Mode preference 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 behavior in a multi-modal environment in which actions can be performed with equivalent effect in speech, keyboard or scroller modes. 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.

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, 8, 4], 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, for example by reducing the effort needed to generate an input.

There is reason to believe that users are quite good at estimating the response characteristics of an interface and can choose an input strategy that optimizes salient aspects of performance, for example decreasing time-to-completion or minimizing task error [5, 9].

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 225 names for the first experiment and was expanded to 240 names for the second experiment.

SYSTEM IMPLEMENTATION

The Personal Information Database (PID) component of the OM system [3, 7] 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 [2] and is capable of speaker-independent continuous speech recognition. The subject interacted with the system through a

NeXT computer which provided attention management [3] as well as application-specific displays. To offload computation, the recognition engine ran on a separate NeXT computer and communicated through an ethernet connection. For the 731-word vocabulary and perplexity 33 grammar used in the first experiment, the system responded in 2.1 times real-time (xRT). 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. 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. 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 would allow then to minimize the number of keystrokes need for a retrieval. Ambiguous search patterns produced a panel of choices; the subject could click on the desired one.
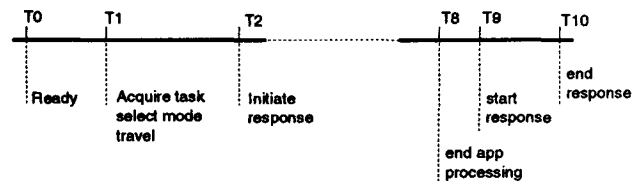
Scroller

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 225 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

Figure 1: *Trial time line, showing events logged by the control program.*

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.

Instrumentation

All applications were instrumented to generate a stream of time-stamped events corresponding to user and system actions. Figure 1 shows the time line for a single trial. In addition to the overall time-line, each mode was also instrumented to generate logging events corresponding to significant internal events. All logged events were time-stamped using absolute system time, then merged in analysis to produce a composite timeline corresponding to the entire experimental session.

The merged event stream was processed using a hierarchical set of finite-state machines (FSMs). Figure 2 shows the FSM for a single transaction with the database retrieval program. Figures 3 show the FSM for the voice mode. During the analysis process, the latter FSM (as well as FSMs for keyboard and scroller) would be invoked within state 1 of the transaction FSM (Figure 2). An intermediate level of analysis (corresponding to conditions) is also used to simplify analysis. Arcs in the FSMs correspond to observable events, either system outputs or user inputs. The products of the analysis include transition frequencies for all arcs in an FSM as well as transition times. The analysis can be treated in terms of Markov chains [6] to compactly describe recognition error, user-mode preferences and other system characteristics.

USER MODE PREFERENCE IN DATA RETRIEVAL

The purpose of the first experiment was to establish what mode-preference patterns users would display when using the PID system. To ensure that subjects

Figure 2: *FSM for a single transaction. From the initial state (0) the subject can click the Next button to move to state 1 at which point the subject has a name to look up and can initiate a query. Queries are described by mode-specific FSMs which are invoked within this state. Figure 3 shows one such FSM. If properly formed, a query will produce a database retrieval and move the transaction to state 4. The subject can opt to enter a response, moving the transaction to state 2 or to repeat queries (by re-entering state 1). At this point, the subject is ready to begin a new trial by transitioning to state 0.*
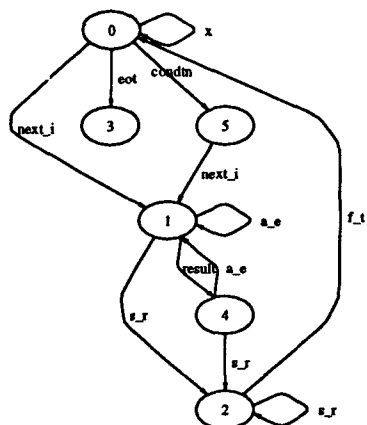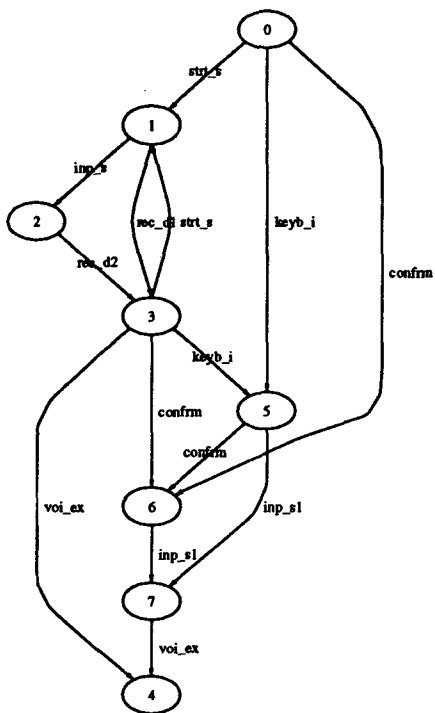

