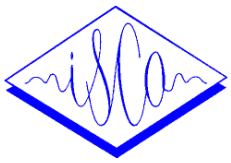


3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

FACTORS AFFECTING CHOICE OF SPEECH OVER KEYBOARD AND MOUSE IN A SIMPLE DATA-RETRIEVAL TASK

Alexander I. Rudnicky

School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213 USA

ABSTRACT
This paper describes some recent experiments that assess user mode selection behavior in a multi-modal environment in which actions can be performed with equivalent effect by speech, keyboard or scroller. Results indicate that users freely choose speech over other modalities, even when it is less efficient in objective terms, such as time-to-completion or input error. Additional evidence indicates that users appear to focus on simple input time in making their choice of mode, in effect minimizing the amount of personal effort expended.

KEYWORDS: Speech recognition, multi-modal systems, user preference.

1. INTRODUCTION

Multi-modal systems allow users to both tailor their input style to the task at hand and to use input strategies that combine several modes in a single transaction. As yet no consistent body of knowledge is available for predicting user behavior in multi-modal environments or to guide the design of multi-modal systems. This is particularly true when interfaces incorporate new technologies such as speech recognition.

For activities in a workstation environment, formal comparisons of speech with other input modes have failed to demonstrate a clear advantage for speech on conventional aggregate measures of performance such as time-to-completion [1, 9, 5], despite a consistent advantage displayed by speech at the level of single input operations. The difference can actually be attributed to the additional incurred costs of non-real-time recognition and error correction. While real-time performance can be achieved, it is unlikely that error-free recognition will be available in the near future. Given these shortcomings, we might ask if speech can provide advantages to the user along dimensions other than task speed and whether such advantages are more salient to the user than overall task time. For example, one such advantage might be a reduction in the effort necessary to generate an input.

There is reason to believe that users are quite good at estimating the response characteristics of an interface and will choose an input strategy that optimizes those aspects of performance that are of interest to them, for example decreasing time-to-completion or minimizing task error [6, 10]. By observing the behavior of users in a situation in which they can freely choose between different strategies, we can gain insight into the factors that govern their preference for different input styles.

A simple data retrieval task was chosen for this study, as the task was one amenable to execution in each of the three modalities that were examined: speech, keyboard and scroller. The database contained information about individuals, such as address, telephone, etc selected from a list of conference attendees. The task consisted of retrieving the record for an individual and recording the last group of digits in their work telephone number (typically of length four). The database contained 240 names.

2. SYSTEM IMPLEMENTATION

The Personal Information Database (PID) component of the OM system [4, 8] served as the database system in this study. Given a search request specified in some combination of first name, last name and affiliation, PID displays a window with the requested information (in this study, the information consisted of name, affiliation and all known telephone numbers). If an unknown name was entered, an error panel came up. If a query was underspecified, a choice panel containing all entries satisfying the query was shown; for example asking for "Smith" produced a panel showing all Smiths in the database. The existing PID was altered to incorporate a scroll window in addition to the already available keyboard and speech interfaces. The remainder of this section provides detailed descriptions for each input mode.

Speech Input

The OM system uses a hidden Markov model (HMM) recognizer based on Sphinx [3] and is capable of speaker-independent continuous speech recognition. The subject interacted with the system through a NeXT computer which provided attention management [4] as well as application-specific displays. To offload computation, the recognition engine ran on a separate computer (an IBM 6000/530, recognition speed was 1.5 xRT) and communicated through an Ethernet connection. Database retrieval was by a command phrase such as `SHOW ME ALEX RUDNICKY`. While subjects were instructed to use this specific phrase, the system also understood several variants, such as `SHOW`, `GIVE (ME)`, `LIST`, etc. which users could and did use on occasion. The input protocol was "Push and Hold", meaning that the user had to depress the mouse button before beginning to speak and release it after the utterance was complete. In those conditions that required all input to be by speech, subjects were instructed to keep repeating a spoken command in case of recognition error, until it was processed correctly and the desired information appeared in the result window.

