A comparison of mouse and speech input control of a text-annotation systems

Institute for Perception Research
PO Box 513, 5600 MB Eindhoven

MMB/mmb 90/03
27.03.1990

Rapport no. 738

A comparison of mouse and speech input control of a text-annotation system

M.M. Bekker

A comparison of mouse and speech input control of a text-annotation system

M. M. Bekker

Institute for Perception Research (IPO)
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

March 1990

Table of contents

Abstract

This report describes an experiment on the use of speech as an input-control medium. A comparison was made between mouse and speech for control of a text-annotationsystem. With this system a document and the annotations added to the document could be displayed on a computer-screen. Because of the complexity of many of the computer systems used nowadays it is interesting to look at the consequences of using various in- and output media when working with a computer. Not only is it interesting to look at the differences between the use of several media, but also looking at the use of more than one medium at the same time can have promising possibilities.

Twenty-four subjects, all experienced secretaries, were asked to participate in the experiment that consisted of three sessions. On the first day two sessions were held; one with mouse- and one with speech control. On the second day they worked with the system while mouse- and speech control could be alternated at will. Their actions were registered and they were asked to fill in a questionnaire. The following aspects were registered: time elapsed between commands, sort of command, command sequence, mouse clicks that did not activate a button and speech commands that could not be fitted to a voice template. The last two aspects are the errors that were registered.

To evaluate both input-control media the following aspects were compared: **speed, efficiency** (minimum number of commands necessary to complete the task / total number of commands), **accuracy** (number of correct commands / total number of commands), working style (processing time / total number of commands), **preference** (expressed in the number of correct commands, when both media can be alternated at will). **Furthermore a subjective evaluation was made based on the opinion of the system given by the subjects**.

The results of this experiment show no significant difference in measures of speed between mouse and speech control. However, mouse subjects have a significantly lower score on the measure of errors and are significantly more accurate and efficient than speech subjects. Furthermore, mouse subjects used more time per command in session 1. In session 2 they used the same amount of time per command as speech subjects used. According to the data of the third session no difference was found in operational preference of mouse and speech commands. The questionnaire shows that 10 subjects prefer mouse control, 9 subjects prefer working with both systems at their disposal and 5 subjects prefer speech control. For speed and ease of use the mouse scores even higher (15 and 17 subjects, respectively).

Conclusions based on the comments given in the questionnaire:

-People, that prefer **speech**, usually think that having your hands free and being able to give a command while making corrections is important;

-And people, that prefer the **mouse**, think it is important to work fast without having to bother about being recognized properly;

-People, that prefer using **both** systems alternately, think that both systems have their advantages and can be combined for greater benefit.

1) Some mainly use the mouse for its speed, ease of use and reliability. They think that speech is faster only in some cases, for instance when two commands have to be given in a row. Or to be able to give a command when they are still making corrections.

2) Some mainly use speech to be able to give a command while still making corrections and not having to change from one device to another. They want to use the mouse in the special occasion that recognition does not function properly or doing a re-

pair action (going directly to the right annotation after several wrong interpretations).
3) Some use one system as much as the other.

However, there is a certain discrepancy between the data of the third session (percentages mouse- and speech commands) and the answers given to the questionnaire. The preference given by the subjects in the questionnaire does not always correspond with their behaviour during the third session. Thirteen subjects know already what they will choose before the third session (3 mouse, 1 speech, 9 both). Eleven subjects use both control systems and choose for one of the two. (7 mouse, 4 speech). It seems that at least some subjects used the third session as another try-out.

1. Introduction

Because of the complexity of many of the computer systems used nowadays it is interesting to look at the consequences of using various in- and output media when working with a computer. Not only is it interesting to look at the differences between the use of several media but also looking at the use of more than one medium at the same time can have promising possibilities. There are several indications that using multi-media can lead to an increase in user productivity. Of course results can be influenced very much by the combinations of tasks that are chosen.

In the framework of the ESPRIT-project (European Strategic Program of Research and Development in Information Technology) some experiments already have been done at the Institute for Perception Research (IPO) to investigate the possibility of using a speech interface. So far the following areas have been looked at :

Speech output:

- provision of spoken information on system control, both before and during task execution: instructions and help messages (Kraak, 1988)
- addition of spoken information such as comments or criticism to other content information which is presented visually (Kraaij, 1987, 1988).

Speech input:

- application of voice commands for system control purposes (Sprenkels, 1988)
  In this experiment speech- was compared to mouse input. Professional secretaries were asked to read a prepared text word-by-word from paper, together with some simple lay-out commands, e.g. 'center'; 'new line'. Special commands as 'text word' were used to escape default interpretation of words such as 'period'. Speech recognition was simulated by the experimenter; he pressed a single key for every correct input, thus displaying already formed words to the subject. He found that ten out of twelve subjects preferred speech over mouse control. In this case subjects used the two media at different sessions.

De Vet and Beuk (1989) compared mouse and speech control of a diary keeping system. During this experiment voice recognition was also simulated. Subjects were novice users. They found that preference given by subjects does not always correspond with their actual behaviour during the experiment. Ten out of twelve subjects indicated that they preferred speech control over mouse control. But two of these ten mainly used mouse commands during the experiment.

In both experiments about speech input voice recognition was simulated. It is interesting to see which input media subjects would prefer when using the technology for voice recognition that is now available. In the experiment described in this report speech- and mouse control of the annotation system could be used alternately (as in the diary keeping system by de Vet and Beuk). The voice recognition was realized by a speaker dependent speech recognition board. These two media were compared on measures of speed, efficiency, accuracy and working style. Furthermore a subjective evaluation is given based on the opinion of the system given by the subjects.

2. Method

2.1 Material

The system that was used is an annotation system that has been designed at the IPO. Annotations (corrections and remarks) that have been added to a document can be stored in a computer and displayed on a computer screen. During the experiment three different documents of 2 pages each were used. More details of the documents are given in appendix A.1.

The instructions for the mouse- and the speech session differed only in instruction about the control-device. The instruction for session three was a shorter version of those used for the first two sessions. The corrections were made on a piece of paper that had the same text printed on it as was presented on the screen. Therefore, the system did not get more complex by adding an editor. Getting used to the editing system would also take extra time. Furthermore lack of experience with the editor might influence processing time in an unknown manner. The task setting of editing on a piece of paper as compared to editing on a computer is similar in the sense that the subject has to use her hands to do the editing. The advantage of speech control of not having to use your hands for the control function can still be tested. A description of the script annotation system can be found in Kraaij (1988).

The following adjustments were made to the original system:

- mouse control was added to the system. The picture of the command menu was adapted (figure 1 and 2): each command can be given by activating the button on the screen. In fact there are now three versions of the annotation system: with speech control, with mouse control and one with both speech- and mouse control;
- the spy-file was adapted. Counters were added to register how many commands were given and how many errors were made with a certain input-device;
- a third document with matching annotations was added to the system. When starting up the system, it is possible to choose which document is displayed on the screen.
  Although document 3 had to be similar to the two documents that were made for the original annotationsystem some differences were built into the annotations that had to be processed. After having some try-outs it was decided that using the footnote annotations again in the new document, subjects might become irritated by having to do exactly the same task again. To make sure subjects would still use the command "extra" two annotations had to be displayed on the screen at the same time to be able to make abbreviations consistent throughout the whole document;
- because of the addition of mouse control the original way of displaying the commands had to be adjusted. To make both systems as much alike as possible the pictures of the menu now show buttons of the commands that can be used (figure 1). In the original system only the words were shown (figure 2).

2.2 Comparing the mouse and the speech version

To be able to let subjects really use both media alternately, without having to change from the 'mouse mode' to the 'speech mode', the layout of the screen had to be the same in the mouse and the speech version. In that case subjects could give a mouse or speech command at any time during session three. The mouse and the speech version differed only in the way the command could be given (respectively saying a command or clicking a mouse-button to activate a command-button on the screen).

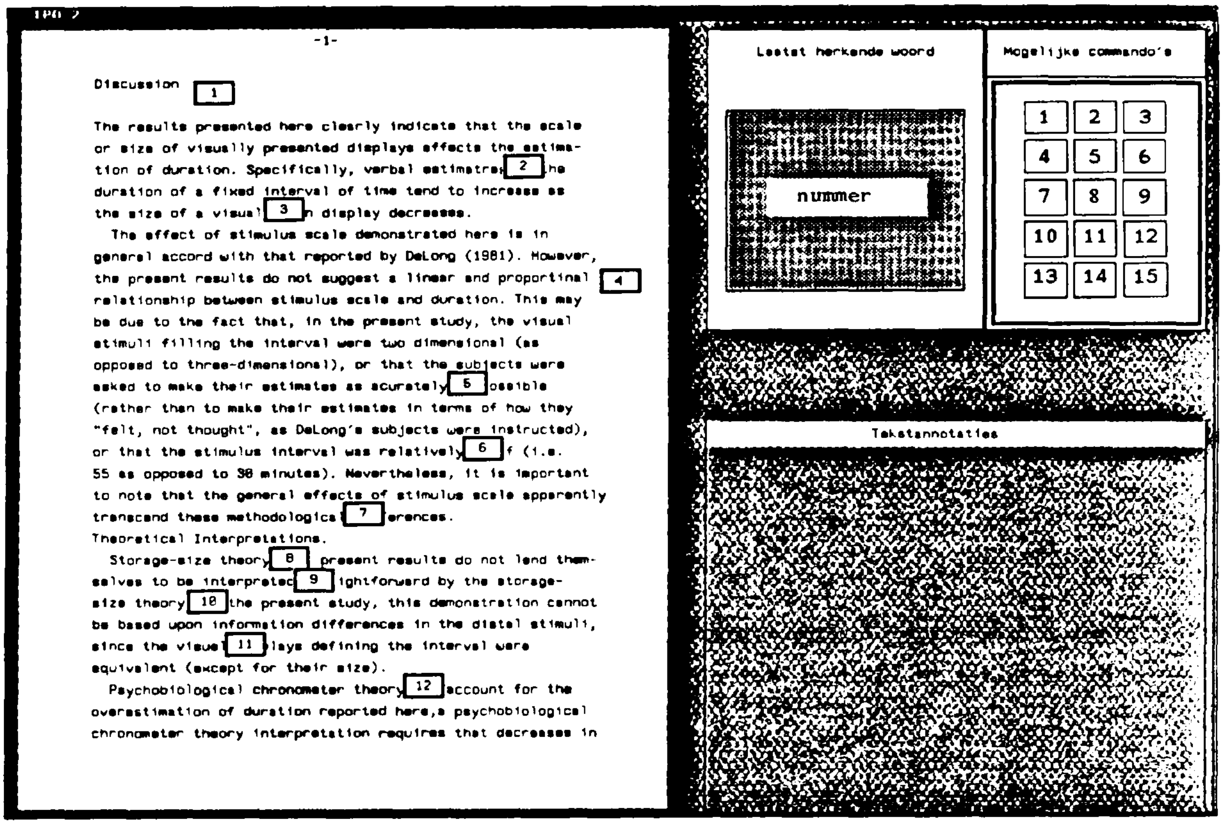


Figure 1: present picture

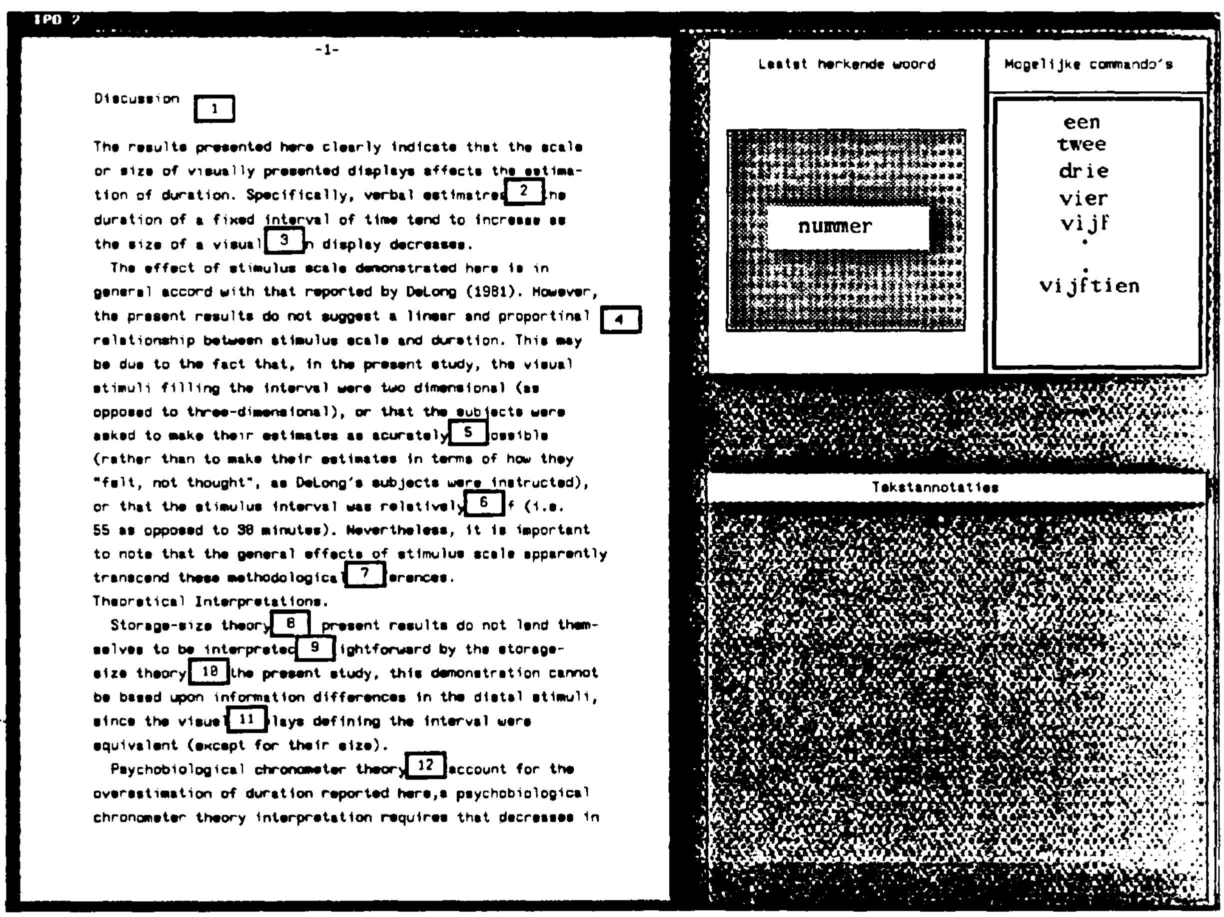


Figure 2: original picture

2.3 Task setting

In the experiment a secretary task was chosen. The setting of the experiment is as follows: an author has written a document, it has been typed on the Word Processor and afterwards the author has entered his annotations on the typed document in a computer. Subjects had to process the annotations of a document in the experimental setup. Each document consisted of two pages, with a total of twenty-two annotations.

For previous experiments two versions of the annotation system were built. In the script version the annotations were given in written form on the screen and in the speech-annotation system the comments were given by voice. Out of the speech- and the script-annotation systems the latter was chosen. When the speech-annotation system is used the subject does not have to look at the screen. It is then possible to completely ignore the feedback given about the last given command. In previous experiments this sometimes caused problems. Sometimes a command was interpreted as the command "abort" and if the subject did not look at the screen, the error made by the system would not be detected. If the next command was erroneously interpreted as "yes" the session would be ended, even though it was not the intention of the subject to abort the system.

It may take years for people to reach a constant work strategy. It is not possible to reach such a state in an experimental setup, where subjects have to work with an unfamiliar system. It is a drawback of the present setup that subjects may not have reached a final strategy yet. It will however be possible to determine the tendency of the work strategy. The results of the experiment can give an indication of the appreciation of using speech as a control device. Furthermore it gives the subjects a possibility to work with more than one modality at their disposal.

2.4 Procedure

Before starting each session the subjects were asked to read carefully the instructions handed to them on paper. After reading the instructions they were asked if they had any questions. After the first two sessions each subject was given only a short questionnaire to get a general idea of the opinion she has of the system. After session three a more specific questionnaire was handed out. After filling in this form the subject was asked about her opinion in more detail, to get a better idea of why a particular control-device was preferred.

The program for the subjects was as follows:

- Mouse session:
  - reading the instructions;
  - experimental task with mouse control;
  - questionnaire.

- Speech session:
  - reading the instructions;
  - training the speech recognizer;
  - experimental task with speech control;
  - questionnaire.

- Mouse- and speech session:
  - reading the instructions;
  - experimental task with mouse- and speech control;
  - questionnaire followed by a short interview.

2.5 Design.
To avoid transfer effects between the three sessions, a counter-balanced design was used. The order of both documents as well as control-devices was rotated. The total number of subjects was divided over twelve groups. Subjects could use only one control-device during sessions one and two. In session three subjects could use either one of the two control-devices. The lay-out of the screen was the same through all three sessions. The total number of subjects were divided into 12 groups:

	Session 1	Session 2	Session 3
group 1:	Speech / doc 1	Mouse / doc 2	Either / doc 3
group 2:	Speech / doc 2	Mouse / doc 3	Either / doc 1
group 3:	Speech / doc 3	Mouse / doc 1	Either / doc 2
group 4:	Speech / doc 1	Mouse / doc 3	Either / doc 2
group 5:	Speech / doc 2	Mouse / doc 1	Either / doc 3
group 6:	Speech / doc 3	Mouse / doc 2	Either / doc 1
group 7:	Mouse / doc 1	Speech / doc 2	Either / doc 3
group 8:	Mouse / doc 2	Speech / doc 3	Either / doc 1
group 9:	Mouse / doc 3	Speech / doc 1	Either / doc 2
group 10:	Mouse / doc 1	Speech / doc 3	Either / doc 2
group 11:	Mouse / doc 2	Speech / doc 1	Either / doc 3
group 12:	Mouse / doc 3	Speech / doc 2	Either / doc 1

Session one and two were held on the same day. Session three was held on the following day. The idea behind this setup was that the influence of preferring a control-device that a subject had used last, was minimized because of the intervening time between the first two and the last session.
In the experiment the following measurements were made: processing time, number of errors, number of commands given. During the three sessions of the experiment all input events, the time between two inputs, the number of inputs given with a certain input-device and the number of times a button was used were registered.

After each session the documents were checked to see whether all the corrections had been made. However, these data have not been used in the analysis.

2.6 Subjects.
To create a real-life setting for the experiment experienced secretaries were asked to be subjects. Conditions to participate in the experiment were:
- knowledge of and experience with the English language
- not having participated before in a similar experiment.

2.7 Apparatus

2.7.1 Hardware:
The program ran on a Sun 3/50 workstation, which has a monochrome monitor of 1152 (horizontally) x 900 (vertically) pixels. . The movements of the mouse on the table are translated in movements of a cursor with the shape of a crosshair on the screen with a multiplication factor of 2.5. The machine runs under a UNIX operating system. For speech recognition a Philips PC NMS 9100 was used, that was extended with a speech recogni-

tion board (Interstate Vocalink). The speech commands were captured by a microphone that was attached to a headset (Shure, SM10A).

2.7.2 Software:

The experiment was controlled by a Prolog program which runs a graphics program, that was designed with the aid of a rapid prototyping system developed at IPO (Van Nes and Beuk, 1987). The experiment was built in the "QUICK" software environment which consists of PROLOG with extended I/O possibilities, i.e. a graphical window for visual output, mouse control and communication with a PC for speech recognition.

3. Results

During all three sessions the measures for processing time, number of commands and number of errors were collected for each subject.

If a subject made errors with the mouse it can have the following reasons:
-the subject gave another command before the system was ready;
-the cursor was not placed on a button.
The first is a shortcoming of the system, while the last is a mistake of the subject.
If a subject made an error with the speech control system, it can have the following reasons:
-the speech-recognizer misinterpreted the given command;
-the speech-recognizer did not recognize the command at all;
-the subject made a noise that was not meant to be received by the voice recognizer;
-the subject gave a command that could not be used at that moment.
The first three are shortcomings of the system, while the last one is a mistake of the subject. The first category is not directly included in the data used to describe efficiency and accuracy, but in appendix B.4.1 the study of the spy-files give an impression of the occurrence of this type of error.

The data of session 1 and 2 were used to compare mouse and speech control on the following aspects:
-speed (processing time);
-efficiency (minimum number of commands necessary to complete the task / total number of commands);
-accuracy (number of correct commands / total number of commands);
-working style (processing time / total number of commands). With this measure the style a subject works in can be described. For example: a subject that has a high processing time can work slow and use a lot of time between commands, while another subject that has a high processing time might have checked out all annotations one more time, thus, also using a high number of commands.
Furthermore some side effects have been looked at:
-learning time (processing time session 1 - processing time session 2);
-learning style (number of commands session 1 - number of commands session 2).
These are used to study the effect, that the order of using the control systems has on working with and understanding the system.
The data of session 3 were used to compare preference (expressed in the number of correct commands with mouse and speech control) for both control systems.

Measurements of these variables were submitted to a statistical analysis. Analyses of variance were used to decide whether differences between both control systems were significant. The data of session two were not used to compare the two media, because the history from the first session may influence the results. For example: difference in learning time of subjects in group 1-6 and group 7-12 may depend on the order the two systems are offered to the subjects. After combining the data of session 1 and session 2 by introducing two extra variables (device order and document order), these data were used to decide whether there was an influence of device and document order, and to verify the conclusions based on the data of session 1. After the third session subjects were asked to fill in a questionnaire. A subjective evaluation was made based on the opinion of the system given by the subjects.

A description of the data can be found in appendix B. In appendix B.1 data about the subjects can be found (age, level of education). Data gathered in session 1, 2 and 3 can be found in appendix B.2. Data gathered in the questionnaire can be found in appendix B.3. An overview of the range and extremes of the measurements is given in appendix B.4. Next to the analysis of variance described in this chapter, the data was submitted to an analysis of regression, which can be found in appendix B.4.
In this chapter tentative conclusions will be given. The conclusions will be discussed in more detail in chapter 4.

Note:
Sometimes the system broke down due to bugs in the software of the annotation system and the UNIX operating system After break-down the system was started up again by the experimenter immediately. One group of subjects (12 secretaries), in a completely counter balanced design (2.5), had sessions in which the system did not break down. Even though no absolute assurance can be given that the subjects were not influenced by a break-down, the interviews and questionnaires did not give the impression that subjects were influenced by the interruption.

Furthermore a comparison was made between the data of a group of twelve subjects without break-downs and the data of all the subjects that did the experiment (24 subjects) to decide whether the outcome would have been influenced by the break-down of the system. The data of the subjects without break-downs tends to lead to the same conclusion on differences between mouse and speech control as the complete data. In some cases these differences are found not to be significant where they are significant in the analysis of the complete data. This is due to the fact that an analysis of data based on 12 subjects has less power (i.e. differences have to be larger to be significant) than one based on 24 subjects .

3.1 Sessions

Group 1 - 6 used speech in session 1 and mouse in session 2, group 7 - 12 used mouse in session 1 and speech in session 2 (see page 7).
An analysis of variance was used to decide whether there was a significant difference between the measures of mouse and speech control that were compared. No statistical analyses were used on the data of session 2 in itself, only the averages of the measures are given of these data.

Because it is hard to compare data of different experiments that already have been done (see Discussion 4.1), it is hard to conclude that one of the devices will be significantly better on one of the aspects that will be tested. Thus, the basic assumption is that there will be no difference in measures between the mouse and the speech version. The probability has to be less than .05 before a significant difference is found. The range and standard deviation of the measurements can be found in appendix B.4 in figures B.1 to B.20 to get a better understanding of why a difference between two measurements is significant or not.

3.1.1 Processing time

Figure 3 shows the processing time of session 1 and 2 averaged over all subjects for mouse (black bar) and speech (striped bar).

<u>Session 1:</u>
The mean processing time for the mouse version (1255 sec) is slightly higher than the

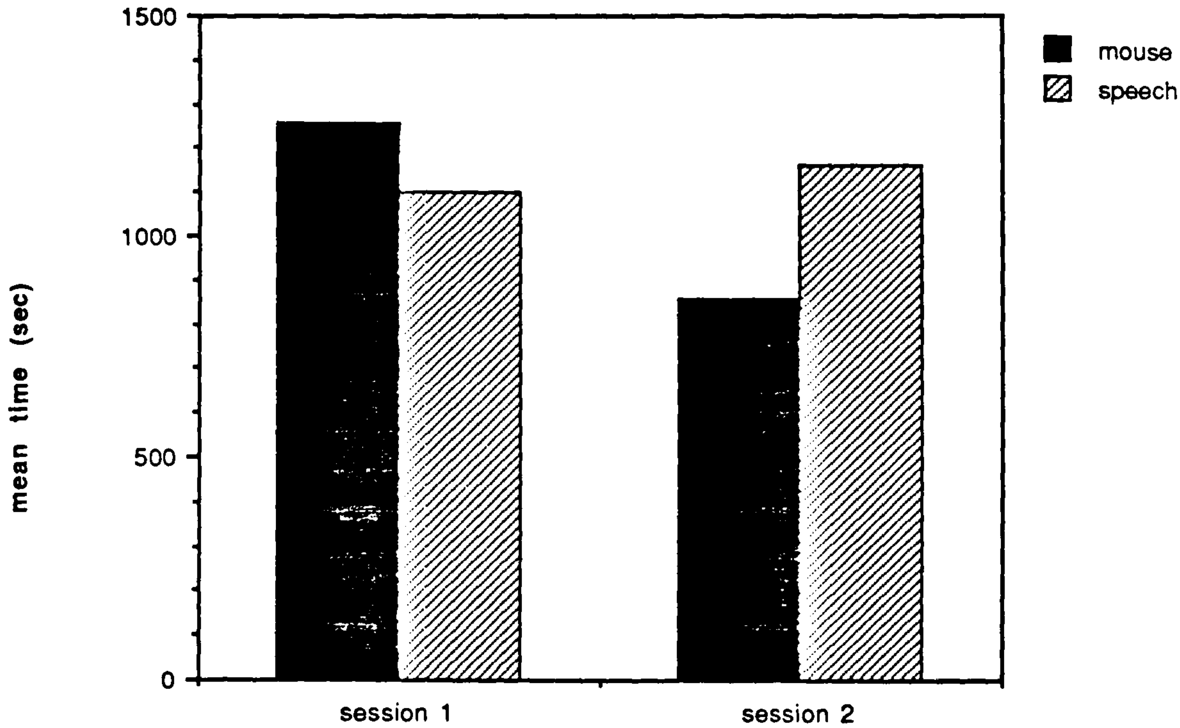


Figure 3: Mean time needed to complete the task by mouse- and speech subjects

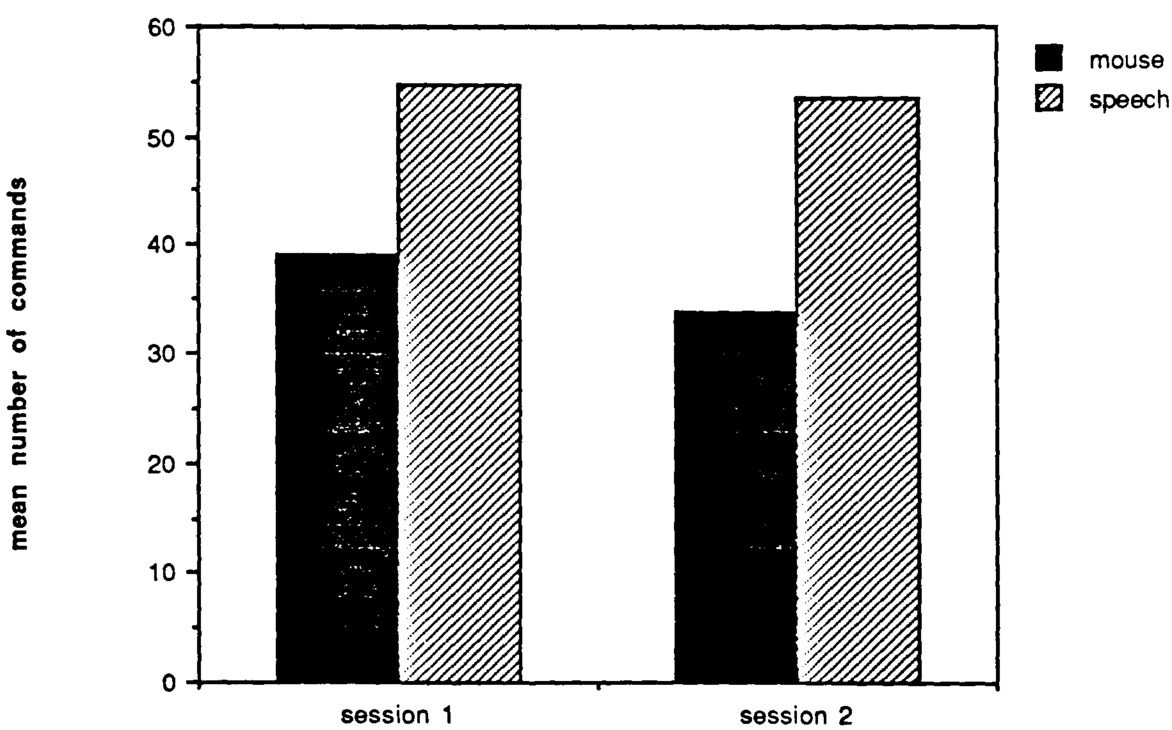


Figure 4: Mean number of commands

mean processing time for the speech version (1099 sec). An analysis of variance showed that this difference is not significant ($F_{1,18} = 1.810$, $p = 0.195$).

