Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems

Clare-Marie Karat, Christine Halverson, Daniel Horn*, and John Karat

IBM T.J. Watson Research Center
30 Saw Mill River Road
Hawthorne, NY 10532 USA
+1 914 784 7612
ckarat, halve, jkarat@us.ibm.com

*University of Michigan
Collaboratory for Research on Electronic Work
701 Tappan Street, Room C2420
Ann Arbor, MI 48109-1234
danhorn@umich.edu

ABSTRACT

A study was conducted to evaluate user performance and satisfaction in completion of a set of text creation tasks using three commercially available continuous speech recognition systems. The study also compared user performance on similar tasks using keyboard input. One part of the study (Initial Use) involved 24 users who enrolled, received training and carried out practice tasks, and then completed a set of transcription and composition tasks in a single session. In a parallel effort (Extended Use), four researchers used speech recognition to carry out real work tasks over 10 sessions with each of the three speech recognition software products. This paper presents results from the Initial Use phase of the study along with some preliminary results from the Extended Use phase. We present details of the kinds of usability and system design problems likely in current systems and several common patterns of error correction that we found.

Keywords

Speech recognition, input techniques, speech user interfaces, analysis methods

INTRODUCTION

Automatic speech recognition (ASR) technology has been under development for over 25 years, with considerable resources devoted to developing systems which can translate speech input into character strings or commands. We are just beginning to see fairly wide application of the technology. Though the technology may not have gained wide acceptance at this time, industry and research seem committed to improving the technology to the point that it becomes acceptable. While speech may not replace other input modalities, it may prove to be a very powerful means of human-computer communication.

However, there are some fundamental factors to keep in mind when considering the value of ASR and how rapidly and widely it will spread. First, speech recognition technology involves errors that are fundamentally different from user errors with other input techniques [2]. When users press keys on a keyboard, they can feel quite certain of the result. When users say words to an ASR system, they may experience system errors – errors in which the system output does not match their input – that they do not experience with other devices. Imagine how user behavior might be different if keyboards occasionally entered a random letter whenever you typed the “a” key. While there is ongoing development of speech recognition technology aimed at lowering error rates, we cannot expect the sort of error free system behavior we experience with keyboards in the near future. How we go from an acoustic signal to some useful translation of the signal remains technically challenging, and error rates in the 1-5% range are the best anyone should hope for.

Second, while we like to think that speech is a natural form of communication [1,9] it is misleading to think that this means that it is easy to build interfaces that will provide a natural interaction with a non-human machine [10]. While having no difference between human-human and human-computer communication might be a laudable goal, it is not one likely to be attainable in the near future. Context aids human understanding in ways that are not possible with machines (though there are ongoing efforts to provide machines with broad contextual and social knowledge) [7]. A great deal of the ease we take for granted in verbal communication goes away when the listener doesn't understand the meaning of what we say.

Finally we argue that it takes time and practice to develop a new form of interaction [4,6]. Speech user interfaces (SUIs) will evolve as we learn about problems users face with current designs and work to remedy them. The systems described in this paper represent the state-of-the-art in large vocabulary speech recognition systems. They provide for continuous speech recognition (as opposed to isolated word recognition), require speaker training for acceptable performance, and have techniques for distinguishing commands from dictation.

Text Creation and Error Correction

We are particularly interested in text creation by knowledge workers – individuals who "solve problems and generate outputs largely by resort to structures internal to themselves rather than by resort to external rules or procedures [5]." Text – in the form of reports or communication with others – is an important part of this output. While formal business communications used to pass through a handwritten stage before being committed to a typed document, this seems to be becoming less frequent. Knowledge workers who used to rely on secretarial help are now more likely to produce their own text by directly entering it into a word processor. We do not have a clear picture of how changes in the processes of text creation have impacted the quality of the resulting text, even though it seems that much of the text produced by knowledge workers – from newspaper articles to academic papers – is now created in an electronic form.

