EYE GAZE AND SPEECH FOR DATA ENTRY: A COMPARISON OF DIFFERENT DATA ENTRY METHODS

Yeow Kee Tan, Nasser Sherkat, Tony Allen

Nottingham Trent University, Department of Computing and Mathematics
Burton Street, Nottingham
NG1 4BU, United Kingdom
[yeow.tan, nasser.sherkat, tony.allen]@ntu.ac.uk

ABSTRACT

In this paper we present a multimodal interface that employs speech recognition and eye gaze tracking technology for use in data entry tasks. The aim of this work is to compare the usability of this multimodal system against other data entry methods (handwriting, mouse and keyboard and speech only) when carrying out the data entry task of filling a form. Discussions regarding the relationships between efficiency, effectiveness, ergonomic quality, hedonic quality, naturalness, familiarity and users preference are presented. The experimental results show that the majority of the users prefer using the proposed eye and speech system compared to the other form-filling methods even though such a method is neither the fastest nor the most accurate.

1. INTRODUCTION

Although research into multimodal interfaces can be dated back to 1980, research into interfaces that combine active (e.g. speech recognition) and passive (e.g. eye gaze) inputs is still in its infancy and needs to be explored further in order to identify new combinations that can improve the human-computer interaction. This type of interface is known as a blended style interface and typical examples include IBM's Manual And Gaze Input Cascaded pointing system [1], that allows the use of mouse and eye input to select the icons on the desktop, and the combination of facial and speech recognition [2] that is able to reduce word error rate by 27% in a noisy environment. These positive results have led us to research combining eye gaze and speech for data entry [3].

An experiment has been carried out in order to compare the usability of the proposed multimodal interface against other data entry methods. The aspects of efficiency, effectiveness and user preference have been evaluated. In addition to these common usability aspects, aspects such as easy to use, fun to use, familiarity and naturalness have also been observed in order to achieve a more extensive usability study.

According to Jordan, "Usability as a concept does not seem to include (positive) feelings such as pride, excitement or surprise" [4]. These positive emotions, mentioned by Jordan, can be classified as Hedonic Quality (HQ). HQ refers to quality attributes with no obvious relation to task/goal-fulfillment, i.e. "original", "innovative", "exciting", or "exclusive". These attributes address the human needs for novelty/change (i.e. excitement) and social power (i.e. status, pride) [5]. It is to be noted that fun to use differs from easy to use. Carroll and Thomas [6] argued that ease of use implies simplicity, which in turn is partly incompatible with fun. For example by making a system easy to use, there is a chance that it will be boring as well.

After years of debate, the Human-Computer Interaction (HCI) research community is now gradually accepting the concept of joy and fun as an important factor in usability [7]. Based on the findings in [8], perceived fun has a stronger effect on user satisfaction than perceived usefulness. It was also observed that user satisfaction will lead to an increased time-spent with a software system. This, in turn, may cause the user to use the system more frequently thereby gaining a better understanding of it (increased familiarity). However by introducing the factor of fun (HQ), the simplicity of a system may also decrease [5]. The question of whether using HQ as one of the main factors in system development is still a debating issue in the HCI community.

In this paper, naturalness is defined as the regular way, by which one human being would pass information to another human being (e.g. Person A tell his/her name to person B). Familiarity, on the other hand, is defined as the most common way a particular task (in this case a data entry task of filling a form) is carried out. Note that naturalness and familiarity are different in this case. Naturalness relates more to whether the user experiences a human-to-human input style whilst interacting with the system. Familiarity is concerned more with the methods the user has previously used for carrying out a particular task.

2. SYSTEM IMPLEMENTATION

The experiment in this paper compares the use of handwriting (HW), mouse and keyboard (MK), speech only (SO) and eye and speech (ES) for a specific data entry task. The ES and SO systems use the same interface layout (figure 1). This allows users to fill-out a form that consists of the 8 fields (television type, title, surname, initials, house number, street name, city

and postcode) found in a Television License application form used in the United Kingdom. The MK system, on the other hand, uses a layout similar to a common online form that requires the user to use mouse to select the required field and use keyboard to enter the required data. The HW system requires the users to fill in a television license application form using a pen. The application form will be recognized later using an offline Cursive Script Recognizer [9].