Keyboard

Subjects were required to click a field in a window then type a name into it, followed by a carriage return (which would drop them to the next field or would initial the retrieval). Three fields were provided: First name, Last Name and Organization. Subjects were provided with some shortcuts: last names were often unique and might be sufficient for a retrieval. They were also informed about the use of a wildcard character which

Table 1. *User mode preference in the Free block. Mixed mode refers to trials in which multiple transactions occurred, not all using the same input mode; for example, the first input might be speech, the second one keyboard.*

Transaction Mode	Mode Choice (%)	Filtered Choice (%)
Scroller	5.8	4.4
Keyboard	14.2	11.3
Voice	74.9	79.9
mixed	5.1	4.4

Table 2. *Times (in sec) (using unfiltered data). The input time for voice is the utterance duration.*

Mode	Transaction	Input Time
Scroller	10.863	4.306
Keyboard	9.560	2.942
Voice	9.463	2.029

would allow then to minimize the number of keystrokes need for a retrieval. Ambiguous search patterns produced a panel of alternatives; the subject could click on the desired one.

Scroller
The scroller window displayed the names in the database sorted alphabetically by last name. Eleven names were visible in the window at any one time, providing approximately 4–5% exposure of the 240 name list. The NeXT scroller provides a handle and two arrow buttons for navigation. Clicks on the scrollbar move the window to the corresponding position in the text and the arrow buttons can be amplified to jump by page when a control key is simultaneously depressed. Each navigation technique was demonstrated to the subject.

Session controller
The experiment was controlled by a separate process visible to the subject as a window displaying a name to look up, a field in which to enter the retrieved information and a field containing special instructions such as `Please use KEYBOARD only` or `Use any mode`. The subject progressed through the experiment by clicking a button in this window labeled `Next`; this would display the next name to retrieve. Equidistant from the the `Next` button were three windows corresponding to the three input modes used in the experiment: voice, keyboard and scroller. All modes required a mouse action to initiate input, either a click on the speech input button, a click on a text input field or button in the keyboard window or the (direct) initiation of activity in the scroller.

All applications were instrumented to generate a stream of time-stamped events corresponding to user and system actions. Logged events were time-stamped using absolute system time, then merged in analysis to produce a composite timeline corresponding to the entire experimental session. Additional details of the logging procedure are provided in [7].

3. EXTENDED EXPERIENCE

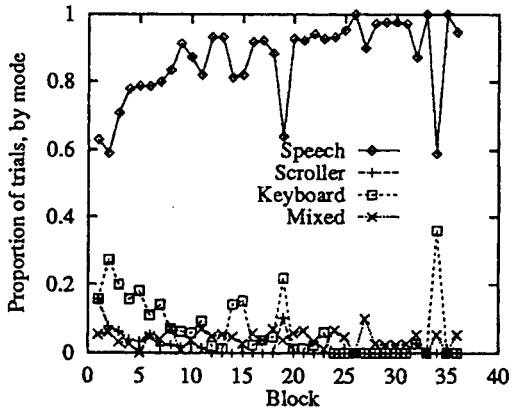
An initial experiment (described in [7]) indicated that users would prefer speech 63% of the time when given free choice among modes. That experiment, however, provided subjects with only a limited exposure to speech input. A possible explanation, therefore, is that the subjects's choice behavior reflected a novelty effect. That is, users displayed a preference for speech input in this task not because of any inherent preference or benefit but simply because it was something new and interesting. Over time we might expect the novelty to wear off and users to refocus their attention on system response characteristics and perhaps shift their preference.

To test this possibility, the current experiment scales up the amount of time spent on the task, by different amounts. Since it was not possible to predict the length of a novelty effect *a priori*, three separate experience levels were examined. A total of 9 subjects participated (4 male and 5 female): 3 did 720 trials, 3 did 1440 trials and 3 did 2160. This is in contrast to the 115 trials per subject in the initial experiment.