Efforts to develop new input technologies continue. ASR is clearly one of the promising technologies. We do not know how a change in modality of entry might impact the way in which people create text. For example, does voice entry affect the composition process? There is some suggestion that it does not impact composition quality [3,8]. Have people learned to view keyboards as "more natural" forms of communication with systems? While people can certainly dictate text faster than they can type, throughput with ASR systems is generally slower. Measures which include the time to make corrections favor keyboard-mouse input over speech – partially because error correction takes longer with speech. Some attempts have been made to address this in current systems, but the jury is still out on how successful such efforts have been.

Error detection and correction is an important arena in which to examine modality differences. For keyboard-mouse entry there are at least two ways in which someone might be viewed as making an error. One can mistype something – actually pressing one sequence of keys when one intended to enter another. Such user errors can be detected and corrected either immediately after they were made, within a few words of entry, during a proofreading of the text, or not at all. Another error is one of intent, requiring editing the text. In both cases, correction can be made by backspacing and retyping, by selecting the incorrect text and retyping, or by dialog techniques generally available in word processing systems such as Find/Replace or Spell Checking. While we do not have a clear picture of the proportion of use of the various techniques available, our observations suggest that all are used to some extent by experienced computer users.

There are some parallels for error correction in ASR systems. By monitoring the recognized text, users can correct misrecognitions with a speech command equivalent of "backspacing" (current systems generally have several variations of a command that remove the most recently recognized text – such as SCRATCH or UNDO. There are ways of selecting text (generally by saying the command SELECT and the string to be located), after which redictating will replace the selected text with newly recognized text. Additionally, correction dialogs provide users with a means of selecting a different choice from a list of possible alternatives or entering a correction by spelling it. These different correction mechanisms provide a range of techniques that map well to keyboard-mouse techniques. However, we do not have evidence of how efficient or effective they are. This study was designed to answer these questions. We were interested in several comparisons – keyboard and speech for text entry, modality effects on transcription and composition tasks, and error correction in different modalities.

SYSTEMS

Three commercially available large vocabulary continuous speech recognition systems were used in this study. All were shipped as products in 1998. These systems were IBM ViaVoice 98 Executive, Dragon Naturally Speaking Preferred 2.0, and L&H Voice Xpress Plus (referred to as IBM, Dragon and L&H below). While the products are all different in significant ways, they share a number of important features that distinguish them from earlier ASR products. First, they all recognize continuous speech. Earlier versions required users to dictate using pauses between words. Second, all have integrated command recognition into the dictation so that the user does not need to explicitly identify an utterance as text or command. In general, the systems provide the user with a command grammar (a list of specific command phrases), along with some mechanism for entering the commands as text. Commands can be entered as text by having the user alter the rate at which the phrase is dictated – pausing between words causes a phrase to be recognized as text rather than as a command.

While all of the systems function without specific training of a user's voice, we found the speaker independent recognition performance insufficiently accurate for the purposes of our study. To improve recognition performance, we had all users carry out speaker enrollment – the process of reading a body of text to the system and then having the system develop a speaker-specific speech model. All products require a 133-166MHz Pentium processor machine with 32MB RAM – we ran our study using 200MHz machines with 64MB RAM.

METHOD

There were different procedures used for the Initial Use and the Extended Use subjects in the study. Although the design of the Initial Use study was constructed to allow for statistical comparisons between the three systems, we report on general patterns observed across the systems as they are of more general interest to the design of successful ASR systems.

Initial Use

Subjects in the Initial Use study were 24 employees of IBM in the New York metropolitan area who were knowledge workers. All were native English speakers and experienced computer users with good typing skills. Half of the subjects were male and half were female, with gender balanced across the conditions in the study. The age range of the subjects was from 20 to 55 years old. An effort was made to balance the ages of the subjects in the various conditions. Each subject was assigned to one of three speech recognition products, IBM, Dragon, or L&H. Half of the subjects completed the text creation tasks using speech first and then did a similar set using keyboard-mouse, and half did keyboard-mouse followed by speech. Subjects received a $75 award for their participation in the three hour long session. All sessions were videotaped.

