic Lateral Sclerosis (ALS) who have lost their ability to control muscle movement. Disabled individuals can communicate with their family and friends, and input characters and commands into a computer by simply gazing at a particular part of the screen. Gaze interaction may also be useful as an input device for other people as well, such as when hands are used for another task performed concurrently. Moreover, there may be some functions for which gaze interaction is in fact faster than conventional interfaces that rely on hands or fingers [e.g., Sibert & Jacob 2000].

Usability is of great importance for a novel interface paradigm such as gaze interaction to be accepted. There are several ways to improve the usability of gaze-based communication systems, for instance designing keys with audio and visual feedback [e.g., Majaranta et al. 2003a; 2003b], using character or word prediction by a language model [e.g., Ward & MacKay 2002], using task models to interpret inaccurate input [e.g., Salvucci & Anderson 2000], or to eliminate the dwell time for each key selection [e.g., Salvucci 1999].

Usability is not a single, one-dimensional property of an interface. Nielsen [1993] suggested it to be associated with the following five attributes: learnability, efficiency, memorability, error, and satisfaction. Among these attributes, efficiency in terms of typing speed – which is referred to as productivity in use after learning the system – has been of particular interest among developers of gaze-based interfaces for the last two decades. But how long time should one actually expect that it might take for a novice to master a gaze communication system – and does the time it takes to reach a stable performance level depend on the interface of the system?

Most conventional gaze communication systems have combined an eye-tracking device with standard on-screen keyboards. However, on-screen keyboards that are designed specifically for eye typing have recently been introduced as a new trend in gaze interaction [e.g., Hansen et al. 2001]. We have been developing "*GazeTalk*" as a gaze-based communication tool, currently for English, Danish, and Japanese [Hansen et al. 2001; 2003]. To make the tool widely available, we plan to use standard consumer cameras to track gaze positions [Hansen & Hansen 2006]. The relatively low resolution of this camera technology requires large on-screen buttons (keys), and therefore only 12 keys can be reliably selected on a standard 15-inch monitor, as will be described in detail in Section 2.2. For Japanese text entry, which we focus on in this study, Hansen et al. [2003] reported a performance of approximately 16 CPM with GazeTalk when using only Hiragana and Katakana, which is equivalent to approximately 8 words per minute (WPM; 1 word = 5 characters, including space) for European languages such as English. *Dasher* is another example of a dedicated gaze typing system. It provides a completely different way of using gaze direction from dwelling and scanning by writing via navigation through a zooming world. After two hours' practice, productivity of 25 WPM has been reported, and the frequency of spelling errors is considered to be negligibly small [Ward & MacKay 2002]. The present study will compare the Dasher approach to a traditional dwell selection of buttons in a hierarchical menu structure that GazeTalk (and many other on-screen keyboards) make use of.

The error rate of gaze typing is often reported to be higher than that of other input modalities [e.g., Ohno 1998; Hansen et al. 2003]. Istance et al. [1996] reported that subjects spent a lot of time for correcting entry errors (e.g., typing the same character twice), and therefore they ended up producing as little as 1 WPM. Such a high error rate with gaze typing is partly caused by the frequently cited obstacle known as the "Midas Touch Problem" [Jacob 1991], describing an unintended gaze activation of a key that may happen just by looking around on a gaze-responsive

[*] e-mail: kenji.itoh@ie.me.titech.ac.jp
[†] e-mail: aoki@ie.me.titech.ac.jp
[‡] e-mail: paulin@itu.dk

Copyright © 2006 by the Association for Computing Machinery, Inc.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions Dept, ACM Inc., fax +1 (212) 869-0481 or e-mail permissions@acm.org.

ETRA 2006, San Diego, California, 27–29 March 2006.
© 2006 ACM 1-59593-305-0/06/0003 $5.00

interface. In a previous study [Aoki et al. 2005], the mean keystroke error rate specific to the Midas Touch Problem was found to be as low as 0.16, after a short practice of 180 character entry.

Efficiency and error are not the only objectives to consider when designing a gaze-based interaction system. Learnability – a metric for ease of learning – is also a critical component of usability. Satisfaction with the system is yet another important attribute for a novel interface. Hansen et al. [2001] listed some of the requirements that a gaze-based system should have in order to be satisfying: the system should be easy to install, maintain, and update; calibrations should be performed easily and quickly; the tracking equipment should not make the user look awkward; and prolonged use should not cause fatigue or cause the eyes to dry out.

The present study evaluates two gaze-based Japanese text-typing systems, each of which applies a different principle of interface design, with various metrics for usability. We conducted an experiment in which subjects performed seven 300-character typing blocks with each system. Several performance indices, including typing speed and error-related frequencies such as over-production rate, and rate of backspacing were calculated from the experimental data. Of the above-mentioned usability attributes, learnability, efficiency, and error were derived from these performance indices for each interface. Subjective satisfaction was also analysed using questionnaire responses collected during the experiment. From the comparative results of usability between the different interfaces examined in this study, we discuss the effects of some design components of gaze-based systems on their usability.

