Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is an author's version which may differ from the publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/58881

Please be advised that this information was generated on 2019-03-21 and may be subject to change.

The effect of prolonged use on multimodal interaction

Janienke Sturm, Ilse Bakx, Bert Cranen, Jacques Terken, Fusi Wang
Dept. Language and Speech, University of Nijmegen, Nijmegen, The Netherlands
Dept. User-Centered Engineering (Technology Management), TU Eindhoven, Eindhoven, The Netherlands

Abstract: In this paper the effect of prolonged use on interaction with a multimodal system is studied. The system accepts spoken input as well as pointing input and provides output both in speech and in graphics. We measured the usability of the system in a pre-test / post-test design and made a detailed analysis of the changes in interaction styles. The results of the study show that with practise users learn to develop interaction styles that ensure reliable and efficient input. This results in decreased dialogue duration and more user satisfaction.

Key words: Dialogue management, multimodal interaction, speech/pen-based interaction

1. INTRODUCTION

With the emergence of networked handheld devices, it has become possible to provide information services on mobile terminals that were up to now only available on desktop computers. However, interaction styles that are natural and easy to use on a desktop computer may easily become cumbersome on miniaturised devices like palmtops or mobile phones. Whereas typing (possibly via a virtual keyboard) or pointing on a screen in general may feel as a natural way to provide input, with small devices typing and pointing may easily become tiresome, due to the absence of hardware keyboards and inherent limitations on the length of menus. In order to fully exploit the capabilities of these handheld devices, an obvious solution seems to deploy multimodal interfaces. One of the most promising extensions to screen I/O alleviating its shortcomings is speech. It is generally assumed that combining elements of spoken dialogue systems and graphical user inter-

faces will enable users to interact with mobile terminals in a more natural way. It cannot be assumed, however, that such an interaction style is fully intuitive, in the sense that it supports users to interact in the most efficient way right from the beginning. Instead, users may take some time to develop a stable interaction pattern that supports efficient use. In the current paper, we investigate the effect of prolonged use on the development of stable and efficient interaction patterns. For this study we used a multimodal interface for obtaining train timetable information that was developed in the MATIS project (Multimodal Access to Transaction and Information Services). The interface accepts both speech-based and pointing input and provides spoken as well as visual feedback.

The interaction with such a multimodal system can be characterised by means of the scheme in Figure 1 (http://www.w3.org). System acts and user acts occur alternately. Each act may consist of several simultaneous sub-acts in which different modalities are used.

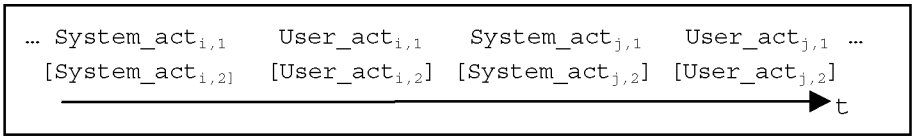

Figure 1 Multimodal interaction

Using this scheme, different types of multimodal interaction can be classified as follows. If User_act$_{i,1}$ and User_act$_{j,1}$ have different modalities, the interaction can be categorised as "sequentially multimodal". Furthermore, if User_act$_{i,2}$ is non-empty, we speak of "simultaneous multimodality". Here, two situations may arise. If User_act$_{i,1}$ and User_act$_{i,2}$ both provide part of a single piece of information, we speak of "coordinated simultaneous multimodality". If User_act$_{i,1}$ and User_act$_{i,2}$ provide distinct pieces of information, we speak of "non-coordinated simultaneous multimodality". Depending on the particulars of the interface, a multimodal interface provides an adaptive interface that enables the user to choose modalities according to his/her preferences and according to the situation.

Sturm et al. (2002) report the results of a usability evaluation of the MATIS interface mentioned above, which was carried out to determine whether providing multiple modalities helps to improve the usability of the system compared to more conventional unimodal systems, such as a spoken dialogue system and a graphical interface (GUI). In this user test only novice users were asked to test the system. This makes sense for public information systems, since those systems should be suitable for use by inexperienced users without training. However, for interfaces running on mobile devices more extensive evaluations are needed. As mobile devices are typically used by the same person for a longer period of time, the user has the opportunity

to gradually develop personal preferences for certain interaction styles while using the interface, so that only monitoring the initial stage of use might lead to inappropriate conclusions. The experiments carried out by Suhm et al. (1999), Petrelli et al. (1997), and Karat et al. (2000) showed that user behaviour indeed changes when users get more experience with a multimodal system. Given these observations, it makes sense to evaluate usability aspects of multimodal interfaces as a function of prolonged use. In the current paper, we therefore focus on the question whether prolonged use of the MATIS system indeed enables users to develop preferred interaction styles, what these interaction styles are in terms of the taxonomy defined earlier, and whether this improves efficiency, effectiveness, and user satisfaction.

