Usability Evaluation of a Dutch Multimodal System for Train Timetable Information

Janienke Sturm*, Ilse Bakx†, Bert Cranen*,

Jacques Terken†, Fusi Wang*

* Department of Language and Speech
Nijmegen, The Netherlands
{Janienke.Sturm,F.Wang,B.Cranen}@let.kun.nl

†Department User-Centered Engineering (Technology Management), TU Eindhoven
Eindhoven, The Netherlands
{I.H.M.Bakx,J.M.B.Terken}@tue.nl

Abstract

In the MATIS project a multimodal system has been developed for train timetable information. The aim of the project was to obtain guidelines for designing multimodal interfaces for information systems. The MATIS system accepts input both in spoken and in graphical mode (no keyboard input) and provides feedback in the same two modes. The user can choose at any time which of the input modalities (s)he prefers to use for a certain action. A user test was carried out in which 25 subjects were asked to evaluate the system. For comparison, users were also asked to test a GUI (Graphical User Interface) version of the train timetable information system as well as a speech-only version of the system. We measured the efficiency and the effectiveness of the interaction and the user satisfaction with all three systems.

1. Introduction

Automatic information services can be implemented in several ways, e.g. as purely graphical user interfaces (GUI) (fill-in forms like the Yellow Pages on the Internet) or as speech-only interfaces (people call and get information via a spoken dialogue). Both types of interfaces have advantages and disadvantages. With respect to transparency, graphical interfaces are clearly superior to speech interfaces. With a graphical interface (if designed properly) the user has few problems in knowing which information needs to be filled in and which information the system already has. Spoken dialogue systems are far less transparent to the user, due to the lack of visual support. With respect to the naturalness and the efficiency of the interaction it has been claimed that speech interfaces are superior to graphical user interfaces (see http://www.bell-labs.com/project/ConC/demo.html and http://www.sls.lcs.mit.edu/ec-nsf/mit-sls.html): pointing at a screen and typing on a (virtual) keyboard may be tiresome, especially on small devices like palmtops or mobile phones, whereas a purely spoken dialogue is a natural and efficient means to convey information (provided that the speech recogniser does not make too many mistakes). Usability research has shown that users' preference for one or the other modality depends strongly on the situation. In fact, users prefer to be able to choose the mode that fits their goals best (Oviatt et al., 2000a). A multimodal system, combining the strengths of both GUI's and spoken interaction could solve the usability problems of both types of interfaces.

The research that is done within the MATIS project (Multimodal Access to Transaction and Information Services) is aimed at finding ways to combine speech and graphical interaction in such a way that the usability is maximised. The research described in this paper is restricted to form-filling interfaces. A prototype system for train timetable information has been developed, that accepts input both in the form of speech and from a graphical interface and provides both spoken and graphical feedback. This paper reports the results of a user test that has been carried out to determine whether providing multiple modalities helps to improve the usability of the system compared to unimodal systems: a spoken dialogue system and a graphical interface. We report on the usability of three systems (speech-only, GUI-only, and multimodal) measured in terms of effectiveness, efficiency, and user satisfaction, and describe to what extent users notice and use the extra interaction facilities that are available in the multimodal system.

We start out with describing the multimodal prototype in Section 2 and the experimental set-up in Section 3. Section 4 contains the actual results of the user test, and in Section 5, the results are discussed and conclusions are drawn.

2. The MATIS system

When maintaining the form-filling metaphor, forms are best represented as visual objects. Therefore, we conceive a multimodal form-filling interface as an enriched GUI where part of or the entire graphical interaction can also be accomplished using speech.

Speech can be elicited in various ways, however. First, one could construct a tap-and-talk interface where people select a field they want to fill by means of a pointing action and then provide a value for that field using speech. Note that such a solution does not require the system to generate any speech. Although such an interface appears to support efficient interaction to experienced users, the interaction facilities created may not be intuitive to a novice user, because it combines concepts from different domains: pointing actions from the event-driven GUI domain and spoken messages from the domain of spoken dialogue systems. Inexperienced users may be helped by a second type of interface, in which a spoken dialogue guides them through the task, while providing support by means of screen input and output. This type of system is

more adaptive in the sense that it allows the user to use combinations of gestures and speech, but it does not force to do so. If desired, the user can complete the dialogue using speech only.