On arrival at the lab, the experimenter introduced the subject to the purpose, approximate length of time, and content of the usability session. The stages of the experimental session were:

1. Provide session overview and introduction.
2. Enroll user in assigned system.
3. Complete text tasks using first modality.
4. Complete text tasks using second modality.
5. Debrief the user.

The experimenter told the subject to try and complete the tasks using the product materials and to think aloud during the session. (While this could cause interference with the primary task, our subjects switched between think aloud and task modes fairly easily.) The experimenter explained that assistance would be provided if the subject got stuck. The experimenter then left the subject and moved to the Control Room. The subject's first task was to enroll in the ASR system (the systems were pre-installed on the machines). Enrollment took from 30 minutes to 1.5 hours for the subject to complete, depending on the system and the subject's speed in reading the enrollment text. After enrollment was completed, the subject was given a break while the system developed a speech model for the subject by completing an analysis of the speech data. After the break, the subject attempted to complete a series of text creation tasks. All text was created in each product's dictation application that provided basic editing functions (similar to Windows 95 WordPad), and did not include advanced functions such as spelling or grammar checkers.

Before engaging in the speech tasks, all participants underwent a training session with the experimenter present to provide instruction. This session was standardized across the three systems. Basic areas such as text entry and correction were covered. Each subject dictated a body of text supplied by the experimenter, composed a brief document, learned how to correct mistakes, and was given free time to explore the functions of the system. During the training session, each subject was shown how to make corrections as they went along as well as making corrections by completing dictation and going back and proofreading. Sample tasks in both transcription and composition were completed in this phase. Each subject was allowed approximately 40 minutes for the speech training scenario. Subjects were given no training for keyboard-mouse text creation tasks.

In the text creation phase for each modality, each subject attempted to complete four tasks - two composition and two transcription tasks. The order of the tasks (transcription or composition) was varied across subjects with half doing composition tasks followed by transcription tasks, and half doing transcription followed by composition. In all, each subject attempted to complete eight tasks – four composition and four transcription, with two of each task type in each modality.

For each composition task, subjects were asked to compose a response to a message (provided on paper) in the simple text entry window of the dictation application. Each of the responses were to contain three points for the reply to be considered complete and accurate. For example, in one of the composition tasks, the subject was asked to compose a message providing a detailed meeting agenda, meeting room location, and arrangements for food. Composition tasks included social and work related responses, and subjects were asked to compose "short replies." The quality of each response was later evaluated based on whether the composed messages contained a complete (included consideration of the three points) and clear (was judged as well written by evaluators) response. All subjects used the same four composition tasks, with an equal number of subjects using speech and keyboard-mouse to complete each task.

For transcription tasks, subjects attempted to complete the entry of two texts in each modality. There were four texts that ranged from 71 to 86 words in length. These texts were drawn from an old western novel. The subjects entered the text in the appropriate modality and were asked to make all corrections necessary to match the content of the original text. The resulting texts were later evaluated for accuracy and completeness by comparing them to the original materials. Evaluators counted uncorrected entry errors and omissions.

In the keyboard-mouse modality tasks, subjects completed composition and transcription tasks using standard keyboard and mouse interaction techniques in a simple edit window provided with each system. Subjects were given 20 minutes to complete the four keyboard-mouse tasks. All subjects completed all tasks within the time limit.

In the speech modality tasks, subjects completed the composition and transcription tasks using voice, but were free to use keyboard and mouse for cursor movements or to make corrections they felt they could not make using speech commands. We intentionally did not restrict subjects to the use of speech to carry out the speech modality tasks, and all subjects made some use of the keyboard and mouse. Subjects were given 40 minutes to complete the four speech tasks.

After each of the tasks (enrollment and eight text tasks), subjects filled out a brief questionnaire on their experience completing the task. After completing the four tasks for each modality, subjects filled out a questionnaire addressing their experience with that modality. After completing all tasks, the experimenter joined the subject for a debriefing session in which the subject was asked a series of questions about their reactions to the ASR technology.