2 Japanese Text Typing by Gaze Interaction

2.1 Text entry in Japanese

Japanese text is primarily produced with three systems of Japanese-specific characters, Hiragana, Katakana, and Kanji (Chinese characters), as well as alphanumeric letters and symbols. The former two character systems are phonetic alphabets, and each system comprises 50 basic characters, plus some additional characters. In general, Katakana characters are used only for writing words of non-Japanese origin such as names and cities in other countries, or modern words developed in western countries, such as computer, interface, engine, system, and human-computer interaction. All Hiragana or Katakana characters are divided into ten groups of consonants – null, “k”, “s”, “t”, “n”, “h”, “m”, “y”, “r”, and “w” – and combine each with five vowels – “a”, “i”, “u”, “e”, and “o”. For instance, the group (line) of “k” includes five phonetic characters, “ka”, “ki”, “ku”, “ke”, and “ko” in Hiragana or Katakana. As such, each Hiragana or Katakana character has a Romanization.

Written text is a combination of Hiragana, Katakana, Kanji, and alphanumeric letters. To write a Japanese sentence in a computer-based text-entry system, we first input Hiragana characters, and then convert them into the corresponding representation of mixed Kanji-Hiragana characters using a Kana-Kanji conversion programme. Thus, the characters included in the mixed Kanji-Hiragana sentence may be slightly different from those made by other individuals even if it is converted from the same Hiragana characters, depending, for instance, on the education level and age of the user (results from adults will differ from those of elementary school children, for instance). In addition, it is important to notice that written sentences that are completely correct in terms of syntax and semantics must be represented with the correct combination of the three character sets, as lack of correctness makes the writing incomprehensible.

2.2 GazeTalk: A hierarchical, static menu system

The original version of GazeTalk [Hansen et al. 2001; 2002] was developed for Danish and English language users. It was equipped with a character prediction function applying a Markov Chain Model, which predicts the six most-likely letters subsequent to the last typed character. The present GazeTalk for Danish and English users is equipped with a word-prediction function in addition to the character prediction. Unlike these language versions, the Japanese GazeTalk furnishes neither character nor word prediction, and consequently no language model is required as this function currently does not seem to work well in Japanese due to very few occasions where the next letter can be predicted by a small number of candidate characters with high probability, since only six buttons are available for character input in each menu of GazeTalk. It is a future challenge to integrate a language model with the Japanese version of GazeTalk.

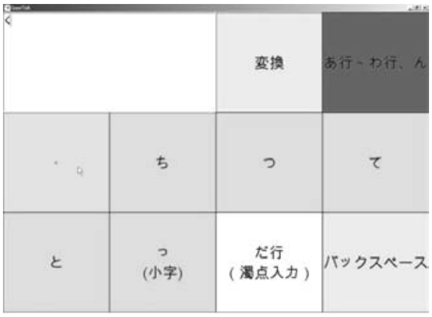

Figure 1: Layout of a menu in the Standard version of the Japanese GazeTalk (character-level menu).

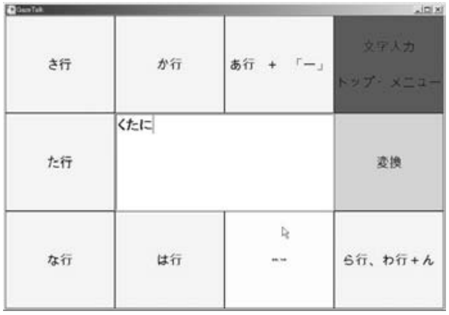

Figure 2: Layout of a menu in the Centre-text version of Japanese GazeTalk (Kana top menu).

Illustrated in Figure 1, each menu in the Japanese version consists of ten on-screen keys (five to eight keys for character input, plus additional keys for Kana-Kanji conversion and backspacing) as well as a text field. A user can activate (or “push”) each of these on-screen keys by gazing at it for a specified “dwell time”. The dwell time is default set at 500 msec, but it can be made shorter or longer depending on the user's preference. As feedback to the user, the present version of the GazeTalk features typeface that changes in size as the user gazes at it (see Figs. 1 and 2). The typeface begins to get smaller when the user first shifts his/her eyes to a key, and it continues to become smaller until the user looks away. The size of the typeface thereby represents the time remaining to activation. If the user allows the typeface to disappear on the key altogether, that key is activated.

Taking into account the characteristics of the Japanese language mentioned previously, we decided to adopt a static, hierarchical menu structure for the typing interface of Japanese GazeTalk. In this study, we implemented two versions for Japanese users to examine the effects of the text field position on gaze-typing usability. One version is the Standard version of (Japanese) GazeTalk (abbreviated S-GazeTalk in this study), in which the position of the text field is the same as in the original Danish/English GazeTalk, that is, in the upper-left corner, occupying a space of two keys (cf. Fig. 1). As its derivative, the Centre-text version of GazeTalk (C-GazeTalk for short), the text field was relocated to the middle of the display so that a user could easily and comfortably check the input of text during typing (cf. Fig. 2).

In the design discussions, we were uncertain if the version shown in Figure 1 allows for a better use of the para-foveal vision, helping the user gain a fast overview of the ever-changing layout, while the centre-text version shown in Figure 2 has a more balanced layout, that some of the system designers preferred for its aesthetic qualities. However, we were unsure if this new layout would be less efficient than the standard layout, and therefore we decided to test, if this could be the case. In a more general sense, the test would also help us develop methods and metrics by which future design discussions, often regarding details of layout or system feedback, could become more qualified.