Figure 1: A field selected in the eye and speech system

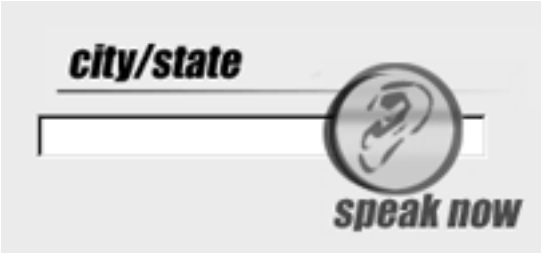

Figure 2: A field selected in the eye and speech system

The SO system utilizes the IBM ViaVoice speech recognizer. Meanwhile the ES system integrates the IBM ViaVoice speech recognizer along with an ASL EyeLink I eye gaze tracking system.

In order to invoke recognition in the ES system, the users are required to look at the field that is to be filled. If the eye gaze system detects that a fixation has occurred it highlights the selected field. The user then speaks the required data. After recognition is complete the system will insert the recognized input into the selected field.

The eye gaze fixation dwell time used is 150 milliseconds and the scan area of each field is 250 x 150 pixel unit. The selection of the fixation dwell time was based on an experiment of comparing a short and a long fixation dwell time carried out earlier.

The microphone is turned off whenever the user's gaze is not within any of the fields in the screen. This provides the system with an unobtrusive on/off microphone switch to prevent accidental speech input that might cause error.

In the SO system, the microphone is switched on all the time the system is not carrying out any speech recognition process. For the SO system, the users are required to say the field name first (e.g. city, house number or title) and then, once the field is selected, user are required to say the data.

If the recognizer is unable to get the correct word (system rejection, substitution or deletion error in either the SO or ES systems), then the users are required to carry out the look and say or speak only process again. If a deletion or rejection error occurs then the users need only speak the data without the need of reselecting the field name. However the user are required to carry out the field selection (look at or say the field name) and then say the data process again if a substitution error occurs.

3. EXPERIMENT

Thirteen (5 non-native and 8 native United Kingdom English speakers) users participated in the experiment. At the beginning of the experiment the users were informed as to the purpose of the experiment and questionnaires were shown so that the users knew the key aspects that needed to be noticed during the experiment. For each system, the users were taught how to interact with the system. This training required less than 2 minutes and the users were shown the interface on the screen and the data entry procedure was explained verbally. At this point the users did not interact with the system.

In the experiment, the users were required to fill in the following fields: television type (lexicon size = 2), title (lexicon size = 5), initial (lexicon size = 676), surname (lexicon size = 1743), house numbers (e.g.132, 55, 900), street name (lexicon size = 1503), city name (lexicon size = 1834) and postcode (all possible combinations).

Users were told to move on to the next field, after five unsuccessful attempts at entering the data in either the SO or ES systems. After this introduction, the users were asked to complete the television license application form using each of the 4 interfaces in turn. Note. Before using the ES system it was necessary to undergo an eye calibration process.

The user's interactions were recorded using a video camera and after completion of the task with each system, the users were requested to fill in the appropriate questionnaire. This process was repeated until all the systems had been evaluated. At the end of the experiment the users were asked to select the interface method they preferred. Whilst making this choice, the users were told to disregard the fact that they were required to wear the eye gaze equipment and undergo a calibration process when using the ES system. Such an assumption is valid as currently there are unobtrusive eye gaze equipment (e.g. ASL Model 504 [10]) currently available and intense research is currently being carried out to achieve free calibration [11].

4. RESULTS

Based on table 1, it can be seen that there is a strong correlation between effectiveness (accuracy) and efficiency (task completion time) for the MK, ES and SO systems. However both of these aspects have a weak correlation with the user's preference. Another weak correlation between EQ and HQ can also be observed in the MK system. However, in both the ES and SO systems, a strong correlation between EQ and HQ can be observed. The experiment was unable to get the user's opinion on the HW system's accuracy. This is because time was required in order to carry out the offline handwriting recognition process.

