Talking to Computers: An Empirical Investigation

Alexander G. Hauptmann
and
Alexander I. Rudnicky
December 16, 1987
Technical Report CMU-CS-87-186(2)

Department of Computer Science
Carnegie-Mellon University
Pittsburgh, PA 15213

Abstract

This paper describes an empirical study of man-computer speech interaction. The goals of the experiment were to find out how people would communicate with a real-time, speaker-independent continuous speech understanding system. The experimental design compared three communication modes: natural language typing, speaking directly to a computer and speaking to a computer through a human interpreter. The results show that speech to a computer is not as ill-formed as one would expect. People speaking to a computer are more disciplined than when speaking to each other. There are significant differences in the usage of spoken language compared to typed language, and several phenomena which are unique to spoken or typed input respectively. Usefulness for work in speech understanding systems for the future is considered.

This research was sponsored in part by the Defense Advance Research Projects Agency (DOD), ARPA Order No. 5167, monitored by the Air Force Avionics laboratory under contract #N00039-85-C-0163.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

Table of Contents

List of Figures

List of Tables

Despite a growing interest in the use of speech recognition for non-trivial tasks, few attempts have been made to empirically investigate how people might prefer to interact with a computer using speech. Existing investigations have limited themselves to evaluating restricted applications for isolated-word speaker-dependent speech recognition systems (Poock, 1982, Morrison, Green, Shaw and Payne, 1984, Poock and Armstrong, 1980). These studies, as well as others (Nye, 1982, Leggett and Williams, 1984), have shown that, while speech input is useful at times, the problems that occur were often different than the ones anticipated in the laboratory. For example, it was found that once a word was misrecognized, attempts at correction such as the use of over-enunciation (extreme stressing) made the situation hopeless for the system. Moreover, irrelevant events, such as lip smacks, inadvertent sighs, coughing, throat clearing, eating, smoking, etc. all created problems for the isolated word systems under study.

Holmgren (1983) examined how people spoke credit card number digits over a telephone line. Holmgren found a fair amount of consistency in how speakers grouped numbers. Many variations were tractable (for example, the interchangeability of *"zero"* and *"oh"* and the use of *"dash"* to separate number groups). Other, fairly infrequent variations (such as *"double three"* or *eighty five"*) disappeared after some training and feedback. Although Holmgren's study is an excellent example of how to study empirically the recognition environment, the limited domain (digit strings) makes extrapolation to larger speech applications difficult.

Some studies have examined larger tasks that might prove to be a good approximation to a realistic environment for speech understanding systems. Gould (1978, 1980) has studied voice input to a dictation machine as well as a spoken interface to the dictation machine using a fairly restrictive editor framework (Gould, Conti and Hovanyecz, 1983). Chapanis (1981) and Grosz (1977), on the other hand, studied communication between people, though not between a person and a computer.

Currently, we are moving into a level of speech understanding, that promises a 2000 word, speaker independent speech understanding system in the next few years. A system like this may be able to overcome the types of difficulties that Gould states in his studies on a "listening typewriter" (Gould *et al.*, 1983). There a 5000 discrete word system was preferred over a 1000 word continuous speech recognition system. On the other hand, in a less ambitious task-oriented setting, much smaller vocabularies have been found to be sufficient (Ford, Weeks and Chapanis, 1980).

There are few good realizations of applications for such capabilities, certainly not the spoken editing of

taped messages (Allen, 1983). As the leading edge of technology advances, engineers have only dealt with idealized grammars and abstract sentences to demonstrate the capabilities of their system (Lowerre, 1976, Woods, Bates, Brown, Bruce, Cook, Klovstad, Makhoul, Nash-Webber, Schwartz, Wolf and Zue, 1976, Erman and Lesser, 1980, Stern, Ward, Hauptmann and Leon, 1987). It is unknown whether the assumptions made about users speaking "good English" hold for real man-computer dialogues. Hayes and Reddy (1983) proposed in a theoretical discussion how integrated communication with computers using speech may work. They have, however, no empirical, human-to-computer speech data to support their speculations. Hayes and Mouradian (1981) also give anecdotal cases of ungrammatical phrases. Chapanis (1981, p. 106), in researching voice as well as written communication, found that "natural human communication is extremely unruly and often seems to follow few grammatical, syntactic or semantic rules."

This statement is confirmed by independent findings of Grosz (1977), whose protocols show incomplete sentences, ungrammatical style, ellipsis, fragments and clarifying subdialogues. Her subjects were cooperatively solving problems with only certain (visual, audio, written) communication channels available. Written communication between people is also quite unlike the way we were all taught to write in school. Typed input to natural language computer systems is very different again. People have been known to revert to "baby talk", that is typing in a very abbreviated style, leaving out many words, or to use incredibly stilted phrases and ignore all punctuation[1].

Therefore we must assume that man-computer interaction using speech also differs significantly from interpersonal spoken dialogue. As a recent dissertation on the subject of on-line help systems (Borenstein, 1985, p. 115) notes, "There is simply no substitute for observing real users."

The goal of this study is to examine how speech input to a computer differs from interpersonal spoken communication, given a situation in which users are fairly free to select a comfortable style of interaction. To highlight the differences, we compare the same task situation (using an electronic mail system) under three different interaction conditions: typing, speaking directly to a computer, and speaking through a human intermediary.

---

[1]Jaime Carbonell, Carnegie-Mellon, personal communication

Method

Subjects

Forty subjects were recruited from a population of electronic mail users in the Computer Science Department at C-MU. Ten of these were used as pilot subjects to test and debug the experimental setup. None of the 30 subjects in the final experiment were classified as naive with respect to their electronic mail experience. However, some had no experience with the particular mail system used in this study, MERCURY (also known as "hg"). Subjects were given $1 ice cream certificates as payment.

Apparatus

The experimental setup included a Digital Equipment Corporation MicroVax II running UNIX, two terminals and a videotape recording device. One room was designated as the subject room, the other as the control room for the experimenter. The two terminals (one in each room) were hooked together to display the same information at all times. Both terminals had their keyboards enabled, such that both the subject as well as the experimenter were able to type commands into the system when necessary. A microphone and video camera in the subject's room allowed speech and terminal display to be recorded (using a recorder located in the experimenter's room). The experimenter observed the subject over the videotape monitor and speaker.