In the next section, we will describe and motivate the system's design. In section 3 we describe the user test that has been carried out to study the effect of experience on the interaction. In section 4 we present the results of the test, and in section 5 we will draw conclusions.

2. THE MATIS SYSTEM

We changed an existing unimodal spoken dialogue system for railway information into a multimodal system, by adding a screen and allowing for both spoken and graphical interaction. Detailed information about the system architecture can be found in Sturm et al. (2001). During the spoken dialogue, the screen shows a graphical representation of the form to be filled in (see Figure 2), and gives feedback on the recognition result and on the current system status. Furthermore, after all fields have been filled in, the screen displays the travel advice.

When the user activates the system, the system identifies itself and sets up a spoken dialogue to collect query parameters. This prompting strategy supports novice users in that it guides them through the task[1] (the system supports mixed initiative behaviour). Moreover, biasing users towards the speech mode is in agreement with the observed preference of users for the speech mode (Bilici et al., 2001 and Suhm et al., 1999). Once the system has initiated a spoken dialogue, switching to another modality requires the user to overrule the system and take over the initiative. We assume that novice users feel uncomfortable doing so, especially when they don't know how to

[1] One could also imagine a system in which speech is elicited by pressing buttons instead of by spoken system prompts. In such a tap-and-talk implementation there would be no spoken dialogue at all, which would make the system faster, but it might also be less suitable for novice users. A comparison between the current implementation and a tap-and-talk implementation is planned for the near future.

use the facilities that are offered to interact with a graphical interface. We also expect, however, that after getting accustomed to the interface, users will get more self-confident and use of the graphical facilities more often.

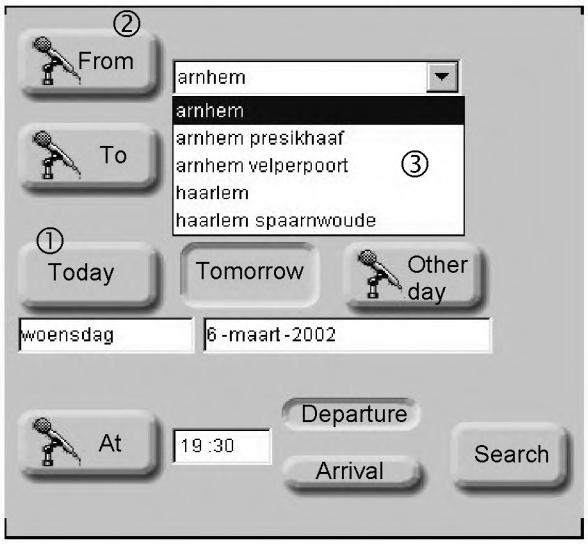

Figure 2 Screen shot of the MATIS interface

The facilities offered to interact in the graphical mode are illustrated in Figure 2. First, the user can press radio buttons (①) to select predefined mutually exclusive values (today/tomorrow or departure/arrival). This facility allows for "sequential multimodality", where a user uses different modalities for subsequent actions, as well as for "non-coordinated simultaneous multimodality", where the user presses a button while also providing an unrelated value in the spoken mode. Second, (s)he can press a microphone button (②) to select a field that (s)he wants to fill by means of speech (e.g. to correct recognition errors - pressing the microphone button will "reset" the field - or simply to speed up the dialogue). In the current implementation, pressing a microphone button for the attribute will trigger a short instruction (e.g. "Say the departure station"), after which the user can enter a value for the field using speech. This allows for "coordinated simultaneous multimodality": the user can exploit two different modalities to specify an attribute-value pair. The input from the two modalities is interpreted by means of late fusion (Kvale, 2001, Oviatt, 2000). Third, in case of a recognition error, users can also select another station name from a drop-down list (③). This would be another form of "sequential multimodality". This options can also be used in an "uncoordinated simultaneous" way, e.g. by selecting an alternative station name while providing unrelated data in the spoken mode. In order to keep the length of the drop-down list limited, it only contains the

recognition alternatives as specified in the N-best list of the speech recogniser, augmented with all alternative stations in the cities that were in the recogniser's N-best list. When the intended station name is not in the drop-down list, the user can clear the field by pressing the microphone button. Finally, speech and pointing gestures may be used simultaneously; for example, while answering a question in the spoken mode, the user can provide a value in the graphical mode by pressing a radio button or by selecting a value from an N-best list. This allows for "non-coordinated simultaneous multimodality".