Extended Use

Subjects in the Extended Use study were the four co-authors of this paper. In this study, the subjects used each of the three speech recognition products for 10 sessions of approximately one hour duration; a total of 30 sessions across the products. During the session the subjects would use speech recognition software to carry out actual work related correspondence. After completing at least 20 sessions, subjects completed the set of transcription tasks used in the Initial Use study. We limit the presentation of the results of the Extended Use phase of the study to some general comparisons with the Initial Use data.

RESULTS

For the analysis of the Initial Use sessions, we carried out a detailed analysis of the videotapes of the experimental sessions. This included a coding of all of the pertinent actions carried out by subjects in the study. Misrecognitions of text and commands and attempts to recover from them were coded, along with a range of usability and system problems. Particular attention was paid to the interplay of text entry and correction segments during a task, as well as strategies used to make corrections. Because of the extensive time required to do this, we completed the detailed analysis for 12 of the 24 subjects in the Initial Use phase of the study (four randomly selected subjects from each of the three systems, maintaining gender balance). Thus we report performance data from 12 subjects, but include all 24 subjects in reporting results where possible. Additionally, we report selected data from the four subjects in the Extended Use phase. The data reported from the three speech recognition systems are collapsed into a single group here.

Typing versus Dictating – Overall Efficiency

Our initial comparison of interest is the efficiency of text entry using speech and keyboard-mouse for transcription and composition tasks. We measure efficiency by time to complete the tasks and by entry rate. The entry rate that we present is corrected words per minute (cwpm), and is the number of words in the final document divided by the time the subject took to enter the text and make corrections. The average length of the composed texts was not significantly different between the speech and keyboard-mouse tasks and was similar to the average length of the transcriptions (71.5 and 73.1 words for speech and keyboard-mouse compositions respectively and 77.8 words for transcriptions). Table 1 below summarizes the results for task completion rates for the various tasks.

	Speech	Keyboard-mouse
Transcription	13.6 cwpm 7.52 min	32.5 cwpm 2.64 min
Composition	7.8 cwpm 9.96 min	19.0 cwpm 4.64 min
Average	8.74 min	3.64 min

Table 1. Mean corrected words per minute and time per task by entry modality and task type (N=12).

Creating text was significantly slower for the speech modality than for keyboard-mouse (F=29.2, p<0.01). By comparison, subjects in the Extended Use study completed the same transcription using ASR in an average 3.10 minutes (25.1 cwpm). The main effect for modality held for both the transcription tasks and the composition tasks. Composition tasks took longer than transcription tasks (F=18.6, p<0.01). This is to be expected given the inherent difference between simple text entry and crafting a message. There was no significant interaction between the task type and modality, suggesting that the modality effect was persistent across task type.

Given this clear difference in the overall time to complete the tasks, we were interested in looking for quantitative and qualitative differences in the performance. There are several areas in which we were interested in comparing text entry through typing to entry with ASR. These included: 1) number of errors detected and corrected in the two modalities, 2) differences in inline correction and proofreading as a means of correction, and 3) differences in overall quality of the resulting document. We consider evidence for each of these comparisons in turn.

Errors detected and corrected

A great deal of effort is put into lowering the error rates in ASR systems, in an attempt to approach the accuracy assumed for users' typing. For text entry into word processing systems, users commonly make errors (typing mistakes, misspellings and such) as they enter. Many of these errors are corrected as they go along – something that is supported by current word processing programs that highlight misspellings or grammatical errors. We were interested in data on the comparison of entry errors in the two modalities, and their detection and correction.

Table 2 presents data summarizing the average number of correction episodes for the different task types and input modalities. A correction episode is an effort to correct one or more words through actions that (1) identified the error, and (2) corrected it. Thus if a subject selected one or more words using a single select action and retyped or redictated a correction, we scored this as a correction episode. A major question is how the number of error correction episodes compares for ASR systems and keyboard-mouse entry.

	Speech	Keyboard-mouse
Transcription	11.3 (7.3)	8.4 (2.2)
Composition	13.5 (6.2)	12.7 (2.4)

Table 2. Mean number of correction episodes per task by entry modality (N=12). Length in steps is in parentheses.