> So: In session 1 no significant difference was found between processing time of mouse- and speech subjects.

Session 2:
The mean processing time for the mouse version (860 sec) is lower than the mean processing time for the speech version (1163 sec).

Session 1 and 2:
No significant difference was found between processing time of subjects using mouse and subjects using speech control ($F_{1,18} = 1.354$, $p = .260$).
Device order had a significant influence on this measure ($F_{1,18} = 6.88$, $p = .017$), however, document order had no significant influence ($F_{2,18} = 1.534$, $p = .243$). The interaction between the factors input device and document order is significant ($F_{2,18} = 7.048$, $p = .005$).

> In session 1 and 2 no significant difference was found for the time that subjects needed to complete the task with mouse and with speech control. Device order has a significant influence on these measures. Note that the significance of input device order on the comparison of processing time with mouse and with speech control shows in the fact that in session 1 mouse subjects are slower than speech subjects, and in session 2 mouse subjects are faster than speech subjects (figure 3).
> The interaction between the factors input device and document order is significant. This means that influence of device order on processing time is not the same for all three document orders. The difference between processing time of mouse and speech subjects with one document order differs significantly from the difference between mouse and speech subjects that had one of the other document orders.
> No significant influence of document order is found, because the differences between influence of document order on processing time averaged over the total of mouse and speech subjects does not differ significantly.

3.1.2 Number of commands
Figure 4 shows the number of commands of session 1 and 2 averaged over all subjects for mouse (black bar) and speech (striped bar).
Session 1:
The mean number of commands for the mouse version (39 commands) is lower than the mean number of commands for the speech version (55 commands). An analysis of variance showed that this difference is not significant ($F_{1,18} = 3.41$, $p = 0.081$).

> In session 1 no significant difference was found between number of commands of mouse- and speech subjects.

Session 2:
The mean number of commands for the mouse version (34 commands) is lower than the

mean number of commands for the speech version (54 commands).

Session 1 and 2:
A significant difference was found between number of commands given by subjects using mouse and subjects using speech control ($F_{1,18}$ = 5.396, p = .032).
Device and document order had no significant influence on this measure ($F_{1,18}$ = 0.227, p = .640 and $F_{2,18}$ = 1.512, p = .247 , respectively). The interaction between the factors input device and document order was not significant either ($F_{2,18}$ = 0.747, p = .488).

In session 1 and 2 subjects use more commands when working with speech control, than when they are working with mouse control. The measure is not significantly influenced by device and document order.

3.1.3 Number of errors

With speech control the following errors can be made:
-the speech-recognizer did not recognize the command at all, because the command could not be fitted to one of the templates, that were saved in the training session;
-the subject made a noise that was not meant to be received by the voice recognizer;
-the subject gave a command that could not be used at that moment;
-the speech-recognizer misinterpreted the given command.
In this experiment only the first three categories of errors are used in the statistical analysis. Because no registration was made of the commands the subjects actually gave, the number of commands that were misinterpreted by the speech-recognizer could not be measured, and thus the last category could not be included in the statistical analysis. The spy-files of the subjects were studied to get an impression of the number of errors in the last category (appendix B.4.1). After further studies of the data it turned out that in some cases the speech-recognizer misrecognized commands very frequently. As a result of this problem the subjects had to restore the action by giving a few extra commands. If for example "previous" was recognized in stead of "next", the subject had to give either twice the command "next" or the commands "number" and "two" to restore the action.

Figure 5 shows the number of errors of session 1 and 2 averaged over all subjects for mouse (black bar) and speech (striped bar).
Session 1:
The mean number of errors for the mouse version (0.83 errors) is lower than the mean number of errors for the speech version (9.25 errors). An analysis of variance showed that this difference is significant ($F_{1,18}$ = 20.61, p = 0.0003).

In session 1 mouse subjects make less mistakes than speech subjects.

Session 1:
The mean number of errors for the mouse version (0.83 errors) is lower than the mean number of errors for the speech version (8.17 errors).

Session 1 and 2:
A significant difference was found between number of errors given by subjects using mouse and subjects using speech control ($F_{1,18}$ = 20.981, p = .000).

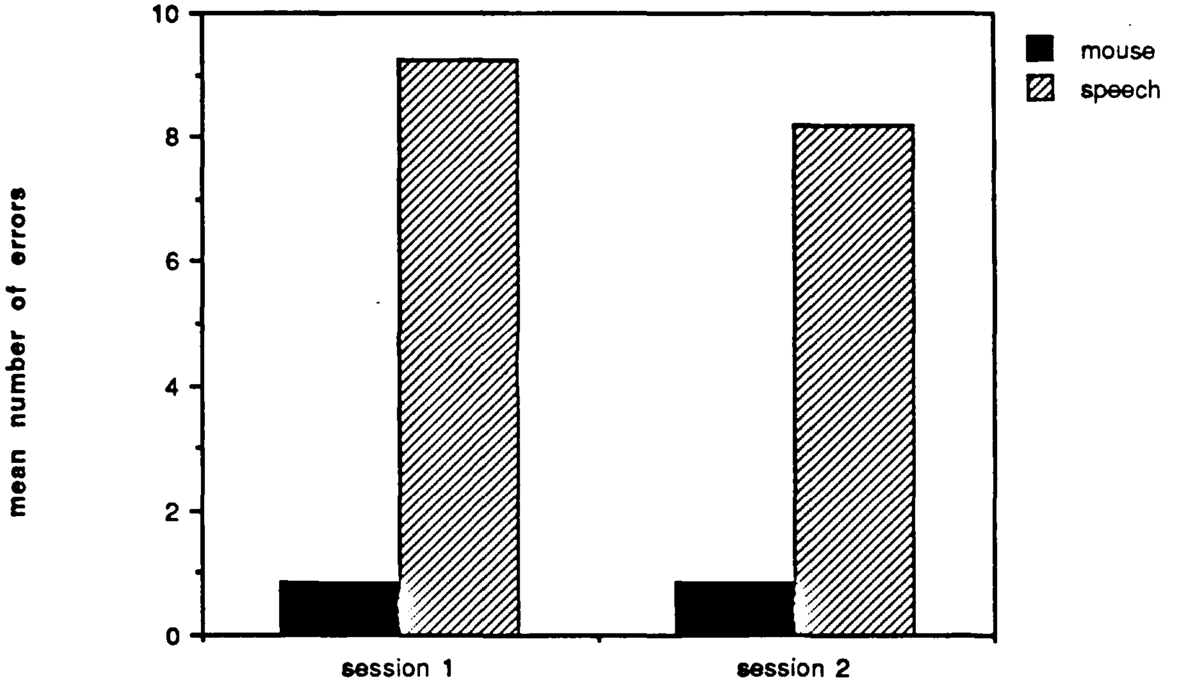


Figure 5: Mean number of errors for mouse- and speech subjects

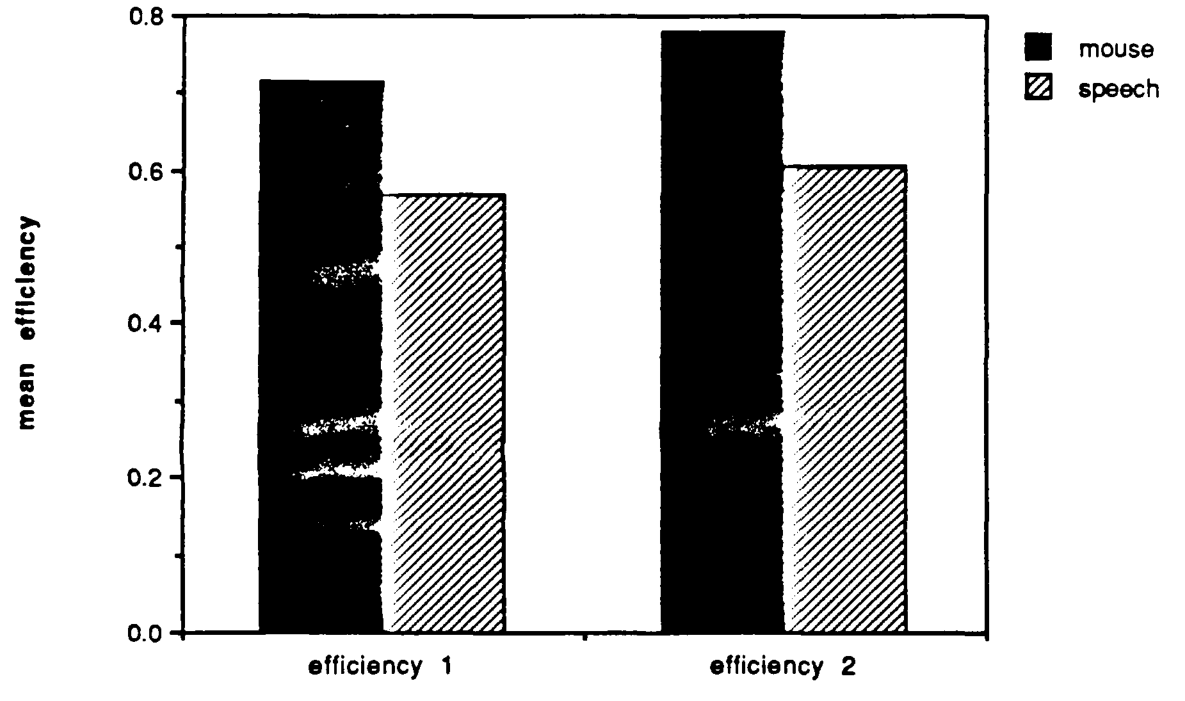


Figure 6: Mean efficiency (minimum number of commands needed to complete the task/total number of commands)

Device and document order had no significant influence on this measure ($F_{1,18}$ = 0.088, p = .771 and $F_{2,18}$ = 1.339, p = .287 , respectively). The interaction between the factors input device and document order was not significant either ($F_{2,18}$ = 1.870, p = .183).

> In session 1 and 2 subjects make less errors when they are using the mouse, than when they are using speech. The measure is not significantly influenced by device and document order.

3.1.4 Efficiency

Efficiency is defined as the minimum number of commands necessary to complete the task divided by the total number of commands given by the subject. Figure 6 shows the efficiency of session 1 and 2 averaged over all subjects for mouse (black bar) and speech (striped bar).

Session 1:

Mouse subjects use an average of 1.5 times the minimum number of commands (efficiency = 0.67), while speech subjects use an average of roughly twice (2.1) the minimum number of commands (efficiency = 0.48) required to perform the task. An analysis of variance showed that this difference is significant ($F_{1,18}$ = 6.42, p =0.021).

> In session 1 a significant difference was found between efficiency of mouse and speech subjects. Mouse subjects are more efficient than speech subjects.

Session 2:

The mean efficiency for the mouse version (0.77) is higher than the mean efficiency for the speech version (0.48).

Session 1 and 2:

A significant difference was found between efficiency given by subjects using mouse and subjects using speech control ($F_{1,18}$ = 12.644, p = .002).

Device and document order had no significant influence on this measure ($F_{1,18}$ = 1.120, p = .304 and $F_{2,18}$ = 1.232, p = .315 , respectively). The interaction between the factors input device and document order was not significant either ($F_{2,18}$ = 1.946, p = .172).

> In session 1 and 2 subjects are more efficient when they are using mouse than when they are using speech control. The measure is not significantly influenced by device and document order.

3.1.5 Accuracy

Accuracy is defined as the number of correct commands (total number of commands - number of errors) divided by the total number of commands.

Figure 7 shows the accuracy of session 1 and 2 averaged over all subjects for mouse (black bar) and speech (striped bar).

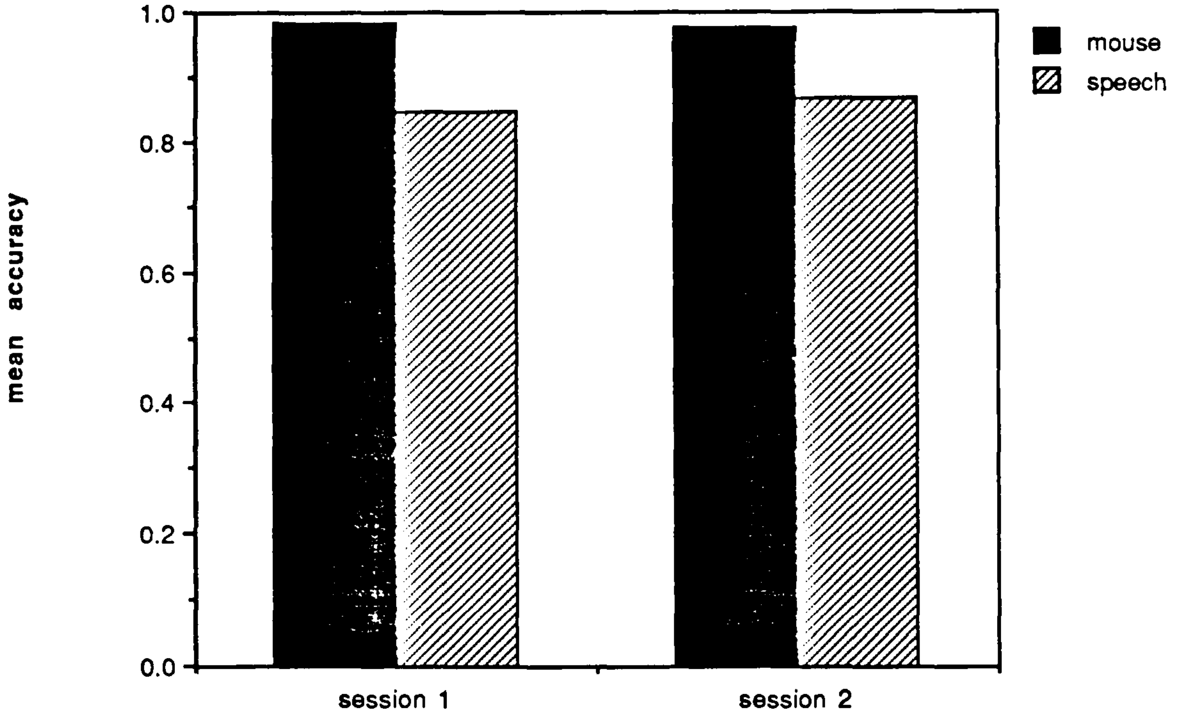


Figure 7: Mean accuracy (number of correct commands/ total number of commands)

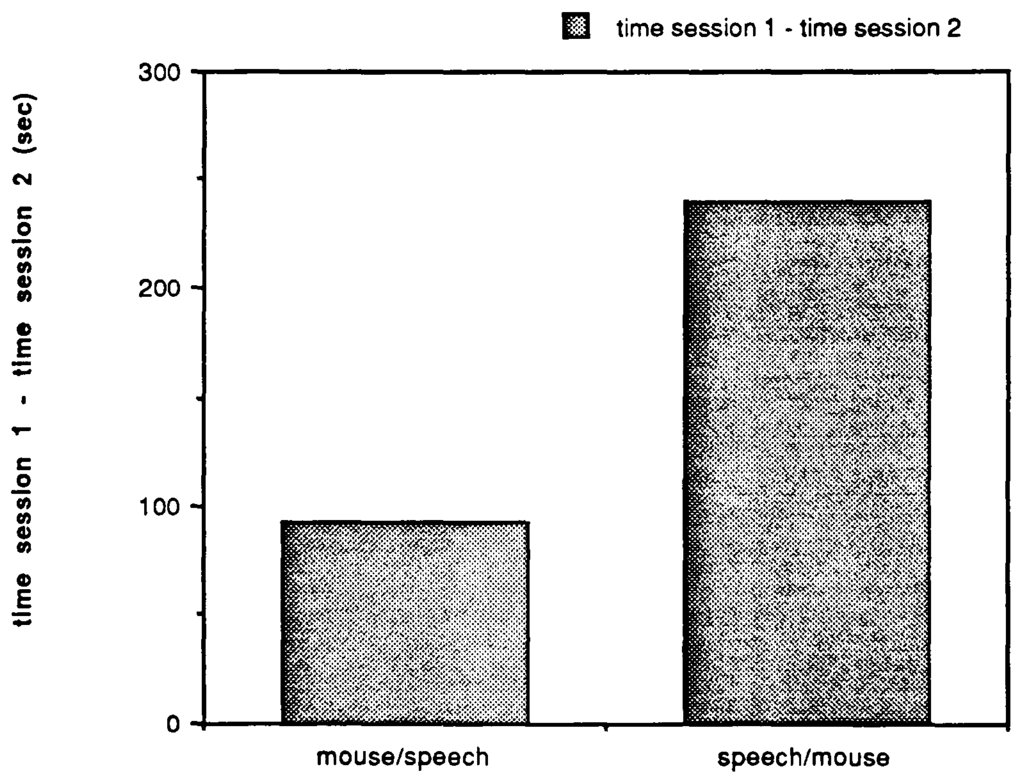


Figure 8: Average difference of time needed to complete the tasks in session 1 and 2

Session 1:
The mean accuracy for the mouse version (0.984) is higher than the mean accuracy for the speech version (0.846). An analysis of variance showed that this difference is significant. ($F_{1,18} = 30.549$, $p = 0.0001$)

In session 1 mouse subjects are more accurate than speech subjects.

Session 2:
The mean accuracy for the mouse version (0.977) is higher than the mean accuracy for the speech version (0.869).

Session 1 and 2:
A significant difference was found between accuracy given by subjects using mouse and subjects using speech control ($F_{1,18} = 46.436$, $p = .000$).

Device and document order had no significant influence on this measure ($F_{1,18} = 0.109$, $p = .745$ and $F_{2,18} = 0.598$, $p = .560$, respectively). The interaction between the factors input device and document order was not significant either ($F_{2,18} = 2.466$, $p = .113$).

In session 1 and 2 subjects are more accurate when they are using mouse than when they are using speech control. The measure is not significantly influenced by device and document order.

3.1.6 Learning time
Figure 8 shows the learning time of mouse/speech and speech/mouse order.
Learning time is defined as the difference between the processing time of session one and the processing time of session two.

The mean processing time of the first session of the mouse/speech order (1255 sec.) is higher than the mean processing time in the second session (1163 sec.).
The mean processing time of the first session of the speech/mouse order (1099 sec.) is higher than the mean processing time in the second session (860 sec.).

The mean learning time of the mouse/speech order (92 sec.) is lower than the mean learning time of the speech/mouse order (236 sec.). An analysis of variance showed that this difference is not significant ($F_{1,18} = 1.354$, $p = .2598$).

Input device order has no influence on the learning time.

3.1.7 Learning style
Figure 9 shows the learning style of mouse/speech and speech/mouse order.
Learning style is defined as the difference between the total number of commands of the first session and the total number of commands of the second session.

The mean number of commands of the first session of the mouse/speech order (39 mouse commands) is lower than the mean number of commands in the second session (54 speech commands).

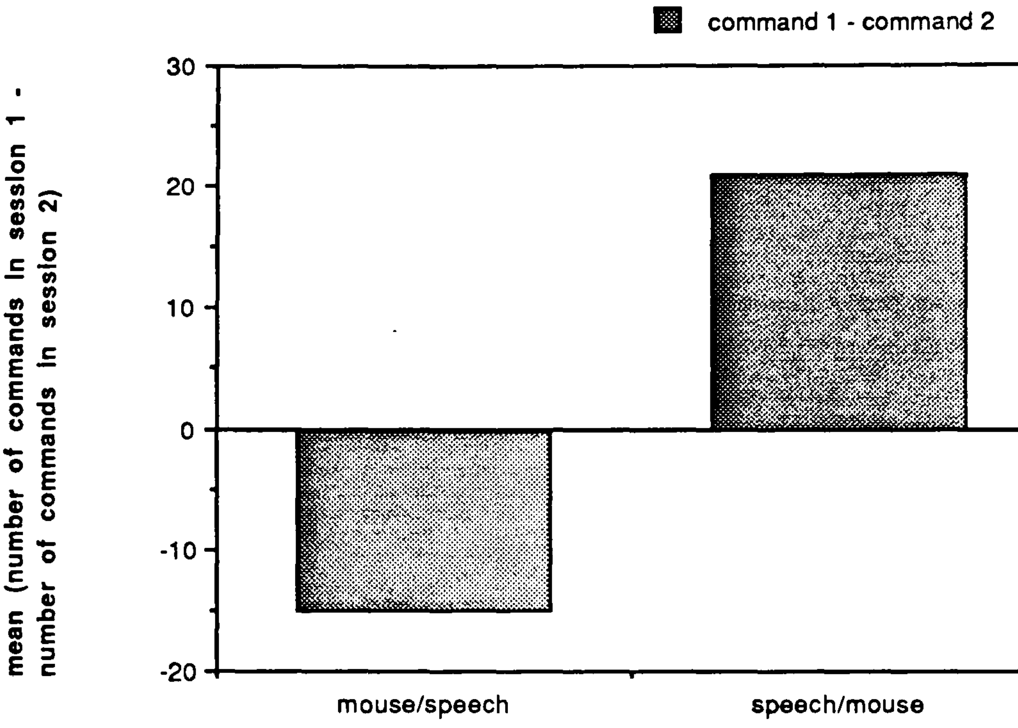


Figure 9: Average difference in number of commands given in session 1 and session 2

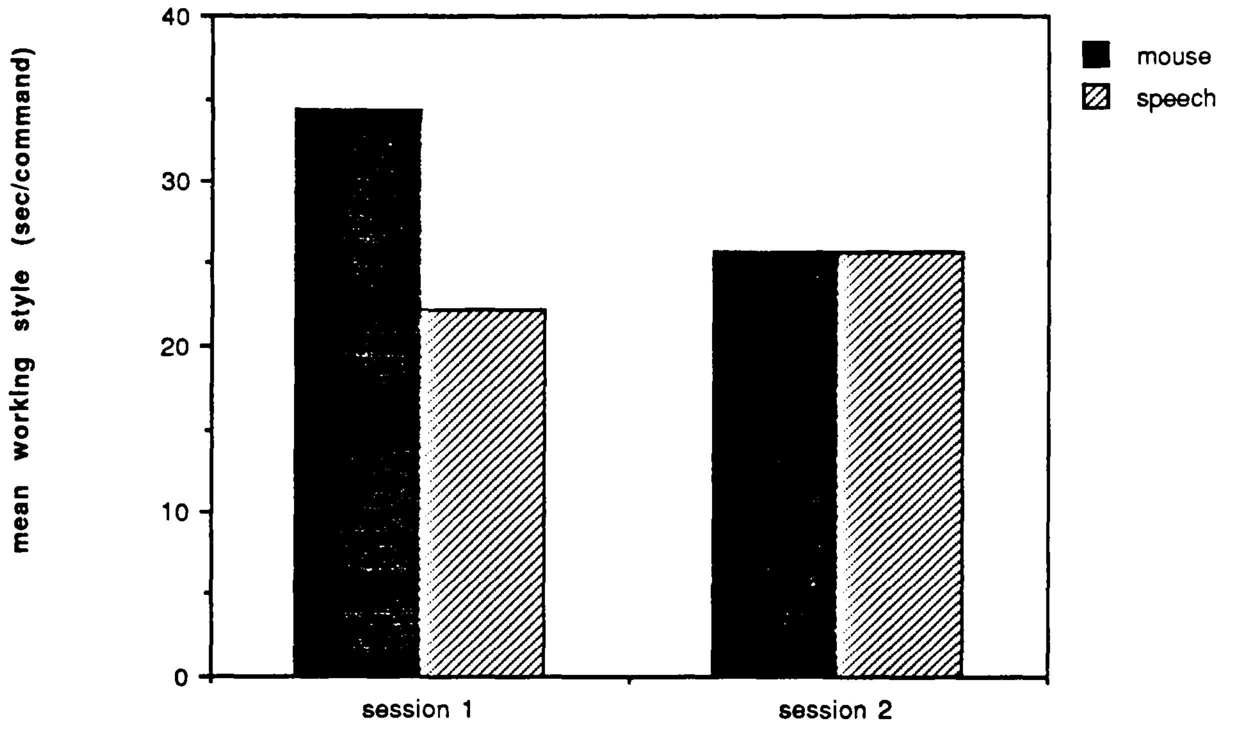


Figure 10: Mean number of seconds needed per command by mouse- and speech subjects

The mean number of commands of the first session of the speech/mouse order (55 speech commands) is higher than the mean number of commands in the second session (34 mouse commands).

The mean learning style of the mouse/speech order (-15 commands) is lower than the mean learning style of the speech/mouse order (21 commands). An analysis of variance showed that this difference is significant ($F_{1,18} = 5.396$, $p = .0321$).

> Input device order has an influence on learning style. Speech subjects used more commands than mouse subjects and thus [mouse commands - speech commands] is a negative and [speech commands- mouse commands] is a positive number of commands.

3.1.8 Working style

Working style is defined as the number of seconds per command (i.e. the processing time divided by the total number of commands).
Figure 10 shows the working style of session 1 and 2 averaged over all subjects for mouse (black bar) and speech (striped bar).

<u>Session 1</u>:
The mean working style for the mouse version (34 sec/command) is higher than the mean working style for the speech version (22 sec/command). An analysis of variance showed that this difference is significant ($F_{1,18} = 12.45$, $p = .0024$).