Materials

The subject's experience with electronic mail was determined through an initial questionnaire (see Appendix I). Each subject was provided with a sheet of background information (see Appendix II). Each subject also was given an instruction sheet, specific to their assigned communication mode (see Appendices III, IV and V). Appendices VIII and IX contain the actual task chores used in the experimental sessions. An attitude questionnaire (in Appendix VI) determined how subjects responded to their interaction mode.

Experimental Design

Ten subjects were randomly assigned to each communication mode. Each subject completed three sessions. This design corresponds to a repeated measures analysis of variance design as described in Myers (1972). The data was analyzed accordingly.

In the **speech-to-computer mode**, subjects were told that the computer could understand them, with *occasional* help of the experimenter (Appendix III). The experimenter was in the adjacent room and transcribed all commands into equivalent system commands, pretending the system itself had understood

the utterance. Whenever the subjects were in an editing mode, speech input was disabled and subjects had to edit manually using the keyboard. Specifically, subjects were asked to speak all hg-mail commands to the system. It was left up to the subjects to choose the most natural way for them to do this.

In the **speech-to-human mode**, the experimenter was sitting in the same room as the subject, obviously translating their utterances into typed commands to the electronic mail system (Appendix IV). The subjects were never deceived about the reality of computer speech understanding. Again, whenever the subjects were in an editing mode, speech input was disabled and subjects had to edit manually using the keyboard. Otherwise this mode was identical to the speech-to-computer mode.

In both speech communication modes, a prompt " `listening:` " appeared on the subjects' keyboard, when the system was ready to accept input. Otherwise a " `processing....` " message appeared, indicating that the system was currently analyzing an utterance. Only the speech that occurred during the listening phase was transcribed and analyzed.

In the **typing-to-computer mode** subjects were led to believe that a computer natural language mail system was interpreting their typing (Appendix V). In effect, this mode was the same as the speech-to-computer mode, without the speech channel. The subjects had to type everything themselves. However, the system, *or rather the invisible experimenter*, was able to process every input intelligently.

Procedure

Upon arrival at the user studies laboratory, subjects were given a questionnaire designed to determine their degree of familiarity with electronic mail (Appendix I). The subjects were given the background information (Appendix II) and an instruction sheet specific to their communication mode. Careful consideration was given to the instructions and the description of the chores/tasks in light of Borenstein's finding that the wording of the instructions and the general conceptual principles of the system are more important than the modalities of communication (Borenstein, 1985).

A total of nine tasks were to be completed. Each task asked the subject to do something with the mail database file she or he was working with. The tasks included replying to mail, locating information about previously sent mail and adding a carbon copy of some new mail to the file. Each subject received the same tasks in the same order. The first of three sessions for each subject was counted as a training session. It was meant to ensure that the equipment was working properly and the subject had understood the basic task. This training session was not used in the final analysis.

After the third session, each subject was given an attitude questionnaire, to assess the subject's opinion of the particular interaction mode (Appendix VI).

During each session, a time stamped screen image together with the voice commands was recorded. The videotape recordings of the two speech modes were transcribed. The typed input from the typing mode provided comparable data. The total time to complete each task in each condition was also recorded. This measure of time from the beginning of a task to the end is somewhat influenced by system response times. However, the system response time was roughly comparable for all subjects in all conditions and should be negligible compared to the total time spent on the task.

Analysis

Each interaction with the system was classified as one utterance. The dependent variables which were examined are classified into four groups: **attitude**, **communication**, **errors**, and **syntax**.

- A questionnaire, adapted with minor changes from Hauptmann and Green (1983), measured the subject's **attitude** towards the experiment. Significant differences in attitude might indicate that factors such as motivation could have determined differences between conditions.
- **Communication** variables reflecting rate, density and vocabulary were measured in the second group of variables. These included as specific scores:
    - The number of utterances spoken or lines typed by a subject during a session.
    - The time the subject took to complete a session.
    - The number of words spoken or typed by the subject during a session. Also the number of words per utterance or line.
    - The number of unique (distinct) words used by a subject during one session. This is effectively the vocabulary that the subject used.
- The third group of dependent variables deals with **errors**. This group of variables reflects problems people had giving instructions to the computer in the various modes.
    - The relative number of false stars or repeated words. In the typed mode, typing errors which were relayed to the system were counted as a false start.
    - The relative frequency of *ahs and hms* per word. These do not include the begin-problem *ahs and hms* below.
    - The number of utterance/lines which contained problems at the beginning. These begin-problems include saying *ah, well, ok, oh, sorry, yes, etc.* at the beginning of an utterance.
    - The relative frequency of lines containing task unrelated phrases other than ahs and ahms.
    - The relative frequency of utterances containing repeated words (false starts), begin-problems, or non-task related information.
- The final group of variables that were analyzed dealt with the **syntactic structure** of the input.
    - The relative number of pronouns used.
    - The relative number of pronouns referring to the dialogue participants, i.e., *I, you, me, your,* etc.

- The relative number of utterances that are syntactic declaratives, questions and commands or syntactically unclassifiable.
- The percentage of perfect hg mail commands. These are commands which could be literally interpreted by the hg program in typed form. For this variable all begin problems and ahs and ahms were ignored.