While the average number of corrections made is slightly higher for the speech tasks than for the keyboard-mouse tasks, the length of the correction episodes is much longer. Interestingly, the improved performance for Extended Use subjects on transcription tasks cannot be accounted for entirely by reduced correction episodes – subjects averaged 8.8 per task. The average number of steps per correction episode is much shorter for the Extended Use subjects – averaging 3.5 steps compared to 7.3 for Initial Use subjects.

In general, the keyboard corrections simply involved backspacing or moving the cursor to the point at which the error occurred, and then retyping (we coded these as a move step followed by a retype step). In a few instances, the user would mistype during correction, resulting in a second retype step. About 80% of the keyboard-mouse corrections were simple position/retype episodes.

For speech corrections there was much more variability. In most cases a misrecognized word could be corrected using a simple locate/redictate command pair comparable to the keyboard-mouse pattern. Such a correction was coded as a voice move, followed by a voice redictate, that was marked to indicate success or failure. Variations include command substitutions such as the sequence voice select, voice delete, and voice redictate. More often the average number of commands required was much greater - generally due to problems with the speech commands themselves that then needed to be corrected, although the overall patterns can still be seen in terms of move to the error, select it and operate on it. Typical patterns included:

1. Simple redictation failures in which the user selected the misrecognized word or phrase (usually using a voice select command), followed by a redictation of the misrecognized word which also was misrecognized. Users would continue to try to redictate, would use correction dialogs that allow for alternative selection or spelling, or would abandon speech as a correction mechanism and complete the correction using keyboard-mouse
2. Cascading failures in which a command used to attempt a correction was misrecognized and had to be corrected itself as a part of the correction episode. Such episodes proved very frustrating for subjects and took considerable time to recover from.
3. Difficulties using correction dialogs in which the user abandoned a correction attempt for a variety of reasons. This included difficulties brought on by mode differences in the correction dialog (e.g., commonly used correction commands such as UNDO would not work in correction dialogs) or difficulties with the spelling mechanism.

High Level Correction Strategies – Inline versus Proofreading Corrections

Another question is whether users employ different correction strategies for the two input modalities. This could be demonstrated in either high-level strategies (such as "correct as you go along" versus "enter and then correct") or in lower-level differences such as the use of specific correction techniques. In Table 3 we present data for the transcription and composition tasks combined, comparing the average number of errors corrected in a task before completion of text entry (Inline) and after reaching the end of the text (Proofreading).

	Speech	Keyboard-mouse
Inline	8.6	8.8
Proofreading	4.2	1.6

Table 3. Average errors corrected per task by phase of entry (N=12).

There are two things to point out in these data. First, there are significantly more correction episodes in inline than in proofreading for both modalities (t=7.18, p<.006 for speech and t=8.64, p<.001 for keyboard-mouse). Performance on the keyboard-mouse tasks demonstrated that subjects are quite used to correcting as they go along, and try to avoid separate proofreading passes. For the speech modality however, subjects still had significant errors to correct in proofreading. In comparison, subjects in the Extended Use study rarely made inline corrections in transcription tasks (less than once per task on average).

Subjects gave us reasons for an increased reliance on proofreading. They commented that they felt aware of when they might have made a typing error, but felt less aware of when misrecognitions might have occurred. Note that in keyboard-mouse tasks, errors generally are user errors, while in speech tasks errors generally are system errors. By this we mean that for keyboard-mouse, systems reliably produce output consistent with user input. A typist can often "feel" or sense without looking at the display when an error might have occurred. For speech input, the user quickly learns that output is highly correlated with speech input, but that it is not perfect. Users do not seem to have a very reliable model of when an error might have occurred, and must either constantly monitor the display for errors or rely more heavily on a proofreading pass to detect them.

Second, the number of inline correction episodes is nearly equal for the two conditions. This suggests a transfer of cognitive skill from the more familiar keyboard and mouse interaction. As in typing, subjects were willing to switch from input mode to correction mode fairly easily and did not try to rely completely on proofreading for error correction.