	Accuracy	Completion time (sec)	Hedonic Quality (HQ)	Ease of Use (EQ)	User's opinion on accuracy	Naturalness	Familiarity
HW	11%	43	40%	85%	-	39%	83%
MK	96%	37	55%	90%	88%	38%	83%
SO	59%	131	71%	72%	63%	60%	34%
ES	72%	105	84%	80%	63%	76%	30%

Table 1: Performance of offline handwriting (HW), mouse and keyboard (MK), Speech only (SO) and Eye and Speech (ES) systems.

From the results obtained, it is obvious that the MK system allows the users to complete the form-filling task in the shortest time with the highest accuracy. The MK system also achieves the highest rating on ease of use (EQ) and users expressed the opinion that MK was the most accurate system among the other methods in the experiment.

However, despite the above achievements only 3 (23%) of the users preferred the MK system. All the users that selected the MK system are non-native speakers that achieved low speech recognition accuracy (25%) when using the ES system. In the worst case, a speaker achieved an accuracy of 0% for getting the correct word on his first attempt for each field. All the users that selected MK system were also found to be unable to complete the tasks of entering all the correct data using either the ES or SO system.

3 users (23%, 1 non-native and 2 native) selected the SO system as their preferred system. The remaining 7 users (54%, 2 non-native and 5 native) preferred using the ES system. This was despite the fact that the ES system only achieving a 72% average accuracy for getting the correct word at the first attempt. All the users were new to using the system that uses eye gaze tracking and therefore they expressed great interest in it.

5. DISCUSSIONS

The result for the MK system shows that by making the interface easy to use (high EQ), user will feel the interface to be uninteresting (low HQ). Such a finding is similar to the findings in [5] where it was argued that if the system is too simple or too complex, the user will feel bored or overloaded. However, for the systems that used speech (SO and ES), a correlation between HQ and EQ can be found. Such a finding differs from the findings in [5] mentioned earlier. This correlation indicates that achieving ease of use through a natural interaction, can lead to a simultaneous increase in HQ and EQ.

Although both the SO and ES system utilized the same speech recognition engine, the majority of users did not prefer the SO system because of the errors that occurred whilst the users were selecting the fields. A total of 14% of errors can be observed during the field selection process with the SO system. Removing the need to speak the field name, in the ES system, reduced the chances of encountering errors and increased the naturalness of the input method.

All the users that found the SO system preferable expressed the opinion that looking at the field and then speaking the data was an awkward action. In addition it was found that these users achieved a high speech recognition accuracy rate (85%), thus causing them to be unappreciative of the advantage of gained in the ES system by removing the need to carry out the field selection process.

Overall then, despite that MK system achieving the highest effectiveness (24% more accurate than ES), efficiency (68 seconds faster than ES) and ease of use, the majority of the users preferred the ES system. In order to investigate the reasons for their selection, the questions in table 2 were asked.

Reasons	MK	SO	ES
Total Users	3	3	7
R1: It's fast?	100%	100%	71%
R2: It's very accurate?	100%	100%	40%
R3: Familiar?	100%	33%	0%
R4: Natural?	0%	0%	57%
R5: Interesting and innovative?	0%	100%	100%

Table 2: Each user's reason for their preferred data entry method.

Based on table 2, it can be seen that the users were able to judge that the ES system is neither the fastest nor the most accurate method. It seems that HQ had a stronger effect on user's preference than performance (accuracy and speed). This is in accord with the findings in [8]. Although all the users agreed that they selected ES because it was interesting (HQ), this does not mean that this is the only reason for their preference. The other reasons did not score as highly as HQ, due to the fact that HQ is probably easier to perceive than the other reasons (R1-R4). Such an incidence was also reported in [5] where HQ aspect is much easier to perceive than EQ.

Based on the results in tables 1 and 2, it is justifiable to presume that the users preferred the ES system because it allows the user to experience a balance of effectiveness, efficiency, hedonic quality and naturalness. In addition, some of the users also expressed the opinion that they preferred to carry out the data entry tasks without the use of hands.

As users find that the ES is preferable, this will encourage them to use such a system more frequently. Although one may argue that HQ will decrease as the usage increases, at the same time this will lead to an increased time spent using the system so allowing the user to gain a better understanding of how the method works (increase familiarity), This, in turn, will allow the user to use the system more productively. When users are able to achieve this balance of naturalness, familiarity, speed and accuracy, this will lead to an overall increase the quality of the ES system.