With the hierarchical-menu structure of the Japanese GazeTalk text is typed as follows. At the Kana top menu (cf. Fig. 2 for the Centre-text version), an entry for each of the Hiragana groups (see Section 2.1) is allocated to each key. When the user activates one of these – say, the key left of the text field in Figure 2 –, then the next-level menu appears, in which five Hiragana characters in the group are included – "ta", "ti", "tu", "te", and "to" in the example case (as can be seen in Fig. 1 for the Standard version). Subsequently, in this "character-level" menu, a Hiragana character can be typed by fixating upon the key one wants to input. Thus, each of most Hiragana characters can be typed by two "gaze" clicks, one in the Kana top menu and the other in the character-level menu. As mentioned previously, Japanese GazeTalk also couples with the Kana-Kanji conversion function to produce a mixed Kanji-Hiragana text. Each menu includes one or more keys relevant to this function. For instance, a key for initiating the conversion is allocated in the right of the text field in both Figures 1 (S-GazeTalk) and 2 (C-GazeTalk).

2.3 Dasher: Text-entry system using continuous navigation

Dasher [Ward 2001; Ward & MacKay 2002] is a text-entry system that has a novel interface incorporating a language model. Text typing is driven by continuous two-dimensional search and navigation with a device such as a mouse, touch-screen, rollerball, breathing device, or eye-tracker. The language model has been trained on example documents, or training corpus, which allows Dasher to predict the probability of each character's occurrence in a given context. The size of the space is allocated for each letter and successive characters according to the predicted probability. A screen shot of Dasher is shown in Figure 3. Ward and MacKay [2002] reported that, when using a mouse as the steering device with Dasher, novice users can be trained to type at more than 25 WPM after one hour of practice, and experts can type at 34 WPM.

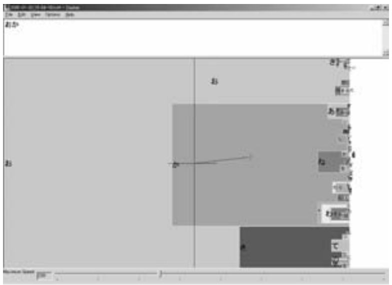

Figure 3: Screen configuration of Dasher.

Dasher works in most languages, and has several derivatives for the Japanese language. Among them, we chose the "Hiragana 60" version of Dasher (version 3.2.11) – which has a set of 60 Hiragana characters, plus numerals and special Japanese symbols – and used it with a 380,000-Hiragana-character training corpus of "everyday phrases" in this study. We did not apply the self-correction option that Dasher has for inaccurate input in this experiment.

3 Experiment

3.1 Subjects

Fifteen Japanese students participated in the experiment. Their mean age was 21 years (range, 19–25 years old). Subjects were divided into three groups, each of which performed experimental trials with only one of the typing systems mentioned in the last section, either S-GazeTalk, C-GazeTalk, or Dasher. For a homogeneous allocation of subjects – in terms of their initial gaze typing skills – to three typing conditions, we conducted a preliminary test with another Japanese gaze-typing system, Hearty Ladder, which was combined with a conventional on-screen keyboard of Hiragana characters, to estimate subjects' initial gaze-typing skills. Each subject was paid 750 Japanese yen (JPY, approximately 6 $) per hour for participation in the experiment. To maintain a high motivation for participation in the four-day typing trials, we informed subjects before starting the experiment that the top two typists for each typing-system group (measured in terms of both typing speed and accuracy) would receive an extra prize: 5,000 JPY (approximately 40 $) to the best typist, and 3,000 JPY (approximately 25 $) to the second-best.

3.2 Task

The experimental task was to perform gaze typing of Japanese phrases and short sentences of daily conversation only in Hiragana characters, resembling ALS patients' use of the system. Previously, we had examined S-GazeTalk when typing Kana-Kanji mixed text [Aoki et al. 2003; 2005]. The task design was based on the fact that daily communication of this kind need not use Kanji characters. Examples of typed sentences (in translation) were: "Give me water", "Turn on the TV set", and "What time is it?" Each sentence was made up of 8–23 characters (15 on average), including Hiragana characters and numbers (in numeric form). The task was performed block by block. Each block comprised 20 sentences, and it contained approximately 300 characters. Subjects were instructed to type these sentences as quickly and accurately as possible. For the first ten sentences of each block, the eye tracker was calibrated using the "standard" calibration procedure to obtain good accuracy of eye tracking, and the other ten sentences were typed with a "deliberate miscalibration" to emulate a low-resolution eye tracker, as will be described in Section 3.4.

The test supervisor would read each sentence aloud for the participants. A longer sentence would be divided into several parts. Subjects were instructed to request repetitions of reading the sentence as necessary. We also instructed them that they need not correct typing errors discovered later at the time of text verification.