Figure 1: Sample Transcript for Speech-to-Human condition, Both Sessions

```
025000 type message one eighty two
043000 list all headers with the word mckean honda
068000 type one eighty three
095000 forward this message to rudnicky on the a
116000 please type message one eighty four
138000 send mail to rudnicky at g
191000 mail it
200000 all done

020000 please type message ninety
039000 list all headers with the word camera
079000 type messages sixty nine seventy seventy one and seventy three
173000 type message ninety one
195000 list all headers with the word ceedee
222000 type message eighty four and eighty five
253000 type message ninety two
278000 copy this message into a file called important dot text
299000 quit
```

Figure 2: Sample Transcript for Speech-to-Computer condition, Both Sessions

```
010000 type message one eighty two
024000 search all for mckean honda
130000 next
156000 forward to rudnicky on a
203000 send
215000 next
229000 mail alex at g
268000 send
280000 the next
288000 bye bye

009000 ninety
029000 search headers for camera
069000 seventy two
095000 back to the unread messages
139000 search bodies of last months mail for ceedees or compact disks
205000 search headers for ceedees
252000 eighty
341000 next new message
361000 save this message in important
381000 next new message
389000 bye bye
```

Apart from attitude, the above variables were hypothesized to differ significantly between the styles of interaction. Some of them were taken from the analyses of the previous studies cited above, while others were suggested by the pilot experiments.

Figure 3: A Sample Transcript for a Type-to-Computer Subject, Both Sessions

```
610000 t 182
653000 does mckean or honda appear in previous messages
746000 next
766000 list it
802000 mail this to rudnicky on a
824000 next
856000 tell alex@g i read it today
904000 send mail to alex@g saying that i read the message on friday
948000 fill in subject with message read on
977000 mail
989000 next
996000 q

1087000 first
1097000 first new
1127000 find messages relating to camera
1185000 list those which have fix stick clean in the bosy
1283000 list those relating to fix stick clean or repair
1388000 just the headers
1404000 next
1513000 messages about cd's ordered by recency
1546000 next
1606000 save message in file that is only readable by me and not changeable
1622000 document
1634000 q
```

Results

A sample transcript for a subject in each mode can be found in Figures 1, 2 and 3. Table 1 summarizes the results of the statistical analyses performed on the quantitative data. A total of 3233 words were spoken/typed by the subjects in 708 utterances/lines. The total vocabulary consisted of 304 distinct words.

- **Attitude.** On the whole, subjects felt positive about the experiment, as indicated by a mean attitude score of 30.3. The expected neutral attitude score would be 0, with possible attitude scores ranging from -76 to +76. There was no significant difference between the three groups of subjects.
- **Communication Variables**
  - The number of utterances per session and the time to completion were not significantly different for the three groups.
  - However, the total number of words used in a session to solve the tasks showed significant differences. The two speech groups (speak-to-computer 60.35 words average; speak-to-human 65.5) both used considerably more words than the typing group (36.8 words average). The two speech groups did not show a significant difference between each other.
  - The utterance length was also significantly different between the respective groups. Speech-to-Computer contained the longest utterances at 6.10 words average. Speech-to-Human contained an average of 5.45 words and typed lines averaged only 3.21 words.
  - The number of unique (distinct) words used was also significantly different in the three

communication modes. The typing condition subjects needed only an average of 23.75 distinct words to complete a session. The speech-to-computer and speech-to-human subjects used 32.7 and 36.65 distinct words to complete a session.

- **Error Variables**
  - The relative percentage of false starts (word repetitions) was not significant between the groups. There simply were too few of these occurring in the corpus of data collected.
  - The relative number of noise words (*ah, ahm, etc.*) was significantly different over the communication modes. The typed mode had 0, the speech-to-computer mode contained about 4 such words per thousand and the speech-to-human mode averaged 15 noise words per thousand words. This particular kind of error was analyzed further.

    Eight of the 20 speakers (6 speaking to a human, 3 speaking to a computer) had some kind of *Ah, uhm* etc. embedded in their utterances other than at the beginning. For the "speaking to a human" group, 23 ahs were collected, as opposed to 8 ahs for the "speaking to computer" group.

    Of these ahs, 18 occurred just before a noun phrase, 2 within a noun phrase, 6 in a prepositional phrase between the preposition and the noun phrase. The ahs in a prepositional phrase have been counted twice. 4 ahs occurred at the clause boundary other than a noun phrase and 7 ahs were spoken as part of a restart in a broken off utterance.

    In terms of cognitive structure, almost all ahs occurred just before a definite reference to a message, file, user name or search field. 7 ahs occurred in restarts and 3 occurred as part of an action specification. The actual data for these can be found in Appendix VII.
  - The percentage of lines that had begin-problems (as defined above), also varied significantly between the groups. The speech-to-computer group had .102 begin problems per line, the speech-to-human group averaged .255 per line and the typed condition had none.
  - The number of lines which contained task unrelated phrases was also significantly different between the 3 groups. The speech-to-computer group had a mean of 0.5 percent unrelated utterances while the speech-to-human group averaged 4 percent. The typed condition had 0 task unrelated utterances. In addition, this was the only significant difference that could be found when just the two speech conditions were compared separately. The difference between the two speech groups alone was significant at $p > .038$.
  - Finally, the difference in the frequency of utterances/lines which had some typical speech problem (begin-problem, ahs or ahms, false starts or task unrelated parts) was also significant. Almost one third of the speech-to-human (31 percent) utterances had something wrong with them in this sense. 16 percent of the speech-to-computer utterances were characterized by at least one of these problems and only 0.7 percent of the typed lines had problems of this type.
- **Syntax Variables**
  - While there was no significant difference between the relative frequency of pronoun usage, the frequency of pronouns referring to the dialogue participants did increase significantly in the speech conditions. The frequency per word spoken went from .048 in the speech-to-computer group and .021 in the speech-to-human group down to .008 dialogue participant referencing pronouns per word in the typing group.
  - There was no significant difference in the relative frequency of questions, commands, declaratives and syntactically unclassifiable sentences.
  - The typing condition showed significantly more literal hg commands with an average

frequency of .691 perfect hg commands per line. The speech-to-computer and speech to human conditions averaged .257 and .168 respectively.