> In session 1 mouse subjects use more time per command than speech subjects do.

<u>Session 2</u>:
The mean working style for the mouse version (26 sec/command) is the same as the mean working style for the speech version (26 sec/command).

<u>Session 1 and 2:</u>
A significant difference was found between working style given by subjects using mouse and subjects using speech control ($F_{1,18} = 7.059$, $p = .016$).
Device and document order had no significant influence on this measure ($F_{1,18} = 1.080$, $p = .312$ and $F_{2,18} = 0.582$, $p = .569$, respectively). The interaction between the factors input device and document order was not significant either ($F_{2,18} = 2.642$, $p = .099$).

> Mouse subjects use more time per command than speech subjects do. The measure is not significantly influenced by device and document order. Note that the difference in tendency between mouse and speech subjects in session 1 (working style of mouse subjects is higher than the working style of speech subjects) and session 2 (working style for both groups is almost the same) does not show in this test, because the differences between sessions are averaged over these two sessions.

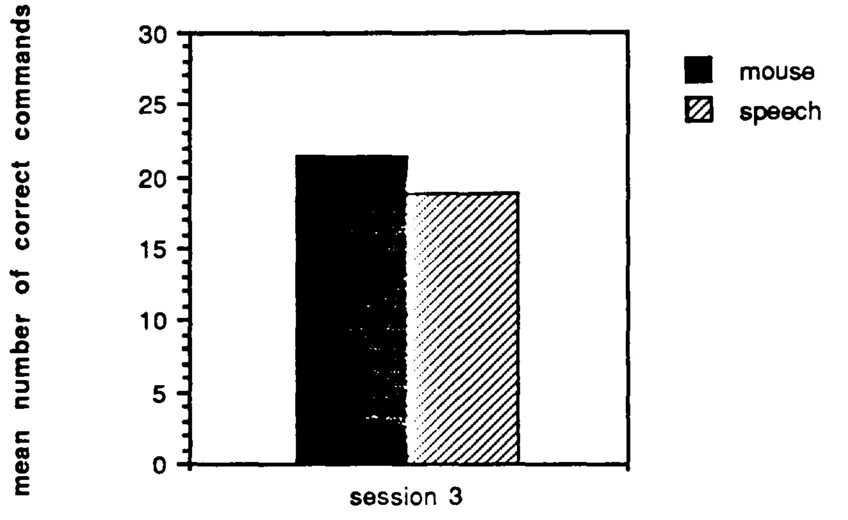


Figure 11: Mean number of correct commands given with mouse and with speech control

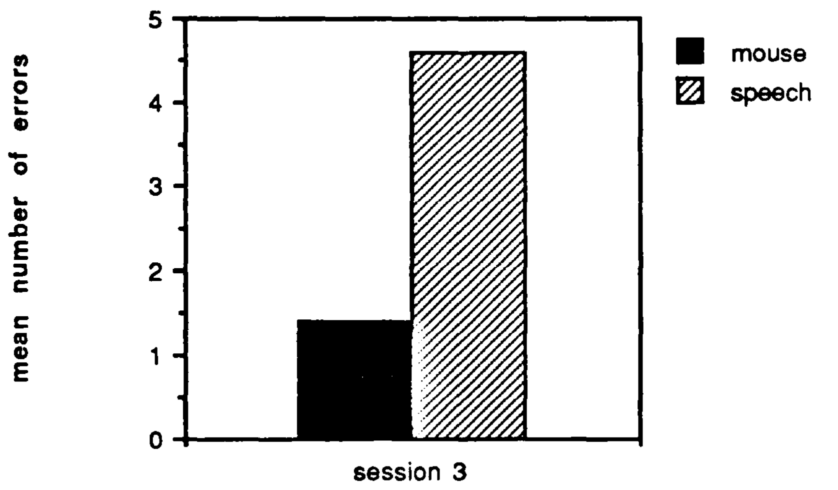


Figure 12: Mean number of errors made with mouse and speech control in session 3

3.1.9 Operational preference

An operational preference exists if subjects in session 3 use one control system significantly more often than the other control system. A comparison was made between the number of correct commands given with mouse and speech control. The number of incorrect commands was not included in the measure that was compared. Otherwise, if a subject made a lot of errors with one device, it might seem that this subject preferred working with this device. Looking at only the correct answers might lead to another conclusion. It is more accurate to use only the correct commands given by subjects, to decide with which control system they prefer to work. Figure 11 shows the number of correct commands in session 3 for mouse (black bar) and speech (striped bar).

The mean number of correct commands for the mouse version (21.5) is higher than the mean number of correct commands for the speech version (18.9). An analysis of variance showed that this difference is not significant ($F_{1,18} = 0.223$, $p = .642$).
Device and document order had no significant influence on this measure ($F_{1,18} = 1.80$, $p = .196$ and $F_{2,18} = 1.823$, $p = .190$, respectively). The interaction between the factors input device and document order was not significant either($F_{2,18} = 0.180$, $p = 0.837$).

> In session 3 no significant difference was found between the number of correct mouse- and speech commands. The measure is not significantly influenced by device and document order.

3.1.10 Number of errors in session 3

Figure 12 shows the number of errors of session 3 for mouse (black bar) and speech (striped bar).

The mean number of errors for the mouse version (1.0) is lower than the mean number of errors for the speech version (4.6). An analysis of variance showed that this difference is significant ($F_{1,18} = 10.428$, $p = 0.005$). Device and document order had no significant influence on this measure ($F_{1,18} = 0.512$, $p = 0.483$ and $F_{2,18} = 1.499$, $p = .250$, respectively). The interaction between the factors input device and document order was not significant either ($F_{2,18} = 0.070$, $p = 0.933$).

> In session 3 subjects made more speech than mouse errors. The measure is not significantly influenced by device and document order.

3.2 Questionnaire

The subjects were asked to fill in a questionnaire after the third session. Group 1 (subjects 1 - 12) used the mouse version in session 1 and the speech version in session 2. Group 2 (subjects 13 - 24) used the speech version in session 1 and the mouse version in session 2.
The questionnaire consisted of:
- a general question about the age and the level of education of the subjects.
- three questions about experience (respectively computer-, mouse - and speech control experience). These data are used to check if subjects in both groups had the same level of experience and to explain certain 'side-effects'.
- three questions to compare the control systems (respectively their choice for prefer-

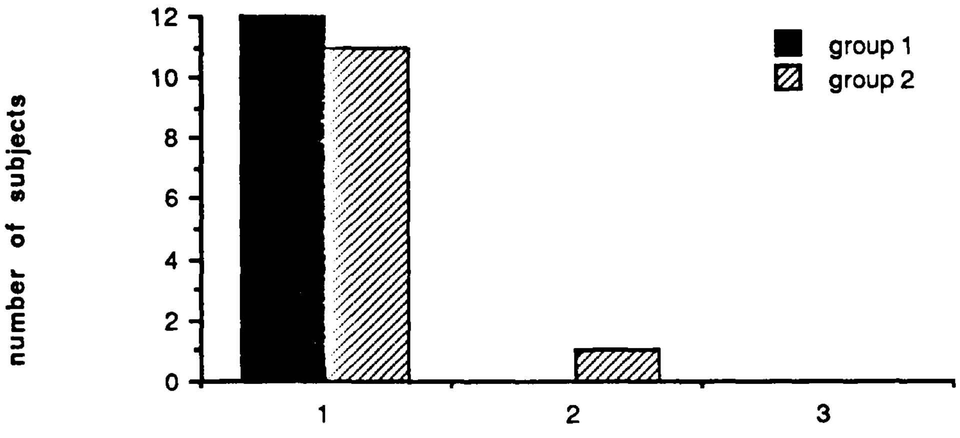


Figure 13 : Level of prior computer experience of the subjects
1=experience, 2=a little experience, 3=no experience
group 1 = mouse/speech order, group 2 = speech/mouse order

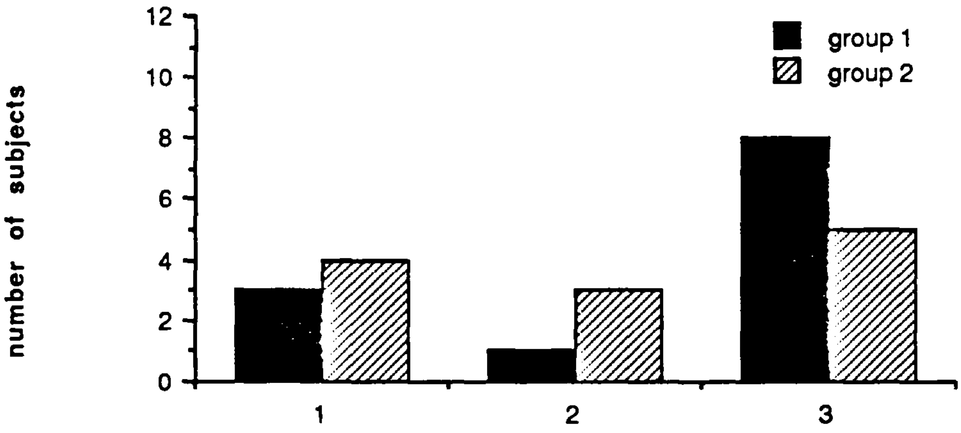


Figure 14: Level of prior mouse experience of the subjects
1=experience, 2=a little experience, 3=no experience
group 1 = mouse/speech order, group 2 = speech/mouse order

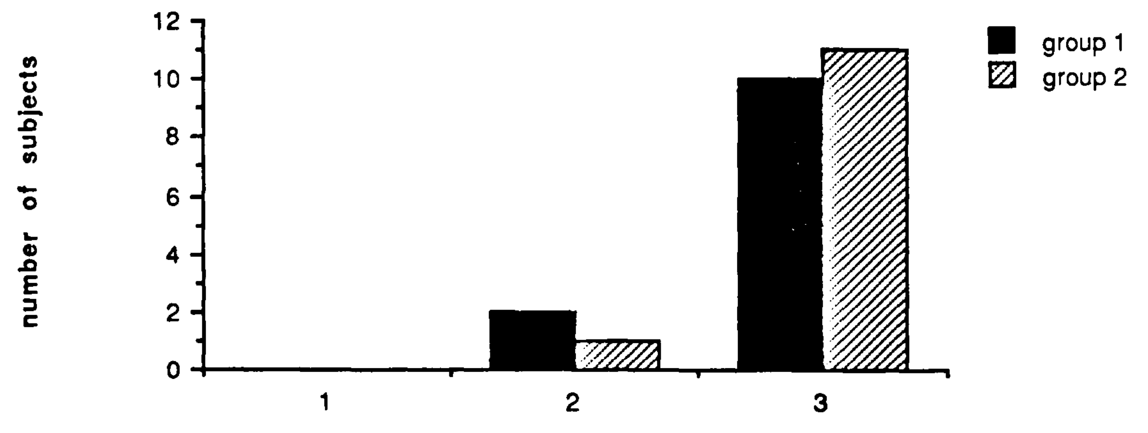


Figure 15: Level of prior speech control experience of the subjects
1= experience, 2=a little experience, 3=no experience
group 1 = mouse/speech order, group 2 = speech/mouse order

ence, speed and ease of use).
- a question about the positive and negative aspects of both control systems.
The data of the general question can be found in appendix B.1 and the data of the other questions can be found in appendix B.3.

3.2.1 Computer experience
Computer experience is defined as the score on the question: "Do you have computer experience?"
The subjects could choose one of the following options:
1: experience
2: a little experience
3: no experience

Figure 13 shows computer experience of group 1 (black bar) and group 2 (striped bar).
The twelve subjects of the first group have more experience than the second group.

3.2.2 Mouse experience
Mouse experience is defined as the score on the question: "Do you have mouse experience?"

The subjects could choose one of the following options:
1: experience
2: a little experience
3: no experience

Figure 14 shows computer experience of group 1 (black bar) and group 2 (striped bar).
Group 2 has more mouse experience than group 1.

3.2.3 Speech control experience
Speech control experience is defined as the score on the question: "Do you have speech control experience?"

The subjects could choose one of the following options:
1: experience
2: a little experience
3: no experience

Figure 15 shows computer experience of group 1 (black bar) and group 2 (striped bar).
The first group of subjects has more speech control experience than the second group.

3.2.4 Preference
Preference is defined as the score on the questions:
-"Which medium do you prefer to control the system with?"
The subjects could choose one of the following options:
1: mouse
2: speech
3: both (working with both control systems, is preferred over working with either mouse or speech control)
-"How strong is this preference?"

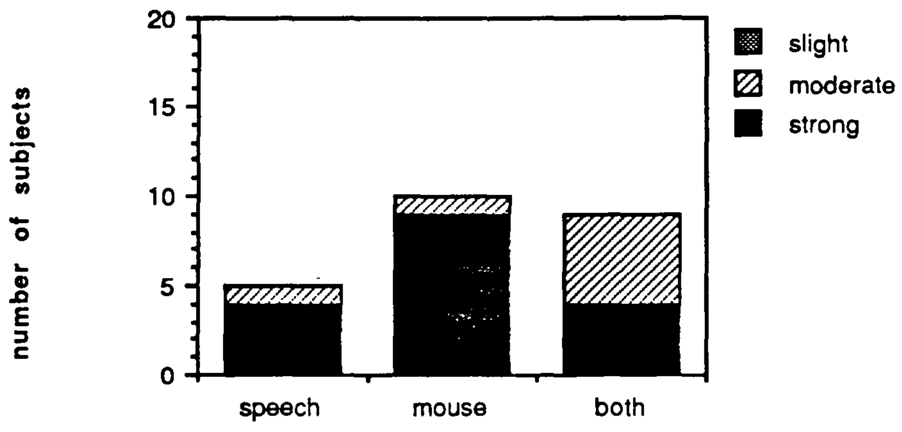


Figure 16: Input device preferred by the subjects

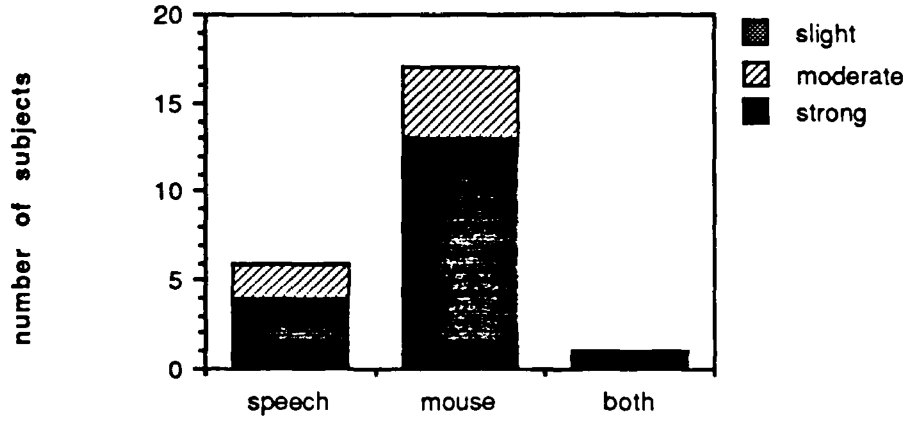


Figure 17: Input device preferred for ease of use by the subjects

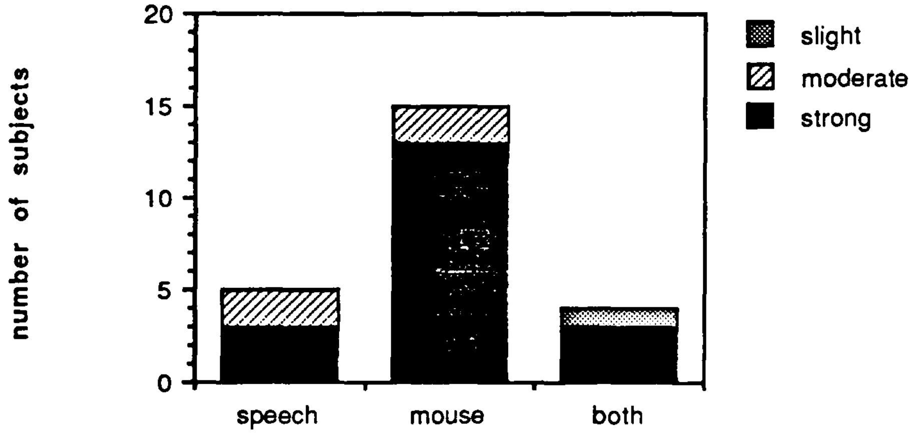


Figure 18 : Input device preferred for speed by the subjects

The subjects could choose one of the following options:
1: strong (black bar)
2: moderate (striped bar)
3: slight (dotted bar)

Figure 16 shows which medium the subjects prefer according to the questionnaire. The mouse is preferred by 10 subjects (42%), to work with both systems is preferred by 9 subjects (37%) and 5 subjects (21%) preferred to use speech control. No subjects had a slight preference.

3.2.5 Ease of use
Ease of use is defined as the score on the questions:
-"Which medium is the easiest to control the system with?"
The subjects could choose one of the following options:
1: mouse
2: speech
3: both (working with both control systems, is preferred over working with either mouse or speech control)
-"How strong is this preference?"
The subjects could choose one of the following options:
1: strong (black bar)
2: moderate (striped bar)
3: slight (dotted bar)

Figure 17 shows which medium the subjects prefer for ease of use according to the questionnaire. The mouse is chosen for ease of use by 17 subjects (71%), to work with both systems is chosen by 1 subject (4%) and 6 subjects (25%) chose for using speech control. No subjects had a slight preference.

3.2.6 Speed
Speed is defined as the score on the questions:
-"Which medium is the fastest to control the system with?"
The subjects could choose one of the following options:
1: mouse
2: speech
3: both (working with both control systems, is preferred over working with either mouse or speech control)
-"How strong is this preference?"
The subjects could choose one of the following options:
1: strong (black bar)
2: moderate (striped bar)
3: slight (dotted bar)

Figure 18 shows which medium the subjects prefer for speed according to the questionnaire. The mouse is chosen for speed by 15 subjects (62%), both systems are chosen by 4 subjects (17%) and 5 subjects (21%) chose for using speech control, for speed.

3.2.7 Positive and negative aspects of mouse- and speech control
The subjects were asked to give their opinion about both control systems. No options

were given; the subjects had to produce positive and negative aspects spontaneously.
These are the positive and negative aspects that were given by the subjects. The number of subjects that gave a specific aspect is listed after each aspect.

Speech:
+ not having to change from one input device to another and thus, not having to stop making corrections (6)
+ fast (6)
+ hands free (3)
+ saves time and actions (3)
+ easy to use (2)
+ exciting (1)
+ do not have to look at the screen (1)
- it is hard to speak consistent all the time (9)
- has to be used in a quiet environment (3)
- too slow (3)
- takes a lot of concentration (2)
- uncertain recognition (1)
- noisy (1)
- hard not to make any extra noises (1)
- it is often necessary to repeat commands (1)
- easier to make mistakes (1)
- headphone has to fit well and should not move with respect to the mouth when user moves (1)
- strange to talk to a machine (1)
- estrangement from the environment, people cannot talk to you anymore (1)
- people still have to get used to it (1)
- unfriendly and not practical as long as the speech-recognizer does not work properly (1)
- headphone (1)

Mouse:
+ fast (13)
+ easy to use (9)
+ certain (4)
+ effective and clear (1)
+ reliable (1)
- you have to stop making corrections to give a command (2)
- needs hands (2)
- no space on your desk if its already full (1)
- gets you out of rhythm if you are used to typing with ten fingers (1)
- would prefer using the keyboard (1)
- you have to get used to working with it (1)
- it should be placed somewhere where the board does not move very easily (1)
- you have to switch your focus point between the screen and the mouse (1)
- needs more skills (1)
-'double commands' like "number-two", "next-extra-next" are easier when using speech (1)

4. Discussion

In the present experiment a comparison between mouse- and speech input was made to decide which of the two input media is faster, more accurate, more efficient and comfortable to work with. Another issue was the preference for either one of the input media if subjects could use them alternately.
It is interesting to see to what conclusions and remarks other research in this field has lead to. This can be found in the first part of this chapter. After that the results of the present experiment will be dealt with. In the third paragraph a discussion is given based on a comparison between the results of the present experiment and the literature discussed in this chapter.

4.1 Aspects of speech and mouse control

First, a brief overview of selected literature shows a number of aspects that are typical of speech and mouse input. Some are general aspects, while some are specifically related to the technical realization of the system.

Foley, Wallace and Chan (1984) made an attempt to provide a systematic structure to aid a designer in his selection of devices and techniques by which the user communicates with a computer. He gathered information about this subject and tried to build a systematic structure. This structure is built on interaction tasks and techniques. He distinguishes six tasks and for each of these tasks 'all' possible techniques (media) are given. An evaluation of these techniques is given based on a comparison of measures of ergonomic quality. Table 4.1 shows the values that are given for these measures of ergonomic quality of the mouse and speech for a selection task. The ratings that are given are relative, based on their readings, experiments and experience.

Techniques	Cognitive load	Perceptual load	Motor load	Visual acquisition	Motor acquisition	Ease of learning	Fatigue	Error proneness	Feedback type
Voice-recognition	L	L	L	L	L	H	L	H	NA
Mouse	M	M	M	M	M	M	L	M	NA

L = low
M = medium
H = high

NA = not analyzed/ not available

Table 4.1: Comparison of interaction techniques

A drawback of these comparisons is that he combines conclusion of several experiments, although they have been run under different circumstances (different tasks, different subjects, etc.) and he does not explain how he combines them.. Furthermore, he does not give a precise definition of the measures of ergonomic quality.

Design guidelines for speech recognition interfaces can be found in an article by **Jones,**

Hapeshi and Frankish (1989). He first lists a number of advantages speech input has over manual methods of entry. Then he gives a list of guidelines to help overcome the areas of potential difficulty of speech recognition. A selection of these advantages and guidelines is given that is especially applicable to speech control.
Advantages:
-Speech may free the hands to undertake other tasks.
-In complex systems, the voice adds another channel through which information can be transmitted. In some cases this means that voice can overlap with other types of input. This will result in improved throughput of information.
-When more sophisticated speech recognition devices become available, the intonation and stress which enrich natural language can be used to advantage. For instance, the same word when pronounced in different ways could be used to produce different consequences.
-Unlike typed commands, spoken commands are never abbreviated or used in acronym form, which means that the degree of naturalness will be even more obvious than for other types of man-machine interaction. However: Any factor which enhances naturalness induces a less cautious and deliberate way of speaking, thereby reducing the likelihood of correct recognition.
-users can be free of the constraints of visual displays.
Guidelines:
-speech input should be used when input is required infrequently. One tactic is to confine speech input to the entry of commands, so that vocal fatigue may be reduced.
-Speech can be combined with other tasks, but only if they are non-verbal. Because: Less interference will be found if, when two tasks are undertaken simultaneously, one task is verbal and the other is spatial.
-A special command vocabulary should be designed for voice input. Otherwise commands may pose special difficulties because they are similar in acoustic terms and hence likely to be confused by the recognizer.

Martin (1989) investigates the following claims through a review of research and through an empirical evaluation of speech input: (1) speech input is faster than typed input; and (2) speech input increases user productivity by providing an additional response channel. He gives an overview of relevant research to conclude that it supports both claims. Further he describes an empirical study, in which the utility of speech input in the context of a VLSI chip design package is evaluated.

A comparison was made between speech-, typed-, full-word commands, mouse commands and to a lesser extent single key presses. Only a restricted set of the commands could be executed by mouse-button presses. Especially the comparison of speech- versus mouse commands is interesting in this context:
-He finds that mouse clicks, and speech input were equally efficient for a restricted set of commands. Note in this experiment there were only two commands that could be given. These mouse clicks should not be compared to mouse clicks used in a more complex menu selection (more menu items or a pop-up menu). In that case, the mouse click operation is likely to be more laborious, since it involves more fine-grained movement and selection from a larger number of alternatives.
-When comparing speech- to mouse- and keyboard input, he says: Traditional response modalities, such as keyboard and mouse input, can be characterized by an ease vs expressiveness trade-off. At one extreme, mouse clicks are extremely easy to execute, but can be used only with a small command vocabulary. At the other extreme, full-word

typed entries are difficult and time consuming to execute, but can be used for a very large command vocabulary. As speech input devices improve and become more widely available, they will offer a way of sidestepping this ease vs expressiveness trade-off. Spoken commands are easy to execute, and this ease only decreases slightly as command vocabulary increases.

De Vet en Beuk (1989) describe an experiment that compares mouse and speech input control of electronic diary keeping. Novice users had to perform tasks during three sessions; the first two sessions were used to learn to work with mouse and speech control. Speech recognition was simulated. In the third session subjects had free choice (mouse or speech or both input media). These are their main conclusions:
-the average number of commands used by the subjects to perform the task was roughly twice the minimum number of commands.
-subjects in the speech mode were faster than subjects in the mouse mode, although they used more speech than mouse commands during the first two sessions.
-most subjects tended to use fewer commands in the second and the third session as compared to the first, as they got acquainted with the command set.
-in the free choice mode ten subjects tried both types of control, of which three subjects used more mouse and seven subjects used more speech commands. Two subjects only used speech commands during this session.
-speech is subjectively faster and can be used with less effort
(no hand-eye coordination problems).
-ten out of twelve speech subjects preferred speech control over mouse control, they also thought it was more comfortable to work with. Nine out of twelve subjects thought that speech control was faster.
-one subject reported to have difficulties when starting to use speech control.
These conclusions have not been based on statistical analysis.

Sprenkels (1988) also describes an experiment that compares mouse to speech control. Subjects had to dictate a text to a 'Voice Actuated Typewriter' (VAT) by speech. Commands could be given either by speech or by mouse input. Speech recognition was simulated. These are his main conclusions:
-the speech control version was preferred by most of the subjects, because of the fact that it seemed to be somewhat faster to them, and that they did not have to search a command button on the screen and move the mouse to the position concerned. For text editing afterwards an additional cursor device such as a mouse, seems to be desirable.
-the average time needed per correct action is significantly shorter in the case of spoken commands.
-the speech version gave rise to more subject errors than the mouse version. The lower number of subject errors in the mouse version may have been caused by the visual support that the command buttons provided.

To compare speech to a manual input medium it is necessary to keep in mind that some aspects are a consequence of the medium itself, while others are consequences of the way the medium is implemented both technically and systematically.

For example: speech in itself may be very fast and easy to learn, because you already are able to talk, but if it is used in a system and has to be recognized by a voice recognizer it is more difficult to use because the user has to know how to make the chance of being improperly recognized as small as possible. Technically this will be influenced by the voice recognizer that is used and systematically by the choice of words in the menu.

Thus, some conclusions may not be generalized, since they are dependent on the design and setup of the experiment.

