Performance Evaluation of Voice Interaction as a Universal Web Interface

Yu-Ming Fei, Chiuhsiang Joe Lin*, Min-Ting Chen, and Chih-Cheng Chiang

Department of Industrial Engineering, Chung Yuan Christian University
200, Chung Pei Rd., Chung Li, Taiwan 32023, R.O.C.
* hsiang@cycu.edu.tw

Abstract. A speech interface sometimes provides easy access to WWW information since it makes browser potentially more friendly and powerful. This study first made a comparison between a voice and a keyboard input task. It then compared the voice interface with the keyboard, mouse, and combination of keyboard and mouse with three WWW browsing tasks. The experimental results showed that the voice interface performed the input task faster than the keyboard. The voice interface reached an input accuracy of 92% after sufficient learning. In the WWW browsing tasks, the voice performed poorly in terms of completion time, compared to the keyboard, mouse, or both. Nevertheless, the voice interface can be used to complete the three assigned tasks without problems. The study showed that voice interaction can be used as a universal web interface, especially for those who are unable to use the mouse and keyboard due to disabilities.

Keywords: Speech interface, Voice interface, IBM ViaVoice, Browse WWW.

1 Introduction

It is very common to find information through internet service. Since internet can supply huge database and can supply a lot of information at any time. Normally, people are familiar with the keyboard and mouse as the major interfaces, but for people such as old people or people with disability, it becomes difficult to access the internet service. For people who cannot use the keyboard and mouse, it is important to develop other methods for them to access the internet.

Software that can convert spoken words into written text (hereafter referred as SWIWT) has been available since the early 1980s. Early continuous speech systems were first introduced in 1994, with the latest commercially available editions having a claimed accuracy of up to 98% of speech recognition at natural speech rates [1]. The IBM software was found to have the lowest mean error rate for vocabulary recognition (7.0 to 9.1 percent) followed by the L&H software (13.4 to 15.1 percent) and then Dragon software (14.1 to 15.2 percent) [2]. In addition, the IBM software was found to have better performances than both the Dragon and the L&H software in the recognition of general English vocabulary and medical abbreviations [8].

* Corresponding author.

C. Stephanidis (Ed.): Universal Access in HCI, Part III, HCII 2007, LNCS 4556, pp. 279–288, 2007.

In spoken words into written text, many researches have been doing in this field to explore the reliability measurement. Wang, Chou, and Chen [6], pointed out that the interface currently provides speak-able commands, bookmarks and links. However there is no general literature pointing out what kind of tasks can be implemented when browsing the web. This study attempts to use an experiment design to verify the task performances on the web using the IBM Via Voice software.

This research first makes a comparison on the difference between SWIWT and keyboard inputs. According to the instruction of the software, SWIWT has its own learning ability. The experiment will only begin after system self-learning and after more voice modules have established. The research then carries out an experiment to view web interface and complete certain tasks by using different input interfaces which include mouse, keyboard, and speech.

2 Method

A two-phase study was designed to explore the performances between speech interface, keyboard, and mouse. The first study was to compare "spoken words into written text" (hereafter referred as SWIWT) with keyboard input. The second study was to compare the performances between speech interface, keyboard, and mouse in browsing the web pages under this experiment. There are 3 male graduates subjects participating in this experiment who study in Chung Yuan Christian University. They had a comparable level of computer literacy and web expertise. The following hardware and software were used: a desktop 3.0-GHz Intel Pentium 4 processor with 256 MB of RAM, noise-canceling headset microphone, and IBM ViaVoice Pro version 10.

2.1 Experiment 1

The main factor was two input text interfaces- SWIWT and keyboard. Each of these three participants had to test each input text interfaces for three times. The sequence of the experiment was nested under each input text interface, and the participants were designed to be randomized block. This study randomly selected an article (398 words) from TQC (Techficiency Quotient Certification) article database for all input tests. The main objective of this study is to compare the difference between SWIWT and keyboard input, and to explore the effects of testing between them.