In addition to the quantitative analysis presented above, the following phenomena were observed:

- In the typed communication mode, subjects tended to abbreviate most commands and some key words. These abbreviations were **never** used in the speech modes. A full 14.5 percent of all words were shortened this way in the type-to-computer mode. The example in the appendix shows the extremely terse style in the typed mode (e.g. *"t 182"*). Most abbreviations are to single letters.
- When complex queries were made, the scope of the conjunctions and quantifiers was often ambiguous, e.g. *"show the headers of messages about compact disks or from Martin Stacey since May"*. Does the *"since May"* describe the messages about compact disks? In queries and unfamiliar situations, subjects used more complete sentences in better grammar.
- Many pauses occurred while subjects were thinking or hesitating, for example, when a specific mail item needed to be described to the system. Such pauses were not noted in the transcript. Occasionally some very long pauses even confused the experimenter into erroneously assuming an utterance was already completed, until the subject suddenly continued speaking.
- Almost all subjects assumed the system could recognize difficult names. "*Rudnicky*" was never spelled out, nor "*McKean Honda*". E.g. in *"remail to Rudnicky at a"* or *"show headers about McKean or Honda"* the spelling of these spoken items is not specified. In all of these instances, the sounds were considered by the subjects to be sufficient to produce a uniquely spelled name.
- Similarly, spoken names of files like *"message dot dat"* were always meant to produce *"message.dat"*. There was never any worry by a subject about confusing punctuation with the name of the punctuation object.
- Plurals were assumed to be eliminated in the query of the database. *"Show me everything about Hondas"* would be intended to mean *search the database for "Honda"* instead of *"Hondas"*.
- Long sentences often contained redundant elements. E.g. *"can you list all the messages that have a header of McKean Honda have in somewhere in the header"* not only contains a restart, but also a redundant prepositional phrase at the end.
- Many subjects wanted to stop the processing of a partial or complete utterance and start over. The attempted ways of dealing with this ranged from *"Oh, no, wait, don't do that"* to *"sorry, ahm, can I do that over"* to a mere restart of the utterance. The restart was usually marked by *"no, no, no, no"* or *"ahm"* or a longer pause.
- The subjects in the speech-to-human group seemed to talk to themselves more. They generally paid less attention to the *listening:...processing...* prompts.
- Relative references were often fuzzy. Was the *last message*, the one that was sent last, that was received last, that was just examined or the last one of the day's new messages which were already read.
- There were many other noises on the video tape which were not transcribed. These include lip smacks, sniffs, heavy breathing, sighing, creaking chairs, shuffling feet, keyboard clicks, pencils or fingers drumming and papers shuffling.

Table 1: Summary ANOVA table.

		Condition		
Factor	***p* value**	**Sp-H**	**Sp-C**	**Typ-C**
Opinion:				
attitude	n.s.	31	26.7	33.1
Communication:				
lines per session	n.s.	12.35	10.4	11.85
time per session	n.s.	335.95	427.9	465.65
words used per session	.007	65.5	60.35	36.8
words per utterance	.005	5.45	6.10	3.21
unique words per session	.010	36.65	32.7	23.75
Errors:				
% lines with repeated words or false starts	n.s.	.55	.7	.1
ahs per word	.006	.014	.004	0
% lines with begin-problem	.015	.255	.102	0
% task-unrelated lines	.007	.041	.005	0
False starts, ahs, ahms or begin-problems/line	.007	.315	.162	.007
Syntax:				
pronouns per word	n.s.	.038	.061	.039
1st and 2nd person pronouns per word	.024	.021	.048	.008
% declaratives	n.s.	.027	.100	0
% questions	n.s.	.029	.092	.011
% unidentifiable syntax	n.s.	1.5	1.9	3.0
% commands	n.s	.839	.625	.727
% perfect hg mailer commands	.009	.257	.168	.691

Notes:

The value reported is the mean for each condition (N = 10). The conditions are: **Sp-H**: Speech to Human, **Sp-C**: Speech to Computer, **Typ-C**: Typing to Computer.

No Task and Task X Mode interactions were significant and are thus not listed. n.s. indicates a not significant comparison ($p > .05$).

Discussion

First, let us compare our findings with the electronic mail task used in the speech recognition project at Carnegie-Mellon (Adams and Bisiani, 1986). The Carnegie-Mellon system uses a lexicon of 325 words for the electronic mail task. Clearly the vocabulary for this task is adequate since our experiment yielded a total of 304 different vocabulary words. However, all the names in the experimental tasks were identical. Thus one would have to add about 50 names to the vocabulary to make the two domains more comparable. Our results thus support the findings by Ford *et. al.* (1980), who argues that a small, but well chosen, set of words is sufficient for limited domain tasks. It cannot be construed as contradicting the large vocabulary preference in Gould *et. al.* (1980) , because the task domain is much more restricted in our case.

The utterances spoken were not nearly as grammatical as we had supposed in the electronic mail task constructed for the C-MU speech system. Subjects tended to be terse, wherever they knew the mailing system allowed incomplete sentences. This terseness had not been anticipated in the C-MU speech system electronic mail grammar. However, when subjects dealt with unfamiliar commands and queries, they resorted to more complete utterances. Clearly, subjects had the expectation that the system could operate in either mode.

Contrary to what Chapanis (1981) and Grosz (1977) found, speaking to a computer is not quite as unruly as interpersonal communication. Somehow, the knowledge that one is dealing with a computer, enabled the subject to perform a well defined task without many of the complexities of discourse between people that are so often reported in the linguistic literature (Biber, 1986).

A number of the phenomena identified by Hayes and Reddy (1983) in their article on interactive systems and speech did not manifest themselves in the present study. Subjects never echoed the systems response, the way they do in interpersonal communication. This echoing or parroting back of information to verify its correct transmission was somehow not relevant in the present environment. Similarly, users did not implicitly or explicitly acknowledge the system's actions. Such acknowledgments were suggested as a possible problem in computer dialogue systems. There were also no cases in which the user gave an indication of incomprehension.

Some of the other phenomena described by Hayes and Reddy did however occur. There were indirect speech acts. Every statement made by the subject was an instruction to the system to do something.

However, a small fraction of them (cf. Table 1) were indirect speech acts in the form of questions or declaratives. Ellipsis was most prominent in response to a system query or after an unsatisfactory answer, for example:

To which file?
message dot text
...
list all messages since May
no applicable messages
how about since January?