In the MATIS project, we chose to explore this second option in further detail. Unpublished data from preliminary user tests showed that the need for graphical input is limited as long as the speech is recognised correctly; subjects used point-and-click input to correct speech recognition errors, but tended to return to speech as soon as these had been solved (see also Bilici et al., 2000). However, one might expect this behaviour to change when people get more experienced in using the system or in using graphical interfaces in general: it is expected that these people are more inclined to combine speech with graphical interaction.

In an attempt to better serve both experienced and novice users, a prototype system was constructed in the following way. A visual component was added to an existing unimodal spoken dialogue system for timetable information on Dutch railway connections (described in more detail in Sturm et al. (2001)). While preserving the spoken dialogue, the system also provides visual feedback about the recognition result, thus giving information on the status and beliefs of the system. The system also allows the user to give graphical input in the form of clicking buttons or selecting from N-best lists. No keyboard in any form is available to the user.

Four human factors experts carried out a heuristic evaluation of this interface in order to identify possible usability problems. The original prototype has been adjusted on the basis of the comments of the experts. The resulting interface is depicted in Figure 1.

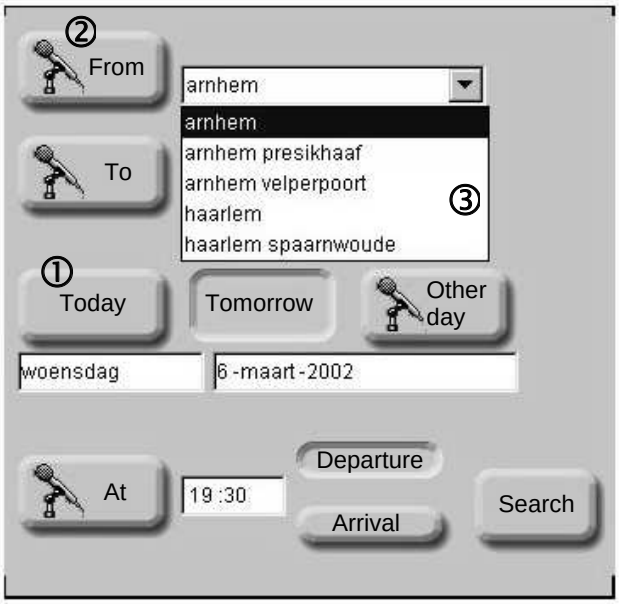

Figure 1 Screen shot of the MATIS interface

When a user calls the system using an ordinary telephone, the system takes the initiative by starting a mixed initiative spoken dialogue, prompting the user to provide values for the fields in the form shown in Figure 1. When the user responds to the system's prompts for information using speech, the system will remain prompting for further information until all fields have been filled. In this way, the interaction can be completed fully through a spoken dialogue. However, the user can influence the course of the interaction through actions in the graphical domain at all times, in several ways. First, (s)he can press radio buttons (①) to select predefined values (today/tomorrow or departure/arrival). 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 or simply to speed up the dialogue). A question is then asked that is triggered by the button that has been pushed (e.g. "Say the departure station"), after which the user can enter a value for the field using speech. Third, in case of a recognition error, (s)he can also select another station name from a drop-down list (③). In order to keep the length of the list limited, it only contains the recognition alternatives as specified in the N-best list of the recogniser, augmented with all alternative stations in the cities that were in the recogniser's N-best list. When the correct station name is not in the drop-down list, the user can clear the field by pressing the microphone button.

Once a value for a field has been filled in using speech[1], the user is free to switch to the graphical mode if that would be more suitable given the situation, e.g. in the case of speech recognition errors. The two input modes may be used simultaneously; for example, while answering a question in the spoken domain, the user can provide a value in the graphical domain by pressing a radio button or by selecting a value from an N-best list. Simultaneous input from the two modalities is interpreted by means of late fusion (Kvale, 2001, Oviatt, 2000b).