The spoken output of the system consists of open questions, instructions, and verification questions. Open questions are asked to fill the slots that have no value yet. Instructions (e.g. "say the arrival station") are triggered by the user when (s)he presses a microphone button indicating that (s)he wants to fill a certain field. Verification questions are asked when the value provided by the user has a confidence score that falls below a pre-set threshold. If the confidence score exceeds the threshold, the value is assumed to be correct and no verification question is asked. Values that are provided through the graphical interaction facilities (② and ③) are always assigned maximum confidence; these are never verified in the spoken dialogue. The spoken output of the system can be interrupted by pressing buttons; barge-in using speech is not possible, however.

The screen always shows the current state of the interaction. Visual feedback about the progress of the dialogue is given by showing the values that are extracted from the spoken replies of the user on the screen. In case a verification question is asked due to a low confidence level of the recognition result, the visual feedback and the spoken verification question are synchronised. Once a radio button has been pressed, it remains in that state until the user presses the alternative option.

When all required information has been provided, a query is sent to the information database, which returns a travel advice. The screen gives all the information in tabular form, whereas the spoken dialogue only gives the main information.

3. EXPERIMENTAL DESIGN

3.1 System

Although the MATIS interface has been designed to operate on small devices such as palmtops or mobile phones, in the user tests it was implemented as a Java-applet on a desktop computer with a touch screen and no

keyboard, for practical reasons. The subjects called the system using an ordinary telephone, equipped with a headset, so that they had both hands free.

3.2 Subjects and tasks

Eight subjects (five male and three female, between 14 and 73 years of age, with mixed educational backgrounds) participated in the test. They were paid for participating. Two subjects had no or very little experience with computers. Two subjects were regular train travellers; the others were only occasional travellers (less than twice a year). To get timetable information most subjects use the booklet supplied by the railway company or ask the person at the ticket counter. Only one subject had used the commercial version of a spoken dialogue system for train timetable information before, and only one subject had ever used another spoken dialogue system before.

The experiment was carried out with a pre-test and post-test (within-subjects) design (Table 1). All sessions were conducted in the home lab of the UCE department of TU Eindhoven, which is furnished as a living room.

Table 1 Design of user test

T1	**T2**	**T3**	**T4**	**T5**	**T6**
Pre-test	Practise Min. 30 min	Practise Min. 30 min	Practise Min. 30 min	Practise Min. 30 min	Post-test

By way of introduction to the pre-test, the test leader explained and demonstrated all possible interaction styles by means of an exercise scenario. Following this explanation, the subjects completed six scenarios. The scenarios were presented graphically in order to avoid influencing the manner in which people express themselves (see Figure 3). To ensure that the test would provide information about how users deal with speech recognition errors, each session contained a couple of scenarios with station names that are highly confusable for the automatic speech recogniser.

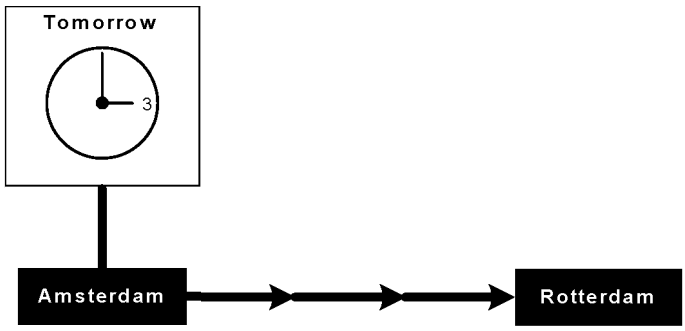

Figure 3 Example of a scenario

After completing the scenarios the subjects completed a questionnaire containing statements concerning different aspects of the system, such as "The combination of speech and graphics is useful" and "The system is slow" (cf. Table 7). The subjects expressed their agreement or disagreement with the statements on a five-point Likert-scale (1 = I strongly disagree, 3 = I agree nor disagree, 5 = I strongly agree).