3.3 Apparatus

The gaze-typing system was run on a personal computer (CPU: 933 MHz) operated by Windows 2000 including the IME Kana-Kanji conversion programme, with a 17-inch colour monitor (1024 x 768 pixels). The viewing distance from the subject to the screen was 70 cm. The dwell time for key activation in both GazeTalk systems initially was set at 500 msec, which was identical to the one in our previous studies [Aoki et al. 2003; Hansen et al. 2003], and the subject could change the dwell time according to his/her preference in any experimental block. Three out of ten subjects made the dwell time slower, i.e., 750 msec in some early blocks, and two tuned it faster, i.e., 400 msec in later blocks. We initially set a speed parameter for Dasher (maximum speed) at 1.5, and the subject could also change that at his/her disposal. Two out of five subjects changed the parameter slower, i.e., 0.9–1.2 in early blocks and two made it slightly faster, i.e., 1.8, in the later stage. A QuickGlance system (EyeTech Digital System, version 3.1) was used as an eye-tracking device with tuning 15 for the update rate and 7 for the smoothing factor.

3.4 Procedure

In this study, we focused specifically on the following three experimental factors: (1) typing systems (between-subject factor), (2) accuracy of eye-tracking systems (within-subject factor), and (3) learning effect (within-subject factor). For the first factor, we used the three gaze-typing systems mentioned in Section 2, S-GazeTalk, C-GazeTalk, and Dasher. The second factor was controlled by arranging a calibration procedure of the eye tracker. High eye-tracking accuracy was produced by calibrating the tracker using the standard QuickGlance procedure. Deliberately low accuracy was achieved by applying a miscalibration procedure in which each fixation point to be calibrated was intentionally distorted slightly from a real calibration target, specifically, with a 2-degree error from the target in one of four randomly selected directions: up, down, right, or left. Table 1 indicates actual errors of calibration at the 16 QuickGlance calibration points in terms of mean, standard deviation, minimum, and maximum values, calculated from the data of 15 subjects in seven experimental blocks. The deliberate miscalibration produced a mean error of 1.07–1.50 degrees (approximately 0.5–0.6 degree standard deviation) from the actual fixation point for each subject.

The learning effect was examined in terms of transition during seven entire experimental blocks, or in terms of the differences noted between several earlier blocks, for performance indices described below. The experiment involved each subject for four days, requiring at most one and a half hours of the subject's time per day. Approximately one week prior to the experimental session, the preliminary test was conducted for the purpose of allocating subjects to typing systems, as mentioned in Section 3.1.

Table 1: Actual error of eye tracking with deliberate mis-calibration at each point (4x4 grid)in a display (in degrees).

μ=1.25	1.07	1.07	1.10
σ=0.51	0.45	0.49	0.54
min=0.30	0.15	0.19	0.21
max=2.82	2.64	2.45	2.58
1.31	1.45	1.39	1.31
0.57	0.54	0.54	0.54
0.00	0.16	0.37	0.27
2.81	2.78	2.68	2.65
1.12	1.38	1.49	1.35
0.61	0.60	0.60	0.49
0.23	0.00	0.00	0.29
2.67	2.69	2.60	2.75
1.35	1.50	1.34	1.32
0.50	0.52	0.51	0.52
0.22	0.29	0.00	0.27
2.85	2.75	2.66	2.70

For aggregated values of 16 calibration points:
μ=1.30,σ=0.55, max=2.85, min=0.00.
Each cell corresponds to a calibration point on the display.

On Day 1, before the experimental session, a subject performed a training session with a typing system that he/she would use in the experiment for approximately 10 minutes. Then he/she performed a one-block experimental session, which included gaze typing ten sentences with high accuracy of eye tracking and ten other sentences with low accuracy, as mentioned above. A short break (approximately five minutes) separated the high-accuracy and low-accuracy sessions. At the end of Day 1, subjects filled in a questionnaire requesting their subjective opinions about the system they used in the experiment.

On Days 2–4, subjects first received a five-minute warm-up trial, and subsequently they performed two blocks of the typing task. On Day 4, they were asked to fill in the same questionnaire as the one used in Day 1.

3.5 Analysed measures

This paper focuses on four of the usability attributes suggested by Nielsen [1993]: learnability, efficiency, error, and satisfaction (memorability not included in the scope of this study). These four attributes and their inter-associations are examined for each gaze-typing interface in terms of usability-related indices derived from data concerning typing speed, errors, and subjectively stated attitudes.

As mentioned previously, typing speed is typically measured in terms of words or characters per minute. Text written in Japanese is best measured in CPM, whereas WPM is applied to European languages. It is difficult to set an exact conversion factor between WPM and CPM, but from our previous experiments with identical text in English and Japanese versions, a factor 2 seems to be reasonable when comparing typing efficiency in CPM with WPM.

Efficiency of an interface can be examined in CPM or WPM performed by experts or skilled users. Efficiency may also be estimated as an optimal speed by a perfect-user simulation or by a learning model. In this study, we evaluated the efficiency of the gaze-typing interfaces in terms of CPM after sufficient trials estimated by learning models.

The learnability of an interface can be evaluated in terms of either a learning factor, identified by applying the “power law of practice” model to experimental data, or in terms of the differential rate of typing speed measured at two different times in an earlier stage of practice. In addition to the learning factor and the differential rate, in this study, mean CPM in several earlier experimental blocks was compared among the gaze-typing systems to evaluate their learnability.