Some of the findings reported in this paper seem also to contradict the preliminary impressions by Werner[2]. He concludes that computer discourse is much less structured than seen in the present experiment. However, this discrepancy is in all likelihood due to the much looser system definition used by Werner in his truck database simulation. The electronic mail system that was simulated here was well defined and the subjects capable of carrying out the task using only their familiar electronic mail system. The constraints that this domain experience added to the discourse situation might be able to account for the discrepancy in findings.

We believe that these different and somewhat contradictory experiences point to the crucial importance of task definition in the success of a speech recognition system. A successful speech recognition application requires careful task analysis, followed by equally careful language and environment design.

Even though people interact with the computer in a more disciplined way, a number of purely speech-related phenomena were still observed. Thus subjects were more likely to stick to their familiar set of commands in the familiar (typed) interaction mode, while they used more *natural* English-like ways of phrasing utterances in the two speech conditions. Some of these differences in communication modes, like the increased use of pronouns in the discourse, represent a quantitative shift in the use of language. The principles of natural language processing systems that can be applied to these phenomena in typed input situations should also be adaptable in the spoken communication mode. This adaptation is by no means trivial, as pointed out by Hayes, Hauptmann, Carbonell and Tomita (1986).

There were, however, several ways in which speech to a computer is completely different from typed input. These mostly seem to originate from the *error* group of variables. In particular, looking at the

---

[2]Philip Werner, Carnegie-Mellon University, unpublished report about a truck database simulation

number of *ahs* and *ahms*, the large number of utterances that are somehow preceded by *"channel checking"* or other noise at the beginning of the utterance, it becomes clear that these will be persistent phenomena any speech application will have to deal with in the future. The large fraction of spoken input that was somehow flawed by a problem at the beginning of the utterance, an *"ah"* or *"hm"* in the middle, a false start (repeated word) or a task-unrelated phrase, will need to be accounted for. Work by Hindle (1983) has already begun to deal with parsing false starts or repeated words under certain, somewhat artificially, flagged conditions.

It is not difficult to imagine a system that filters out a certain set of channel-checking words at the beginning of the utterance. Even in the face of uncertain word recognition, this set of words remains limited and could be eliminated through some parsing strategy. This strategy would specialize in the beginning of the utterance and discard the irrelevant words at the beginning. In effect, this adaptation merely changes the starting point of the utterance to begin after the noise words.

A system that can deal with an arbitrary *ah* or *ahm* at any place in the utterance is much more difficult to imagine. If the input were always perfectly recognized, as Hindle assumed, it would be trivial to skip all these words. However, given an uncertain word recognition system and huge sets of alternatives, the branching factor would be increased significantly if we allow a noise word at any point in the sentence. The analysis of the *ahs* (see results above), shows that noise words are not arbitrarily placed in the utterance. Subjects would never say *show ah me this message*, but might say *show me ah message five*. If one could construct a model of where these noise words might occur, the recognition task would be greatly simplified. In its simplest form the model could be based on syntactic structures (58 percent of all embedded *ahs* can be accounted for this way). It seems more likely that a better model would also take into account the degree of cognitive load involved (67 percent of embedded *ahs* from our data). Most of the noise words occurred when a definite reference needed to be specified to the computer and the user had to pause and decide exactly what he was referring to.

Other problems, such as spelling of ambiguously pronounced names, will also need to be addressed in the future. These seem more tractable in the sense that one could give feedback to the user about which names are under consideration for this pronunciation and letting him/her choose.

The present study contains a number of flaws:

- The design should have tested subjects in all 3 communication modes, to reduce inter-subject variability. As it was, we inflated the amount of variation that is always created by

individual differences. A better design was outlined in Borenstein (1985).

- Some subjects were completely deceived by the setup, whereas others could not be fooled. Thus the realism of the speech-to-computer simulation was fairly limited.
- Originally, the speech-to-human mode was intended to provide a control situation, mimicking natural interpersonal communication. Due to perceived experimenter expectancy effects, subjects nevertheless pretended to talk to a computer. The experimental setup also focused the subject on the computer terminal and microphone, rather than on the interpersonal communication in the speech-to-computer communication mode.
- We had intended to provide some measure and comparison for grammaticality, but found this impossible to establish reliably. Our subjects were too familiar with electronic mail and preferred using the standard mail system commands, no matter how unnatural they seemed.
- Finally, many unfilled pauses went unnoticed due to the transcription method. In this experiment we only transcribed spoken input, including ahs which fill pauses, but ignore the long pauses which were not filled with any sounds. Whether this distinction will later be important remains an unanswered question.

Spoken input introduces additional complexity in addition to the problems encountered in understanding typed natural language input. This study has identified what we believe are the most important problems in human/computer interaction. Our efforts should focus on the elimination of speech-specific problems, such as the end-point problem (knowing when an utterance actually begins or ends), correctly modeling the occurrence of "irrelevant" events like filled pauses (i.e., *"ah"* or *"hm"*), task unrelated phrases and false starts, repeated words and broken off utterances. Without these phenomena filtered out, even perfectly recognized spoken words could not be understood by a natural language processor for typed input.

A further challenge will be to apply such techniques to the problem of understanding speech in real time, given an only imperfect recognition system: the true speech recognition scenario. Here the ability to skip phrases, predict *ahs* and *hms* and filter out begin problems must be used as an added constraint both at the parsing level and as feedback to the lexical access level of the system. The situations where it is legal to ignore a portion of speech must be carefully delineated and used to prune the search space.

Empirical research of the type described here will help direct the efforts of speech understanding projects towards the development of realistic application interfaces.

References

Adams, D.A. and Bisiani, R. (1986). The Carnegie-Mellon University Distributed Speech Recognition System. *Speech Technology, 3*, 14 - 23.

Allen, R.B. (1983). Composition and Editing of spoken letters. *International Journal of Man-Machine Studies, 19*, 181 - 193.

Biber, D. (1986). Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings. *Language, 62*, 384 - 414.

Borenstein, N.S. (1985). *The Design and Evaluation of On-line Help Systems.* Doctoral dissertation, Carnegie-Mellon University, Tech Report CMU-CS-85-151.