Once the subjects had done the pre-test, they practised with the system during three or four sessions of at least 30 minutes each (one session a day). During these sessions they either used scenarios offered by the experimenter or they devised their own scenarios. The subjects were asked explicitly to try out all the different interaction facilities offered by the system. After having successfully completed 30 to 40 dialogues, the subjects were asked whether they thought they had developed stable interaction patterns. If so, the post-test was carried out, if not, they practised for another 30 minutes. In the post-test the subjects were asked to carry out the same six scenarios that were used in the pre-test and to complete the same questionnaire, once again.

3.3 Data capture and evaluation metrics

Speech and clicking actions of all dialogues were automatically logged (including time stamps). Additionally, all dialogues were videotaped. Based on this information, detailed analyses were made of the interaction patterns to find out whether prolonged use affects the way people interact with the MATIS system. Also, effectiveness and efficiency were measured for the pre-test as well as for the post-test. Effectiveness was defined in terms of the number of dialogues completed successfully (the dialogue success rate). Efficiency was defined as task completion time (i.e. the time span between the start of the first user answer and the moment at which the query is sent to the information database). User satisfaction was measured using Likert-scales.

4. RESULTS AND DISCUSSION

In total, 48 dialogues were recorded both for the pre-test and the post-test. In section 4.1 we consider the question whether users changed interaction styles as a function of prolonged use. In section 4.2 we consider the question whether this affected the usability of the system in terms of effectiveness, efficiency, and user satisfaction. All analyses in these two sections are based on the successfully completed dialogues only (42 in the pre-test and 45 in the post-test).

4.1 Interaction styles

4.1.1 Speech vs. pointing input

The MATIS interface offers two modalities: speech and pointing. Table 2 shows the average number of actions per dialogue, split up into speech acts and pointing acts. Table 2 shows that, both in the pre-test and the post-test, most of the interaction is done using speech. This is not surprising, as the fields for departure station, arrival station, time, and day (if day is not today or tomorrow) can only be filled using speech.

Table 2 Distribution of speech input and pointing input per dialogue

Modality	Pre-test	Post-test
Speech	6.3 (73%)	4.7 (68%)
Pointing	2.3 (27%)	2.2 (32%)
Total	8.6 (100%)	6.9 (100%)

As can be seen, the total average number of actions per dialogue decreases substantially (from 8.6 to 6.9). This decrease is mostly due to a decrease in speech actions.

As has been shown by others (Karat et al, 2000), users may learn to adjust their speaking style to the capabilities of the system. A change in speaking style may manifest itself in various ways. First, one might expect a decrease of the number of recognition errors. However, the number of misrecognitions (counting substitution errors only) remained constant from pre-test to post-test (cf. Table 4). Another effect of changed speaking styles may be that the confidence level of the recognised words increases, which would mean that less verification questions have to be asked. Also, subjects may have learned to use the mixed initiative capabilities of the system and provide more data in one utterance. More detailed analyses of the data are needed to establish if speaking style really changed and to what extent this can account for the decrease in number of speech acts. This is planned for the near future.

Clearly, not only changes in speaking style, but also a different usage of the multimodal interface during the post-test may account for the decreased number of speech acts, as we will show later. In the next sections we will take a closer look at the data in Table 2 and describe in more detail in which situations user preferences for certain interaction patterns changed and how this may also explain why fewer speech actions were needed to accomplish the same task.

4.1.2 Choice of modality

A first piece of evidence concerning the changed user preferences for particular interaction patterns comes from the use of radio buttons. A number of pre-defined values can be filled in both by speech and by pressing a radio button (today/tomorrow and arrival/departure, cf. Figure 1). Table 3 shows how often subjects used radio buttons rather than speech to provide such a value. The percentages in Table 3 are based on the total number of times this value had to be provided in the successful dialogues: arrival/departure had to be provided in all six scenarios, today/tomorrow occurred in four out of six scenarios.

Table 3 Use of radio buttons

Radio button	Pre-test	Post-test
Today / Tomorrow	16/30 (53.3%)	23/31 (74.2%)
Arrival / Departure	28/42 (66.7%)	35/45 (77.8%)
Overall	44/72 (61.1%)	58/76 (76.3%)

As can be seen, subjects preferred using radio buttons to using speech for those values that could be provided in both ways. In the pre-test 61.1% of the values were provided using radio buttons, increasing to 76.3% in the post-test. A McNemar test for the significance of changes showed that this increase is significant (χ^2=12.07, $p < .01$). The preference for gestures rather than speech may be accounted for both by the reliability of radio buttons compared to speech and by the minimal effort of pressing a radio button (cf. Bilici et al., 2000). This partly explains the decreased number of speech acts in the post-test (cf. Table 2).