As error-related measures, several indices such as over-production rate, rate of backspacing, and minimum string distance (MSD) [Soukoreff & MacKenzie 2003], have been suggested in addition to the rate or frequency of errors per character or unit time. The MSD is a sentence-based error measure based on how many key-manipulation steps one needs to take to obtain a target sentence from the sentence actually typed (including errors). In the performance data collected from our experiment, almost all the sentences were error free (with or without correction) during gaze typing, and therefore it would be impractical to apply MSD to our data. The over-production rate is referred to as the rate of the actual number of (gaze) selections over the optimal (least) number of selections needed for constructing a given sentence. This index is particularly useful in examining the frequency of mistyping when using a hierarchical-menu system like Japanese GazeTalk. It is not possible to measure the number of single selections in Dasher, since it is operated by continuous navigation and not by single selections, but the rate of backspacing can be calculated by dividing the total number of backspace-key activations (or the total number of characters erased prior to the cursor position) by the total number of typed characters. We primarily used the over-production rate and the rate of backspacing to evaluate the error attribute of systems’ usability.

From subjective opinions of interface design, we evaluated not only subjective satisfaction, but also task-performance aspects of speed and error rate that are perceived by users. Question items related to these aspects were described in a five-point semantic differential (SD) scale – a pair of terms having opposite meanings such as “very fast” and “very slow”. The following subjectively rated items were included in the questionnaire: perceived typing speed, perceived likelihood of error, interface preference, satisfaction with the system, perceived fatigue, and uncomfortable feeling of motion sickness.

4 Results

4.1 Typing speed

Experimental data on the typing speed (in CPM) were analysed with a 3-way ANOVA, with the typing systems (S-GazeTalk, C-GazeTalk, and Dasher), tracking accuracy (low and high), and learning effect (Blocks 1–7) as the independent variables. Subjects were treated as repetitions. A result of the ANOVA is shown in Table 2. There was a significant difference in CPM between the three typing systems. The standard version of GazeTalk exhibited significantly better performance in typing speed compared with the other two systems. There was no significant difference between C-GazeTalk and Dasher. The grand mean of typing speed was 23.3 CPM, S.D. = 4.0. Mean CPM for S-GazeTalk was 24.2 (S.D. = 3.9), 22.7 CPM for C-GazeTalk (S.D. = 2.9) and 23.0 CPM for Dasher (S.D. = 5.1).

Table 2: Result of ANOVA on CPM.

Factor	s.s.	d.f.	V	F_0
System (A)	84.8	2	42.4	3.49*
Accuracy (B)	8.5	1	8.5	0.70
Block (C)	1000.2	6	166.7	13.73**
A×B	3.7	2	1.9	0.15
A×C	62.0	12	5.2	0.43
B×C	146.9	6	24.5	2.02
A×B×C	219.9	12	18.3	1.51
Error	2039.6	168	12.1	
Total	3565.6	209		

s.s.: sum of squares, d.f.: degree of freedom; V: mean square
* $p < 0.05$, ** $p < 0.01$.

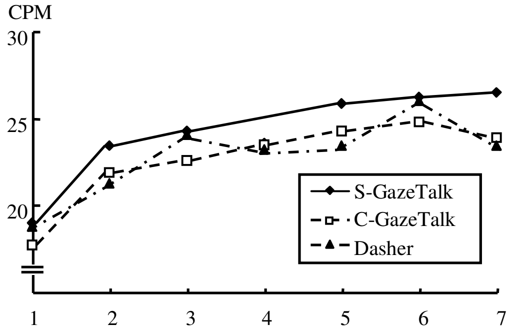

Figure 4: Transitions of CPM per experimental block for each typing system.

A significant learning effect was also observed between the seven blocks. The learning effect for each typing system is depicted in Figure 4. The typing speed increased with each block for each system.

Quantitative estimation of the learning effect on typing speed was made by applying the power law of practice learning model [Card et al. 1983; pp.57–59] to data from individual subjects. This model can estimate the time to perform the *n*th trial as a power law: $T_n=T_1n^{-\alpha}$, where α is a constant, representing a learning factor. Table 3 indicates the results of parameter estimation of the learning model, based on the typing systems, for all subjects. We observed that the power law of practice seemed well-fitted to any subject in the S-GazeTalk and the C-GazeTalk group whereas there were only two out of the five Dasher users whose learning effect could be explained well by this model. The learning factors – which themselves may represent a measure of learnability – are alike and reasonably high (about 1.6 on average) for the two versions of GazeTalk. On the other hand, those Dasher users for whom the power law of practice was well-fitted to their CPM data were very high, even compared with the best-learned GazeTalk users.

Table 3: Parameter estimation of the power law of practice on CPM for each typing system.