Figure 3: *FSM used for the analysis of voice input.*

were equally familiar with each of the input modes, the experiment was divided into two parts (although it was run as a single session, without breaks). In the first part, subjects were asked to perform 20 retrievals using each mode. Initial testing determined that this was sufficient to acquaint the subjects with the operation of each mode. In the second part, they were instructed to use "any mode", with the expectation that they would choose on the basis of their assessment of the suitability of each mode. A total of 55 entries were presented in the second part.

The same sequence of 60 entries was used for the familiarization stage for all subjects. However, the order in which the subject was exposed to the different modes was counter-balanced according to a Latin square. Three different blocks of test items (each containing 55 entries) were used, for a total of nine different combinations.

Details about the operation of the different modes as well as the experiment controller were explained to the subject during a practice session prior to the experiment proper (a total of four practice retrievals were performed by the subject in this phase).

Subjects

Nine subjects participated in this study, 7 male and 2 female. All had had some previous exposure to speech systems, primarily through their participation in ongoing speech data collection efforts conducted by our research group. This prior exposure ensured that the subjects were familiar with the mechanics of using a microphone and of interacting with a computer by voice. No attempt was made to select on demographic characteristics or on computer skills. The group consisted primarily of students, none of whom however were members of our research group.

Results and Analysis

A finite state machine (FSM) description of user behavior was used to analyze session data. Separate FSMs were defined for condition, transaction, sequence and intra-modal levels and were used to tabulate metrics of interest.

Table 1 shows the durations of transactions for each of the modes during the familiarization phase. A transaction is timed from the click on the Next button to the carriage return terminating the entry of the retrieved telephone number. Speech input leads to the longest transaction times. Input time measures the duration between the initiation of input and system response (note that these times include recognition time, as well as the consequences of mis-recognition,

Table 1: *Times (in sec) for the familiarization blocks in the first experiment.*

Mode	Transaction	Input	Utterance duration
Scroller	13.623	4.917	—
Keyboard	14.526	5.371	—
Voice	15.041	5.593	2.464

Table 2: *User mode choices in the Free block (trials 61-115).*

Transaction Mode	Choice (%)	First Choice (%)
Scroller	14.3	14.7
Keyboard	21.8	22.4
Voice	48.3	62.8
mixed	15.5	—

Table 3: *User mode preference in the Free block of the second experiment.*

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

Table 4: *Times (in sec) for the second experiment (using unfiltered data). The input time for voice is the utterance duration.*

Mode	Transaction	Input
Scroller	10.863	4.394
Keyboard	9.560	3.035
Voice	9.463	2.078

i.e., having to repeat an input). Here speech is also at a disadvantage (though note that the duration of a single utterance is only 2.464 sec). Transaction durations for modes are statistically different $(F(2,14) = 5.54, MS_{err} = 0.836, p < 0.05)$, though in individual comparisons only voice and scroller differ $(p < 0.05$, the Neuman-Keuls procedure was used for this and all subsequent comparisons). Order of presentation was a significant factor $(F(2,14) = 8.3, p < 0.01)$, with the first mode encountered requiring the greatest amount of time.