A second piece of evidence concerning a change in users' preferences for specific modalities stems from the way they deal with speech recognition errors. When a wrong value is filled in in one of the fields, the interface offers several facilities to correct this error. First, subjects can correct the value by means of a spoken reaction (e.g. "*no not from Amsterdam but from Rotterdam*"). Second, subjects can press the microphone button to clear the field and immediately fill in a new value using speech. Third, if the misrecognition concerns a station name, subjects can choose the correct value from a drop-down list. Table 4 shows the relevant data.

Table 4 Preferred action types for correction of recognition errors

Action type	Pre-test	Post-test
Spoken dialogue	10 (50%)	5 (23%)
Microphone button	10 (50%)	14 (64%)
Drop-down list	0 (0%)	3 (13%)
Total	20 (100%)	22 (100%)

Table 4 shows that, whereas in the pre-test people used the spoken dialogue and the microphone buttons equally often (ten times each) to correct recognition errors, in the post-test the spoken dialogue was continued in only five cases and the microphone buttons were used in fourteen cases. Unfortunately, the N-best list appeared not very useful: the drop-down menu was only used three times in the post-test. This is caused by the fact that often the recogniser did not find any sufficiently likely recognition alternatives, so that no N-best list was available. Also, if the N-best list is available, there is no guarantee that it contains the correct value.

4.1.3 Simultaneous multimodality

Because radio buttons are always active, they can be used not only to provide input in a mode that is more reliable than speech and takes less effort, but also as a way to make the interaction faster. Users may press a button in response to a system prompt asking for that value. Alternatively, they may press a radio button providing that value while at the same time answering an unrelated system prompt ("non-coordinated simultaneous multimodality"). An additional advantage of using the system this way is that it precludes the system from having to ask for this value later on. An important distinction that can be made, then, is one between cases where buttons are pressed as an answer to a system question, and cases where buttons are pressed while the user is doing other things. Table 5 shows the distribution of the use of radio buttons over these two categories.

Table 5 Timing of radio buttons

Moment	Pre-test	Post-test
As answer to system question	21/44 (48%)	14/58 (24%)
During other actions	23/44 (52%)	44/58 (76%)

Table 5 shows that in the pre-test the radio buttons are spread evenly over the two categories, whereas in the post-test 76% of the buttons are pushed while the user is engaged in another action. This indicates that with practice the subjects changed from a sequential multimodal interaction pattern to a simultaneous multimodal interaction pattern, therewith speeding up the dialogue. The arrival/departure button, for example, was pressed often while providing the time, and this would prevent a subsequent question from the system concerning the arrival/departure attribute.

Another way to prevent the system from asking questions or avoiding the need to reply to questions, therewith speeding up the interaction, is by pressing the microphone buttons and the Search button. When a microphone button is pushed, the current system question is suppressed and the focus of

the interaction is directed to corresponding field. As a side effect, verification questions concerning values that were previously filled in are placed on a stack. When all values have been filled in and the Search button is pressed, the system will start querying the information database. So, by pressing the Search button before all fields have been verified, the user implicitly answers all verification questions remaining on the stack, thereby speeding up the interaction even more. The Search button was used equally often in the pre-test and in the post-test (14 times), and the percentage of verification questions that was skipped by doing so was equal in pre-test and post-test as well.

4.1.4 Inter-subject variation

There were clear differences between the interaction patterns of different users. Two users who used only speech in the pre-test, still preferred to be guided by the spoken system questions in the post-test. They left the initiative completely to the computer and took over the initiative only once or twice to correct a speech recognition error by pressing the microphone button. This interaction pattern strongly contrasts with that of the other subjects, who preferred to keep the initiative themselves and use the system more in a tap-and-talk manner, pressing buttons to fill in values and skip verification questions as much as possible.

4.2 Usability

4.2.1 Effectiveness

The effectiveness of the interface is defined in terms of the number of successfully completed dialogues. Both in the pre-test and the post-test, 48 dialogues were recorded.