3.1. Method

The experiment was divided into blocks of 60 trials. Each block consisted of 15 required trials in each of the three input modes (speech, keyboard, scroller) followed by 15 "free" trials during which subjects could choose the input mode. The required trials ensured that subjects would continue to be aware of the characteristics of each mode over the course of the experiment. The order in which modes were set was counterbalanced over blocks.

Figure 1. *User preference over blocks (filtered data). Note that the spikes at blocks 19 and 34 are due to equipment failure.*

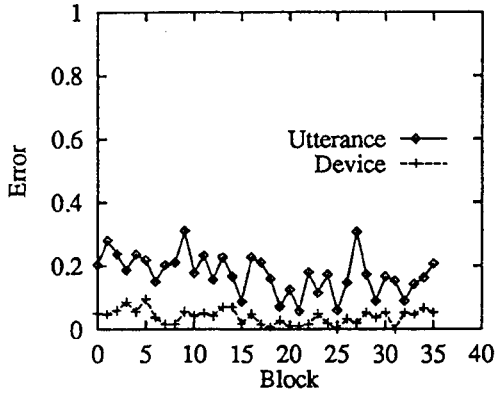


3.2. Results and Analysis

Recognition performance for the system used in this experiment was estimated from the correction behavior exhibited by subjects; if a transaction was repeated, this implied that on the original attempt the system misrecognized the utterance. Using this definition recognition error was 20%, which corresponds to approximately a 5% word error rate. In addition, it was possible to identify 3.8% of speech inputs as containing "device errors", meaning that subjects did not correctly follow the push-and-hold protocol. Error rates did not change substantially over the course of the experiment, though a trend towards decreasing recognition error can be seen (see Figure 2).

The mean preference for different modes in this experiment is shown in Table 1. Subjects display a strong bias in favor of voice input (74.9%). Preference for voice across individual subjects ranged from 28% to 91% with all but one subject (S3) showing preference levels above 70% (the median prefer-

Figure 2. *Utterance and device error over the course of the experiment.*



ence is 82.5%). Differences in mode preference are significant ($F(2,16) = 34.6, MS_{err} = 0.037, p < 0.01$) and the preference is greater ($p < 0.01$, by Newman-Keuls) for voice than for either of the other input modes.

Since some of the names in the database were difficult to pronounce, we also tabulated choice data excluding such names. Nineteen names (about 8% of the database) were excluded on the basis of ratings provided by subjects.[1] The data thus filtered are shown in Table 1; in this case (for names that subjects were reasonably comfortable about pronouncing) preference for speech rises to 79.9% (median of 86.1%). Clearly, subjects did not show a blind preference for speech input but were acting to optimize the overall efficiency of their performance.

The aspect of performance optimized by subjects, however, may not correspond to overall task time. Table 2 shows the mean transaction and input times for the second experiment, computed over subjects. Transaction times are significantly different ($F(2,16) = 16.8, MS_{err} = 0.327, p < 0.01$), with scroller times longer than keyboard or speech times ($p < 0.01$) which in turn are not different. If subjects were attending to the time necessary to carry out the task, keyboard and voice should have been chosen with about equal frequency. Nevertheless, he subjects in this experiment decisively chose speech over keyboard (and scroller) input.

The time needed to carry out the task decreased by almost a half over the course of the experiment (see Figure 3), showing the classic linear decrease in log time (see, e.g., [2, p.57]). The decreases in transaction time came from shorter lags in initiating responses and in the time necessary for entering the retrieved telephone number. We can assume that over the course of the experiment subjects developed and refined a variety of strategies for carrying out their assigned task; the preference for speech input was maintained over this interval.

[1]Participants in this experiment rated each name in the database prior to the experiment itself. A name was presented to the subject, who was asked to rate on a 4-point scale their lack of confidence in their ability to pronounce it. They then heard a recording of the name pronounced as expected by the recognizer and finally rated the degree to which this canonical pronunciation disagreed with their own expectation. A conservative criterion was used to place names on the exclusion list: any name for which both ratings averaged over 1.0 (on a 0–3 scale) was excluded.