Lower-level strategies for error correction

Almost all keyboard-mouse corrections were made inline and simply involved using the backspace key or mouse to point and select followed by typing. In comparison, the voice corrections were much more varied. This is undoubtedly due to the wide range of possible errors in the ASR systems compared to keyboard-mouse entry. The major classes of possible errors in ASR include:

- **Simple misrecognitions** in which a single spoken word intended as text is recognized as a different text word.
- **Multi-word misrecognitions** in which a series of words are recognized as a different series of words.
- **Command misrecognitions** in which an utterance intended as a command is inserted in the text.
- **Dictation as command misrecognitions** in which an utterance intended as dictation is taken as a command.

All of these occurred in all of the systems in the study. In addition, subjects did some editing of content in their documents. Because errors in ASR are correctly spelled words it is difficult to separate edits from errors in all cases. In what follows these are treated the same since both use the same techniques for correction.

Methods of making corrections in the two modalities can be compared. For example, keyboard-mouse corrections could be made by making a selection with the mouse and then retyping, by positioning the insertion point with cursor keys and then deleting errors and retyping, or by simply backspacing and retyping. These segment into two categories: deleting first then entering text or selecting text and entering over the selection. In speech, these kinds of corrections are possible in a variety of ways using redictation after positioning (with voice, keyboard or mouse). In addition, there is use of a correction dialog which allows spelling (all systems) or selection of an alternative word (in two of the three systems). Table 4 summarizes the techniques used by subjects to make corrections in the texts.[1]

The dominant technique for keyboard entry is to erase text back to the error and retype. This includes the erasure of text that was correct, and reentering it. For speech, the dominant technique was to select the text in error, and to redictate. In only a minority of the corrections (8%) did the subjects utilize the systems' correction dialog box. Almost a third of the corrections were to correct problems created during the original correction attempt. For example, while correcting the word "kiss" to "keep" in "kiss the dog", the command "SELECT kiss" is misrecognized as the dictated text "selected kiss", which must be deleted in addition to correcting the original error.

	Speech	Keyboard-mouse
Select text then reenter	38%	27%
Delete then reenter	23%	73%
Correction box	8%	NA
Correcting problems caused during correction	32%	NA

Table 4. Patterns of Error Correction based on overall corrections (N=12).

Low use of the correction dialogs may be explained by two phenomena. First, correction dialogs were generally used after other methods had failed (62% of all correction dialogs). Second, 38% of the time a problem occurred during the interaction inside the correction dialog with 38% of these resulting in canceling out of the dialog. Understanding more fully why the features of the correction dialogs are not better utilized is an area for future study.

Overall Quality of Typed and Dictated Texts

There are two areas in which we tried to evaluate the relative quality of the results of text entry in the two modalities. For transcription tasks, we evaluated the overall accuracy of the transcriptions – that is, we asked how many mismatches there were between the target document and the produced document. For composition tasks we asked three peers, *not part of the study*, to evaluate several aspects of the messages produced by subjects. These judges independently counted the number of points that the message covered (there were three target points for each message). We also asked for a count of errors in the final message and for an evaluation of the overall clarity. Finally, for each of the four composition tasks, we asked the judges to rank order the 24 messages in terms of quality from best to worst.

In Table 5 we summarize the overall quality measures for the texts produced. These measures include average number of errors in the final products for both the transcription and composition tasks, and the average quality rank for texts scored by three judges for the composition tasks.

There were many more errors in the final transcription documents for the speech tasks than for the keyboard-mouse tasks. The errors remaining in the final documents were broken into three categories: wrong words (including misspellings), format errors (including capitalization and

[1] Only one of the 12 subjects used an explicitly multi-modal strategy for correction. That subject relied on the keyboard to move to the error and switched to speech to select and redictate the text.

punctuation errors), and missing words. The average number of wrong words (F=25.4, p<.001) and format errors (F=12.6, p<.001) were significantly lower for keyboard-mouse compared with speech tasks. There was no difference in the number of missing words.