The overall effectiveness, measured over all scenarios, increased from 87.5% in the pre-test to 93.8% in the post-test. In the pre-test 6 dialogues failed, whereas in the post-test only 3 dialogues failed. Three of the six failures in the pre-test were caused by the fact that the subjects did not notice that wrong values were filled in. In all other cases the subject ended the dialogue because of persistent recognition errors. The success rate increased most for the most difficult scenario (scenario 6), with 3 failures in the pre-test and 1 failure in the post-test. Since the success rate was already very high in the pre-test, this leaves room for a small improvement only. We therefore refrain from evaluating this difference statistically, but only note that the difference is in the right direction.

4.2.2 Efficiency

Results on the efficiency of the dialogues are shown in Table 6. For each scenario the mean duration of a dialogue is shown in seconds measured from the start of the first user utterance to the query to the information database. The scenarios are presented in order of increasing difficulty.

Table 6 Average dialogue duration (in seconds)

Scenario	Average dialogue duration	
	Pre	Post
1	66.8	37.5
2	52.1	30.6
3	69.8	50.1
4	65.5	35.9
5	134.5	45.5
6	115.2	83.8
Mean	79.8	46.5

Table 6 shows that on average dialogues are completed faster in the post-test than in the pre-test: the mean duration decreased from 79.8 seconds in the pre-test to 46.5 seconds in the post-test, with reductions ranging from 20 seconds to 89 seconds. Two analyses of variance were conducted, one with Pre-Post and Dialogues as main factors, one with Pre-Post and Subjects as main factors. The effect of Pre-Post was significant in both analyses ($F_{1,7}$=16.76, p=.005 and $F_{1,5}$=18.29, p=.005, respectively). No interactions were significant. From this we conclude that the effect of difference between pre-test and post-test is robust, and that the reduction in average duration from pre-test to post-test is not significantly different for different subjects (across dialogues) or different dialogues (across subjects). As can be seen, in the pre-test the two most difficult scenarios resulted in durations that are substantially longer than those for the other four scenarios. In the post-test, scenario 6 still has the longest duration, but the duration of scenario 5 has decreased with 89 seconds to a level that is below the average duration. As this is a scenario where many speech recognition errors were made, obviously subjects succeeded in dealing with these errors more efficiently in the post-test. Both in the pre-test and in the post-test scenarios 3 and 6 yield relatively long dialogues, which can be explained by the fact that in these scenarios people had to ask for a day other than today or tomorrow, so that radio buttons could not be used to provide the value for "Day".

4.2.3 User satisfaction

Table 7 shows a summary of the answers to the Likert-scale statements. For the negative statements 4 ("The system was slow") and 11 ("I was distracted by the display") the scores have been inverted, so that high scores denote the positive end of the scale.

Table 7 Answers to questionnaire (1 = disagree, 3 = agree nor disagree, 5 = agree)

Statement	Rating	
	Pre	**Post**
1. The system was easy to use	3.9	4.6
2. I always understood what was expected from me	4.5	4.8
3. Correcting errors was easy	2.9	3.4
4. The system was *not* slow	1.6	1.5
5. Speech and graphics were well tuned to one another regarding the contents	3.9	4
6. Speech and graphics were well tuned to one another regarding the timing	3.3	4.3
7. I liked being able to use speech as well as the touch screen	4	4.6
8. The system reacted adequately to the combined input	3.9	4
9. The length of the spoken utterances was good	4.3	4
10. Visualising the fill-in form was useful	4.5	4.8
11. I was *not* distracted by the display	2.7	3.2
12. Visualising the travel advice was useful	4.9	5
13. Giving the travel advice in spoken form was useful	3.4	3.6
14. After a while I started using the system differently	2.9	3.8
15. I used the touch screen more often as I got more experienced	3.5	4.0

Statements 1 through 13 in Table 7 are related to usability aspects. For 11 out of 13 questions the average score in the post-test is higher than that in the pre-test. In a sign test a score of 11 out of 13 in the right direction is significant ($z = 2.63; p < .01$). Thus, we conclude that the usability is judged higher in the post-test than in the pre-test.