Some of such examples for both speech and mouse input have been collected from the literature mentioned before:

General aspects of speech as an input medium :

-speech may free the hands to undertake other tasks (Jones et al.,1989)
-speech adds another channel for information throughput (Jones et al.,1989, Martin ,1989)
-speech is very natural, no abbreviations or coded words are necessary (Jones et al., 1989)
-speech can be combined with other tasks, but only if they are non-verbal (Jones et al., 1989)

Implementation-specific aspects of speech as an input medium:

-tolerance to variability in speech is relatively poor for machine recognition (Jones et al., 1989)
-a special command vocabulary should be designed to enhance proper recognition (Jones et al., 1989)
-there are limitations to the size of command vocabulary (Jones et al.,1989)
-spoken commands are easy to execute, and this ease only decreases slightly as command vocabulary increases (Martin, 1989)

General aspects of the mouse as an input medium:

-you need your hands
-a menu has to be (at least partially) visible on the screen
-mouse clicks are extremely easy to execute (Martin,1989)
-hand-eye coordination is important (De Vet and Beuk, 1989)

Implementation-specific aspects of the mouse as an input medium:

-item order (alphabetical, frequency of use or logical) and target size of the command buttons (Foley et al.,1984)
-position (static or moving), visibility (always visible or sometimes visible) and organization (horizontally or vertically, single level or hierarchical) of the menu (Foley et al., 1984)
-the amount of space it takes on the screen depends on the number of commands and the size and sort of menu presentation (pop-up or menus that are always visible) (Martin,1989)
-mouse commands can only be used in small vocabularies (Martin, 1989)

4.2 Discussion of the experiment

The results of chapter 3 will now be discussed and combined. The following aspects that were used to evaluate and compare the use of speech versus the mouse as a control medium will now be discussed:

-speed (processing time);
-accuracy (number of correct commands / total number of commands);
-efficiency (minimum number of commands necessary to complete the task / total number of commands);

-working style (processing time / total number of commands).
The data of session two were not used to compare the two media, because the history from the first session may influence the results. For example: difference in learning time of subjects in group 1-6 and group 7-12 may depend on the order the two systems are offered to the subjects. The data of session 1 and session 2 are used to look at side effects as influence on learning and understanding of the system and to decide whether there was an influence of device and document order.

4.2.1 Sessions

4.2.1.1 Processing time

The processing time consists of the time needed to give the commands and the time needed to process the annotations. Theoretically the time needed for reading the annotations and making the corrections is the same in the mouse as in the speech version. Because in the speech version commands can be given when a subject is still making corrections, it is necessary to include the time needed for reading and correcting in the measure that is used to compare the speed of working with mouse and speech control.

There is no significant difference between the time that subjects use to complete the task with mouse or with speech control.

In session 1 speech subjects are (not significantly) faster than mouse subjects, however, in session 2 mouse subjects are (not significantly) faster than speech subjects.
The fact that mouse subjects are faster in session 2 and speech subjects are faster in session 1 may be caused by either one, or more than one, of the following aspects:
-the group of subjects that use speech in session 1 and mouse in session 2 are faster than the other group of subjects;
-the order of devices in which the subjects have to work with the system influences the time that subjects need to finish their task;
-learning to work with the system interacts with learning to work with the mouse, in which case subjects working with the mouse in session 1 may be slower than speech subjects in session 1 and mouse subjects in session 2 may be faster than speech subjects in session 2.
The measure is significantly influenced by device order and by the interaction of device and document order. This means that the history of already having worked with one input device and one document influences the outcome of this test. This means that the data of session 2 cannot be used to verify the conclusions based on the data of session 1.

There is no significant correlation between *processing time* and *total number of commands*. Neither is the *processing time* significantly correlated with *total number of errors* for both versions.

4.2.1.2 Efficiency (minimum number of commands necessary to complete the task / total number of commands)

Mouse subjects use an average of 1.5 times the minimum number of commands, while speech subjects use an average of roughly two times the minimum number of commands required to perform the task. This difference is significant. The measure is not significantly influenced by device and document order.

The minimum number to complete the task is exactly the minimum number of commands necessary to have a look at all annotations in a row and to quit the program. (12 annotations, page, two, 10 annotations, quit, yes). A study of the spy-files of the sessions leads to a better understanding of the measure found for efficiency. First of all efficiency is influenced by the number of errors the subject made. But also subjects sometimes have a look at an annotation for a second time to check whether they have made the correction properly. Sometimes subjects do not realize that they have finished all annotations on a page and ask for the next annotation, thus, displaying the first annotation once more.

A significant correlation was found between *efficiency* and *total number of errors* for mouse and speech subjects (R = .882 and R = .876, respectively).

The number of commands is naturally higher if a subject makes a lot of errors. This means that for both mouse and speech control efficiency rises if few mistakes are made. If each of these two devices would have lead to a different working style, subjects might have used a higher number of commands with one device than with the other, but not have made a higher number of errors. So it seems that this is not the case.

For example: if subjects using the mouse did not want to move the mouse a lot, they would only use the command "next", while with speech it does not take more effort to use different commands. Thus, with speech, subjects might be more willing to use many different commands. Of course, in this experiment it was not really necessary to use many different commands.

No significant correlation was found between *efficiency* and *time* for mouse subjects, whereas there was for speech subjects (R = .242 and R= .75 respectively). This means that speech subjects that score high on efficiency (use little commands) also use little time.

This means that if speech subjects make more errors it also takes more time for them to finish their task. This is not the case for mouse subjects. A lot of errors made with speech lead to a wrong interpretation by the speech recognizer. The result is that the subject has to make a repair action, to get to the annotation she first planned to go to. In the case of the mouse subjects errors occur if the cursor is not placed on the button. In this case nothing happens, so no repair action is necessary, and thus, it does not take as much time to correct an error.

4.2.1.3 Accuracy (number of correct commands / total number of commands)

Mouse subjects make significantly less errors per number of commands than speech subjects and thus, are more accurate. The measure is not significantly influenced by device and document order.

If a subject made errors with the **mouse** it may have the following reasons:

-the subject gave another command before the system was ready;

Sometimes when a subject thinks that the first command she gave was not executed, she gives many commands in a row. The spy-file will register these commands (apart from the first one) as errors, when the system changes from one state to the next after the first command and the cursor is not placed on a button any more, because there is no button on the same place as in the previous state.

-the cursor was not placed on a button.

The first is a shortcoming of the system, while the last is a mistake of the subject.

If a subject made an error with the **speech** control system, it can have the following reasons:

-the speech recognizer did not recognize the command at all, thus, no command was exe-

cuted. The system gave a tone as feedback that a noise had been registered, but no command had been executed;
-the speech recognizer misinterpreted the given command. Another command than the given command was executed.
-the subject made a noise that was not meant to be received by the speech recognizer;
-the subject gave a command that could not be used at that moment.
The first three are shortcomings of the system, while the last one is a mistake of the subject. Of course, the attribution of an error to the system or the user depends on what understanding and cooperation is expected of the user.

Most errors that were made in the speech version are because of shortcomings of the system. It is not always possible to see in the spy-files which of the two types of mouse errors occurred. So it is hard to tell whether the errors made in the mouse version were caused by shortcomings of the system or the subjects. For example: one subject gave 5 mouse commands in a row, while the cursor was not placed on a button. Maybe she did not look at the screen, and thus, did not notice the error, or she gave the commands before the system was ready.

There was no significant correlation present between *accuracy* and *time*. Subjects that work faster are not necessarily less accurate.

4.2.1.4 Working style (processing time / total number of commands)

Mouse subjects use significantly more time per command than speech subjects do in session 1. However, in session 2 there is hardly any difference. This is shown in figure 10. If the data about the processing time are studied more closely, it is possible to see that mouse subjects use (not significantly) more time than speech subjects in the first session, although they use (not significantly) less commands. In the second session mouse subjects use (not significantly) less time than speech subjects, and significantly less commands. So it seems that subjects that have to learn to work with the system, when using mouse control, need a lot of extra time to get used to this input device. The measure is not significantly influenced by device and document order.

The significant positive correlation (for speech subjects only) suggests that subjects who use more *time per command* will also make more *errors*. This is not the case with mouse subjects. If an error takes less time than a command, that is given to look at an annotation, subjects that make a lot of errors will use less time per command. For speech subjects a positive correlation is found between time per command and errors. So it seems that either making an error with speech control takes more time than correcting an average annotation, or that subjects that make a lot of errors use more time to correct an annotation than subjects that make few errors.

4.2.1.5 Operational preference (expressed in number of correct commands)

In session 3 no significant difference was found between the number of correct commands with the mouse and with speech. Thus, it seems there is no operational preference for either one of the control systems. The measure is not significantly influenced by device and document order.

4.2.1.6 Regression analysis on the data of session 3

A description is given of interesting conclusions based on the regression analysis of session 3 (appendix B.4). These conclusions can be used to get a better understanding of the way subjects worked with both systems.

The most interesting conclusion is that subjects that use *a lot of mouse commands* will also use *a lot of speech commands.*

Subjects who are *fast* use a *small amount of commands* and a *small amount of speech commands* compared to subjects that are slow. They do not necessarily use a relatively small amount of mouse commands. Note that number of commands and number of speech commands are also correlated, therefore the correlation between processing time and number of speech commands may be a consequence of the fact that processing time and total number of commands are correlated.

Furthermore subjects that are *slow* will use a *large amount of commands* compared to subjects that are fast, but subjects that are slow will not necessarily make a large amount of errors. But subjects that use a *large amount of commands* will have made a *large amount of errors.*

Subjects that use a *large amount of (total) commands* compared to subjects, that use a small amount of commands, make a *large amount of mouse errors (incorrect mouse commands)*, but not necessarily use a large amount of mouse commands.

And subjects that make a *large amount of errors* will also make a *large amount of speech errors*, but not necessarily a large amount of mouse errors.

There is an *efficiency / time trade off*, but no accuracy / time trade off. Since the time spent using a control system is influenced by the length of the annotation and thus, by the time needed to make a correction, it is not possible to compare the time used with speech and the time used with mouse commands. Therefore, these measures are made over the complete session and are not separated into mouse and speech effects.

4.2.2 Questionnaire

4.2.2.1 Preference, ease of use and speed

Figure 16 shows that nine subjects have a strong preference for the mouse. Five subjects have a moderate preference to use both systems. Many of these subjects choose the mouse or speech for ease of use and speed. This can easily be seen in figures 17 and 18. The bar for both has been reduced and especially the mouse has gained a lot of support. This can be explained if the positive and negative aspects given by the subjects are compared with their choices for preference, ease of use and speed (4.2.2.2). The number of subjects that choose speech is rather constant for all three questions.

These are the numbers of subjects (percentages of subjects) that chose a specific medium:

Preference : 10 mouse, 5 speech, 9 both
(42%) (21%) (37%)

Ease of use : 17 mouse, 6 speech, 1 both
(71%) (25%) (4%)

Speed : 15 mouse, 5 speech, 4 both
(62%) (21%) (17%)

Conclusion:
Depending on how subjects values the positive and negative aspects they will decide which device or combination she will prefer.
In this experiment the following choices were made:
-15 subjects were very consistent in their appreciation of a device:
they chose the same device for all three aspects we questioned them about
(10 mouse, 4 speech,1 both)

-the other 9 subjects chose different devices for these three aspects
(8 preferred using both and 1 preferred using speech).
Depending on how they value the most important characteristics they make their decision.

4.2.2.2 Positive- and negative aspects of mouse- and speech control

If the positive and negative aspects are categorized according to the choice subjects made for preference it is possible to get a better understanding why people prefer working with both control systems (Appendix B.3.3). It shows that some subjects have different priorities than others. Subjects that choose speech sometimes give the same positive aspects for the mouse as subjects that choose mouse control. Obviously they think that the combination of positive and negative aspects they give for speech outweigh the positive and negative aspects they give for mouse control. Subjects that choose mouse control give a higher priority to other aspects and come to a different conclusion.

Generally speaking:
-People that prefer **speech** usually think that speech is fast and being able to give a command while processing an annotation is important.
-And people that prefer **mouse** control think it is important to work fast without having to bother about being recognized properly.
-People that choose for **both** think that both systems have their advantages and can be combined for greater benefit. Roughly speaking three groups can be distinguished:
1) Some mostly use the mouse for its speed, ease of use and reliability. They think that speech is faster only in some cases, for instance when two commands have to be given in a row. Or to be able to give a command when they are still making corrections.
2) Some mostly use speech to be able to give a command while still making corrections and not having to change from one device to another. They want to use the mouse in the special occasion that recognition does not function properly or to be able to do a repair action (going directly to the right annotation after several wrong interpretations).
3) Some use one system as much as the other.

4.2.3 Comparison of session 3 and questionnaire

Some subjects use the third session to decide how they would prefer to control the system. A difference is found between the percentage of commands they use of each device and the answers they give to the questionnaire.
4 know already what they will choose before the third session (3 mouse, 1 speech)
9 use both and choose for both of the devices
(tendency in session 3 : 3 mouse, 3 speech, 3 both)
11 use both and choose for one of the two devices (7 mouse, 4 speech)
(tendency in session 3 : 3 mouse, 2 speech, 5 both, 1 both/mouse)
Roughly half of the subjects use the last session to decide which way of controlling the system they prefer; using the mouse, speech or both control systems. So it seems that the number of correct commands given in session 3 with a particular control system is not a proper measure for preference. Furthermore, this measure does not include preference to work with both control systems alternately.

subjects	mouse commands (%)	speech commands (%)
1	97	3
2	6	94
3	41	59
4	23	77
5	12.5	87.5
6	43	57
7	28	72
8	54	46
9	91	9

Table 4.2 : Percentage of mouse- and speech commands used by subjects that according to the questionnaire prefer to work with both control systems.

In the present experiment the percentage of mouse commands of the nine subjects that preferred to work with both control systems varied from 6 to 97 % (table 4.2). It is obvious that some of the subjects that choose both mostly use speech while others mostly use the mouse.

A comparison can be made between the **speech errors subjects made in sessions 1 to 3** and the **choice of input device they made in the questionnaire.**

In session one 4 subjects made more than the average of 9 speech errors:
-2 of these subjects chose mouse control in the questionnaire
-1 subject chose to control the system with both control systems
-1 subject chose working with speech for preference, working with both for ease of use and for speed.
This means that two of the four subjects that made a lot of errors still want to have speech at their disposal. The other two subjects used some speech commands in the third session, but chose mouse control in the questionnaire.

In session two 4 subjects made more than the average of 8 speech errors:
-3 subjects chose mouse control in the questionnaire
-1 subject chose speech control
This means that one of the four subjects that made a lot of errors wants to have speech at her disposal. The other three subjects used some speech commands in the third session, but chose mouse control in the questionnaire.

In session three 5 subjects made more than 10 speech errors:
-1 chose mouse control in the questionnaire
-1 chose speech control
-1 chose speech control for preference, mouse control for ease of use and control with both systems for speed
-1 chose for control of the system with both input devices
-1 chose control with both input devices for preference, mouse control for ease of use

and for speed.
This means that even though some subjects have problems with being recognized properly they still want to be able to use speech.

A longer period of time to work with both control systems might influence the working method and preference of the subjects. Because of a better understanding of the working of the speech and the mouse system they might give a different priority to the positive and negative aspects of each of these systems. Thus, the number of correct commands after such a short time is not a good measure to decide which of the two devices subjects prefer.

4.3 Comparison of the results of the experiment and the literature
In this paragraph a comparison is made between the result of the described experiment and the literature discussed in paragraph 4.1. Some of the compared aspects are implementation dependent, while others are general aspects. This comparison leads to a general discussion at the end of the paragraph.

According to **Foley, Wallace and Chan (1984)**:
-Ease of learning is high for speech and medium for mouse control.
In the present experiment it seems that mouse subjects have more trouble getting used to the system than speech subjects do. In the long run mouse subjects use less time but during the first session they used more time (although not significantly more) than speech subjects did.
-Error proneness is high for speech and medium for mouse control.
Speech subjects working with the annotation system made significantly more errors than mouse subjects did.
The other factors in table 4.1 can not be directly compared to the results of the present experiment.

According to **Jones, Hapeshi and Frankish (1989)** less interference will occur, if when two tasks are undertaken simultaneously, one task is verbal and the other is spatial. This would mean that in such a case subjects would work faster and make less errors than if these two tasks would interfere. Typing, writing and giving speech commands are verbal tasks, while giving mouse commands is a spatial task. This means that combining mouse commands with writing (or typing) would lead to less interference than combining speech control with writing (or typing).
In the present experiment no significant difference was found between processing time of mouse and speech subjects. Speech subjects make significantly more errors than mouse subjects. There are too many other factors that influence the number of errors that are made (for example: recognition rate of the speech recognizer) to attribute this difference to the interference of these two tasks.

Martin (1989) found that mouse clicks and speech input were equally efficient. He defines efficiency as the time to complete a command sequence. He emphasizes the fact that the comparison was made for a restricted set of commands where no fine-grained movements were necessary. In the present experiment no significant difference was found between processing time of mouse and speech subjects either. In the experiment by **De Vet and Beuk (1988)** speech subjects turned out to be faster than mouse subjects. However, in this experiment the menu was more complex, the total of commands consisted of roughly 50 commands. Thus, more movement was necessary for giving a mouse command

than with as small a menu as in the annotation experiment.
Martin (1989) also points out that mouse control is most profitable with relatively small vocabularies and that speech control hardly becomes more difficult to execute when using a large vocabulary.

In the experiment by **de Vet and Beuk (1988)** subjects used roughly twice the amount of commands necessary to complete the task. In the annotation experiment mouse subjects used 1.5 and speech subjects used 2.1 times the minimum number of commands necessary to complete the task.

In the diary experiment subjects in the speech mode were faster than in the mouse mode, although they used more commands than subjects in the mouse mode. These conclusions were not based on a statistical analysis. In the annotation experiment there is no significant difference between the speed and the number of commands that mouse and speech subjects use.

Subjects working with the diary system used fewer commands in the second and the third session as compared to the first session. In the present experiment 63 % of the subjects used less commands in the second session as compared with the first session and 46 % of the subjects used less commands in the third session as compared to the second session. The percentage of subjects that use less commands in the third session as compared to the first session is 54 %.

While in the diary experiment subjects had a tendency to use more speech commands, in the annotation experiment the difference between both input media is not as obvious (table 4.3).

use of input medium	diary system (%)	annotation system (%)
mouse	0	13
speech	17	4
both : mostly mouse	25	29
mostly speech	42	33
equally both	16	21

Table 4.3 : Percentage of subjects that use a certain control system in the free choice session

Table 4.4 to 4.6 show that in the diary experiment subjects tend to prefer speech. Table 4.4 shows that in the annotation experiment speech has the lowest percentage of subjects for preference. To control the system with the mouse (42 %) or with both types of input media alternately (37%) is preferred by more subjects than to control the system with speech input (21 %). Some subjects that preferred to work with both control systems at their disposal, think that, although it is faster and easier to work with just mouse control, speech has other advantages that are worth having it at their disposal (e.g. hands free).

choice of input medium for preference	diary system	annotation system
mouse	17 %	42 %
speech	83 %	21 %
both	--	37 %

Table 4.4 : Percentage of subjects that chose the various input control media for **preference.**

choice of input medium for speed	diary system	annotation system
mouse	25 %	62 %
speech	75 %	21 %
both	--	17 %

Table 4.5 : The percentage of subjects that chose the various input media for **speed**

choice of input medium for ease of use	diary system	annotation system
mouse	17 %	71 %
speech	83 %	25 %
both	--	4 %

Table 4.6 : The percentage of subjects that chose the various input media for **ease of use**.

In the experiment by **Sprenkels (1988)** speech was preferred by most subjects. In his experiment subjects had to read a text out loud and give commands either by mouse or by speech; voice recognition was simulated.

In the VAT experiment by Sprenkels more errors were made when using speech than when using the mouse. This may have been due to the visual support the command buttons provided in the mouse version. In the present experiment speech subjects also made more errors, however the visual support was the same for the mouse and the speech version. Most of the errors occurring in the speech version were caused by wrong interpretations by the speech recognizer.

In the VAT experiment the percentage of errors was 6 and 10 % for the low and the high error frequency, respectively (the speech recognition errors were 'programmed' into the simulated speech recognition). In the annotation experiment the mean percentage of errors in the speech version was 17 % and 15 % (in session 1 and 2, respectively). Because the percentage of errors in the annotation experiment was higher than in the VAT experiment, the strain caused by the possibility that a command would not be recognized properly was probably bigger in the case of the annotation experiment than in the VAT experiment. This may be a reason why less subjects preferred working with speech in the annotation- than in the VAT experiment.

General discussion

A general conclusion of the comparison of the results of the present experiment with the literature is that the results can be highly influenced by the setting (the implementation-specific aspects of the experiment) that is chosen for a certain experiment. The structure that Foley, Wallace and Chan (1984) gives for a selection of input devices and -techniques is an attempt to define at least a part of the setting that may influence an evaluation. They distinguishes six tasks and nine measures of ergonomic quality. There are, however, more aspects to be considered that are important. Some were already evident in the comparison between the results of the annotation experiment and the literature that has been discussed in 4.1 :

-task setting; the total set of tasks that have to be performed.
Interference may occur on cognitive, perceptual and motor levels, thus, influencing measures as task completion time, number of errors and number of commands.

-complexity of the system (the number of states or modes and the number of possible commands that can be given in a certain state). The more complex a system, the more effort needed to work with it. The amount of effort needed, may have different effects on the performance for different media. For example, if the number of commands is high, mouse input will become more difficult because more movements may be necessary to find the right command, while speech input may become more difficult because more commands have to be remembered and the chances of bad recognition by the speech recognizer are bigger, because it is harder to design a command vocabulary that is easy to distinguish.

-implementation of the system (for example: level of visual support, type of speech recognition board, possibilities for repair actions)

To be able to minimize the research needed to give a thorough evaluation it is important to understand in what way the results are influenced by changes in the setting.

5. Conclusions

The comparison of the data of the subjects that worked with the mouse and speech controlled annotation systems leads to the following conclusions.
No significant difference was found between mouse and speech subjects in processing time in session 1. Speech subjects, however, make significantly more errors. Mouse subjects are more efficient and more accurate than speech subjects. In session 1 mouse subjects used more time per command than speech subjects. In session 2, however, this tendency is changing and mouse subjects used almost the same amount of time per command as speech subjects used.
The difference between session one and two is probably caused by the following factors:
-the difference in history; speech subjects have already worked with the mouse version, while mouse subjects have already worked with the speech version. In the case of processing time, the influence of device order is significant. The other measures were not significantly influenced by device order.
-subjects are still getting more experienced when working with the system; if the time needed to complete a task would be drawn in a graph, the learning curve would not have reached the flat part yet.
Mouse subjects make less errors than speech subjects. The errors in the speech version are mostly caused by shortcomings of the system (wrong recognition), while in the mouse version errors are mostly caused by mistakes of the subjects themselves (not placing the cursor on the button).
If it would have been more natural to give a command with speech (provided it is recognized properly) than with the mouse, subjects might have had a different working strategy with speech than with mouse control. They might have used the 'number' command more often, because it might have been more natural than using the 'next' command all the time (e.g. like sellers in a bakery that call out numbers to find out whose turn it is, they do not say "next please" all the time). However, this does not seem to be the case; the different types of control do not seem to lead to a different working method.

The data of the free choice session (number of correct commands with a control medium) seem to show that there is no preference for one of the control systems over the other. When a comparison is made between the data of the third session and the data of the questionnaire a difference is found. *A majority of subjects prefer to control the system with the mouse for speed and ease of use. For general preference, however, a number of subjects switch from the mouse to both control systems.*
There are at least two reasons for this difference:
-subjects decide which control medium they prefer to work with after they have finished the third session.
-the third option in the questionnaire, to prefer to control the system with both input media, is not included in the statistical analysis. A large percentage of subjects (37 %), however, prefer to control the system with both control systems at their disposal.

It is interesting to see that even though the number of errors with the speech-system is higher than that with the mouse-system, subjects think that speech is a valuable input medium. *A third of the subjects prefer to control the system with both input media*, to be able to combine the advantages of both input media.
The priority that is given to the positive- and negative aspects of both control systems

differs over the subjects:
-People that prefer **speech** usually think that speech is fast and being able to give a command while making a correction is important.
-And people that prefer the **mouse** think it is important to work fast without having to bother about being recognized properly.
-People that choose for **both** think that both systems have their advantages and can be combined for greater benefit. Roughly speaking, three groups can be distinguished:
1) Some mostly use the mouse for its speed, ease of use and reliability. They think that speech is faster only in some cases, for instance when two commands have to be given in a row. Or to be able to give a command when they are still making corrections.
2) Some mostly use speech to be able to give a command while still making corrections and not having to change from one device to another. They want to use the mouse in the special occasion that recognition does not function properly (give the same command with the mouse after some recognition errors without consequences) and to be able to do a repair action (going directly to the right annotation after several recognition errors that resulted in executions of a the wrong commands).
3) Some use one system as much as the other.

It seems that these three sessions were not sufficient to ascertain that the subjects had already reached their final decision on preference and the flat part of their learning curve.
There are several factors that might influence the results of the experiment if subjects would be able to work with the system for a greater period of time:
-better understanding of the speech recognition board. For example, in some cases subjects tended to speak louder if a command was not recognized properly, however the error had occurred because the command was not pronounced in the same way as in the training session. Working for a greater period of time with the system enhances the understanding of the errors that may occur. This may also lead to more acceptance of an error that has occurred. For example, if the cursor of the mouse was not placed on the command button, it is easier to accept that an error has occurred than if you give a spoken command and the computer 'all of a sudden' executes another command.
-getting more faith in the speech recognizer and the way to handle errors that may occur and thus needing less concentration for giving spoken commands.

As shown in the discussion, the setting of an experiment can highly influence its outcome. I want to stress the importance of the fact that the conclusions of this experiment may not be generalized to the use of mouse and speech control in other situations. If an overall evaluation of speech control has to be given, further research has to be done on the influences of the setting on the results of the experiment. That way all aspects that may influence the evaluation can be looked at, before a general conclusion is given.

5.1 Recommendations

- Have subjects work with the system for a longer period of time.

If they have worked with the system long enough, they will not change in the way they work anymore. In that case the last session can be a valid measure for preference, if a way is found to include a measure for preference to work with both systems.

- Determine which aspects of a setting influence the results of an evaluation of in- and output media.
- Evaluate the use of speech as a control medium in a different setting (another choice for task, complexity of the system or experience of the user).

Acknowledgement

I would like to thank Floris van Nes for his extensive comments on earlier versions of this report and all other members of the 'Information Ergonomics' and 'Cognition and Communication' groups for their support and suggestions.

References

Jones, D. , Hapeshi, K. and Frankish, C. Design guidelines for speech recognition interfaces, Applied Ergonomics, 1989, 20.1, 47- 52.

Foley, J. D. , Wallace, V.L. and Chan P. The human factors of computer graphics interaction techniques, IEEE Computer Graphics and Applications, November, 13-48, 1984.

Kraak, L.J. , Speech is more effectively processed than text. Report no. 678, Institute for Perception Research, Eindhoven, 1988.

Kraaij, W. Reception of voice- and script annotations for correcting documents. Report no. 599, Institute for Perception Research, Eindhoven, 1987.

Kraaij, W. A voice controlled annotation system for correcting documents, Report no. 638, Institute for Perception Research, Eindhoven, 1988.

Martin, G.L., The utility of speech input in user-computer interfaces, Int. J. Man-Machine Studies, 1989, 30, 355-375.

Nes, F.L. van and Beuk L.G.M. Rapidly developing user interfaces with dialogue control modules, Report no. 584, Institute for Perception Research, Eindhoven, 1987.

Nes, F.L. van and Heijden, J. van der , The use of computers by ordinary people, IPO Annual Progress Report, 13, 102-107, 1978.