System	Sub.	a†	b††	R²	F₀
S-GazeTalk	S_1	0.047	-0.159	0.761	15.89*
	S_2	0.044	-0.122	0.985	332.86**
	S_3	0.056	-0.121	0.598	7.43*
	S_4	0.047	-0.195	0.892	41.53**
	S_5	0.063	-0.202	0.931	67.97**
Mean of learning factor			-0.160		
C-GazeTalk	S_6	0.052	-0.119	0.720	12.88*
	S_7	0.051	-0.178	0.891	40.97**
	S_8	0.057	-0.149	0.668	10.04*
	S_9	0.052	-0.119	0.644	9.06*
	S_{10}	0.058	-0.229	0.754	15.35*
Mean of learning factor			-0.159		
Dasher	S_{11}	0.079	-0.332	0.884	38.01**
	S_{12}	0.041	-0.066	0.241	1.59
	S_{13}	0.064	-0.289	0.697	11.48*
	S_{14}	0.040	-0.049	0.231	1.50
	S_{15}	0.059	-0.045	0.257	2.08
Mean of learning factor			-0.156		
			-0.310‡		

†, ††Parameters of the power law of practice: $y=ax^b$, where x=blocks practiced (ca. 300 characters/block), and y=typing time per character = 1/CPM (min./character).
‡Mean obtained only from subjects having significant effect.
*$p < 0.05$, **$p < 0.001$.

Table 4: Estimated CPM of the best-learned user for each typing system at various time points.

	Experiment		Estimated by models			
	n=1	*n*=7	*n*=7	*n*=10	*n*=20	*n*=50
S-GazeTalk	15.6	23.6	23.5	25.3	29.1	35.0
C-GazeTalk	15.4	24.9	26.7	29.0	34.0	41.9
Dasher	12.8	21.0	24.0	27.1	34.1	46.2

n: number of blocks practiced (each block includes 300 characters), e.g., *n*=50 indicates CPM after 15,000 characters typed.

The typing speed after a certain number of trials (i.e., the number of blocks) – related to efficiency – can be estimated by the power law of practice models. Table 4 shows the estimated CPM of the best-learned subject – the one having the greatest learning factor – from each typing-system group at the time of the 7th, 10th, 20th, or 50th block (one block includes 300 characters). The typing speed of the best Dasher subject was estimated to catch up with that of the best S-GazeTalk user at approximately 34 CPM after a 20-block typing practice, which would be after typing about 6,000 characters by gaze, corresponding to a total of nearly four hours of gaze typing. In the subsequent trials, the typing speed with Dasher is then expected to outperform that with S-GazeTalk. The learning model estimated an increased typing speed with Dasher at 46 CPM, which might be equivalent to approximately 23 WPM for European languages (as mentioned previously, we used a conventional factor of exchange between Japanese CPM and English WPM of roughly 2). The learning model also estimated an increased typing speed with C-GazeTalk at 42 CPM after a 50-block practice (i.e., 15,000 characters entered).

4.3 Typing errors

(1) Over-production rate

The result of a 3-way ANOVA for the over-production rate is shown in Table 5. This index can not be applied to Dasher, as mentioned in Section 3.5, and therefore only two versions of GazeTalk were analysed. The only significant difference was observed between the blocks, and there were no significant effects for any other factors. As depicted in Figure 5, the over-production rate of each version of GazeTalk gradually decreased with blocks. In particular, the learning effect on this index is seen until Block 5, and subsequently the rate seems to become constant.

Table 5: Result of ANOVA on the over-production rate.

Factor	s.s.	d.f.	V	F₀
System (A)	0.014	1	0.014	0.61
Accuracy (B)	0.013	1	0.013	0.54
Block (C)	0.514	6	0.086	3.69*
A×B	0.027	1	0.027	1.17
A×C	0.177	6	0.030	1.27
B×C	0.215	6	0.036	1.54
A×B×C	0.233	6	0.039	1.67
Error	2.603	112	0.022	
Total	3.800	139		

* $p < 0.05$.

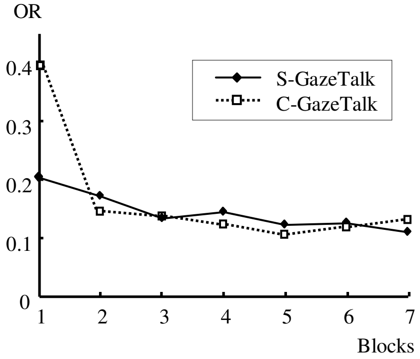

Figure 5: Transitions of over-production rate with experimental blocks for the two versions of GazeTalk.

(2) Rate of backspacing

The result of ANOVA for the rate of backspacing with the same three factors is shown in Table 6. For this error-related index, a significant difference was observed between the three typing systems. In particular, as can be seen in Figure 6, the rate of backspacing with Dasher was significantly far higher than with each version of GazeTalk. The grand mean of the rate of backspacing was 0.037 (per typed character), S.D. = 0.032. The mean rate for S-GazeTalk was 0.029 (S.D. = 0.021), 0.028 for C-GazeTalk (S.D. = 0.021) and 0.053 for Dasher (S.D. = 0.041). There was no significant difference between the two versions of Japanese GazeTalk.

Table 6: Result of ANOVA on the frequency of using backspace.