Therefore the model for the nested design is

$$Y_{ijk} = \mu + \tau_i + \alpha_{j(i)} + \beta_k + \varepsilon_{ijkl} \begin{Bmatrix} i = 1,2 \\ j = 1,2,3 \\ k = 1,2,3 \end{Bmatrix} \tag{1}$$

$$\text{Input text interface: SWIWT and keyboard} \quad \tau_i \rightarrow i = 1,2 \tag{2}$$

$$\text{Sequence} \quad \alpha_{j(i)} \rightarrow j = 1,2,3 \tag{3}$$

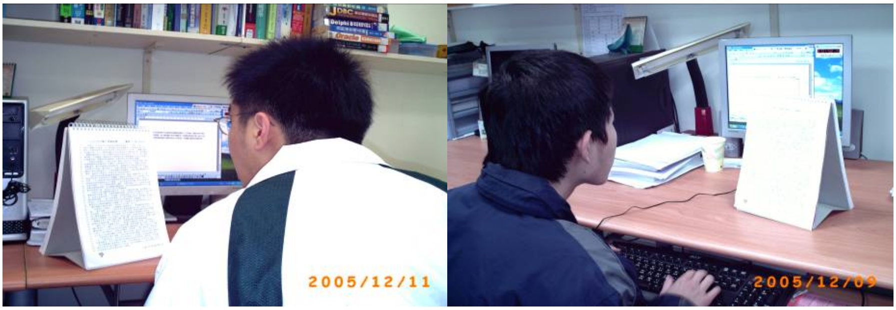

Fig. 1. Figure 1: The participants used SWIWT and keyboard to input texts

$$\text{Participants} \quad \beta_k \rightarrow k = 1,2,3 \tag{4}$$

When using the SWIWT interface, it was different in the speech model training each time. In the first time, every participant must create a base of personal speech model, which took a long time. Before the second SWIWT input task, each participant had to spend about 35 minutes to read an article from IBM ViaVoice system to enhance the established personal speech model. Similarly, before the last SWIWT input task, another 20 minutes or so were used by each subject to enhance the personal speech model again.

This experiment recorded the completion time and the error words of the input tasks. The study also observed the effect of self-learning with the IBM ViaVoice system.

2.2 Experiment 2

This research then started the second experiment on web control. There were three tasks assigned in the experiment, and they were all universal access activities when browsing the web. The second experiment included two factors. The first factor is the tasks in browsing the web pages, containing the following three tasks: (1)Task one requires the subject to find out designated web and click on it from My Favorite. (2)Task two must trigger the required action on the web by moving the cursor. (3)The objective of the third task is to input data into appointed location on the web.

The second factor is the interface used to brow the web pages. There are four interfaces in this factor, 1) use keyboard and mouse, 2) just use keyboard, 3) use mouse only, 4) use speech interface. The three participants were designed to be the randomized block. In this experiment the time spent for each task was observed and recorded from the start to the end.

The statistical model for this design is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma_k + \varepsilon_{ijk} \begin{Bmatrix} i = 1,2,3 \\ j = 1,2,3,4 \\ k = 1,2,3 \end{Bmatrix} \tag{5}$$

$$\text{Tasks:} \quad \alpha_i \rightarrow i = 1,2,3 \tag{6}$$

$$\text{Interfaces:} \quad \beta_j \rightarrow j = 1,2,3,4 \tag{7}$$

$$\text{Participants:} \quad \gamma_k \rightarrow k = 1,2,3 \tag{8}$$

2.3 Task and Instructions

The participants were told prior to experiments as to the nature of the study that we were interested in determining whether the tasks are better in terms of time to complete them. However, they were not told what types of results were expected. Participants have one hour practicing before carrying out the experiment, and they are requested to be familiar with all the voice commands.

Task One. The objective of task one is to find out designated web and click on it from My Favorite. The procedure is as follows.

1. Open browser and use my favorite to go to the page, e.g. “健康醫網” (“Health Web”).
2. Find out the “News” of the health link and go to the fist news link. Visit the page to the bottom, and then use the “Back” to go to “News”.
3. Find out the first news on “Health of residence”, and visit the page to the bottom and then say the last sentence.

Task Two. The objective of task two is to trigger the action on the web by moving the cursor, as described in the following.

1. Use my favorite to go to the page, e.g. “花花世界” (“The Flower World”), and then go into the main page from the door picture in the center of web page.
2. Slide the cursor to move about the four flower pictures, and make them to change to another picture. After flower pictures changed, keep sliding the cursor to point to the “花之物語” (“Symbolization of Flowers”) button underneath the page.
3. Find out the symbolization for tomato in the “Symbolization of Flowers” page, and then move the cursor to point to the “走馬看花” (“Poems of Flowers”) button underneath the page.
4. Browse through the “Poems of Flowers” page to the bottom and then say the last sentence.

Task Three. The objective of task three is to input data into appointed location on the web.

1. Use my favorite to open the page, e.g. “意見調查” (“The suggestion”), and then keep the cursor on the form fill-in.
2. Type “王小明” (“Wang Xiao-Ming”) for name information, type “54321” for password, then select “Female” for sex, and type “網頁製作精美，我非常喜歡。” (“A beautiful web design and I like it very much.”).
3. Press the “Send” button when finished with the typing task.