The spoken output of the system consists of open questions and verification questions. Open questions are asked during the normal slot-filling dialogue flow to fill the slots that have no value yet, or in reaction to a user pressing a microphone button to indicate 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 the threshold. Values that are provided using the graphical interaction facilities (① and ③) are always assigned maximum confidence; these are never verified in the spoken dialogue. Furthermore, the spoken dialogue gives only a summary of the travel advice; the complete travel advice is shown on the screen.

There is no other coordination of output modes than synchronization of spoken and visual output in case a verification question must be asked due to a low confidence level of the recognition result. The spoken output of the system can be interrupted only by pressing buttons; barge-in using speech is not possible.

3. Experimental set-up

3.1. Systems

In order to assess the presumed benefit of combining multiple input and output modalities, we compared the performance of the multimodal system with the performance of two unimodal train time table information services: a graphical user interface accessible via the Internet - the NS-Reisplanner (GUI), and a purely spoken dialogue system, accessible via the telephone (Speech-only). This spoken dialogue system is essentially the same

[1] Note that the values "today / tomorrow" and "departure / arrival" can be filled in without using speech, whereas for filling in all other values speech is required.

as the MATIS system without the screen. However, in the spoken dialogue system each user answer is verified, regardless of the confidence level of the recognised utterance. Furthermore, in the speech-only system, the complete travel advice is given in spoken form.

Although the MATIS system has been designed to operate on small devices such as palmtops or mobile phones, in the present experiment, for practical reasons, the system was implemented on a desktop computer without a keyboard, but with a touch screen that displayed the fill-in form. To start using the system subjects had to call in using an ordinary telephone equipped with a headset. The Speech-only system was operated by an ordinary telephone. For the GUI system, a normal desktop computer was used with a keyboard and a normal screen on which the GUI fill-in form was shown.

3.2. Subjects and tasks

Twenty-five subjects (fifteen male and ten female, between 19 and 28 years of age) took part in the test. The subjects were all students who travel by train regularly. They mostly use the Internet to get timetable information; only a few subjects had ever used a commercial spoken dialogue service providing timetable information, or any other spoken dialogue system. All subjects were experienced users of computers.

All subjects tested all three systems. Different groups of subjects used the systems in different orders, to avoid confounding effects of order of presentation (see Table 1).

	Series 1	**Series 2**	**Series 3**
Group 1	MATIS	Speech-only	GUI
Group 2	Speech-only	GUI	MATIS
Group 3	GUI	MATIS	Speech-only

Table 1 Experimental design

After a short introduction, subjects were asked to complete three scenarios with each system. The scenarios were presented graphically in order to avoid influencing the manner in which subjects express themselves (see Figure 2). Different scenarios were created for each system, in order to circumvent any learning effect. Furthermore, to ensure that the test would provide information about how users deal with speech recognition errors, each series of scenarios contained at least two station names that are highly confusable for the automatic speech recogniser.

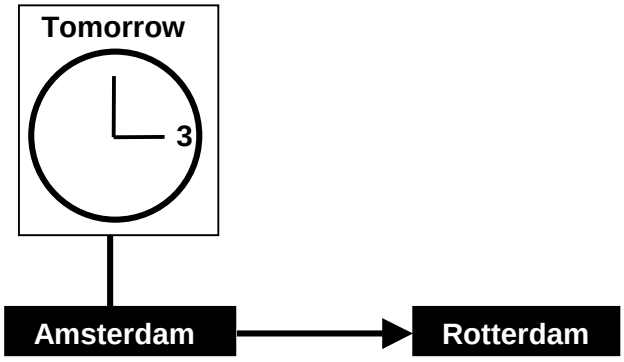

Figure 2 Example of a test scenario

Subjects started each test series with an exercise scenario allowing them to get used to the system. Assistance was given after the completion of this scenario, but only if the user had not been able to complete the exercise. In case of the MATIS system, if the subjects did not use any of the graphical interaction possibilities in the exercise scenario, the test leader would show how the display could be used, without explicitly encouraging the subject to use the display.