	Speech	Keyboard-mouse
Transcription errors	3.8 errors	1.0 errors
Composition errors	1.8 errors	1.1 errors
Composition (rank)	13.2	11.4

Table 5. Mean quality measures by modality (N=24).

Composition quality showed a similar pattern. Errors in composition included obviously wrong words (e.g., grammar errors) or misspellings. There were fewer errors in the keyboard-mouse texts than in the speech texts (F=7.9, p<0.01). Judges were asked to rank order the texts for each of the four composition tasks from best (given a score of 1) to worst (given a score of 24). While the mean score was lower (better) for keyboard-mouse texts than for speech texts, the difference was not statistically significant.

Transcription versus Composition

For both the keyboard-mouse modality and the speech modality, composition tasks take longer than transcription tasks. We did not find significant differences in the length or readability of texts composed in the two modalities. Additionally, topics such as correction techniques or error frequencies did not seem to vary between modalities and task types.

Subjective Results – Questionnaire Data

Subjects (N=24) in the Initial Use study consistently report being dissatisfied with the ASR software for performing the experimental tasks. When asked to compare their productivity using the two modalities in the debriefing session, subjects gave a modal response of "much less productive" for speech on a 7-point scale ranging from "much more productive" to "much less productive", and 21 of 24 subjects responded "less" or "much less productive". Subjects' top reasons for their ratings, (frequency of response in parentheses summed across several questions) were:

- Speech recognition is unreliable, error prone (34).
- Error correction in speech is much too hard – and correction can just lead to more errors (20).
- Not knowing how to integrate the use of speech and keyboard-mouse efficiently (19).
- Keyboard is much faster (14).
- Command language problems (13).
- It is harder to talk and think than to type and think (7).

Additionally, when asked if the software was good enough to purchase, 21 of 24 subjects responded "No" to a binary Yes/No choice. The three subjects that reported a willingness to purchase the software all gave considerable qualifications to their responses. When asked for the improvements that would be necessary for ASR technology to be useful, subjects' top responses included:

- Corrections need to be much easier to make (27).
- Speech recognition needs to be more accurate (25).
- Need feedback to know when there is a mistake (8).
- Command language confusion between command and dictation needs to be fixed (8).

DISCUSSION

There are many interesting patterns in the data presented above. Early speech recognition products varied in the strategies of error correction that they encouraged for users. For example, IBM's VoiceType system encouraged users (in documentation and online help) to dictate first and then switch to correction mode, while Dragon Dictate encouraged users to make corrections immediately after an error was dictated. To a large extent these strategies were encouraged to have user behavior correspond to system designs, and not because of a user driven reason. The systems in the current study all accommodate inline correction and post-entry correction equally well. One thing that the results of the Initial Use study point to is the general tendency for subjects to make corrections as they go along, rather than in a proofreading pass. Table 2 shows that subjects made many more corrections inline than they did after completion of entry in both the speech and keyboard-mouse conditions.

When subjects made errors in keyboard-mouse text entry, they tended to correct the error within a few words of having made it. In contrast, some subjects made specific mention of not being as aware of when a misrecognition had occurred and needing to "go back to" a proofreading stage for the speech tasks. Taken together with the tendency toward inline correction, this suggests supporting users in knowing when a misrecognition has occurred.

Misrecognition Corrections

The most common command used in any of the systems is the command to reverse the immediately preceding action. While each of the systems has multiple variant commands for doing this (some mixture of UNDO, SCRATCH, and DELETE), users generally rely on a single form that they use consistently. However, the command variants have subtle distinctions that were frequently lost on the subjects in this study. Many of the usability problems with respect to these commands appear related to the users strategy of relying on a single form for a command, even if it was not appropriate for the tasks at hand. Developing more complex strategies for selecting between command forms seems to require additional expertise. We do not observe these confusions at this level in the Extended Use study.

Quality Measures

Attempts to compare the composition quality of texts produced by speech input and more traditional input generally predate the existence of real systems for ASR [e.g., 3,8]. The current study shows no statistical difference between the quality of the texts composed using speech recognition as compared to keyboard and mouse.