3 Results

This section presents and discusses the results of the two experiments. For each experiment, the data were assessed as to the fit of the normality assumption, the independence assumption, and the equality of variance. Then, summary measures and tests for main effects and interactions are presented on completion time of tasks across levels of experimental factors. The data analysis was performed on a Pentium 4 PC using Minitab release 14.

3.1 Experiment 1

The ANOVA summary table for data on completion time (Table 1) indicates a significant main effect for input text interface (F = 57.25, P = .000). However, there was no significant main effect for sequence (F = .48, P = .749). The results suggest that SWIWT (M=144 seconds, SD=7.06 seconds) used less completion time than keyboard to input texts (M=688 seconds, SD=238.3 seconds). The SWIWT showed a more consistent response in task speed among the participants as reflected by the smaller standard deviation.

The ANOVA summary table for error words (Table 2) indicates a significant main effect for input text interface (F = 19.77, P = .001). However, there was no significant main effect for sequence (F = 2.26, P = .135).

A good way to further analyze the data is the analysis of means (ANOM). This test not only answers the question of whether or not there are any differences among

Table 1. Analysis of variance for completion time, using adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Input text interface	1	1331712	1331712	1331712	57.25	0.000
Sequence	4	44781	44781	11195	0.48	0.749
Participant	2	180652	180652	90326	3.88	0.056
Error	10	232609	232609	23261		
Total	17	1789754				

S = 152.515 R-Sq = 87.00% R-Sq(adj) = 77.91%.

Table 2. Analysis of variance for error words, using adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Input text interface	1	5477.6	547736	5477.6	19.77	0.001
Sequence	4	2505.6	2505.6	626.4	2.26	0.135
Participant	2	398.8	398.8	199.4	0.72	0.510
Error	10	2770.6	2770.6	277.1		
Total	17	11152.4				

S = 16.6450 R-Sq = 75.16% R-Sq(adj) = 57.77%.

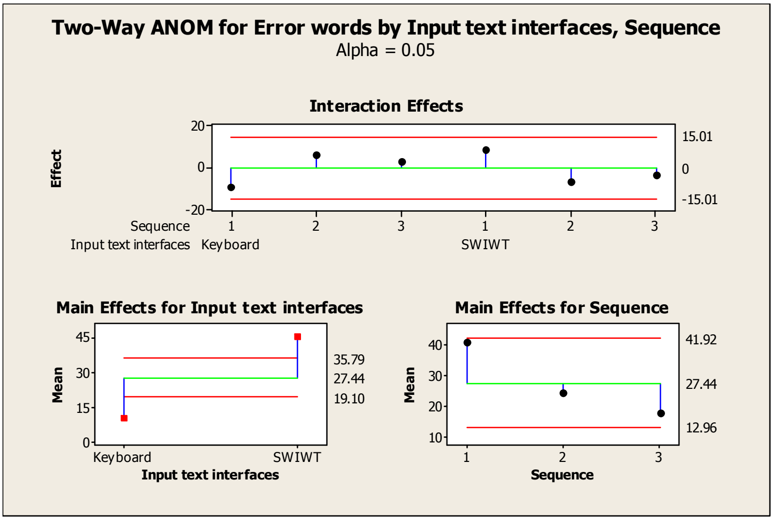

Fig. 2. Two-way ANOM for error words by Input text interface, Sequence

the factor levels, but also tells us which levels are better and which are worse when there are differences [3]. For the sake of providing visual picture of the distinction between factor levels, consider the graphic representation in Figure 2. The number of error words between keyboard and SWIWT were strikingly different, with the error words of keyboard fewer than SWIWT. Since the self-learning ability of SWIWT was most concerned in this study, we observed a decreasing trend in the number of error words for sequence, as can be seen in Figure 2. Since in the original model of ANOVA, sequence was nested under interface, no interaction between input text interface and sequence was assumed. If we only looked at the sequence effect under the SWIWT interface, Table 3 further shows the mean error rate for sequence. Clearly, the data show that the time spent in enhancing the personal speech model had a positive effect on participants performance to reduce error rate. The result indicate that the self-learning feature of the IBM ViaVoice system could improve accuracy rate to above 92%.

Table 3. The mean error rate of sequence in SWIWT

	First time	Second time	Third time
Error rate	16.83%	8.83%	8.00%

3.2 Experiment 2

In this experiment we wish to investigate the effect of single and multimodal interfaces, which combine two modes from the keyboard, mouse, or speech for browsing web page tasks. But in the study, task two (moving the cursor) could not be performed by using the keyboard only. Therefore, the analysis was divided into two parts. The first analysis leaves out the data on task two, 'moving the cursor'. The second analysis leaves out the data on the interface, 'just use keyboard'.