Chapanis, A. (1981). Interactive Human Communication:Some Lessons learned from laboratory experiments. In Shackel, B. (Ed.), *Man-Computer Interaction: Human Factors Aspects of Computers and People*. Rockville, MD: Sijthoff and Noordhoff. 65 - 114.

Erman, L.D. and Lesser, V.R. (1980). The Hearsay-II Speech Understanding System: A Tutorial. In Lea, W.A. (Ed.), *Trends in Speech Recognition.* Englewood Cliffs, NJ: Prentice-Hall. 340 - 360.

Ford, W.R., Weeks, G.D. and Chapanis, A. (1980). The effect of self-imposed brevity on the structure of dyadic communication. *Journal of Psychology, 104*, 87 - 103.

Gould, J.D. (1978). How experts dictate. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 648 - 661.

Gould, J.D. (1980). Experiments on Composing Letters: Some Facts, some Myths, some Observations. Gregg, L.W. and Steinberg, E.R. (Eds.), *Symposium on Cognition: Cognitive processes in writing.* Carnegie-Mellon University, Hillsdale, NJ: L. Erlbaum, 97 - 127.

Gould, J.D., Conti, J. and Hovanyecz, T. (1983). Composing letters with a simulated listening typewriter. *Journal of the Association for Computing Machinery, 26*, 295 - 308.

Grosz, B.J. (1977). *The representation and use of focus in dialogue understanding* Technical Note No. 151). Stanford, CA: SRI Stanford Research Institute,

Hauptmann, A.G. and Green, B.F. (1983). Comparing Command, Menu and Natural Language Computer Systems. *Behaviour and Information Technology, 2*, 163 - 178.

Hayes, P. J., Hauptmann, A. G., Carbonell, J. G., and Tomita, M. (1986). Parsing Spoken Language: a Semantic Caseframe Approach. *Proceedings of COLING-86.* Bonn, Germany: Association for Computational Linguistics.

Hayes, P.J. and Mouradian, G.V. (1981). Flexible Parsing. *American Journal of Computational*

Linguistics, 7, 232 - 241.

Hayes, P.J. and Reddy, D.R. (1983). Steps toward graceful interaction in spoken and written man-machine communication. *International Journal of Man-Machine Studies, 19*, 231 - 284.

Hindle, D. (1983). Deterministic Parsing of Syntactic Non-fluencies. *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 123 - 128.

Holmgren, J.E. (1983). Toward Bell System Applications of Automatic Speech Recognition. *Bell System Technical Journal, 62*, 1865 - 1880.

Leggett, J. and Williams, G. (1984). An empirical investigation of voice as an input modality for computer programming. *International Journal of Man-Machine Studies, 21*, 493 - 520.

Lowerre, B. T. (1976). *The HARPY speech recognition system.* Doctoral dissertation, Carnegie-Mellon University, Computer Science Department,

Morrison, D.L., Green, T.R.G., Shaw, A.C. and Payne, S.J. (1984). Speech Controlled Text-editing: effects of input modality and of command structure. *International Journal of Man-Machine Studies, 21*, 49 - 64.

Myers, J.L. (1972). *Fundamentals of Experimental Design.* Boston, MA: Allyn and Bacon.

Nye, J.M. (1982). Human Factors Analysis of Speech Recognition Systems. *Speech Technology, 1*, 50 - 57.

Poock, G.K. (1982). Voice Recognition boosts Command Terminal Throughput. *Speech Technology, 1*, 36 - 39.

Poock, G.K. and Armstrong, J.W. (1980). Effect of Operators Task Load on Performance of a Voice Recognition System. *Perceptual and Motor Skills, 51*, 506.

Stern, R.M., Ward, W.H., Hauptmann, A.G. and Leon, J. (1987). Sentence Parsing with Weak Grammatical Constaints. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 380 - 383.

Woods, W. A., Bates, M., Brown, G., Bruce, B., Cook, C., Klovstad, J., Makhoul, J., Nash-Webber, B., Schwartz, R., Wolf, J., and Zue, V. (1976). *Speech Understanding Systems - Final Technical Report* (Tech. Rep. 3438). Cambridge, MA: Bolt, Beranek, and Newman, Inc.,

I. Questionnaire evaluating experience with Electronic Mail

Subject #: Age: Date:

How often do you use electronic mail? (x-times daily/weekly/monthly)

When did you first start to use electronic mail?

How many hours of your life have you spent using electronic mail? Check one:
- less than 10
- between 10 and 100
- between 100 and 500
- over 500

What is the approximate number of electronic letters you type each week?

What is the approximate number of physical mail letters you type each week?

Compared to others in this department, do you consider yourself to be:
- expert electronic mail user
- intermediate electronic mail user
- novice electronic mail user

If you have ever used the electronic mail program called "hg" (MERCURY, also known as "RDMAIL" on some systems)

when have you first used it?

how often do you use it? (x-times daily/weekly/monthly)

how many total hours have you spent using it?
- less than 10 hours
- between 10 and 100 hours
- between 100 and 500 hours
- more than 500 hours

II. Background Instructions given to all Subjects

BACKGROUND OF THE MAIL PROGRAM

This program is designed to process mail messages.
It allows you to mail messages to other people on other machines,
read the new mail for you as well as search through messages that
you have received previously.

Each message consists of 7 parts:

message number - this is a number given to the message when
you receive it. You can refer to any message with this number.
It is suggested that you write down the message numbers of
the instruction messages that you will need to read today.

message sender - this is a person-id on some computer. You can
use this name when referring to one or more messages.
This is normally marked as the "From:" field in the message.

message date - the date the message was sent. You may refer to
one or more messages using the message dates.

carbon copy recipient - a person-id who will get a copy of the message
This name us usually marked as the "Cc:" field in the message.

recipient - the person-id for whom this message is intended.
This name usually is marked by the "To:" field in the message.

message body - The actual text of the message. You can search for
and refer to messages that contain specific words in the body.

message subject - The subject of the message. You can search for and
refer to messages that contain specific words in the header.
This field is usually marked as "Subject:"

A header of a message is a one-line summary of the message:

90 - 9 Jun 86 Ellen.Siegel@G.CS.CMU.EDU Re: Dinner Invitation (346)

The header above tells you this is message number 90 in your file
It was sent on June 9th, 1986 by Ellen.Siegel on the computer named
G.CS.CMU.EDU. The subject of the message was "Dinner Invitation".
The body of that message is 346 letters long.