Sprenkels, H.A.A. Some ergonomic aspects of speech recognition systems with large vocabularies, Report no.640, Institute for Perception Research, Eindhoven, 1988.

Vet, J.H.M. de , Beuk, L.G.M. Comparing mouse and speech input control for an office application, Report no.696, Institute for Perception Research, Eindhoven.

Appendix A.1 Documents

-1-

SCORING AND STATISTICAL PROCEDURES

For all scoring methods misspellings were not counted wrong provided the criteria of the particular scoring method were met. However, disagreement in tense or number, except for GC scores, was considered an error. FORTRAN programs were written for the IBM 1620 to calculate Verbatim, Synonym, and Redundancy scores. Subjects' responses to items were pinched on IBM cards according to specific directions (Appendix B). Phonetic spellings and other spellings that made it clear what response was intended were corrected before being punched. Otherwise, responses were punched as they appeared. Where a response was illegghible, the letter B was punched. A limit of eight letters in length was imposed on responses. In the few cases where responses exceeded this length, only the first eight letters of the word were punched but in every case eight letters proved sufficient to identify a correct match with the originally deleted word. Where a subject responded with more than one word, the first word was taken as his answer.

To test wheter cloze scores ranked passages in terms of reading difficulty, an external criterion of passage reading difficulty was required. Such a criterion was provided by having four experienced reading teachers rank the three passages in order of difficulty. Cloze procedure as a measure of reading difficulty was then tested by correlating cloze rankings with judges' rankings. Judges' rankings as a criterion may be criticized on the grounds of subjectivity.[1] However, the appropriateness of readability formulae to assess the difficulty level of reading materialks in Papua New Guinea is probably more open to question and besides, in the present case, the four judges showed perfect agreement

1) Bermuth,K,Cloze tests as Measures of Readability and Comprehension Ability, Proceedings of the International Reading Association,228-232,4,nr.2,1966.

Document 1, page 1

by ranking the three passages from easiest to most difficult in the order A, B, C.

RESULTS

A chi- square value of 130.453 with 57 degrees of freedom obtained by a Friedman's two-way analysis of variance interpreted as a t ratio after Guildford has a probabilty of less than 0.01. The conclusion was that cloze scores discriminated between pupils in the experiment.

Average scores on each of the three passages expressed as percentages for the Verbatim, Synonym, Alternative Response, Grammatical Class, and Redundancy scoring methods are shown in table 4.1. It is clear that all five scoring methods caused a similar passage ranking. The Synonym score for each passage differed little from the Verbatim score whereas the Alternative response and Grammatical Class scores were higher than both but, it is to be noted, by an almost constant quantity.

Table 4.1. Relative Difficulties of the Three Passages.

	Passages		
Scoring Procedure	A %	B %	C %
Verbatim	58.63	40.95	17.64
Synonim	59.64	40.95	17.64
Alternative Resp.	64.03	45.91	21.02
Grammatical Class	55.49	57.28	26.19
Redundancy	62.33	42.39	19.20

Document 1, page 2

Discussion

Of the six tested hypotheses using three-dimensional analysis of variance, only the hypothesis concerning the mean difficulty levels of Passages A, B, and C was not true. For standard- and exact-length blanks the passages were ramked in the order of A, B, C, thus confirming the main finding of Experiment 1. The maintenance of cloze rankings of difficulty over Experiments 1 and 2 using different samples of subjects indicated in some measure the readability of cloze procedure as a measure of readability.

The principal finding concerned differences between the mean score for standard-length blanks and the mean score for exact- length blanks. Although the latter yielded slightly higher scores than the former for passages at three levels of difficulty, this difference was not significant.

The proportion of the total difficulty associated with each passage showed quite close agreement in passage dicrimination under two kinds of blank length. Further, the proportions of total difficulty obtained in Experiment 2 showed close agreement with the proportions of total difficulty obtained in Experiment 1 under five different scoring procedures.

There was evidence thet the use of standard- and exact-length blanks ranked subjects similarly. This finding together with the non-significant difference between mean scores for length of blank suggested that it for subjects of similar ability and for passages dimilar in difficulty to those in the present experiment, little matters whether blanks in cloze test be of a standard length or the same length as the deleted word.[1] Some evidence of the

1) A.J. Kingston and W.W.Weaver,Recent developments in Readability Appraisal, Journal of Reading, 11, (1967) : 44-47.

Document 2, page 1

concurrent validity of cloze test as measures of general reading comprehension was yielded by the correlations (averaged over three passages) between standard- and exact-length versions of the cloze test with the Watts Test. The obtained correlstaion coefficients of 0.61 and 0.75 respectively were good considering the shortness of the cloze test. If the Watts Test has some validity as a measure of reading comprehension with the present subjects, then so have cloze tests too.

EXPERIMENT 3: Only responses corresponding to the deletion of the thirty-second and sixty_fourth word were scored since there items were common to the five deletion rates. Each subject's scores were summed over the three passages he completed, giving an individual maximim score of six. The scores of the two missing cases were estimated by calculating the mean of the remaining observations in the same cell and rounding this to the nearest integer. The assumption of random sampling from a homogeneous population was satisfied.

Table 5.6. Three-dimensional Analysis of Variance

Source of Variation	df	SS	MS	F	p
Deletions (A)	4	16.03	29.01		
Difficulties	1	88.50	88.50		
Levels	2	92.60	46.30		
Interaction(AxB)	4	27.33	43.34	45.77	<0.01
Interaction(AxL)	8	16.48	20.61	2.160	<0.05
Interaction(BxL)	2	1.405	0.703	0.737	NS

Document 2, page 2

-1-

Discussion

The results presented here clearly indicate that the scale or size of visually presented displays affects the estimation of duration. Specifically, verbal estimatres of the duration of a fixed interval of time tend to increase as the size of a visual given display decreases.

The effect of stimulus scale demonstrated here is in general accord with that reported by DeLong (1981). However, the present results do not suggest a linear and proportinal relationship between stimulus scale and duration. This may be due to the fact that, in the present study, the visual stimuli filling the interval were two dimensional (as opposed to three-dimensional), or that the subjects were asked to make their estimates as acurately as possible (rather than to make their estimates in terms of how they "felt, not thought", as DeLong's subjects were instructed), or that the stimulus interval was relatively brief (i.e. 55 as opposed to 30 minutes). Nevertheless, it is important to note that the general effects of stimulus scale apparently transcend these methodological differences.

Theoretical Interpretations.

Storage-size theory.[1)] The present results do not lend themselves to be interpreted straightforward by the storage-size theory. In the present study, this demonstration cannot be based upon information differences in the distal stimuli, since the visuel displays defining the interval were equivalent (except for their size).

Psychobiological chronometer theory. To account for the overestimation of duration reported here,a psychobiological chronometer theory interpretation requires that decreases in

Document 3, page 1

-2-

visual display size results in increases in the activity of the chronometer. The present experiment did not include direct measures of psychological processes, and it is not obvious whether, or how, variations in visual display size relate to change in heart rate, brain chemistry, body temperature, etc. However it doesn't seem very likely that variations in the present display sizes differentially affected the attentional resources devoted to the psychobiological chronometer. As noted earlier, there was no difference in task difficulty for the various experimental conditions. Also, more recent evidence tends to discount the notion that display size diferences, per se, affect duration estimates via workload differences. Bobko, Bobko and Davies (1985) compared duration estimates from a small and a large display (0.13- and 0.58-m diagonal) placed at looking distances of 0.91 and 2.73 m, respectively.

Table 1. Analysis of Variance for the Effects of Screen Size and Sex on Duration Estimates

Source	Sum of Squares	d.f.	Mean Square	F ratio
Screen size	12.40	2	6.36	5.81 *
Sex	4.08	1	4.08	3.73 **
Interaction	0.29	2	0.15	0.13
Residal	72.26	66	1.09	
Total	89.03	70		

* $p < 0.005$
** $p < 0.06$

Document 3, page 2

Appendix A.2 Annotations

During the experiment three different documents were used. In this appendix the annotations are given. Each document has the same number of annotations and also the same number of annotation in each category of annotations. The annotations are divided into the following categories:
- lay-out
- letter level
- word level
- sentence level
- reference
- numbers

Document 1, page 1.

Lay-out

1 Onderstreep dit kopje

Letter level

4 punched
5 illegible
9 whether
11 materials

Word level

3 GC volledig uitschrijven: Grammatical Class
6 Voeg 'or omitted' toe.
7 Maak hiervan: 'the word originally deleted.'
10 Verander 'to assess' in 'for assessing'.

Sentence level

8 Voeg de volgende zin in:
Marker or analyst reliability was perfect in the case of the four scores obtained by computer.

Reference

2 Voeg na 'method' een verwijzing naar een artikel in, in de vorm van een voetnoot. Het artikel is van J.Anderson, de titel is "A Scale to Measure the Reading difficulty of Children's Books". Het artikel komt uit het University of Queensland Faculty of Education Paper, Volume 1, nummer 6 uit 1967. Paginanummers 326 t/m 342. Wat betreft de notatie van deze verwijzing gelden dezelfde opmerkingen die ik gemaakt heb over de verwijzing die er al staat onder aan deze bladzijde, in annotatie 12.

12 De notatie en volgorde voor deze literatuurverwijzing moeten veranderd worden: {Voorletter(s)},{Achternaam},"{titel artikel}",{Naam

tijdschrift},vol.{nr},nr.{nr},{jaar}, p.{..-..}.

Document 1, page 2.

Letter level
3 probability
7 Response ook met een hoofdletter
9 Synonym

Word level
Niet Guildford maar Gray
4 Mean i.p.v. average
5 apparent in plaats van clear.
6 Vervang 'caused a similar passage ranking' door: 'ranked the passages in the same way'.

Sentence level
1 De eerste zin van deze paragraaf moet herschreven worden in:
Friedman's two-way analysis of variance resulted in a chi-square value of 130.453 with 57 degrees of freedom which interpreted as a t ratio ....

Numbers
8 Dit getal moet vervangen worden door 58.92
10 De eerste twee getallen achter Grammatical Class moeten vervangen worden door:
62.51 46.45

Document 2, page 1.

Lay-out
1 Dit kopje moet in hoofdletters.

Letter level
4 ranked
7 Dit moet zijn: reliability
9 that
11 similar

Word level
2 six hypotheses tested
3 vervang 'not true' door 'rejected'.
6 Vervang different door independent.
10 Het woordje 'it' moet verplaatst worden tot na de eerstvolgende komma:
...experiment, it matters little whether......

Sentence level

8 Voeg de volgende zin toe:
The cloze procedure gives a worse measure than the method developed by Watts but we preferred it for its simplicity.

Reference

5 Hier moet een verwijzing komen naar het artikel van Chris Rankin jr., het heet "The cloze procedure, a survey of research". Het artikel staat in het " Fourteenth Yearbook of the National Reading Conference". Volume 14, de pagina's 133 t/m 150 , uitgegeven in 1965. In annotatie 12 heb ik al aangegeven hoe een verwijzing er uit moet komen te zien.

12 In dit rapport moet een literatuurverwijzing als volgt genoteerd worden:
{achternaam},{voorletters},({jaartal}),
"{titel van het artikel}",{naam v.h. tijdschrift}, vol.{nr},p.{..-..}.
Dus de naam en het jaartal op een regel en op de volgende regel eerst inspringen en dan de rest.

Document 2, page 2.

Letter level

1 correlation
5 Niet 'there' maar 'their'
6 maximum

Word level

2 vervang good door satisfactory
3 Maak hiervan: "then so too have cloze tests."
4 Voeg hier tussen: 'For each passage'
9 Achter 'Difficulties' en 'Levels' missen respectievelijk nog de index (B), en de index (L).

Sentence level

7 Vervang de zin: The assumption .... satisfied. door:
This resulted in a loss of two degrees of freedom for total mean square and a similar loss for error mean square.

Numbers

8 Dit getal moet vervangen worden door 14.74
10 De drie getallen achter 'Interaction(AxB)', namelijk de waarden voor de condities SS, MS en F moeten

veranderd worden in 14.55, 52.89 en 36.31 .

Document 3, page 1.

Lay-out
1 zet dit kopje in het midden boven de tekst

Letter level
2 estimates i.p.v. estimatres
4 proportional in plaats van proportinal
5 niet acurately maar accurately
11 visual i.p.v. visuel

Word level
3 vervang 'a visual given' door: 'a given visual'
6 vervang very door relatively
7 voeg hier tussen: and experimental
9 to straightforward interpretation i.p.v. to be interpreted straightforward

Sentence level
10 voeg de volgende zin toe:
Such an interpretation would require a demonstration that subjects who view a small screen retrieve more information than do subjects who view large screen.

Reference
8 Voeg achter "Storage-size theory" de afkorting tussen haakjes toe. In annotatie 12 heb ik al aangegeven hoe dat moet en welke andere gevolgen het heeft.
12 Voeg de afkorting toe. Afkortingen worden in hoofdletters gegeven, zonder achter elke letter een punt te zetten. En vervang telkens verderop in de tekst het volledige begrip door de afkorting. In dit geval dus (PCT)

Document 3, page 2.

Letter level
2 niet psychological maar physiological
6 niet diferences maar differences
9 residual

Word level
1 niet 'results in' maar 'produce'
3 vervang obvious door clear
4 verander 'doesn't seem very likely' in 'seems unlikely'

7 vervang looking door viewing

Sentence level

5 vervang de zin: there ...... condtions. door:
measures of task difficulty employed in this study
did not vary across experimental conditions.

Numbers

8 dit getal moet worden vervangen door 12.73

10 De twee getallen achter Total moeten vervangen
worden door: 89.36 71

Appendix A.3 Instructions

This appendix contains the instructions (in Dutch) that were given to subjects that participated in the experiment. Subjects that were going to work with speech control were given the instructions (1), (2A), (3), (4A), (5) and (6). Subjects that were going to work with speech control were given the instructions (1), (2B), (3), (4B), (5) and (6). Before starting with session 3 they received the shorter version of the instructions (7).

Experiment besturing van het tekstannotatiesysteem

(1) Inleiding

U gaat werken met een tekstannotatiesysteem. Met dit systeem kan commentaar dat op artikelen of andere documenten is gegeven in een computer opgeslagen worden. Deze annotaties kunnen dan weer opgeroepen worden en de verbeteringen of veranderingen kunnen worden aangebracht.. De lengte van het experiment bedraagt ongeveer anderhalf uur.

(2A) Programma

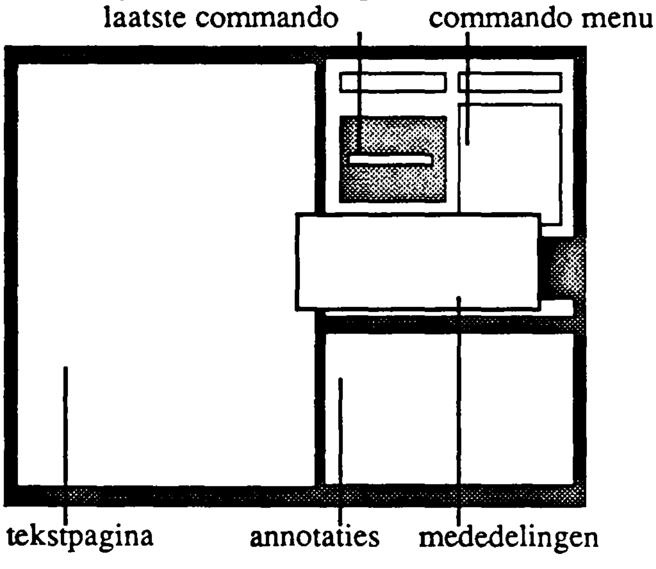
- Experiment .
- Vragenlijst.

(2B) Programma

- Trainen van de spraakherkenner.
- Experiment .
- Vragenlijst.

(3) Handleiding voor het tekstannotatiesyteem.

De verschillende pagina's tekst en annotaties kunnen zichtbaar worden gemaakt door het geven van commando's.Rechtsboven op het scherm is steeds het menu met de op dit moment herkenbare muiscommando's te zien.
Links van dit menu kunt u zien wat het laatste gegeven commando is.
Eventuele mededelingen van het syteem verschijnen onder het commandomenu.
De tekstannotaties verschijnen rechtsonder op het scherm wanneer ze geselecteerd zijn.



(4A) Uitleg over het gebruik van de muisbesturing :

- U kunt de cursor over het scherm bewegen door de muis over de metalen plaat te schuiven.
Wanneer U de muis optilt, blijft de cursor op het scherm op zijn oude plaats staan.
U kunt de cursor gemakkelijker over een grote afstand bewegen, als U de muis optilt en vanaf een rand over de metalen plaat schuift.

- Een button is een rechthoek waarin een bepaald woord staat.
U kunt een button uitkiezen door de cursor in deze rechthoek te zetten. Wanneer u de muisknop heeft ingedrukt wordt het commando uitgevoerd.

(4B) Uitleg over het gebruik van de spraakbesturing.

- Houd er rekening mee, dat het systeem wat traag reageert op de commando's.
- Probeer het commando uit te spreken zoals U het uitsprak bij het trainen van de spraakherkenner.
- Wees met name bij korte commando's als: "ja", "een", "twee", "drie", "tien", enz. consequent in de uitspraak.
- Wanneer een commando niet herkend wordt, kunt U het beste proberen om opnieuw het commando in te spreken, zonodig een aantal malen.
- Maak zo weinig mogelijk bijgeluiden.
- Wanneer U toch nog problemen ondervindt, kunt U de proefleider raadplegen door middel van de intercom.

(5) Commando's

- door commando's te geven kunt U de verschillende pagina's en tekstannotaties zichtbaar maken.

volgende:
-U kunt de annotaties kiezen in de volgorde waarin deze, zichtbaar als tekentjes, op het scherm staan. Bij het begin van een nieuwe pagina en bij het begin van een sessie verschijnt de eerste annotatie.
- Na het bekijken van de laatste annotatie van een pagina, zal met dit commando de eerste annotatie van dezelfde pagina op het scherm verschijnen.

vorige:
- U kunt met dit commando de annotaties kiezen in de volgorde als waarin deze, zichtbaar als tekentjes, op het scherm staan.

nummer:
- Wanneer U dit commando geeft, verschijnen de nummers van de annotaties op het scherm.
- Daarna kunt U een annotatie kiezen door het nummer te geven.
- Wanneer de annotatie met dit nummer niet bestaat, komt U terug bij het begin van de pagina waarmee U bezig was.

extra:
- U kunt een extra annotatie zichtbaar maken, terwijl de vorige op het scherm blijft staan.
- Voor iedere extra annotatie die U wilt kiezen is het nodig het commando "extra"te geven,

voorbeeld:
- U wilt annotaties 1, 2 en 4 tegelijk zichtbaar maken.
- Dit kunt U doen door achtereenvolgens de volgende commando's te geven: "nummer", "een", "extra", "volgende", "extra", "nummer", "vier".
- De annotaties 1, 2 en 4 verdwijnen van het scherm, wanneer U een nieuwe annotatie kiest zonder het commando"extra" te geven.

pagina:
- Dit commando kunt U gebruiken om van pagina te wisselen.
- U kiest een pagina door hierna de button met het nummer van de pagina te geven.

einde:
- Met dit commando beeindigt U het programma.

(6) Taakomschrijving

Dit systeem bevat twee pagina's uit een Engels geschreven rapport. We veronderstellen even dat de pagina's door u zijn uitgetypt. De auteur heeft de tekst op type- en andere fouten gecorrigeerd en op enkele plaatsen aanvullingen en/of veranderingen aangebracht.
Deze correcties zijn in het systeem opgeslagen in de vorm van tekstannotaties.

Het is de bedoeling dat u de annotaties bij elke bladzijde leest en verwerkt, door de voorgestelde verbeteringen aan te brengen in de gedrukte teksten (die identiek zijn aan de pagina's tekst op het beeldscherm) die bij dit experiment worden uitgereikt. Wanneer er niet voldoende ruimte is op het papier, kunt u eventueel uw correcties op een apart blaadje schrijven en deze correcties merken met b.v. een *.
Overigens is het niet erg wanneer u de tekst inhoudelijk niet begrijpt, daar gaat het tijdens dit experiment niet om.

Werk zo goed en zo snel mogelijk uw taak af.
Succes!

(7) Experiment muis- en spraakbestuurd annotatiesysteem

<u>Inleiding</u>

Net als bij de twee voorgaande experimenten gaat u weer werken met het tekstannotatiesysteem.
Bij dit experiment heeft u echter de vrije keuze tussen muis- en spraakbediening voor het systeem.
De lengte van het experiment bedraagt ongeveer anderhalf uur.

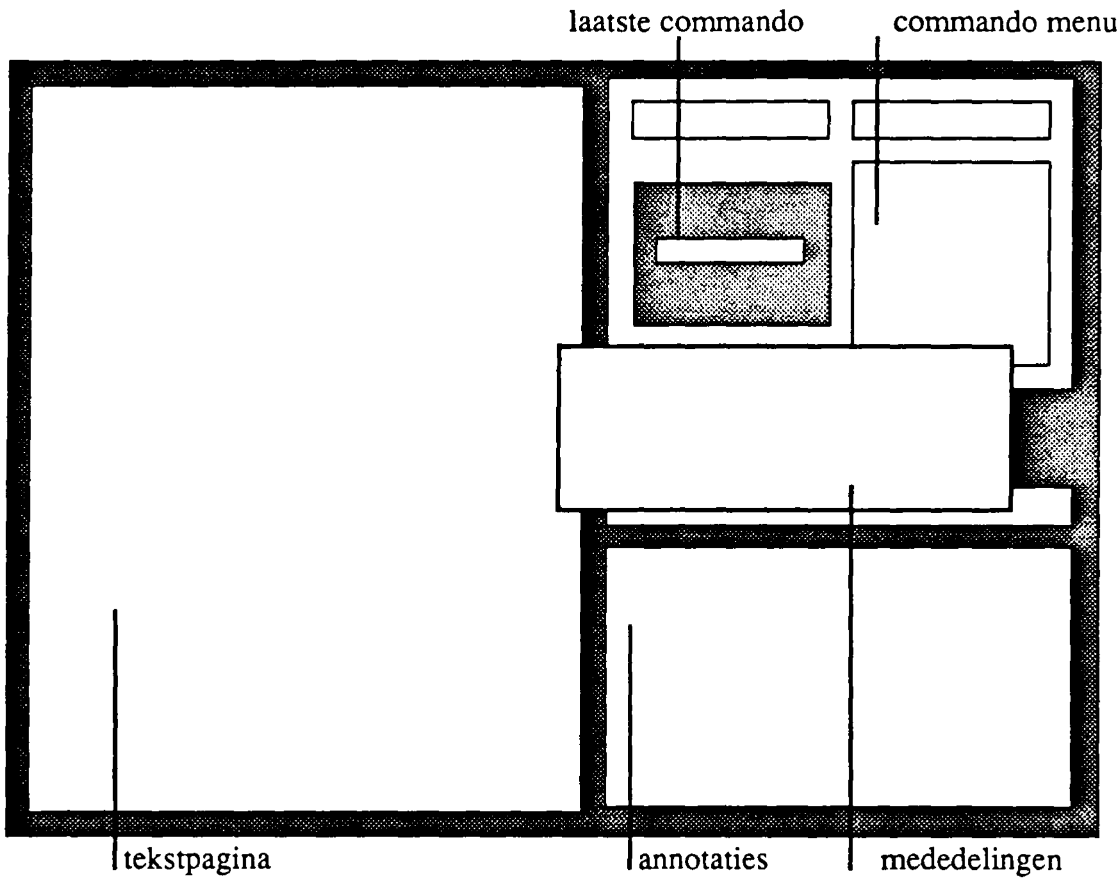
<u>Programma</u>

-Trainen van de spraakherkenner

-Experiment
-Vragenlijst en interview

Handleiding

De verschillende pagina's en annotaties worden zichtbaar gemaakt door muis- of spraak-besturing.
Indeling van het scherm:



Commando's

volgende, vorige :
-u kunt met deze commando's de annotaties in de volgorde waarop ze op het scherm staan selecteren.
nummer:
-met het commando nummer kunt u een annotatie direct selecteren door het nummer te geven.
extra:
-voor iedere *extra* annotatie die op het scherm wilt kiezen moet u het commando extra geven.
pagina:
-selectie van een pagina.

einde:
-beeindiging van het programma

Instrukties voor de bediening naar keuze

De vorige experimenten hebt u uitgevoerd door commando's te geven via de muis of de spraakherkenner. In dit laatste experiment bent u vrij in de manier waarop u de toetsen op het scherm bedient, of u kunt de toets met de muis activeren of u kunt het commando inspreken. Beide manieren van bedienen kunnen door elkaar gebruikt worden. Let er op dat alles wat u zegt door de spraakherkenner kan worden opgemerkt en mogelijk een toets kan activeren.
Steeds als een toets is geactiveerd is dat te zien in het speciale kader voor het laatst herkende commando.

Taakomschrijving

Het systeem bevat twee pagina's uit een in het Engels geschreven rapport.
De auteur heeft de tekst gecorrigeerd. De correcties zijn opgeslagen in de vorm van annotaties. Het is de bedoeling dat u de annotaties leest en verwerkt, door de voorgestelde verbeteringen aan te brengen in de gedrukte teksten die bij het experiment worden uitgereikt. Werk zo goed en zo snel mogelijk uw taak af.

Succes !

Appendix A.4 Questionnaire

This appendix shows the questionnaire (in Dutch) that was used after the third session

Vragenlijst experiment muis- en spraakbestuurd annotatiesysteem.

Tot slot verzoeken wij u de volgende vragenlijst in te vullen.

Leeftijd : .... jaar
Geslacht : (vrouw / man)
Beroep : ....
Opleiding : ....

1) Heeft u vaker met computers gewerkt?.............................. (ja / nauwelijks / nee)
Zo ja, hoe vaak?...

2) Heeft u eerder met muisbediening gewerkt?........................ (ja / nauwelijks / nee)

3) Heeft u eerder met spraakbediening gewerkt?..................... (ja / nauwelijks / nee)

4) Wat is uw kennis van de Engelse taal?................................(goed / matig / slecht)
....

5) Waarmee zou het systeem het liefst willen bedienen? (muis / spraak / beide)
Hoe sterk is deze voorkeur?.. (sterk / matig / zwak)

6) Waarmee zou u het gemakkelijkst kunnen werken?........ (muis / spraak / beide)
Hoe sterk is deze voorkeur?.. (sterk / matig / zwak)

7) Waarmee zou u het snelst kunnen werken? .... (muis / spraak / beide)
Hoe sterk is deze voorkeur?.. (sterk / matig / zwak)

8) Hoe vond u het om de muis te bedienen?..
....

9) Hoe vond u het om tegen het apparaat te spreken?...
....

10) Wilt u in het kort weergeven wat u de positieve- en de negatieve kanten vindt van de beide bedieningssystemen .
....
....
....

Op de achterkant van dit formulier kunt u commentaar (suggesties, vragen, opmerkingen) schrijven. Bedankt voor uw medewerking.