Results by Excluding the Task of Moving Cursor. The results on the performance of completion time are presented in Table 4. Analysis of variance (ANOVA) showed a significant effect of task, F(1,14)=158.89, p < .001; a significant effect of interface, F(3,14) = 145.16, p < .001; and an interaction between task and interface, F(3,14) = 29.81, p < .001. A more detailed understanding of the relationship between these effects can be gained from Figure 3. Using speech interface to browse through the web was the most difficult and wasted a lot of time than the other interfaces.

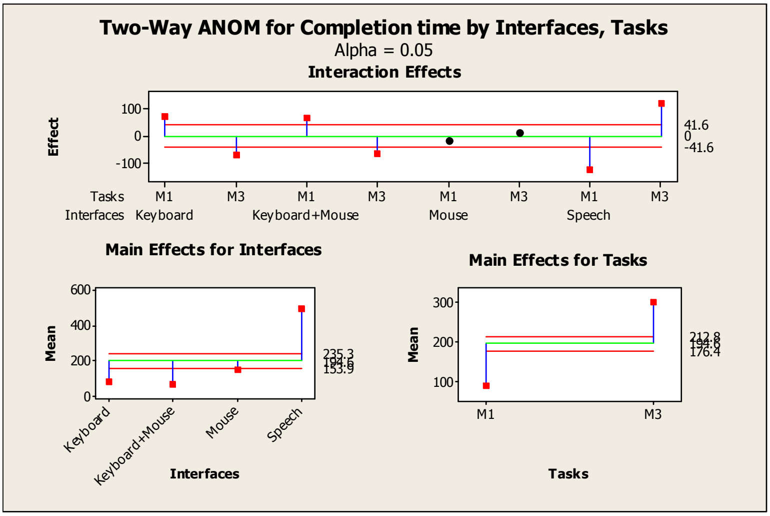

Fig. 3. Two-way ANOM for completion time by interface and task (M1: Task one, M2: Task two, M3: Task three.)

Table 4. Analysis of variance for completion time, using adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Task	1	267126	267126	267126	158.89	0.000
Interface	3	732147	732147	244049	145.16	0.000
Task*Interface	3	150378	150378	50126	29.81	0.000
Participant	2	4666	4666	2333	1.39	0.282
Error	14	23537	23537	1681		
Total	23	1177854				

S = 41.0028 R-Sq = 98.00% R-Sq(adj) = 96.72%.

Results by Excluding the Only Keyboard Interface. As shown in Table 5, analysis of variance (ANOVA) showed a significant effect of task, F (2, 16) =368.00, p < .001; a significant effect of interface, F (2, 16) = 113.05, p < .001; and a significnat interaction between task and interface, F (4, 16) = 22.37, p < .001. A more detailed understanding of the relationship between these effects can be gained from Figure 4. The speech interface to browse through the web was the most difficult and wasted a lot of time than the other interfaces.

Table 5. Analysis of variance for completion time, using adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Task	2	992544	992544	496272	368.00	0.000
Interface	2	304900	304900	152450	113.05	0.000
Task*Interface	4	120653	120653	30163	22.37	0.000
Participant	2	8348	8348	4174	3.09	0.073
Error	16	21577	21577	1349		
Total	26	1448022				

S = 36.7229 R-Sq = 98.51% R-Sq(adj) = 97.58%.

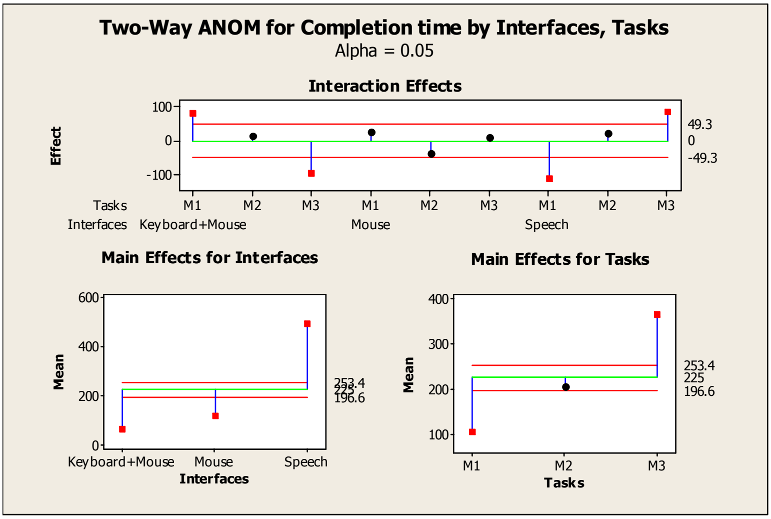

Fig. 4. Two-way ANOM for completion time by interface and task (M1: Task one, M2: Task two, M3: Task three.)