Table 2 shows choice of mode in the Free block. The mixed mode line refers to cases where subjects would first attempt a lookup in one mode then switch to another (for example because of misrecognition in the speech mode). The right-hand column in the table shows the first mode chosen in a mixed-mode transaction. In this case, voice is preferred 62.8% of the time as a first choice. The pattern of choices is statistically significant $(F(2,14) = 6.31, MS_{err} = 288, p < 0.01)$, with speech preferred significantly more than either keyboard or scroller$(p < 0.05)$.

This experiment suggests that speech is the preferred mode of interaction for the task we examined. This is particularly notable since speech is the least efficient of the three modes offered to the user, as measured in traditional terms such as time-to-completion. Most previous investigations (see, e.g. the review in [4]) have concentrated on this dimension, treating it as the single most important criterion for the suitabil-

ity of speech input. The present result suggests that other aspects of performance may be equally important to the user.

EXTENDED EXPERIENCE

One possible explanation of the above result is that it's due to 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, we performed a second experiment, scaling up the amount of time spent on a 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 first experiment.

Method

Based on observations made during the first experiment, several changes were made to the system, primarily to make the speech and keyboard inputs more efficient. Recognition response was improved from 2.1 xRT to 1.5 xRT by the use of an IBM 6000/530 computer as the recognition engine. Keyboard entry was made more efficient by eliminating the need for the user to clear entry fields prior to entry. These changes

resulted in improved transaction times for these two modes relative to the scroller, which was unchanged except for a slight reduction in exposure (this due to an increase of the number of entries to 240, done to facilitate details of the design).

Figure 4: *User preference over blocks (filtered data). Note that the spikes at blocks 19 and 34 are due to equipment failure.*
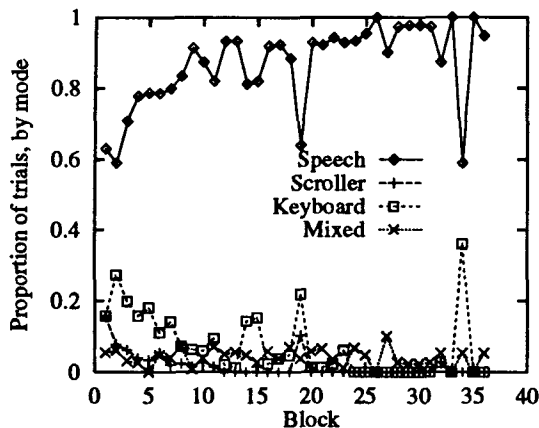
Results and Analysis

The mean preference for different modes in this experiment is shown in Table 3. 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 preference 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 $(p < 0.01)$ 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 3; in this case (for names that subjects were reasonably comfortable about pronouncing) preference for speech rises to 79.9% (median of 86.1%).

---

[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 the 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 4 shows the mean transaction and input times for the second experiment, computed over subjects. Compared to the first experiment, these times are faster, probably reflecting the greater amount of experience with the task for the second group of 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. The subjects in this experiment nevertheless chose speech over keyboard (and scroller) input.

Figure 4 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 sufficient training.

As can be seen in Figure 4 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. Data from followup experiments not reported here suggest that this may be the case: increasing the duration of the query utterance decreases the preference for speech.

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 spend-

ing 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.

This paper also sketches an analysis technique based on FSM representations of human–computer interaction that permits rapid automatic processing of long event streams. The statistical properties of these event streams (as characterized by Markov chains) may provide insight into the types of information that users themselves compute in the course of developing satisfactory interaction strategies.

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

[3] 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.

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

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

[6] RUDNICKY, A. I., AND HAUPTMANN, A. G. Models for evaluating interaction protocols in speech recognition. In *Proceedings of CHI* (New Orleans, Louisiana, April 1991), ACM, New York, 1991, pp. 285–291.

[7] 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.

[8] 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.

[9] 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.