Appendix B.1 Subjects

Subjects	age	level of education
1	33	LHNO
2	40	mulo-A
3	55	non-universatary
4	44	MULO
5	22	MEAO
6	36	HBO
7	23	HEAO
8	53	MMS, HBO
9	36	MULO, secretary course
10	27	MAVO secretary course
11	48	MULO
12	43	HEAO
13	34	HAVO, Schoevers
14	23	HAVO, Schoevers
15	28	HAVO, secretary course
16	47	MMS, secretary course
17	35	HBS
18	39	MULO
19	26	HBO
20	26	MEAO
21	37	MMS
22	28	Schoevers
23	27	HTS
24	19	MAVO, Schoevers

Table B.1.1 : Information about the subjects

Appendix B.2 Experimental data

Session 1 Subjects	Document	Device	Time (sec)	Commands Mouse	Commands Speech	Errors Mouse	Errors Speech
1	1	M	1492	33		0	
2	2	M	1544	38		1	
3	3	M	1000	35		0	
4	1	M	769	35		0	
5	2	M	1257	30		1	
6	3	M	919	39		3	
7	1	S	949		38		4
8	2	S	1905		71		7
9	3	S	710		36		3
10	1	S	832		34		4
11	2	S	878		63		14
12	3	S	980		47		13
13	1	M	1416	33		0	
14	2	M	1511	83		5	
15	3	M	869	30		0	
16	1	M	1392	33		0	
17	2	M	1664	33		0	
18	3	M	1238	47		0	
19	1	S	985		48		9
20	2	S	1086		57		18
21	3	S	1174		44		4
22	1	S	1053		28		1
23	2	S	1830		149		29
24	3	S	811		43		5

Table B.2.1 : Data of session 1

Session 2 Subjects:	Document	Device	Time (sec)	Commands Speech	 Mouse	Errors Speech	 Mouse
1	2	S	1103	31		1	
2	3	S	1253	43		1	
3	1	S	846	32		3	
4	3	S	776	69		17	
5	1	S	948	32		8	
6	2	S	1131	73		18	
7	2	M	901		30		0
8	3	M	758		30		0
9	1	M	821		29		0
10	3	M	503		30		0
11	1	M	713		31		1
12	2	M	1263		32		1
13	2	S	1202	41		4	
14	3	S	901	44		1	
15	1	S	1063	46		11	
16	3	S	708	33		1	
17	1	S	2110	162		29	
18	2	S	1510	36		4	
19	2	M	794		38		6
20	3	M	524		38		0
21	1	M	1288		34		1
22	3	M	906		37		1
23	1	M	1059		35		0
24	2	M	785		41		0

Table B.2.2 : Data of session 2

Session 3	Document	Time	Number of commands			Number of errors		
Subjects:		(sec)	Total	Mouse	Speech	Total	Mouse	Speech
1	3	729	31	24	7	0	0	0
2	1	1542	33	32	1	1	0	1
3	2	968	31	3	28	1	1	0
4	2	1372	76	33	43	14	0	14
5	3	990	36	36	0	1	1	0
6	1	747	34	29	5	2	1	1
7	3	702	27	14	13	0	0	0
8	1	1101	35	35	0	0	0	0
9	2	839	34	34	0	1	1	0
10	2	935	49	27	22	8	1	7
11	3	576	32	21	11	4	0	4
12	1	1228	82	19	63	14	2	12
13	3	772	36	10	26	4	0	4
14	1	858	37	20	17	5	1	4
15	2	850	41	17	24	4	0	4
16	2	1449	37	0	37	1	0	1
17	3	1156	58	9	49	12	0	12
18	1	1585	88	11	77	14	9	5
19	3	644	44	9	35	6	1	5
20	1	643	34	27	7	5	1	4
21	2	1311	34	31	3	1	1	0
22	2	1178	81	41	40	16	0	16
23	3	955	56	36	26	14	2	12
24	1	878	51	22	27	7	3	4

Table B.2.3 : Data of session 3 (free choice mode).

Appendix B.3 Questionnaire data

Subjects	computer experience	mouse experience	speech experience
1	yes	no	no
2	yes	no	hardly
3	yes	yes	hardly
4	yes	no	no
5	yes	no	no
6	yes	yes	no
7	yes	no	no
8	yes	no	no
9	yes	yes	no
10	yes	no	no
11	yes	no	no
12	yes	hardly	no
13	yes	no	no
14	yes	no	no
15	yes	no	no
16	hardly	hardly	hardly
17	yes	hardly	no
18	yes	no	no
19	yes	yes	no
20	yes	yes	no
21	yes	no	no
22	yes	hardly	no
23	yes	yes	no
24	yes	yes	no

Table B.3.1 : Data about the experience of the subjects

Subjects	preference	ease of use	speed
1	both /strong	mouse /moderate	mouse /strong
2	speech/strong	speech/strong	speech/strong
3	mouse /strong	mouse /strong	mouse /strong
4	mouse /strong	mouse /strong	mouse /strong
5	mouse /strong	mouse /strong	mouse /strong
6	mouse /strong	mouse /strong	mouse /strong
7	both /moderate	mouse /moderate	mouse /moderate
8	mouse /strong	mouse /strong	mouse /strong
9	both /moderate	mouse /moderate	mouse /strong
10	mouse /strong	mouse /strong	mouse /strong
11	mouse /strong	mouse /strong	mouse /strong
12	speech/strong	mouse /moderate	both /slight
13	both /strong	speech/moderate	both /strong
14	both /strong	speech/moderate	speech/moderate
15	mouse /moderate	mouse /strong	mouse /strong
16	speech/strong	speech/strong	speech/strong
17	speech/moderate	speech/strong	speech/strong
18	both /moderate	mouse /strong	mouse /strong
19	mouse /strong	mouse /strong	mouse /strong
20	mouse /strong	mouse /strong	mouse /strong
21	speech/strong	speech/strong	speech/moderate
22	both /moderate	mouse /strong	mouse /moderate
23	both /strong	both /strong	both /strong
24	both /moderate	mouse /strong	both /strong

Table B.3.2 : Choices for type of control system for preference, ease of use and speed by the subjects.

B.3.3 Answers to the questionnaire categorized according to preference of the subjects

A number between brackets stands for the number of the subject that gave the response. The number before each answer stands for the number of subjects that gave the response.

Five subjects chose speech:

Speech pro
3 : do not have to stop making corrections (16,17,21)
3 : faster (2,12,17)
1 : saves time and actions (21)

Speech con
2 : uncertain recognition (2,21)
1 : takes a lot of concentration (2)
2 : hard to keep the right intonation (12,21)
1 : it's more difficult to help other people with questions (12)
1 : has to be used in a quiet environment (17)

Mouse pro
1 : certain (12)
2 : easy to use (2,12)
1 : effective and clear (12)
3 : fast (12,16,21)

Mouse con
1 : takes more space on your desk (2)
3 : have to stop making corrections to give a command (16,17,21)

Ten subjects chose mouse:

Speech pro
2 : fast (# 3 if used without head microphone, # 15 if 100% correct recognition)
2 : easy to use (# 3 if used without head microphone, # 15 if 100% correct recognition)
1 : exciting (5)
1 : hands free (11)
1 : do not have to look at the screen (11)

Speech con
8 : difficult to keep being recognized properly
(same intonation, bad recognition)
(3,4,5,7,8,15,19,20)
2 : too slow (4,11)
2 : noisy (10,20)
1 : strange to talk to a machine (8)
1 : estrangement from the environment, people can not talk to you anymore (10)
1 : people still have to get used to it (15)
1 : headphone (20)

Mouse pro:
8 : fast (3,4,5,6,8,10,11,15)
3 : easy to use (6,19,20)
1 : more certain (5)
1 : noiseless (8)
1 : efficient(8)
1 : can be used in any kind of environment (20)

1 : you know where the command area will be on the screen next so you can anticipate and already move to that area. Psychological effect: while waiting for the next menu to appear on the screen you can move the mouse. It seems that you're busy when you move the mouse. When using speech you have to sit 'idle' while waiting (11)

Mouse con

1 : no space on your desk if it is already full (2)
1 : would prefer using the keyboard (4)
1 : you have to get used to working with it (5)
1 : it should be placed somewhere where the board does not move very easily (8)
1 : you have to switch your point of focus between the screen and the mouse (11)

Nine subjects chose for controlling the system with both devices at their disposal:

Speech pro:

3 : hands free (1,14,22)
2 : it's possible to give a command while making corrections (9,24)

Speech con

4 : it's hard to speak consistent all the time(7,13,18,22)
1 : hard not to make any extra noises (1)
1 : it is often necessary to repeat commands (9)
2 : easier to make mistakes (14,23)
1 : headphone has to fit well and not move when user moves (14)
1 : has to be used in a quiet environment (24)

Mouse pro

5 : fast (7,9,13,23,24)
4 : easy to use (9,18,22,23,24)

Mouse con

2 : you need your hands (1,14)
1 : needs more skills (14)
1 : 'double commands' like "number-two", "next-extra-next" are easier when using speech (23)

Appendix B.4 Data description

To get a better understanding of the gathered data it is described in several ways.
Unexpected events may have occurred that can be detected by examining data that looks out of place. For example: a subject with completely different results as all the other subjects. The following methods have been used:
-Calculations of range, mean, standard deviation of all measurements to get a better understanding of what happened during the sessions.
-Submitting the data to an analysis of regression.

B.4.1 Range, mean and standard deviation of the data
Figure B.4.1 to figure B.4.20 give a description of the data of the three sessions. The order of subjects for session 1 and session 2 is the same in these figures; for session 1 mouse subjects are listed first, while for session 2 speech subjects are listed first.

All data that is gathered is described by means of range, mean and standard deviation (S.D.). To decide whether there are extreme values in the data that have been measured, an interval is calculated that stands for 95% of the population (between mean - 2 {standard deviation} and mean + 2 {standard deviation}). Values that are not part of the interval have to be studied in more detail to understand why an extreme value was found. The range and standard deviation for the speech version is generally bigger than for the mouse version.

An asymmetrical distribution of the data may be caused by an extreme value. A large standard deviation is caused by large differences between subjects.
For the measures of efficiency, accuracy, errors, time and commands the standard deviations of the mouse sessions are smaller than the standard deviations of the speech sessions. This means there is less difference in measures between mouse subjects than between speech subjects. Especially the difference between the standard deviation in number of errors of mouse and speech subjects is very high (S.D. speech subjects = 5 x [S.D. mouse subjects]).

Session 1:
<u>Total of session 1:</u>
Extreme values (the values that are no part of the interval that stands for 95% of the population):
-time (1905 sec)
-commands(149)
-errors(29)
-efficiency(0.17, caused by an extreme number of commands [149])
-accuracy: none
-working style (50.4 sec/command, caused by a relatively high processing time [1664 sec] compared to the number of commands [33] given during this time).
Three subjects in the **speech version** had difficulty with being properly recognized. After further studies of the data it turned out that in these three cases the speech-recognizer recognized wrong commands very frequently. As a result of this problem the subjects had to restore the action by giving a few extra commands. If for example "previous" was recognized in stead of "next", the subject had to give either twice the command "next" or the

commands "number" and "two" to restore the action. The extremes in commands given, efficiency and in processing time were a consequence of this problem.
Another subject had problems with being recognized at all. She had a total of 29 errors in the **speech version.** The extreme value for working style is found for a subject that compared to other subjects used a relatively long time to complete the task with a relatively small number of commands.

Mouse session 1:
Extreme values:
-time : none
-commands (83)
-errors (5)
-efficiency (0.31, caused by an extreme number of commands [83])
-accuracy (0.92 , caused by a relatively large number of errors [3] compared to the number of commands [39] that were given)
-working style: none
One subject had a total of 83 commands given in the **mouse version.** After further studies of the data, it turned out that she finished making corrections after 46 commands and used the remaining commands to look at all the annotations once more.
She also made 5 errors. It seems that the cursor was not on a button on the screen, when she pressed the mouse button. No further explanation can be given. The extreme value for efficiency is caused by the extreme value for commands.

Speech session 1:
Extreme values:
-time (1905 sec)
-commands (149)
-errors (29)
-efficiency: none
-accuracy: none
-working style (37.6 sec/command, caused by a relatively high processing time [1053 sec] compared to the number of commands [28] given during this time)
The extremes in time, number of commands and errors are the same as those described in the total of session 1. The extreme value for working style is found for a subject that compared to other subjects used a relatively long time to complete the task with a relatively small number of commands.

Session 2:
Total of session 2:
Extreme values:
-time (2110 sec)
-commands (162)
-errors (29)
-efficiency (0.21, caused by an extreme number of commands [162])
-accuracy: none
-working style: none
The extremes in this session are caused by the same problem as in session one:
the speech recognizer recognized a different command than the one that was uttered. The commands that were given to restore the action lead to a high number of commands and a

high processing time.
This subject also had problems with being recognized at all (total number of errors: 29). Note that this subject is not the same one as the speech subject in session 1 who made 29 errors.

Mouse session 2:
Extreme values:
-time: none
-commands: none
-errors (6)
-efficiency: none
-accuracy (0.84, caused by a relatively large number of errors [6] compared to the number of commands [38] that were given)
-working style: none
One subject gave 6 mouse commands in a row while the cursor probably wasn't on the button on the screen.

Speech session 2:
Extreme values:
-time (2110 sec)
-commands (162)
-errors (29)
-efficiency (0.16, caused by an extreme number of commands [162])
-accuracy: none
-working style: none
The extremes are the ones that were described in the total of session 2.

Session 3:
Total of session 3:
Extreme values:
-time: none
-commands (88)
-errors: none
-efficiency: none
-accuracy: none
-working style (46.7 sec/command, caused by a relatively high processing time [1542 sec] compared to the number of commands [33] given during this time)
Note: Since the time spent using a control system is influenced by the length of the annotation and thus, by the time needed to make a correction, it is not possible to compare the time used with speech and the time used with mouse commands. Therefore, these measures are made over the complete session and are not separated into mouse and speech effects.

One subject has given an extreme number of commands. She also gave the extreme number of speech commands and mouse errors. It seems as though she finished making all corrections after 50 commands and used the remaining 38 commands to check all the annotations one more time. Note: this is not the same subject as the subject that checked all annotations once more in session 1. The extreme value for working style is found for (another) subject that compared to other subjects used a relatively long time to complete the task with a relatively small number of commands.

Mouse session 3:
Extreme values:
-time : not available
-commands (88)
-errors (9)
-efficiency: not available
-accuracy (0.18, caused by a relatively large number of errors [9] compared to the number of commands [11] that were given and infinite => no errors and no commands)
-working style: not available
One subject made 9 errors when using mouse. All these errors were made in a row. She waited for a while after the first time and then hit the button for a couple of times in a row.
A possible explanation is that she didn't notice that the cursor was not placed on the button properly. The extreme value for accuracy is caused by the fact that this subject used a very small number of mouse commands and a large percentage of these mouse commands were errors.

Speech session 3:
Extreme values:
-time : not available
-commands (77)
-errors (16)
-efficiency: not available
-accuracy (infinite => no errors and no commands)
-working style: not available
One subject gave 77 speech commands. As explained in the total of session 3 she checked all the annotation once again after 50 commands of which 39 were speech commands.
One subjects had problems to be recognized properly. There was one command that caused problems (" next ").

Learning time
One subject needed an extreme period of time to complete the task in session 1 with speech. In session 2 she did not need an extreme time to complete the task. The extreme value lead, however, to an extreme value on learning time. The reason for the high processing time is already discussed in speech-session 1.

Learning style
Two subjects needed an extreme number of commands to complete the task in the speech session. One had speech control at the first session and the other had speech control at the second session. After further studies of the data it turned out that in these cases the speech recognizer recognized wrong commands very frequently. Thus, a lot of extra commands were needed for repair actions.

Operational preference
One subject gave an extreme number of correct speech commands [72], she checked all the annotations once again after a total of 50 commands of which 39 were speech commands.

Number of errors in session 3
One subject made 9 mouse errors and another subject made 16 speech errors.

Conclusion:
Some subjects had problems with being recognized properly. Because of this problem they used more time and commands to complete the task. Some subjects had problems to be recognized at all, thus the total number of commands increased considerably. Some subjects checked all the annotations for a second time, thus increasing the total number of commands and decreasing efficiency. Some mouse subjects made a large percentage of errors compared to other mouse subjects, and thus had a low score on accuracy. Some subjects worked slower than others, using a relatively small amount of commands compared to the time they used. For the measures of efficiency, accuracy, errors, time and commands the standard deviations of the mouse sessions are smaller than the standard deviations of the speech sessions. This means there is less difference in measures between mouse subjects than between speech subjects.

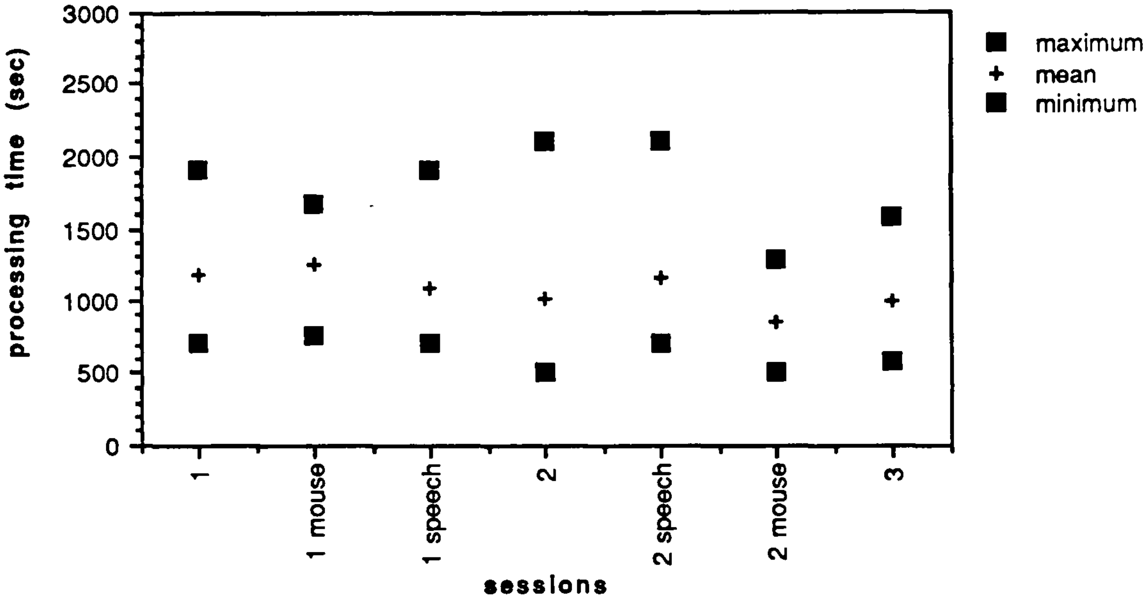


Figure B.4.1 : Range of data on processing time for sessions 1, 2 and 3 Note: the measure of session 1 is based on 24 subjects and the measure of 1 mouse and 1 speech is based on 12 subjects each

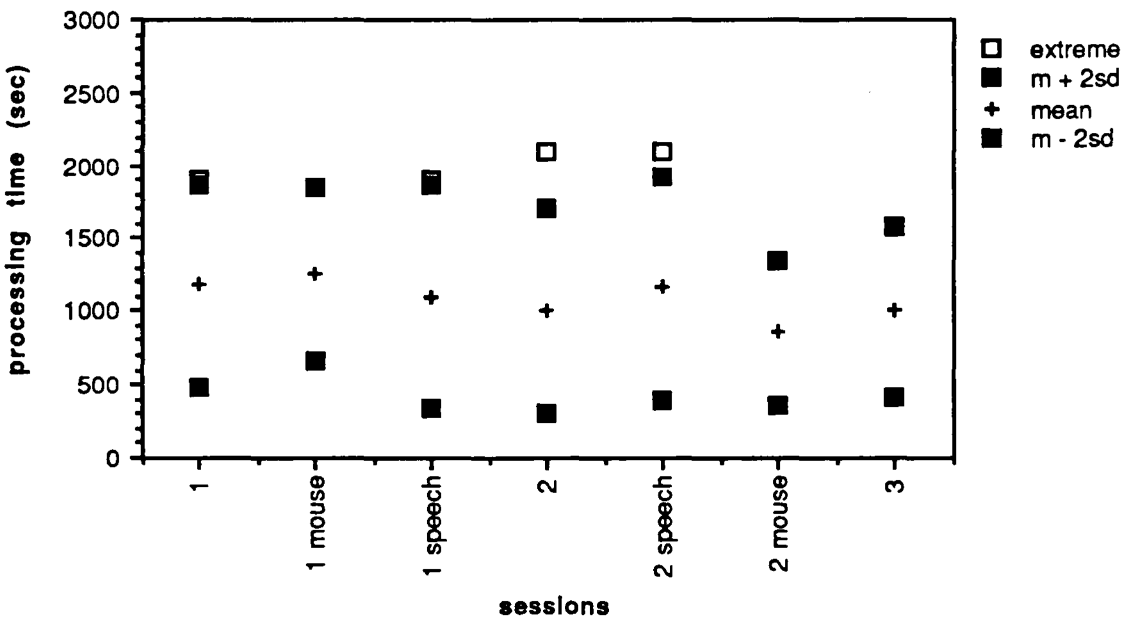


Figure B.4.2: Data description of processing time by means of standard deviation and extremes

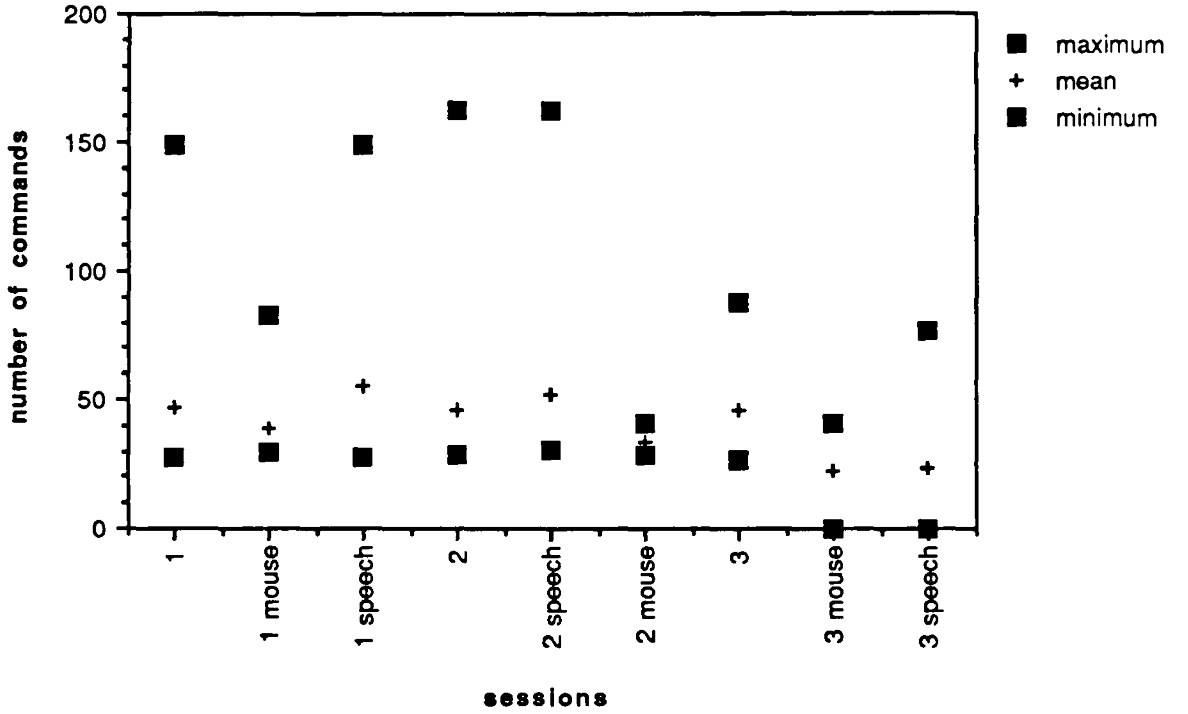


Figure B.4.3: Range of data on number of commands

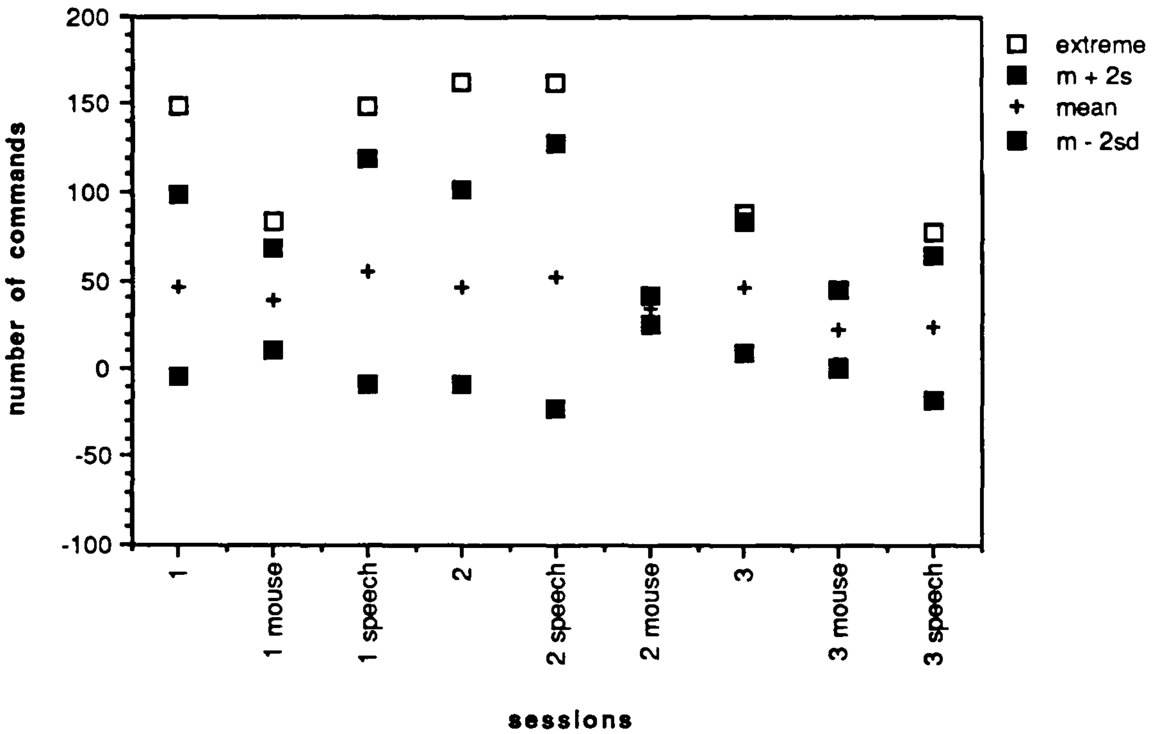


Figure B.4.4: Data description of number of commands by means of standard deviation and extremes