Factor	s.s.	d.f.	V	F_0
System (A)	0.0276	2	0.01380	15.24**
Accuracy (B)	0.0034	1	0.00340	3.76
Block (C)	0.0045	6	0.00075	0.82
A×B	0.0026	2	0.00132	1.46
A×C	0.0112	12	0.00093	1.03
B×C	0.0049	6	0.00082	0.91
A×B×C	0.0057	12	0.00048	0.53
Error	0.1521	168	0.00091	
Total	0.2120	209		

** $p < 0.01$.

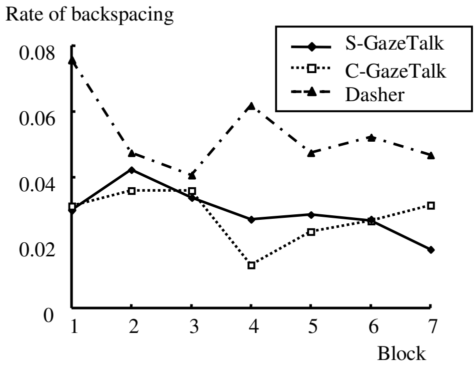

Figure 6: Transitions of the rate of backspacing with experimental blocks for three typing systems.

4.4 Subjective ratings

The results of subjective ratings at two different occasions during system usage were summarised in Table 7 in terms of percentage agreements of each self-reported item, one in an early stage (after the first experimental block), and the other at the end of the experiment (after seven blocks). The percentage agreement was calculated based on the typing systems as a rate of subjects who agreed (strongly or slightly) with the left-hand side term in each SD scale, e.g., “very fast”, “unlikely to make error”, or “felt uncomfortable due to motion sickness”. As an overall trend, no remarkable difference was identified from the subjective responses between the typing systems for most items. Differences between the typing systems and between the time instances were 20% for most items, which could be made by only a single subject’s response – as only five subjects were included in each system group. In addition, there might have been an individual difference in the subjective criteria of rating for each item. Therefore, we cannot derive a sound conclusion about subjective satisfaction with each typing interface from these results. However, it seems that none of the systems induce motion sickness.

Table 7 Percentage agreements of subjective ratings at the beginning and the end of the experiment

Items	S-Ga.	C-Ga.	Dash.
Typing speed	20%	40%	20%
(very fast --- very slow)	20%	20%	40%
Interface preference	80%	20%	80%
(sophisticated --- difficult)	60%	40%	20%
Error	80%	60%	60%
(unlikely --- likely to make error)	60%	40%	20%
Satisfaction with system	60%	60%	80%
(very satisfied --- very dissatisfied)	80%	40%	60%
Fatigue	60%	60%	80%
(very tired --- not tired at all)	80%	40%	60%
Motion sickness	0%	0%	0%
(felt bad --- did not feel bad at all)	0%	0%	0%

Upper row: responses after Block 1
Lower row: responses after Block 7

5 Conclusion and future work

In this study, we conducted usability evaluations of two different gaze-typing systems that run in Japanese, representing different approaches to character selection, either by dwell clicking (in GazeTalk) or by continuous navigation (in Dasher). We made two versions of the Gazetalk system, one with the text field in the upper-left corner (S-GazeTalk) and one with the text field in the centre (C-GazeTalk) of the menu. During learning S-GazeTalk was slightly more efficient than the other systems. The clear separation of the text field from the key area in the display and the possibility to overlook all keys in the parafoveal vision may explain its superiority when compared to C-GazeTalk. The use of a well-established button metaphor and a well-known hierarchical structure may explain its initial superiority when compared to Dasher.

Typing speed seems to be sensitive to the small variations in designs between the S- and the C-version of GazeTalk, even though the differences could not be revealed by subjective ratings. Some designers in our team had argued that C-GazeTalk would probably be more efficient than S-GazeTalk as the user could verify typed characters in the text field with shorter saccades from the key area – or possibly just do it by parafoveal vision. The results presented suggest that ease of verification is of less importance than maintaining an overview of an ever-changing character layout. In future experiments, we would like to investigate the relative importance of saccade lengths and parafoveal vision, for instance by comparing typing speed and error measures on different sizes of the same keyboard.

A higher learning factor was obtained for some Dasher users than for S- and C-GazeTalk subjects when estimated by the power law of practice. Dasher was predicted to become more efficient for these users than any of the two versions of GazeTalk after some amount of practice. Consequently, we would like to encourage user of GazeTalk to try out if Dasher will work for them by providing a direct access to Dasher from within future versions of GazeTalk.

Contrary to our expectations, there was no significant effect of eye-tracking accuracy on any of the performance measures examined in this study. This is a promising result for future development of low-resolution eye trackers [Hansen & Hansen 2006]. The method of miscalibration can possibly be applied on every typing system to identify the threshold of tracking errors that cause poor performance on it. For Dasher and GazeTalk the thresholds are higher than 1 degree. Traditional qwerty on-screen keyboards with a large number of small keys will presumably be more sensitive to inaccuracies in tracking.

In conclusion, the present study confirms that gaze typing can be learned within some hours of practice and become almost error-free even when novel typing principles, like the one of Dasher, is applied. We regard a productivity of more than 20 CPM to be acceptable; in fact it is slightly higher than the typing speeds that has been reported previously for gaze typing in English [Majaranta & Räihä 2002].