Table 3. *Times (in sec) and choice for lengthened input utterances. Speech Advantage notes the difference between speech and keyboard input (positive differences represent a speech advantage).*

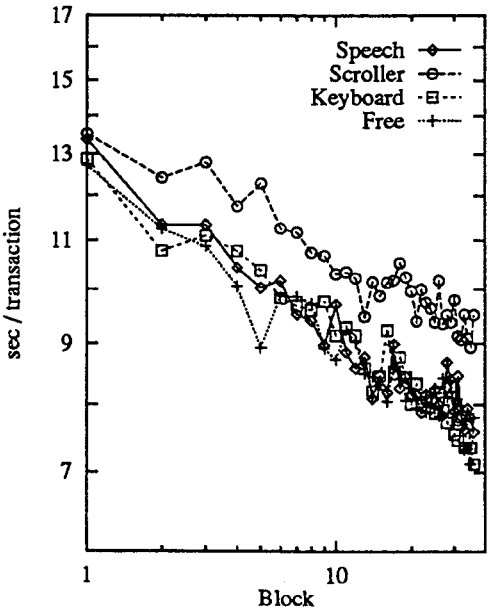
Input Duration	Speech (sec)	Speech Advantage (sec)	Choice (%)
Short	2.029	0.913	81
Medium	2.525	0.448	63
Long	3.002	−0.457	14

Figure 1 shows preference for voice input over the course of the experiment. Preference for speech increases over time, and begins to asymptote at about 10–15 blocks (representing about 250 utterances). This phenomenon suggests that speech input, while highly appealing to the user requires a certain amount of confidence building, certainly a period of extended familiarization with what is after all a novel input mode. Additional investigation would be needed, however, to establish the accuracy of this observation. In any case, this last result underlines the importance of providing users with sufficient training when introducing a new input technology.

As can be seen in Figure 1 that preference for speech shows no sign of decreasing over time for the duration examined in this experiment. Preference for voice input appears to be robust. The 36 block version of the experiment took on the average 8–9 hours to complete, with subjects working up to 2 hours per day.

A possible explanation for this finding may be that, rather than basing their choice on overall transaction time, users focus on simple input time (in both experiments voice input is the fastest). This would imply that users are willing to disregard the cost of recognition error, at least for the error levels associated with the system under investigation.

Figure 3. *Transaction time for required and free portions, across blocks.*



4. FACTORS GOVERNING CHOICE

The results of the above experiment suggest that preference for speech might be attributable to some inherent predisposition to use speech. However, it would be preferable to link this preference to some external, observable correlate of the situation experienced by the subject. Table 2 shows the input times for the three modes in the previous experiment. Input time is the actual time taken to complete an input, i.e., uttering a command, keying in a name or scrolling to the right name. Note that the time needed for the input of a single speech command is appreciably lower than for either scroller or keyboard. We might conjecture that subjects in this experiment attended to simple input time rather than total transaction time in choosing a preferred input mode. Although speech input is errorful, it nevertheless offers the chance to complete the transaction with less effort than by say keyboard.

To test this possibility, the previous experiment was replicated, but with the (time) cost of using speech increased by extending the length of the carrier phrase used in the voice command (for example, PLEASE SHOW ME ALEX RUDNICKY and PLEASE SHOW ME THE RECORD FOR ALEX RUDNICKY). The mean duration of the input utterance increases correspondingly, as can be seen in Table 3. The original experiment was replicated, using two new groups of 9 subjects each but otherwise keeping the design identical. Each subject completed 24 blocks of the task. The resulting choice behavior is also shown in Table 3. As can be seen, choice follows the relative advantage for speech on the dimension of input duration.