III. Instructions for Subjects in the Speech-to-Computer Mode

This is an experiment to find out about people using natural language speech-understanding computers to process electronic databases and mail. You will be asked to participate in 2 sessions of about 15 to 20 minutes each, during which you will be using the mail processing program "hg" (mercury, also known as RDMAIL).
We have had some success with speech understanding computers here at CMU, but the current system is experimental and occasionally requires human help. The experimenter in the next room will provide this help, invisible to you.

The exact questions will be specified in the first piece of mail that you will receive. The database consists of pieces of mail that others have sent to the message file you are currently using and which has been loaded into your program.
To familiarize yourself with the procedure, you will practice on 3 training messages which will not be evaluated.
Please write down the numbers of the messages that you have not yet read so that you can return to them later.
After you have read all of the new messages and followed their instructions you should exit the mail program.

The system is designed to accept free, natural language speech. Therefore you should not feel restricted to use the unnatural command language implied by the "hg" mail processing program.

While using the system, you should remember two things:
1. Clearly SPEAK all the commands to the system into the microphone. Speak in the way that seems natural to you.

2. TYPE everything you want to do in the EMACS editor.

This means that you can ask the system something, or tell it to do something verbally. But you must still type yourself, when you want to input or edit text. Thus "Dear John, I am fine, how are you. Love Bobbie-Jean" would all need to be typed by you. You can speak to the program when you want to tell it to retrieve a certain message or to delete a message, etc.

Before and after the experiment you will answer some questions about your background with electronic mail and your opinions on this particular system. If you have any questions at this point, feel free to ask the experimenter.

IV. Instructions for Subjects in the Speech-to-Human Mode

This is an experiment to find out about people communicating by speech to process electronic databases and mail.
You will be asked to participate in 2 sessions of about 15 to 20 minutes each, during which you will query a database using the mail/database program "hg" (mercury).

Rather than querying the computer directly, you will speak to the experimenter.
He will act as your interpreter and translate what you say into a mail processing command or database query, as appropriate.

The exact questions will be specified in the first piece of mail that you will receive. The database consists of pieces of mail that others have sent to "bovik" and which have been loaded into your program.
To familiarize yourself with the procedure, you will practice on 3 chores which will not be evaluated.

Since you are talking to a person,
you should not feel restricted to use the unnatural command language implied by the "hg" mail/database program.

Before and after the experiment you will answer some questions about your background with electronic mail and your opinions on this particular system.

V. Instructions for Subjects in the Type-to-Computer Mode

This is an experiment to find out about people using natural language understanding computers to process electronic databases and mail.
You will be asked to participate in 2 sessions of about 15 to 20 minutes each, during which you will query a database using the mail/database program "hg" (mercury).
We have had some success with language understanding computers here at CMU, but the current system is experimental and occasionally requires human help. The experimenter in the next room will provide this help, invisible to you.

The exact questions will be specified in the first piece of mail that you will receive. The database consists of pieces of mail that others have sent to "bovik" and which have been loaded into your program. To familiarize yourself with the procedure, you will practice on 3 chores which will not be evaluated.

The system is designed to accept free, natural language typed input. Therefore you should not feel restricted to use the unnatural command language implied by the "hg" mail/database program.
While using the system, you should remember to
TYPE everything you want to do in the way that seems natural to you.

Before and after the experiment you will answer some questions about your background with electronic mail and your opinions on this particular system.

VI. Attitude Questionnaire

Please give your frank opinions in answering the following questions. The next x pages have statements of opinion, followed by the words "AGREE" and "DISAGREE" separated by dashes and colons. The dashes correspond to the different degrees of intensity of agreement and disagreement. Between each pair of them, put one (and only one) check mark where it most accurately reflects your opinion.

For example, if you were to see the statement "It is cold outside" and you QUITE agree with that statement, you would mark the sheet in the following way:

It is cold outside.
STRONGLY AGREE :____:_XX_:____:____:____:____:____: STRONGLY DISAGREE

However, if you think you completely disagree, mark it this way:

STRONGLY AGREE :____:____:____:____:____:____:_XX_: STRONGLY DISAGREE

If you feel you neither agree nor disagree, put a check mark on the middle dash of that line.

IMPORTANT!
1) Please respond to every set of statements, even if some don't seem to describe your opinion exactly. DO NOT skip any.
2) Put each check mark on the dashed line BETWEEN colons.
3) Remember to use only one check mark for each pair of choices.

I was easily able to do what I needed with the program.
STRONGLY AGREE :____:____:____:____:____:____:____: STRONGLY DISAGREE

The system was unforgiving of mistakes.
STRONGLY AGREE :____:____:____:____:____:____:____: STRONGLY DISAGREE

The system makes simple tasks difficult.
STRONGLY AGREE :____:____:____:____:____:____:____: STRONGLY DISAGREE

The time spent learning how to use this system is worthwhile, considering what you can do with it.
STRONGLY AGREE :____:____:____:____:____:____:____: STRONGLY DISAGREE

I would rather handle my mail by hand or by conventional electronic mail, than use this system.
STRONGLY AGREE :____:____:____:____:____:____:____: STRONGLY DISAGREE

I was satisfied with the way I was able to use this system.
STRONGLY AGREE :____:____:____:____:____:____:____: STRONGLY DISAGREE

It was easy to find and correct mistakes.
STRONGLY AGREE :____:____:____:____:____:____:____: STRONGLY DISAGREE

I was poorly prepared to do the tasks with the system.
STRONGLY AGREE :____:____:____:____:____:____:____: STRONGLY DISAGREE