Acknowledgements

This research was partly supported by Grant-in-Aid for Scientific Research (A), No. 14208040, the Japan Society for the Promotion of Science, and by the COGAIN Network of Excellence, IST IU 6, Contract Number 511598. We are grateful to David J. MacKay, University of Cambridge, for his insightful guidance in the experiment with Dasher and for his suggestion to deliberately miscalibrate trackers. We would like to acknowledge Hiroaki Kodera, Tokyo Institute of Technology, for his cooperation with data collection in the experiment.

References

AOKI, H., ITOH, K., and HANSEN, J. P. 2005. Learning to Type Japanese Text by Gaze Interaction in Six Hours. In *Proceedings of the 11th International Conference on Human-Computer Interaction*, Las Vegas, NV, July (CD ROM).

AOKI, H., ITOH, K., SUMITOMO, N., and HANSEN, J. P. 2003. Usability of Gaze Interaction Compared to Mouse and Head-Tracking in Typing Japanese Texts on a Restricted On-screen Keyboard for Disabled People. In *Proceedings of the 15th Triennial Congress of the International Ergonomics Association*, Seoul, Korea, Vol. 1, 267–270, August.

CARD, S. K., MORAN, T. P., and NEWELL, A. (1983). *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum Associates, Hillsdale, NJ.

HANSEN, J. P., and HANSEN, D. W. 2006 (to appear). Eye Typing with Consumer Cameras. Poster presented at *Eye Tracking Research & Applications Symposium 2006*, San Diego, CA, March.

HANSEN, J. P., HANSEN, D. W., and JOHANSEN, A. S. 2001. Bringing Gaze-based Interaction Back to Basics. In *Proceedings of Universal Access in Human-Computer Interaction (UAHCI 2001)*, New Orleans, LA, 325–333, August.

HANSEN, D. W., HANSEN, J. P., NIELSEN, M., JOHANSEN, A. S., and STEGMANN, M. B. 2002. Eye Typing Using Markov and Active Appearance Models. In *IEEE Workshop on Applications on Computer Vision*, 132–136.

HANSEN, J. P., JOHANSEN, A. S., HANSEN, D. W., ITOH, K., and MASHINO, S. 2003. Command Without a Click: Dwell Time Typing by Mouse and Gaze Selections. In M. Rauterberg et al. (Eds.), *Human-Computer Interaction – INTERACT´03*. IOS Press, 121–128.

ISTANCE, H. O., SPINNER, C., and HOWARTH, P. A. 1996. Providing Motor Impaired Users with Access to Standard Graphical User Interface (GUI) Software via Eye-based Interaction. In *Proceedings of 1st European Conference on Disability, Virtual Reality and Associated Technology,* ECDVRAT, U.K., 109–116.

JACOB, R. K. 1991. The Use of Eye Movements in Human-Computer Interaction Techniques: What You Look at Is What You Get. *ACM Transactions on Information Systems*, 9(3), 152–169.

MAJARANTA, P., and RÄIHÄ, K. -J. 2002. Twenty Years of Eye Typing: Systems and Design Issues. In *Proceedings of the Symposium on ETRA 2002: Eye Tracking Research & Applications Symposium 2002*, New Orleans, LA, 15–22.

MAJARANTA, P., MACKENZIE, I. S., AULA, A., and RÄIHÄ, K. -J. 2003a. Auditory and Visual Feedback During Eye Typing. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems CHI 2003*. ACM, New York, 766–767.

MAJARANTA, P., MACKENZIE, I. S., and RÄIHÄ, K.-J. 2003b. Using Motion to Guide the Focus of Gaze during Eye Typing. In *Proceedings of the ECEM12*, Dundee, U.K., August.

NIELSEN, J. 1993. *Usability Engineering*. Academic Press, San Diego, CA.

OHNO, T. 1998. Features of Eye Gaze Interface for Selections Tasks. In *Proceedings of the 3rd Asia Pacific Computer Human Interaction – APCHI'98*. IEEE Computer Society, 1–6.

SALVUCCI, D. D. 1999. Inferring Intent in Eye-Movement Interfaces: Tracing User Actions with Process Models. In *Human Factors in Computing Systems: CH1 99 Conference Proceedings*, Pittsburgh, PA, ACM Press, 254–261.

SALVUCCI, D. D., and ANDERSON, J. R. 2000. Intelligent Gaze-Added Interfaces. In *CH1 2000 Conference Proceedings*, The Hague, The Netherlands, ACM Press, 273–280.

SIBERT, L. E., and JACOB, R. J. K. 2000. Evaluation of Eye Gaze Interaction. In *Human Factors in Computing Systems: CHI 2000 Conference Proceedings*, The Hague, The Netherlands, 281–288.

SOUKOREFF, R. W., and MACKENZIE, I. S. 2003. Metrics for Text Entry Research: An Evaluation of MSD and KSPC, and a New Unified Error Metric. In *Proceedings of the Conference on Human Factors in Computing Systems*, 5(1), 113–120.

WARD, D. J. 2001. Adaptive Computer Interfaces. Doctoral Dissertation, University of Cambridge, U.K.

WARD, D. J., and MACKAY, D. J. C. 2002. Fast Hands-Free Writing by Gaze Direction. *Nature* 418, p. 838, August 22.