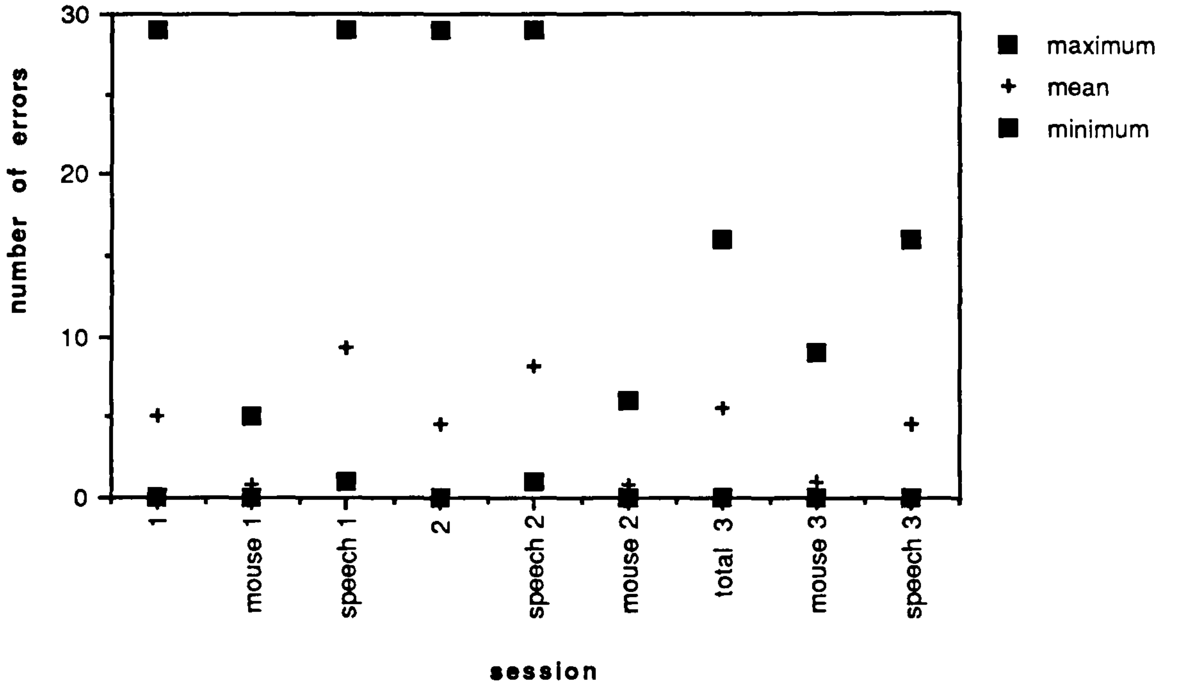


Figure B.4.5 : Range of data on number of errors

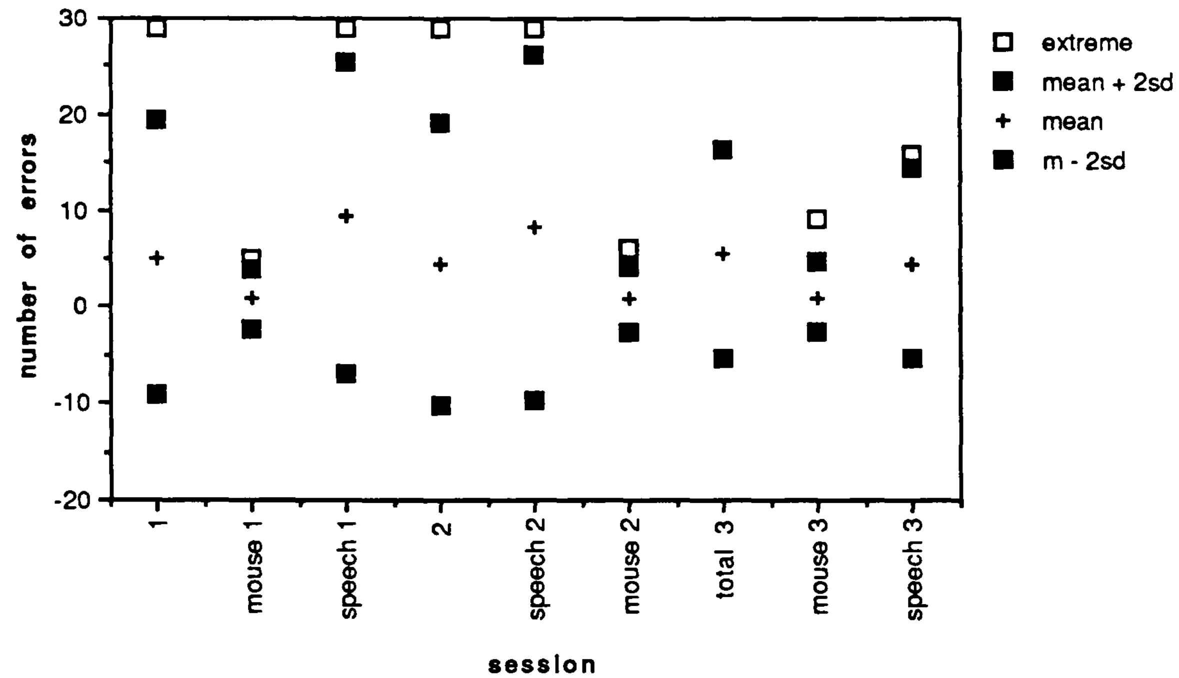


Figure B.4.6: Data description of number of errors by means of standard deviation and extremes

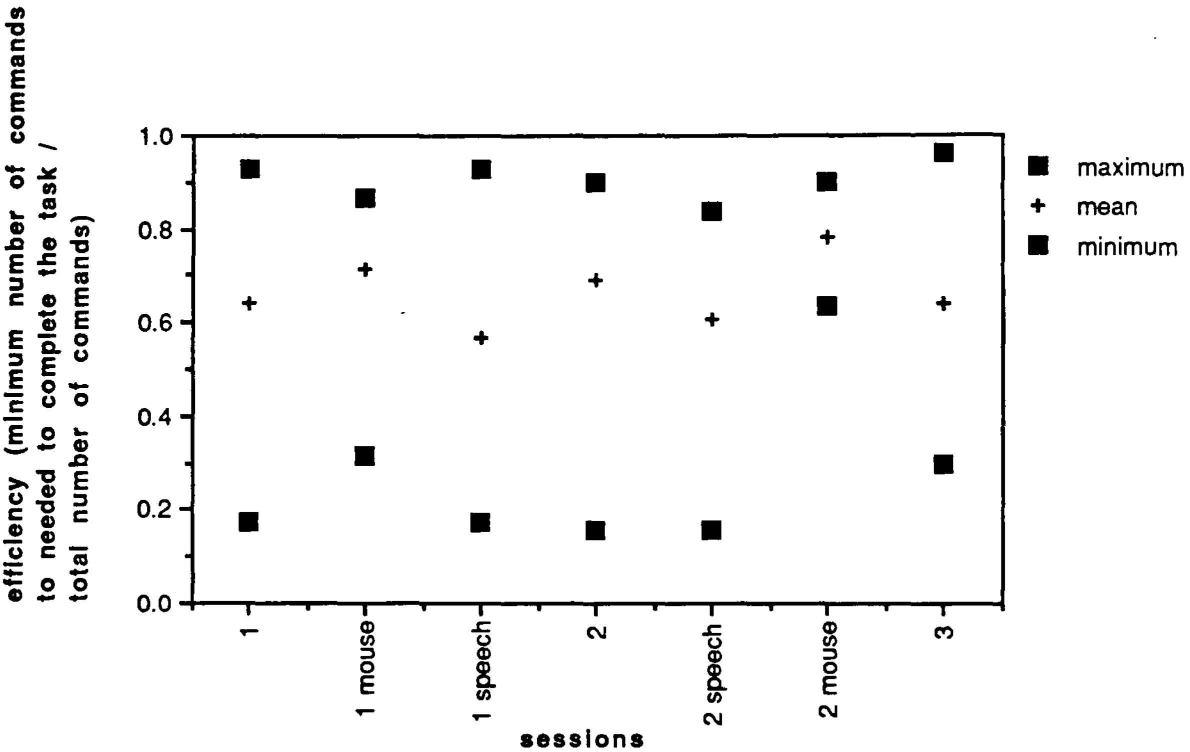


Figure B.4.7 : Range of data on efficiency (minimum number of commands needed to complete the task / total number of commands)

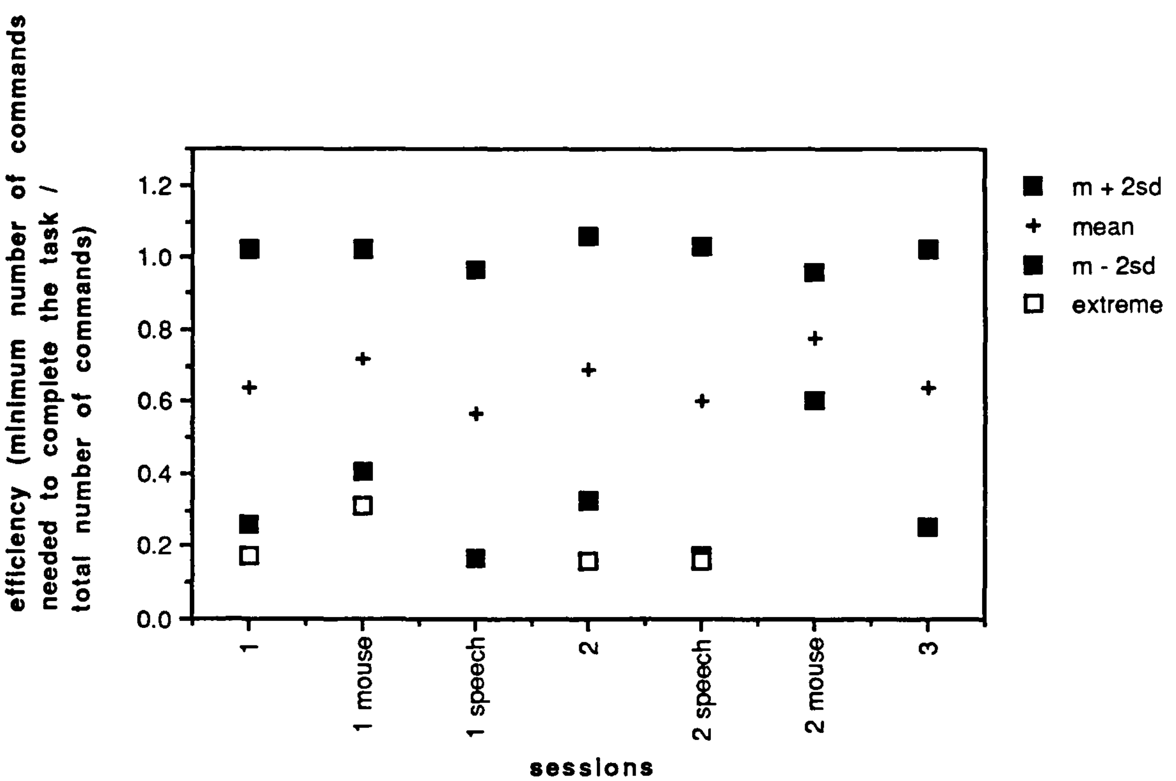


Figure B.4.8: Data description of efficiency by means of standard deviation and extremes

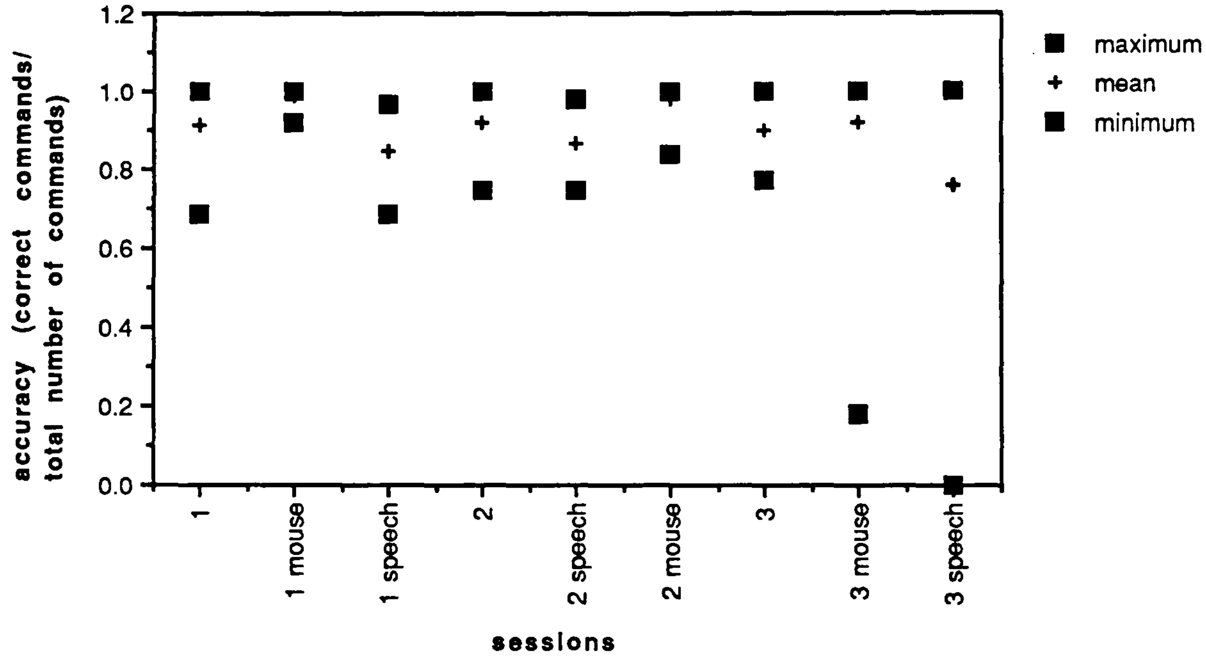


Figure B.4.9 : Range of data on accuracy (correct commands/total number of commands)

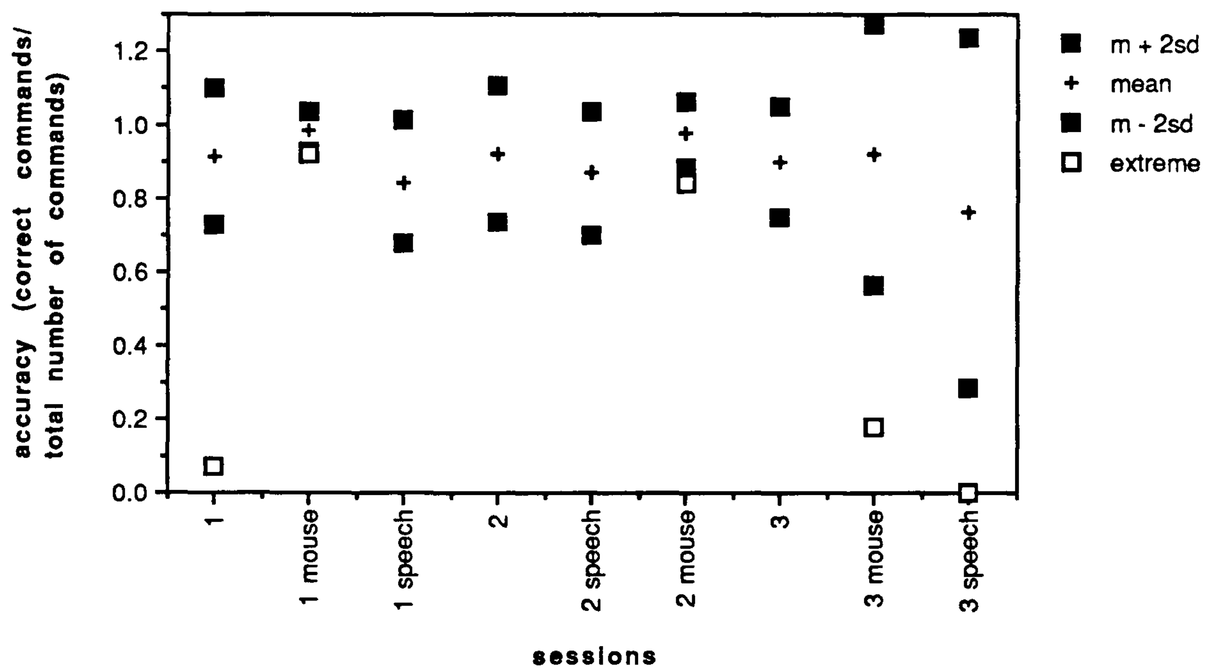


Figure B.4.10 : Data description of accuracy by means of standard deviation and extremes

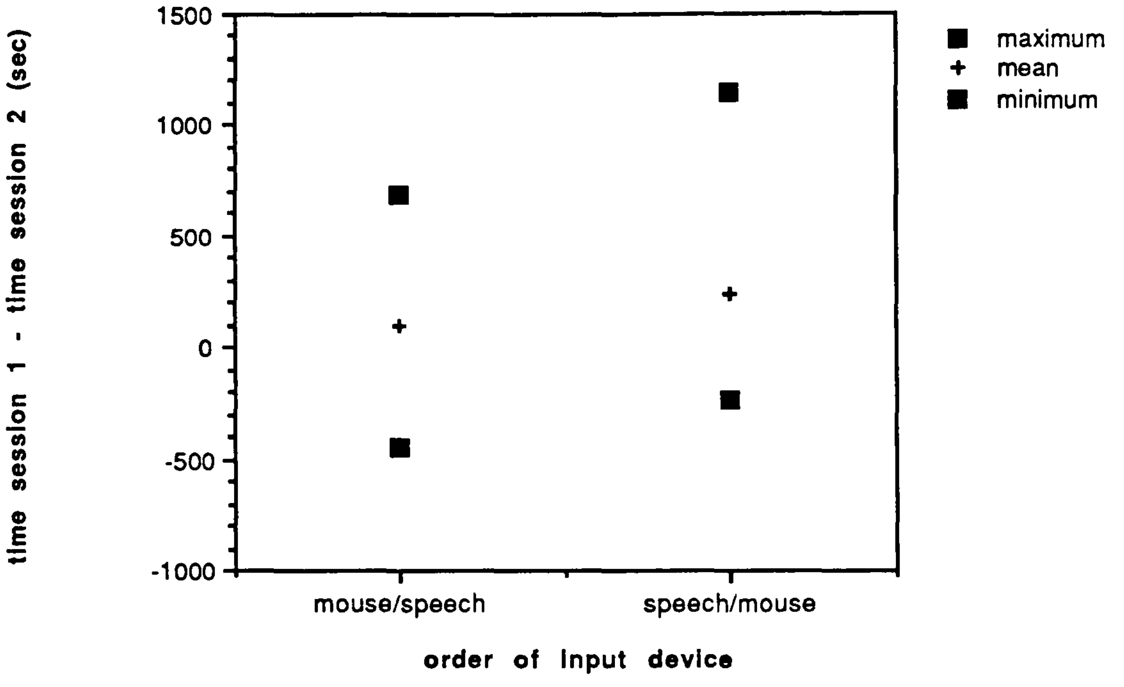


Figure B.4.11 : Range of data on learning time

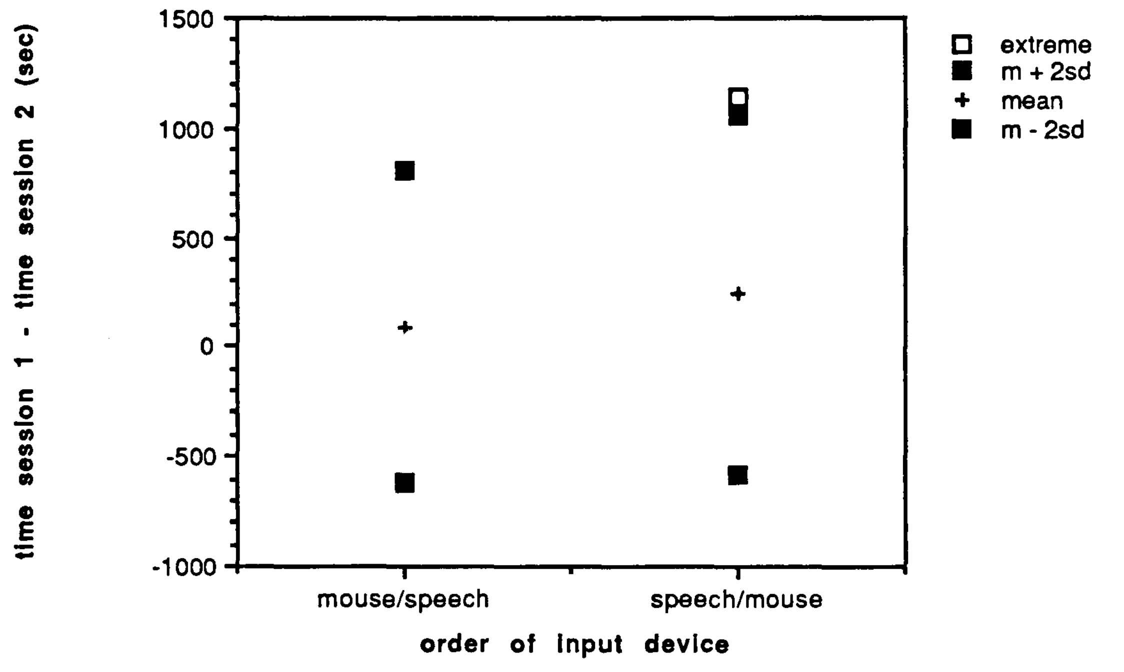


Figure B.4.12 : Data description of learning time

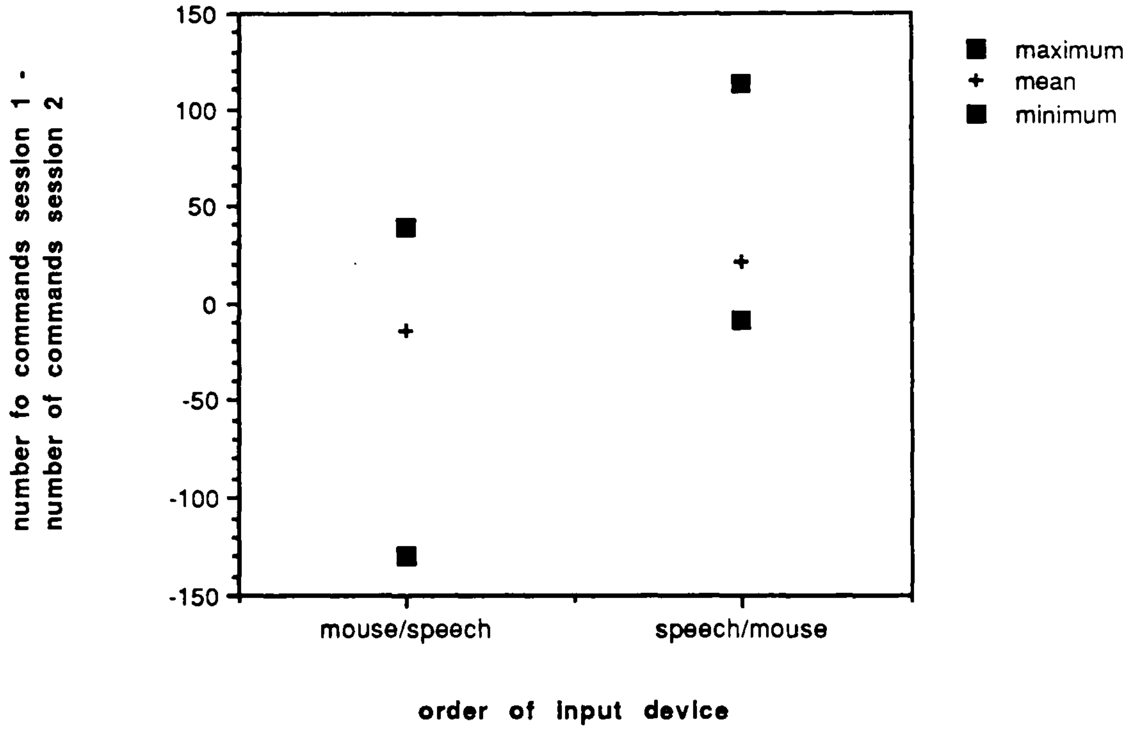


Figure B.4.13 : Range of data on learning style (time session 1 - time session 2)

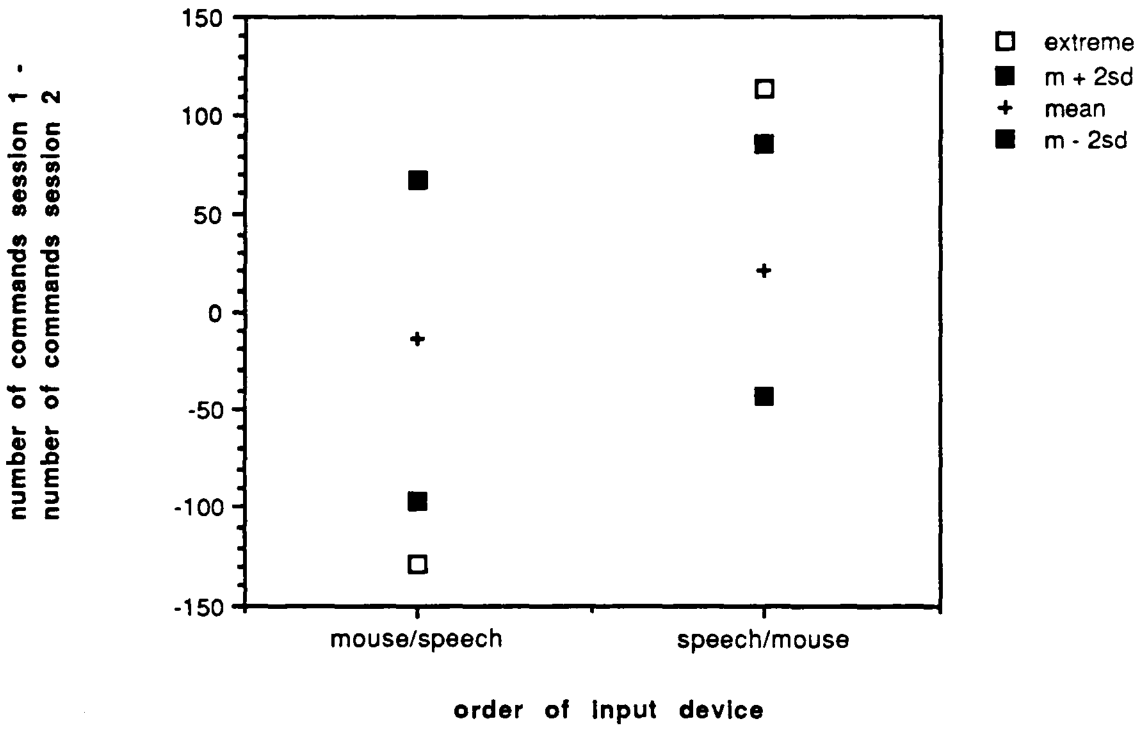


Figure B.4.14 : Data description of learning style by means of standard deviation and extremes