Figure 4. *Preference for speech with Short, Medium and Long input utterances.*

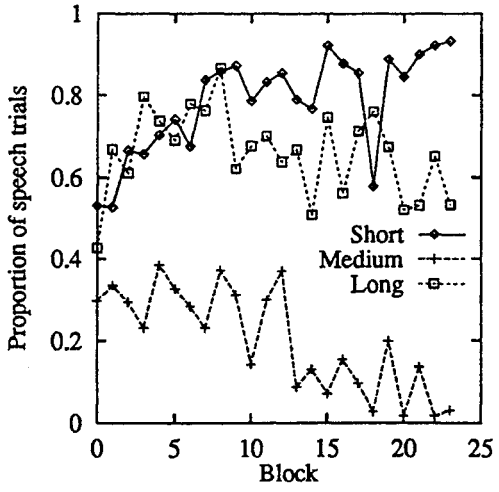


Figure 4 shows the percent of free transactions for Short, Medium and Long input utterances. The Medium group shows less overall preference for voice, though the level of preference appears to hold over the duration of the experiment. The Long group shows an initial growth in preference, followed by a decline. The initial component appears to follow the familiarization behavior noted previously.

It is clear that in this experiment subjects functioned in a rational manner, they chose the input mode that optimized the criteria that were important to them (and there appears to be remarkable unanimity on what these are), they based their choice on evidence accumulated over time and they tailored their strategy to the characteristics of the task (e.g., difficult to pronounce names were typed in, despite an overall preference for speech). We should expect that in other situations users will also go through a stage of rational evaluation. The success of a speech interface will depend on the outcome of such an assessment. The current study provides some insight into the factors that users take into consideration.

5. CONCLUSION

The study reported in this paper indicates that users show a preference for speech input despite its inadequacies in terms of classic measures of performance, such as time-to-completion. Subjects in this study based their choice of mode on attributes other than transaction time (quite possibly input time) and were willing to use speech input even if this meant spending a longer time on the task. This preference appears to persist and even increase with continuing use, suggesting that preference for speech cannot be attributed to short-term novelty effects.

REFERENCES

[1] BIERMANN, A. W., FINEMAN, L., AND HEIDLAGE, J. F. A voice- and touch-driven natural language editor and its performance. *International Journal of Man-Machine Studies 37* (1992), 1–21.

[2] CARD, S. K., MORAN, T. P., AND NEWELL, A. *The psychology of human-computer interaction.* Erlbaum, Hillsdale, N.J., 1983.

[3] LEE, K.-F. *Automatic Speech Recognition: The Development of the SPHINX System.* Kluwer Academic Publishers, Boston, 1989.

[4] LUNATI, J.-M., AND RUDNICKY, A. I. The design of a spoken language interface. In *Proceedings of the Third Darpa Speech and Natural Language Workshop* (Hidden Valley, June 1990), Morgan Kaufmann, San Mateo, CA, 1990, pp. 225–229.

[5] MARTIN, G. The utility of speech input in user-computer interfaces. *International Journal of Man-Machine Studies 29* (1989), 355–376.

[6] RUDNICKY, A. System response delay and user strategy selection in a spreadsheet task. CHI'90, invited poster, April 1990.

[7] RUDNICKY, A. I. Mode preference in a simple data-retrieval task. In *Proceedings of the Arpa Workshop on Human Language Technology* (Princeton, NJ, March 1993), Morgan Kaufmann, San Mateo, CA, 1993, p. in press.

[8] RUDNICKY, A. I., LUNATI, J.-M., AND FRANZ, A. M. Spoken language recognition in an office management domain. *Proceedings of ICASSP* (May 1991), 829–832.

[9] RUDNICKY, A. I., SAKAMOTO, M. H., AND POLIFRONI, J. H. Spoken language interaction in a spreadsheet task. In *Human-Computer Interaction – INTERACT'90*, D. Diaper et al., Eds. Elsevier, 1990, pp. 767–772.

[10] TEAL, S. L., AND RUDNICKY, A. I. A performance model of system delay and user strategy selection. In *Proceedings of CHI* (Monterey, CA, May 1992), ACM, New York, 1992, pp. 295–206.