After each series of dialogues with one of the systems, subjects completed a questionnaire in which they expressed their agreement or disagreement with statements on a five point Likert-scale (1 = "I strongly disagree", 3 = "I agree nor disagree", 5 = "I strongly agree"). Five statements were the same for each of the systems:

G1. I consider the system easy to use
G2. I always understood what was expected from me
G3. I found it easy to correct errors
G4. I thought the system was slow
G5. I thought the travel advice was clear

Ten additional statements specifically concerned the MATIS system:

M1. Speech and graphics were well tuned to one another regarding the contents
M2. Speech and graphics were well tuned to one another regarding the timing
M3. The length of the spoken utterances was appropriate
M4. The combination of speech and graphics was useful
M5. I was distracted by the display
M6. Visualising the travel advice was useful
M7. Visualising the filling form was useful
M8. I liked using speech besides the touch screen
M9. I used the touch screen more often as I got more experienced
M10. The system reacted adequately to the combined input

Finally, once all three systems had been tested, subjects gave preference judgments by rank ordering the systems on a number of aspects, such as "Which system did you consider easiest to use?" and "Which system would you prefer to use in the future?".

3.3. Data capture and evaluation metrics

Clicking actions and utterances of all dialogues with the multimodal system and the speech-only system were automatically logged (including time stamps). Additionally, all dialogues were videotaped. As clicking and typing actions with the GUI system could not be logged automatically, the data for this system were obtained from the videotapes.

The usability of the systems is measured in terms of effectiveness, efficiency, and user satisfaction. The effectiveness of the systems is measured as the number of successfully completed dialogues (the dialogue success rate). The efficiency of each of the systems is measured as task completion time (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 is measured through Likert-type scales and user preferences concerning relevant aspects of the systems.

4. Results

4.1. Effectiveness and efficiency

The effectiveness of the systems is shown in Table 2 as the number of successfully completed dialogues per scenario per group. The first column (*Success*) contains the number of successfully completed dialogues (dialogues where the user got the correct travel advice). The second column (*Wrong data*) contains the number of dialogues where users got the wrong travel advice, e.g. because (s)he provided input that differed from the instructions in the scenario. The third column (*Failed*) shows the number of dialogues where no travel advice was given at all, because the subject ended the dialogue, e.g. because of persistent recognition errors. The fourth column shows the total number of dialogues, and the fifth column (*Success rate*) contains the percentage of successfully completed dialogues (dialogues where the user got the wrong travel advice have been omitted when computing the success rate, because it is not immediately clear what caused the wrong travel advice, inaccurate reading of the scenario or lack of attention to the feedback of the system).

		Success	**Wrong data**	**Failure**	**Total**	**Success rate**
MATIS	1	24	0	0	24	100%
	2	24	0	0	24	100%
	3	21	5	3	29	88%
GUI	1	25	0	0	25	100%
	2	24	0	0	24	100%
	3	24	0	0	24	100%
Speech-only	1	23	0	0	23	100%
	2	19	0	6	25	76%
	3	16	1	15	32	50%

Table 2 Dialogue success rate per scenario

In total 273 dialogues were recorded. 43 Of these dialogues are not included in Table 2 and have been omitted from all further calculations. It concerns dialogues that ended prematurely, due to technical problems: a bug in the system caused the system to hang up when the user started to press buttons before the welcome message had ended or when the user kept silent after the first system utterance. The bugs were solved in the course of the experiment. After encountering such an error, most subjects called again to redo the scenario. However, some subjects were confronted with the same error the second time as well, after which they gave up. As a consequence, some subjects did not succeed to complete the scenario at all (this happened 10 times); this explains why the total number of dialogues in each row is sometimes less than the total number of subjects (= 25). Two dialogues are missing for the GUI system as well, due to the fact that the computer that was used crashed.

Furthermore, a number of subjects did not succeed to complete a scenario the first time (e.g. because of recognition errors) and hung up. Then they called the system again; this explains that the total number of dialogues (Column 5) may be larger than 25. Using the speech-only system, 5 subjects called in twice to complete Scenario 2, whereas 9 subjects called in twice to complete Scenario 3. Using the MATIS system, 8 subjects called in twice to complete Scenario 3. All other tasks were completed in a single dialogue.