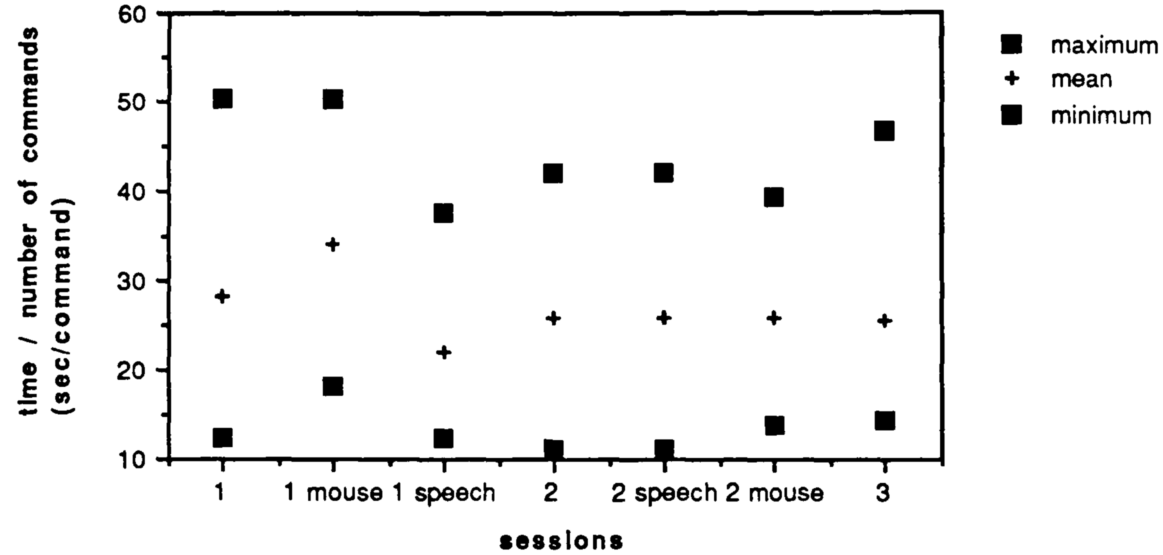


Figure B.4.15 : Range of data on working style

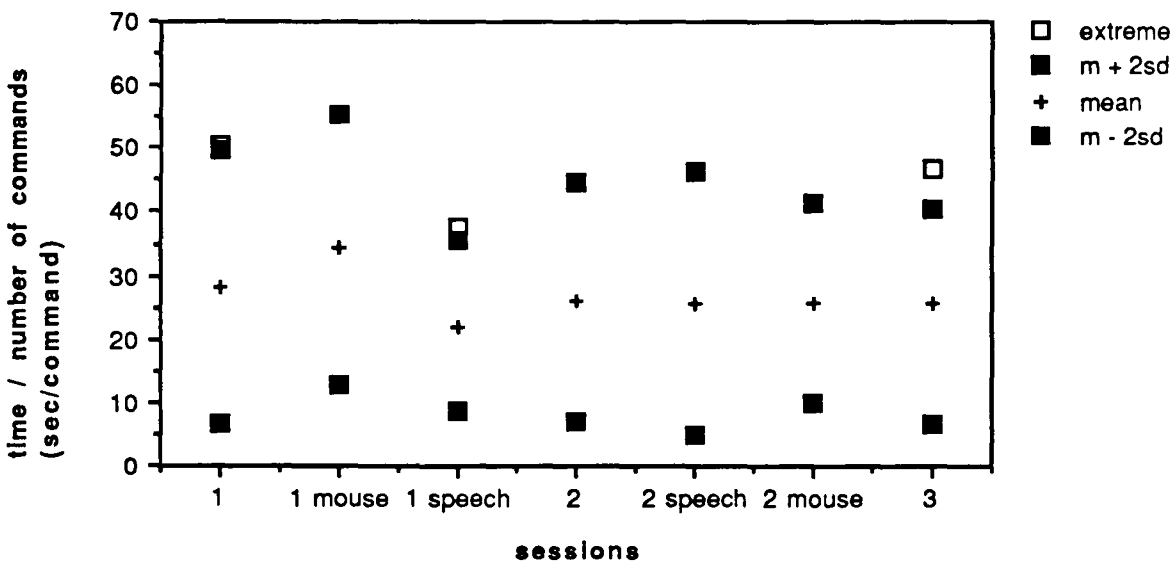


Figure B.4.16 : Data description of working style by means of standard deviation and extremes

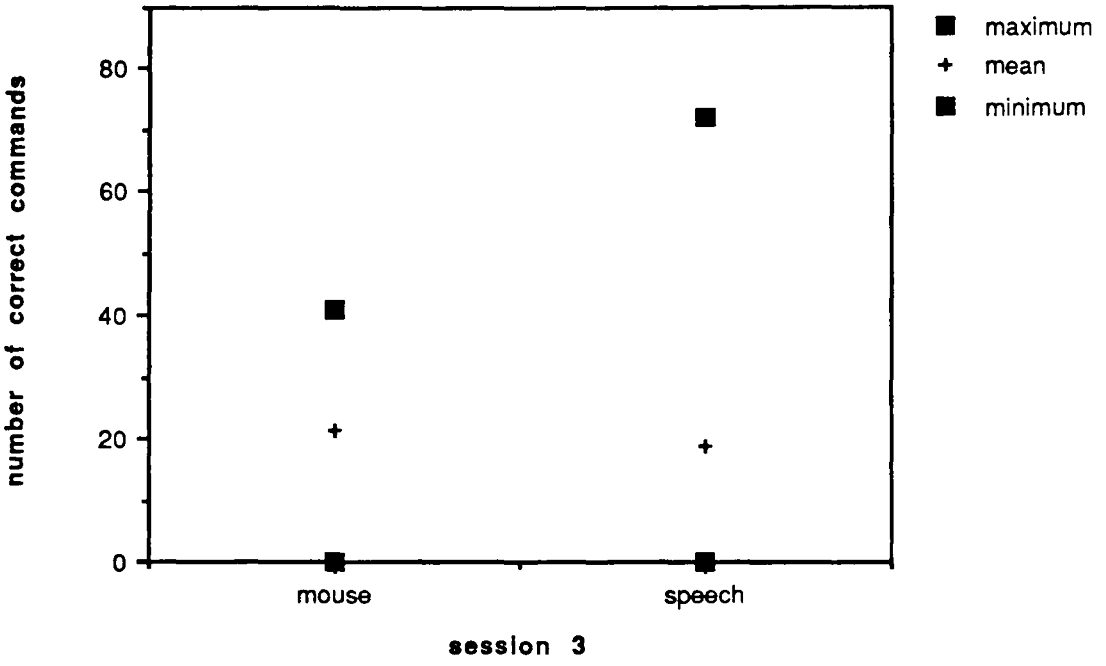


Figure B.4.17: Range of number of correct commands with mouse and speech control

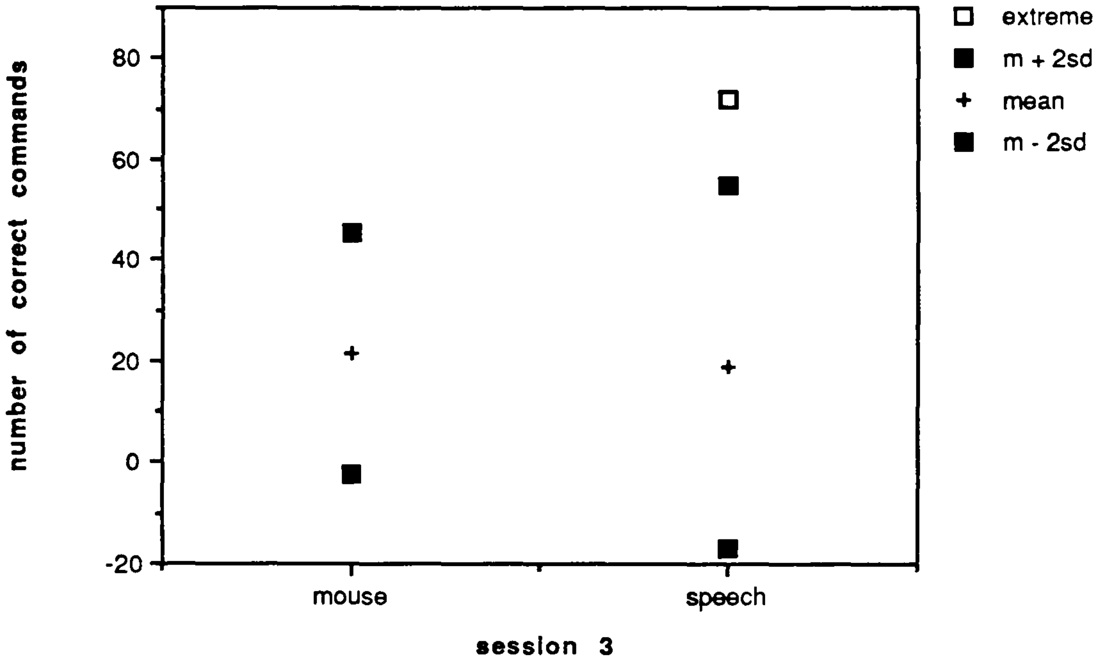


Figure B.4.18: Data description of number of correct commands in session 3 by means of standard deviation and extremes

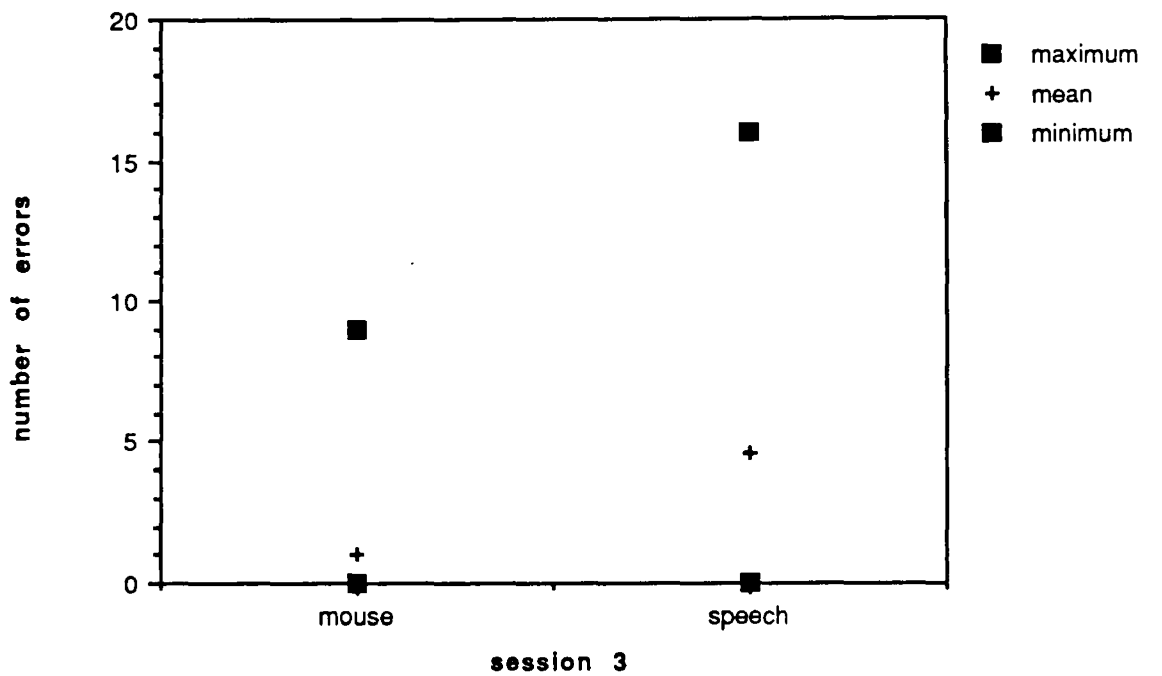


Figure B.4.19 : Range of data on number of errors in session 3

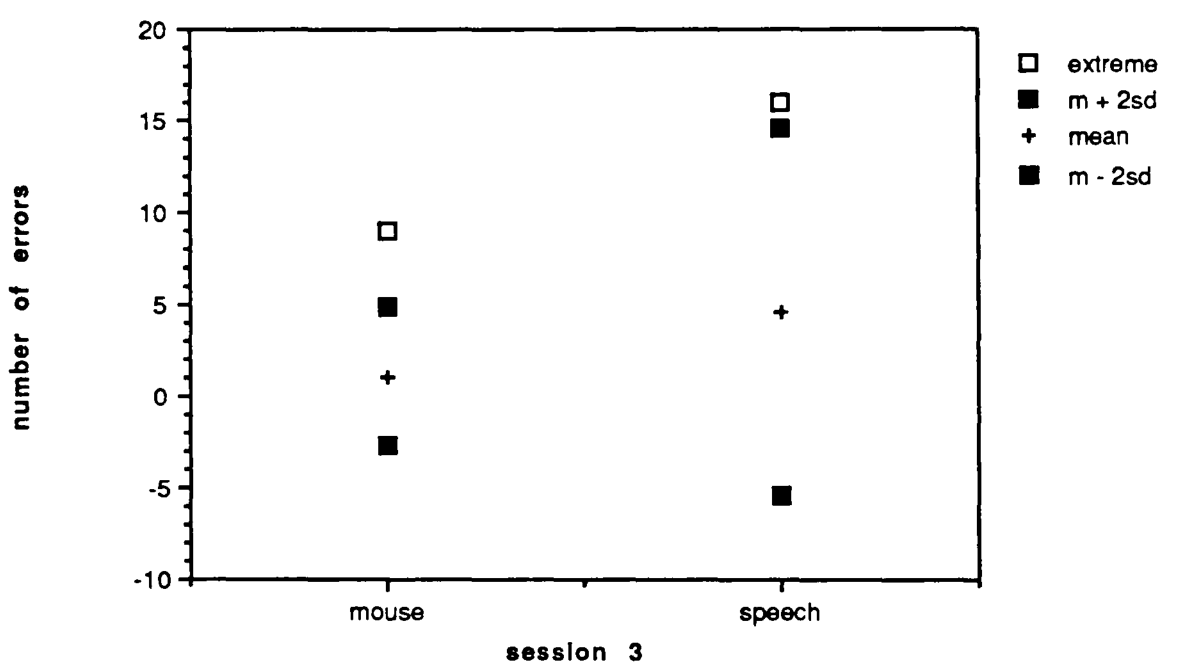


Figure B.4.20: Data description of number of errors by means of standard deviation and extremes

B.4.2 Statistical analysis without extreme values

To reduce the effects on the results of extreme values of the measurements that were made, the analyses of variance were done on the data without these extreme values.

B.4.2.1 Processing time

Session 1:

The mean processing time for the mouse version (1255 sec) is slightly higher than the mean processing time for the speech version (1026 sec). An analysis of variance showed that this difference is not significant ($F_{1,17} = 3.684$, $p = 0.0719$). (1 extreme value)

Session 2:

The mean processing time for the mouse version (860 sec) is slightly higher than the mean processing time for the speech version (1076 sec). (1 extreme value)

Session 1 and 2:

The mean processing time for the mouse subjects (1057 sec) is slightly lower than the mean processing time for the speech subjects (1051 sec). No significant difference was found between processing time of subjects using mouse and subjects using speech control ($F_{1,16} = 0.161$, $p = .693$). (2 extreme values)

B.4.2.2 Number of commands

Session 1:

The mean number of commands for the mouse version (35 commands) is lower than the mean number of commands for the speech version (46 commands). An analysis of variance showed that this difference is significant ($F_{1,16} = 24.812$, $p = 0.0001$). (2 extreme values)

This is a different conclusion than the conclusion based on all the data (including the extremes).

Session 2:

The mean number of commands for the mouse version (34 commands) is lower than the mean number of commands for the speech version (44 commands). (1 extreme value)

Session 1 and 2:

The mean number of commands for the mouse subjects (34 commands) is lower than the mean number of commands for the speech subjects (45 commands).
A significant difference was found between number of commands given by subjects using mouse and subjects using speech control ($F_{1,15} = 15.435$, $p = .001$). (3 extreme values)

B.4.2.3 Number of errors

Session 1:

The mean number of errors for the mouse version (0.46 errors) is lower than the mean number of errors for the speech version (7.45 errors). An analysis of variance showed that this difference is significant ($F_{1,16} = 28.991$, $p = 0.0001$). (2 extreme values)

Session 2:

The mean number of errors for the mouse version (0.36 errors) is lower than the mean number of errors for the speech version (6.27 errors). (2 extreme values)

Session 1 and 2:

The mean number of errors for the mouse subjects (0.41 errors) is lower than the mean number of errors for the speech subjects (7.05 errors). A significant difference was found between number of errors given by subjects using mouse and subjects using speech control ($F_{1,14}$ = 28.046, p = .000). (4 extreme values)

B.4.2.4 Efficiency

Session 1:

Mouse subjects use an average of 1.3 times the minimum number of commands (efficiency = 0.75), while speech subjects use an average of 1.7 times the minimum number of commands (efficiency = 0.57) required to perform the task. An analysis of variance showed that this difference is significant ($F_{1,17}$ = 16.85, p =0.0009). (1 extreme value)

Session 2:

Mouse subjects use an average of 1.3 times the minimum number of commands (efficiency = 0.77), while speech subjects use an average of roughly 1.5 times the minimum number of commands (efficiency = 0.65) required to perform the task. (1 extreme value)

Session 1 and 2:

The mean efficiency for mouse subjects (0.77) is higher than the mean efficiency for speech subjects (0.61). A significant difference was found ($F_{1,16}$ = 6.397, p = .001). (2 extreme values)

B.4.2.5 Accuracy

Session 1:

The mean accuracy for the mouse version (0.990) is higher than the mean accuracy for the speech version (0.846). An analysis of variance showed that this difference is significant. ($F_{1,17}$ = 32.08, p = 0.0001). (1 extreme value)

Session 2:

The mean accuracy for the mouse version (0.989) is higher than the mean accuracy for the speech version (0.869) . (1 extreme value)

Session 1 and 2:

The mean accuracy for the mouse version (0.984) is higher than the mean accuracy for the speech version (0.858). An analysis of variance showed that this difference is significant. ($F_{1,16}$ = 36.377, p = 0.000). (2 extreme values)

B.4.2.6 Learning time

The mean learning time of the mouse/speech order (93 sec.) is lower than the mean learning time of the speech/mouse order (157 sec.). An analysis of variance showed that this difference is not significant ($F_{1,17}$ = .598, p = .4498) .(2 extreme values)

B.4.2.7 Learning style

The mean learning style of the mouse/speech order (14 commands) is lower than the mean learning style of the speech/mouse order (21 commands). An analysis of variance showed that this difference is significant ($F_{1,17}$ = 2.016, p = .1748).(1 extreme value)

B.4.2.8 Working style

Session 1:

The mean working style for the mouse version (34 sec/command) is higher than the mean working style for the speech version (21 sec/command). An analysis of variance showed that this difference is significant($F_{1,17}$ = 15.29, p = .0011). (1 extreme value)

Session 2:

The mean working style for the mouse version (25.775 sec/command) is higher than the mean working style for the speech version (25.779 sec/command). (0 extreme values)
Session 1 and 2:
The mean working style for the mouse subjects (30 sec/command) is higher than the mean working style for the speech subjects (23 sec/command). A significant difference was found between working style by subjects using mouse and subjects using speech control ($F_{1,17}$ = 7.504, p = .014). (1 extreme value)

B.4.2.9 Operational preference
The mean number of correct commands for the mouse version (21.458) is higher than the mean number of correct commands for the speech version (18.875). An analysis of variance showed that this difference is not significant ($F_{1,18}$ = 0.223, p = .642).

B.4.2.10 Number of errors in session 3
The mean number of errors for the mouse version (0.727) is lower than the mean number of errors for the speech version (4.045). An analysis of variance showed that this difference is significant ($F_{1,18}$ = 9.209, p = 0.008).

Conclusion: the conclusions based on all the data hardly differ from the conclusions based on the data without the extreme values. The only difference that was found (difference in the number of commands between mouse and speech subjects is significant in the data without extremes and not significant in all the data), has no influence on the total conclusions, since this tendency also showed in the analysis of the data of session 1 and session 2.

B.4.3 Regression analysis
Session 1:
The regression between the *number of commands* and the *number of errors* was significant for both the mouse version (R = .822 , Z = 3.47, => α << .0004) and the speech version (R = .876, Z = 3.04 => α = .0024)

This means that subjects that use more commands also make more errors than subjects who use less commands.

The regression between *processing time* and *total number of errors* was not significant for both the mouse version (R = .106, Z = 1.3, α = .1936) and the speech version (R = .499, Z = 1.64 , α = 0.101).

No significant correlation is present between speed and number of errors.

The regression between *processing time* and *number of commands* was not significant for both the mouse version (R = .242 , Z = .741 , α = .4592)and the speech version (R = .75, Z = 1.91, α = .0562).

No significant correlation is present between processing time and number of commands.

The regression between *efficiency* and *errors* was significant for both the mouse version

(R = .822 , Z = 3.49 , α = .0004) and the speech version (R = .876, Z = 4.07, α << .0004).

This means that subjects who are highly efficient make less errors that subjects that are less efficient.

The regression between *efficiency* and *time* was not significant for the mouse version (R = .242 , Z = .741, α = .4592) ,whereas it was for the speech version (R = .75, Z = 2.91, α = .0036).

Speech subjects who are more efficient use less time than speech subjects who are less efficient.
For mouse subjects no significant correlation was present between these measures.

The regression between *working style* and *errors* was not significant for the mouse version (R = .525 , Z = .1.749, α = .0802) ,whereas it was for the speech version (R = .718, Z = 2.71, α = .0068).

Speech subjects who use more time per command will make more errors than speech subjects who use less time per command.
No significant correlation was present between these measures for mouse subjects..

The regression between *accuracy* and *time* was not significant for both the mouse version (R = .016 , Z = .099, α = .92) and the speech version (R = .016, Z = .048, α = .9802).

No correlation was present between these measures.

Session 2:
The regression between *the number of commands* and *the number of errors* was not significant for the mouse version(R = .328 , Z = 1.021 , => α = .3078), though it was for the speech version (R = .896 , Z = 4.351 , => α << .0004).

This means that speech subjects that use more commands will also make more errors than subjects that use less commands. For mouse subjects no significant correlation was present for these measures.

The regression between *processing time* and *number of errors* was not significant for both the mouse version (R = .073 , Z = .219 ,= > α = .8258) and the speech version (R = .549 , Z = 1.851, =. α = .0644).

No correlation was present between processing time and number of errors made.

The regression between *processing time* and *number of commands* was not significant for the mouse version (R = .027 , Z = .081 => α = .9362), whereas it was for the speech version (R = .681 , Z = 2.491 => α = .0142).

This means that speech subjects who are fast use less commands than slow speech subjects. This significant correlation is not present for the mouse subjects.

The regression between *efficiency* and *errors* was not significant for the mouse version (R = .328 , Z = 1.02 , α = .3078) ,whereas it was for the speech version (R = .896, Z = 4.354, α << .0004).

This means that speech subjects who are more efficient make less errors than subjects who are less efficient. For mouse subjects no significant correlation was present between these measures.

The regression between *efficiency* and *time* was not significant for the mouse version (R = .027 , Z = .08, α = .9362) ,whereas it was for the speech version (R = .681, Z = 2.49, α = .0128).

This means that speech subjects who are more efficient use less time to complete the task than speech subjects who are less efficient. No significant correlation was present between these measures for mouse subjects.

The regression between *working style* and *error* was not significant for both the mouse version (R = .144 , Z = .435, α = .66) and the speech version (R = .546, Z = 1.838, α = .0654).

No correlation was present between these measures.

The regression between *accuracy* and *time* was not significant for both the mouse version (R = .256 , Z = .0814, α = .418) and the speech version (R = .352, Z = 1.103, α = .2704).

No significant correlation was present between these measures.

Session 1-2:
The regression between *learning time* and *learning style* was significant for both mouse/speech order (R = .687 , Z = .2476, => α = .0132) and for speech/mouse order (R = 598 , Z = 2.07 , α = .0382)

This means that subjects who have a greater difference between processing time of session 1 and session 2 will also have a greater difference between numbers of commands in session 1 and session 2 than subjects who have a smaller difference between the two processing times and the numbers of commands in session 1 and session 2.

Session 3:
The regression between *processing time* and *number of commands* was significant.
(R = .519 , Z = 2.597 => α = .0094)

The regression between *processing time* and *the number of speech commands* was signifi-

cant
($R = .445$, $Z = 2.19$, $\alpha = .0286$)

The regression between *processing time* and *number of mouse commands* was not significant
($R = .034$, $Z = .034$, $\alpha = .8808$)

> This means that subjects that use more speech commands will also use more time than subjects who use less commands. However, such a significant correlation was not present for mouse commands. The correlation found between processing time and total number of commands might be a consequence of the correlation found between speech commands and processing time.

The regression between *processing time* and *the total number of errors* was not significant. ($R = .326$, $Z = 1.55 \Rightarrow \alpha = .1212$)

The regression between *processing time* and *the number of speech errors* was not significant.($R = .23$, $Z = 1.07 \Rightarrow \alpha = .2846$)

The regression between *processing time* and *the number of mouse errors* was not significant.($R = .33$, $Z = 1.57 \Rightarrow \alpha = .1164$)

> This means that there is no significant correlation between processing time and number of errors what so ever.

The regression between *number of commands* and *number of speech commands* was significant.($R = .836$, $Z = 5.539 \Rightarrow \alpha << .0004$)

The regression between *number of commands* and *number of mouse commands* was not significant. ($R = .104$, $Z = .313 \Rightarrow \alpha = 07566$)

> This means that subjects that use a greater amount of commands will also use more speech commands than subjects who use less commands, whereas this is not the case for mouse commands.

The regression between *number of mouse commands* and *number of speech commands* was significant. ($R = .459$, $Z = 2.273 \Rightarrow \alpha = .0232$)

> This means that subjects using a greater amount of mouse commands will also use more speech commands than subjects who use less mouse commands.

The regression between *number of commands* and total *number of errors* was significant
($R = 939$, $Z = 7.932$, $\alpha << .0005$)

The regression between *number of commands* and *number of speech errors* was significant ($R = .824$, $Z = 5.361$, $\alpha << .0005$)

The regression between *the number of commands* and *number of mouse errors* was significant (R = .522 , Z = 2.655, α = .0078)

> This means that subjects that use a greater amount of commands will also make more errors than subjects who use less commands.

The regression between *the number of mouse commands* and *number of mouse errors* was not significant (R = .145 , Z = .6688 , => α = .5028)

The regression between *the number of speech commands* and *number of speech errors* was significant (R = .64 , Z = 3.472 => α => .0006)

> This means that subjects that use more speech commands will also make more errors than subjects that use less commands with the speech control. No significant correlation was present for these measures with mouse control. This might be explained by the following factors: 1) that there is no significant difference in number of correct commands given with mouse and speech control, 2) that the average number incorrect speech commands is much higher than the average number of incorrect mouse commands. The total number of speech commands might be influenced directly by the number of incorrect speech commands, while this is not the case for mouse control.

.

The regression between *the number of speech commands* and *the number of mouse errors* was significant (R = .546 , Z = 2.806 , => α = .005)

The regression between *the number of mouse commands* and *the number of speech errors* was not significant (R = .173, Z .8004 , => α = .4238)

> This means that subjects that use more speech commands will make more mouse errors that subjects using less speech commands, though this significant correlation is not present between mouse commands and speech errors.

The regression between *total number of errors* and *number of speech errors* was significant. (R = .930 , Z = 7.938 => α << .0005)

The regression between *total number of errors* and *number of mouse errors* was not significant.(R = .393, Z = 1.904 => α = .0524)

> This means that subjects that make more errors also make more speech errors. This correlation is not present between mouse errors and the total number of errors.

The regression between *the number of mouse errors* and *the number of speech errors* was not significant (R = .053 , Z = .243 , => α = .8104)

> This means that no significant correlation is present between the number of mouse errors and the number of speech errors.

The regression between *efficiency* and *time* was significant (R = .519, Z = 2.587, α =

.0097).

> This means that subjects that are more efficient use less time to complete the task than subjects who are less efficient.

The regression between *accuracy* and *time* was not significant (R = .067, Z = 0.302, α = .7842).

> No significant correlation was present between these measures.