I feel very confident about speaking to a computer.
STRONGLY AGREE :____:____:____:____:____:____:____: STRONGLY DISAGREE

I would use this sort of speech understanding program, if it were available.
STRONGLY AGREE :____:____:____:____:____:____:____: STRONGLY DISAGREE

Please describe what you think of such speech understanding system:

STIMULATING	:____:____:____:____:____:____:____:	DULL
RANDOM	:____:____:____:____:____:____:____:	PREDICTABLE
DIFFICULT	:____:____:____:____:____:____:____:	EASY
COMPLICATED	:____:____:____:____:____:____:____:	SIMPLE
ACCEPTABLE	:____:____:____:____:____:____:____:	UNACCEPTABLE
HINDERING	:____:____:____:____:____:____:____:	HELPFUL
OBEDIENT	:____:____:____:____:____:____:____:	BOSSY
DESIRABLE	:____:____:____:____:____:____:____:	UNDESIRABLE
IMPATIENT	:____:____:____:____:____:____:____:	PATIENT
DEPENDABLE	:____:____:____:____:____:____:____:	UNDEPENDABLE
FLEXIBLE	:____:____:____:____:____:____:____:	RIGID
FRUSTRATING	:____:____:____:____:____:____:____:	SATISFYING
EFFECTIVE	:____:____:____:____:____:____:____:	INEFFECTIVE
EFFICIENT	:____:____:____:____:____:____:____:	INEFFICIENT
ENJOYABLE	:____:____:____:____:____:____:____:	UNPLEASANT
HOSTILE	:____:____:____:____:____:____:____:	FRIENDLY

What did you like about this type of (speech) interaction?

What did you dislike about this type of interaction?

What suggestions do you have that would make this system better?

Can you list any particular problems you had during the experiment?

Where could you see a speech interface being useful?

VII. Utterances Containing Embedded AHs

1 how about AH with just honda in the header
2 ok can i AH see message number one eighty three
3 ok id id like to see AH all messages from martin stacey that that has AH ceedee or compact disk in the title
4 ok ahm id like to read AH message eighty three
5 ok ahm i would like to write out message number ninety two to a file and call it AH important dot document
6 send mail to AH alex at g
7 hm display all messages da AH can i start sorry
8 ok forward AH that last message to rudnicky on the a
9 ok lets see AH message one eighty four
10 are there any AH messages about camera repair
11 ok print AH sixty nine through seventy two
12 ok did martin stacey send anything about AH ceedees
13 the last AH eight months
14 ok ah ok copy that last message to a file and put it in AH user b l h archive
15 ok show me the messages or all messages with AH honda in the subject field
16 ah call it just AH message dot text or something
17 can i see all the messages which contain in their bodies the word AH mckean
18 type AH eighty nine
19 ok answer AH message one eighty two
20 oh sorry AH send this
21 ahm print out the messages which contain AH camera repair in the header
22 display message one eighty AH or excuse me one AH ninety one
23 display the message from martin stacey AH concerning ceedees
24 ahm we want to include this message in some mail that we are about to send AH i want to
25 append this to AH temp AH temp two
26 forward forward one eighty three to AH rudnicky on a
27 ah next no wait i take that back AH new headers
28 forward eighty six AH wait a second no no no no no dont do that header one eighty two

VIII. Task Messages for Session A

```
---- Message 182 (311 chars) is ----
Received: from SPEECH2.CS.CMU.EDU by CAD.CS.CMU.EDU; 15 Jun 86 20:16:25 EDT
Date: 15 Jun 1986 20:15-EDT
From: Alexander.Hauptmann@speech2.cs.cmu.edu
To: cousin@cad
Subject: newtask.1

I intend to buy a new Honda from a dealer called "McKean Honda."
Is there any information in the database that is relevant?


---- Message 183 (321 chars) is ----
Received: from SPEECH2.CS.CMU.EDU by CAD.CS.CMU.EDU; 15 Jun 86 20:17:34 EDT
Date: 15 Jun 1986 20:16-EDT
From: Alexander.Hauptmann@speech2.cs.cmu.edu
To: cousin@cad
Subject: newtask.2

Please make sure that Rudnicky on the A gets to see this message
on his terminal.
(You must find a way to do that with this program)


---- Message 184 (285 chars) is ----
Received: from SPEECH2.CS.CMU.EDU by CAD.CS.CMU.EDU; 15 Jun 86 20:18:42 EDT
Date: 15 Jun 1986 20:16-EDT
From: Alexander.Hauptmann@speech2.cs.cmu.edu
To: cousin@cad
Subject: newtask.3

Let alex at G know (electronically) on what day of the week
you were able to read this message.
```

IX. Task Messages for Session B

```
---- Message 90 (288 chars) is ----
Date: Thu, 10 Jul 86 15:43:44 EDT
From: User.Studies.Lab@ZOG.CS.CMU.EDU
To: User.Studies.Lab@ZOG.CS.CMU.EDU
Subject: newtask.4

My camera needs to be cleaned or fixed. It occasionally sticks when
you press the shutter release.

Are there any recommendations about this in the database?


---- Message 91 (364 chars) is ----
Date: Thu, 10 Jul 86 15:45:38 EDT
From: User.Studies.Lab@ZOG.CS.CMU.EDU
To: User.Studies.Lab@ZOG.CS.CMU.EDU
Subject: newtask.5

I am thinking of ordering some compact disks by mail.
Somehow, I seem to remember that Martin Stacey has put something
in the database about CD's. In particular, you should make sure
 you
don't miss any recent information about this.


---- Message 92 (424 chars) is ----
Date: Thu, 10 Jul 86 15:48:32 EDT
From: User.Studies.Lab@ZOG.CS.CMU.EDU
To: User.Studies.Lab@ZOG.CS.CMU.EDU
Subject: newtask.6

This is an important document. Make sure it will not be destroyed or
deleted if your mail file is ever lost/deleted/destroyed or otherwise
corrupted.

Congratulations, now all you have to do is answer a final
few questions with pencil and paper.
Thank you for being so patient and cooperative.
```