4 Discussion and Conclusions

From the results of the experiment, this study has found out that apparently less time was spent on the text input when using the voice interface in comparison with typing (non-voice interface). In aspect of error rates, although there were more errors on the text recognition system when initially started using the voice interface, incorrect texts were obviously reduced after voice training was given. According to the result of the experiment, voice interface did have the learning ability and its mean accurate rate can reach up to 92% especially in later stage of the voice interface. Even thought error rate of voice interface is higher than typing, the speed of text input is quicker, and the voice interface is still very practical.

The error rate was decreased through system learning by enhancing personal speech models. In 1999, a study of ASR (Automatic speech recognition) performance on broadcast news showed wide variations, from a maximum of 88% words correctly recognized to a minimum of 35%, with a mean of 67% [7]. A recently research of ASR performance in the word error rate of the overall integrated system could reduce to 9.26% to 9.61% [4]. Up to this point, these results show ASR still advancement. An additional result was 165.8 words per minute in this study. It may be of interest for future research on how to increase speech speed and accuracy.

To summarize the salient features of the analysis, several findings are of interest. This study revealed that improving method of operation appears to be an important variable in browser WWW. The speech interface could cost a lot time even though it can do most of job.

For the assigned tasks in this study, it is more difficult to control cursor and to browse web when using the voice interface. However, it does provide an advantage for disabled users. The performance evaluation of voice interaction as a universal web interface can help those who are unable to use the mouse and keyboard to achieve the same result as using the mouse and keyboard.

For WWW applications, the continuous reaction time must be immediate and smooth. Participants were anxious when they used the IBM ViaVoice to shift the cursor. Often users like to find some interesting part to browse around a web page. Using speech to browse a text link had fast response time, but moving the cursor to point to a link took a long response time, because there was a need to keep using the voice to shift the cursor even though some commands could be retrieved fast. Shneiderman, and Plaisant, [5] pointed out that an appropriate response time on typing, cursor motion, or mouse selection was around 50-150 milliseconds. It seems that there is still a big gap for the voice interface to be realistically used in the future.

Another point is that a popular web needs to be designed for universal web Interface. The study found that there was no way to move the cursor on a web page if usuing the keyboard only. Voguish web design uses a lot of skills and techniques, but usually does not consider what possible universal interfaces will be used. For popular web pages, designing should be suitable for all multimodal interfaces (keyboard, trackball, speech, pen, touch…). Further WWW design issues should be pursued toward a handicap-friendly place on the Internet.

Although the sample in the study was small, the following recommendations could serve as some general principles for researchers who would like to experiment with ASR in similar contexts. The results of the present study may be concluded by pointing out that ASR could be applied to Web Interface but still needs some pratical improvements.

References

1. Al-Aynati, M.M., Chorneyko, K.A.: Comparison of Voice-Automated Transcription and Human Transcription in Generating Pathology Reports. Archives of Pathology and Laboratory Medicine 127(6), 721–725 (2003)

2. Devine, E.G., Gaehde, S.A., Curtis, A.C.: Comparative Evaluation of Three Continuous Speech Recognition Software Packages in the Generation of Medical Reports. Journal of the American Medical Informatics Association 7, 462–468 (2000)
3. Nelson, P.R., Coffin, M., Copeland, K.A.F.: The Analysis of Means. Introductory Statistics for Engineering Experimentation, pp. 250–260. Elsevier Academic Press, North-Holland, Amsterdam (2003)
4. Schwenk, H.: Continuous space language models. Computer Speech and Language. In Press, Corrected Proof, Available online October 9, 2006 (2006)
5. Shneiderman, B., Plaisant, C.: Designing The User Interface: Strategies For Effective Human-Computer Interaction, 4th edn. p. 473. Addison Wesley, USA (2005)
6. Wang, H.M., Chou, Y.H., Chen, B.: Browsing The Chinese Web Pages Using Mandrin Speech. International Journal of Computer Processing of Oriental Languages 13(1), 33–50 (2000)
7. Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F., Singhal, A., SCAN,: Designing and evaluating user interfaces to support retrieval from speech archives. In: Proc. ACM SIGIR ’99 (1999)
8. Zhang, J., Zhao, J., Bai, S., Huang, Z.: Applying Speech Interface to Mahjong game. Multimedia Modelling Conference. In: Proceedings. 10th International, pp. 86–92 (2004)