Subjective Results

The majority of subjects felt that they would be less or much less productive with speech recognition than with keyboard and mouse using the current products. They provided some clear insights into where efforts need to be made to improve these systems in order for them to be useful and usable. Top concerns include the performance of the systems and several key user interface issues. There is a critical need to learn about people's performance and satisfaction with multi-modal patterns. The field needs to better understand the use of commands and people's ability and satisfaction with natural language commands. Also, there are intriguing issues to be researched regarding cognitive load issues in speech recognition and how to provide feedback to users. The subjects said that they were excited about the future possibility of using speech to complete their work. They were pleased with the feeling of freedom that speaking allowed them, and the ease and naturalness of it.

CONCLUSIONS

It is interesting to note that several of the Initial Use subjects commented that keyboard entry seemed "much more natural" than speech for entering text. While this seems like an odd comment at some level, it reflects the degree to which some people have become accustomed to using keyboards. This relates both to the comfort with which people compose text at a keyboard and to well learned methods for inline error detection and correction. Speech is also a well learned skill, though as this study shows, the ways to use it in communicating with computers are not well established for most users. There is potential for ASR to be an efficient text creation technique – the Extended Use subjects entered transcription text at an average rate of 107 uncorrected words per minute – however correction took them over three times as long as entry time on average.

When desktop ASR systems first began appearing about 5 years ago, it was assumed that their wide-scale acceptance would have to await solutions to "mode problems" (the need to explicitly indicate dictation or command modes), and the development of continuous speech recognition algorithms which were sufficiently accurate. While all of the commercial systems evaluated in this study have these features, our results indicate that our technically sophisticated subject pool is far from satisfied with the current systems as an alternative to keyboard for general text creation. They have given a clear prioritization of changes needed in the design of these systems. These changes merit significant attention.

It is possible – though we do not think it is very likely – that less skilled computer users would react to the software more positively. The methods for error correction, and the complexity that compound errors can produce, leads us to believe that decreased rather than increased performance would have to be tolerated by any users – even those with limited typing skills. While this might be acceptable for some populations (RSI sufferers or technology adopters), wide scale acceptance awaits design improvements beyond this current generation of products.

REFERENCES

1. Clark, H. H. & Brennan, S. E. (1991). Grounding in communication. In J. Levine, L. B. Resnick, and S. D. Behrand (Eds.), Shared Cognition: Thinking as Social Practice. APA Books, Washington.
2. Danis, C. & Karat, J. (1995). Technology-driven design of speech recognition systems. In G. Olson and S. Schuon (eds.) Symposium on designing interactive systems. ACM: New York, 17-24.
3. Gould, J. D., Conti, J., & Hovanyecz, T. (1983). Composing letters with a simulated listening typewriter. Communications of the ACM, *26*, 4, 295-308.
4. Karat, J. (1995). Scenario use in the design of a speech recognition system. In J. Carroll (ed.) Scenario-based design. New York: Wiley.
5. Kidd, A. (1994). The marks are on the knowledge worker, in *Proceedings of CHI '94* (Boston MA, April 1994), ACM Press, 186-191.
6. Lai, J. & Vergo, J. (1997). MedSpeak: Report Creation with Continuous Speech Recognition, in *Proceedings of CHI '97* (Atlanta GA, March 1997), ACM Press, 431 - 438.
7. Laurel, B. (1993). Computers as Theatre. Adison Wesley, New York.
8. Ogozalek, V.Z., & Praag, J.V. (1986). Comparison of elderly and younger users on keyboard and voice input computer-based composition tasks, in *Proceedings of CHI '86*, ACM Press, 205-211.
9. Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. Computer Speech and Language, *9*, 19-35.
10. Yankelovich, N., Levow, G. A., & Marx, M. (1995). Designing SpeechActs: Issues in speech user interfaces, in *Proceedings of CHI '95* (Denver CO, May 1995), ACM Press, 369-376.