Substantial improvements (i.e. an improvement of 0.5 or more) are observed for those aspects where we expected training to be of influence. In the post-test, subjects found the system easier to use (s1) and they were more satisfied about being able to use both speech and graphics (s7). Furthermore, during the post-test subjects judged speech and graphics to be better tuned to one another in time (s6) than during the pre-test (although objectively the timing was the same), and they felt they were less distracted by the display (s11). Apparently, while practising with the system, the subjects developed a mental model of the system from which they could understand and anticipate

the system's behaviour. A number of aspects were already rated very highly in the pre-test, such as the visualisation aspects (s10 and s12) and the transparency of the interface (s2). For these statements no substantial changes were observed or expected. Other aspects of the system were rated less favourably both in the pre-test and the post-test and need to be improved. Although subjects considered correcting errors to be easier in the post-test than in the pre-test (s3), this aspect is judged to be poor in general. Obviously, the lack of possibilities to switch to another modality for specifying values that are poorly recognised is considered a major flaw. Furthermore, the speed is judged to be poor (s4), and the spoken travel advice is not considered a useful addition to the visually displayed advice (s13). Finally, although people indicate that they appreciate the visualisation of the fill in form, they somewhat surprisingly indicate that they feel distracted by the display (s11), even if this is less so in the post-test than in the pre-test. At this moment it is not quite clear how to interpret this result. More extended interviews are needed to clear up this issue.

Statements specifically dealing with changes in interaction styles (s14 and s15) were rated substantially higher in the post-test than in the pre-test, indicating that subjects themselves perceived an effect of prolonged use.

5. CONCLUSIONS

The experiments carried out show that prolonged use of the interface indeed enables users to exploit the opportunities offered by the system and interact with the system in a more efficient way. From detailed analyses of the results the major factor appears to be a shift in the direction of non-coordinated simultaneous multimodality, with the effect that the system is precluded from prompting for information later on and asking verification questions. As a result, a substantial decrease was observed in the number of speech acts in the post-test, the net effect being a gain in efficiency: the successfully completed dialogues in the post-test were completed in almost half of the time that was needed in the pre-test (Table 6). Another factor that may account for the decrease in the number of speech acts and the associated gain in efficiency is a change in speaking style. Up till now we only investigated a few of the different ways in which a change of speaking style can manifest itself, and these could not be shown to have an effect. However, until all details have been analysed we are not able to establish the size of the effect of speaking style on the efficiency. The gain in efficiency is associated with an increase in perceived usability as evident from the subjective ratings and a subjectively perceived increase of multimodal use (Table 7).

We observed interesting differences between subjects. In the pre-test all subjects were clearly looking for the easiest way to interact with the system. After about two hours of training all subjects had developed stable interaction patterns to which they would stick as much as possible. However, this pattern was not the same for different subjects. Whereas most subjects preferred to keep the initiative thus striving for more efficiency, two subjects preferred to maintain their initial behaviour and to be guided by the system in the spoken dialogue. (Surprisingly, these were not the ones with the least computer experience). We conclude that, even though the interface design can be improved in several ways, it has been successful in that it accommodates different types of users, enabling them to sort out their preferred interaction pattern. For each subject, the preferred interaction pattern seems to be the result of the perceived optimal balance between the effort (s)he has to put in the interaction and the efficiency with which the interaction takes place.

6. ACKNOWLEDGMENT

The MATIS project is funded by the Dutch Ministry of Economic Affairs through the Innovation Oriented Programme Man-Machine Interaction.

7. REFERENCES

V. Bilici, E. Krahmer, S. Te Riele, R. Veldhuis (2000), "Preferred modalities in dialogue systems". *Proceedings of ICSLP2000.*

J. Karat, D. Horn, C. Halverson, C. Karat (2000), "Overcoming Unusability: developing efficient strategies in speech recognition systems". *Proceedings of CHI-2000.*

S. Oviatt, A. De Angeli, K. Kuhn (1997), "Integration and synchronization of input modes during multi-modal human-computer interaction". *Proceedings of CHI-97.*

S. Oviatt (2000), "Taming recognition errors with a multimodal interface". *Communications of the ACM, vol. 43, no. 9, 45-51.*

D. Petrelli, A. De Angeli, W. Gerbino, G. Cassano (1997), "Referring in Multimodal Systems, the importance of user expertise and system features". *Proceedings of ACL-EACL'97.*

J. Sturm, B. Cranen, F. Wang (2001), "Adding extra input/output modalities to a spoken dialogue system". *Proceedings of Eurospeech2001.*

J. Sturm, I. Bakx, B. Cranen, J. Terken, F. Wang (2002), "Usability evaluation of a Dutch multimodal system for railway information". *Accepted for LREC2002.*

B. Suhm, B. Myers, A. Waibel (1999), "Model-based and empirical evaluation of multimodal interactive error correction". *Proceedings of CHI-99.*