All the users that selected MK did so because they achieved an unacceptable performance from the speech recognition engine. The incapability of the speech recognition also caused the effects of HQ and naturalness of the ES system to be weakened. Since the users were not able to experience a good balance of accuracy, speed, natural, hedonic quality and ease of use, they turned to the MK system that did posses the attributes of effectiveness, efficiency, ease of use and familiarity.

However, those users that did select the MK system also expressed the opinion that they would be willingly to select ES as their preferred form filling method, if the speech recognition component was able to present them with a better recognition performance. A few of the users also expressed the need for a better error-recovery method. This was because it was not always possible for them to recover from errors using re-speaking without elimination method. Currently, this is the only error recovery method used in the ES and SO systems.

6. CONCLUSION

Based on the experiment results, the majority of the users (7 out of 13) preferred the eye gaze and speech interface method even though such a method is neither the fastest, most accurate or the method with which they are most familiar. Based on the results obtained, it was found that users prefer a data entry method that posses a balance of accuracy, speed, naturalness, ease of use and hedonic quality compared to a method that is the fastest, most accurate, easy to use and that they are familiar with. However the effect of this balance on the user's preference can be weakened if an unbalance within the mixture occurs. For instance all the users that selected the mouse and keyboard method did so because the speech accuracy presented in eye and speech system was unacceptable thereby causing those users to be unable to complete their data entry task.

As with other research findings, it was found that an increase in ergonomic quality (easy to use) causes hedonic quality to decrease for the mouse and keyboard method. However different findings were observed in the systems that utilized speech (ES and SO). It was found that by implementing a natural input method into the eye and speech system, both ergonomic and hedonic quality improved at the same time.

Although the proposed eye and speech system presented in this paper presents positive results, further evaluation still needs to be carried out on a larger population of users with better demographic distribution. In addition, work on investigating new ways of using eye gaze and/or speech for error detection and error recovery in order to increase the success rate for non-native speakers when used for form-filling tasks needs to be carried out.

7. ACKNOWLEDGEMENTS

We would like to thank all the subjects, especially the researchers in the Faculty Research Institute and the GreenHat group and, who volunteered for the experiment. We would also like to express our deepest gratitude to Dr Jean Underwood and the technician from the Psychology Department for loaning us the eye gaze equipment.

8. REFERENCES

[1] Zhai, S., Morimoto, C., & Ihde, S. (1999). Manual and gaze input cascaded (MAGIC) pointing. Proc. of the CHI'99. ACM Press: New York, 246-253.

[2] Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., & Zhou, J., Audio-visual speech recognition. Technical Report WS00AVSR, Johns Hopkins University, CLSP, 2000.

[3] Tan, Y. K., Sherkat, N., Allen, T, Speech & Telephone Keypad for Data Entry Tasks, Proc. of AVIOS 2002, p. 191-210.

[4] Jordan, P. (1998) Human Factors for pleasure in product use. Applied Ergonomics, 29[1], p. 25-33.

[5] Hassenzahl M., Platz, A., Burmester, M & Lehner, K. Hedonic & Ergonomic Quality Aspects Determine a Software's Appeal, Proc. of CHI 2000.

[6] Carrol, J. M. & Thomas, J. C. Fun. SIGCHI Bulletin 19 3 (1988), 21-24.

[7] Computers & Fun, York, December 2000, http://www-users.york.ac.uk

[8] Igbaria, M., Schiffman, S. J. & Wieckowski, T. J. The respective roles of perceived usefulness & perceived fun in the acceptance of microcomputer technology. Behaviour & Technology 13 6 (1994), 349-361.

[9] Evans, R G. Sherkat, N. Whitrow, R J. Holistic Recognition of Static handwriting Using Structural Features. Document Image Processing and Multimedia 99, IEE Colloquium 99/041, P. 121-124.

[10] http://www.a-s-l.com/504_home.htm

[11] Stiefelhagen, R., Jie Yang, & Waibel, A. "Simultaneous tracking of head poses in a panoramic view", International Conference on Pattern Recognition 2000.