The failures for the second and third scenarios using the speech-only system and the MATIS system were mainly caused by the fact that users were unable to correct misrecognised station names, as these scenarios contained the most confusable station names. Although the station names used in the MATIS scenarios were just as difficult to recognise, these caused less dialogue failures. It must be noted in this respect, that the system would never end the dialogue by itself; in the failed dialogues, it was always the user who decided to hang up the phone.

Table 3 shows for each system the mean duration of a dialogue per group. The figures are based on successfully completed dialogues only and have been averaged over the three scenarios.

	Duration (in seconds)			
	Group 1	Group 2	Group 3	*Mean*
MATIS	61	65	59	*62*
GUI	42	43	48	*48*
Speech-only	72	76	68	*73*
Mean	*59*	*65*	*61*	

Table 3 Task completion time per group

As can be seen from Table 3, dialogues were shortest using the GUI system. A two-way analysis of variance was conducted to evaluate the significance of the differences between means, with Groups as a between-subjects factor and Systems as a within-subjects factor, collapsing across the three scenarios per subject per system (means were not weighted, as it has been assessed in advance that there was no correlation between the mean per subject per system and the number of failed dialogues per subject per system). There was a main effect of systems ($F(2,44) = 15.72$, $p < .01$); the effect of Groups and the GroupsxSystems interaction were not significant. Posthoc comparisons (pairwise t-tests with Bonferoni correction of significance levels) showed that mean durations for all three systems were significantly different. That is, interaction with the GUI system is significantly faster than interaction with the MATIS system ($t(24) = 3.29$; $p < .02$), and the dialogues with the MATIS system are significantly shorter than those with the speech-only system ($t(24) = 2.77$; $p < .02$). A detailed analysis of interaction styles with the latter two systems is required to determine whether this difference in task completion time is caused by the additional interaction facilities in the MATIS system or simply by the fact that each utterance is verified in the speech-only system, whereas in the MATIS system only utterances with low confidence are verified. This analysis is planned for the near future.

Table 3 does not show separate figures for each of the three scenarios. Inspection of the data showed that difficult scenarios (i.e. those with station names that are hard to recognise) resulted in longer dialogues both in the MATIS system and in the speech-only system. The general advantage of the MATIS system over the speech-only system also holds for the difficult dialogues: these dialogues were completed faster in the MATIS system than in the speech-only system. This indicates that solving

speech recognition errors can be accomplished more efficiently in the MATIS system than using speech only. Again, further analysis of interaction styles is needed, because the difference may again be caused by the fact that less verification questions are asked by the MATIS system.

4.2. User satisfaction and preferences

Table 4 shows a summary of the answers to the five Likert-scale statements that concerned all three systems. (Scores for the negative statement (G4) have been inverted so that high values denote the positive end of the scale).

Statement	Rating		
	MATIS	GUI	Speech-only
G1. System is easy to use	3.1	4.8	3.2
G2. Clear what is expected	4.0	4.9	4.0
G3. Easy to correct errors	3.2	4.8	3.2
G4. System is not slow	1.5	3.8	1.3
G5. Travel advice is clear	4.2	4.7	2.9

Table 4 Results of the Likert-scales for the general statements (1 = "disagree", 5 = "agree")

Table 4 shows that users rated the GUI system substantially higher than the speech-only system and the MATIS system in all respects. A two-way analysis of variance per question was conducted to evaluate the significance of the differences between the ratings, with Groups as a between-subjects factor and Systems as a within-subjects factor. There was a main effect of systems for all questions. Pairwise t-tests with Bonferroni correction of significance level showed that the ratings for GUI were significantly higher on the first four statements and that ratings for MATIS and Speech-only grouped together. For the final statement ("I thought the travel advice was clear") both the MATIS system and the GUI were rated significantly higher than the speech-only system ($F(2,42) = 20.30$; $p < .01$): people appreciated the textual version of the travel advice more than the spoken version alone. Although objectively it appears that recovering from speech recognition errors is easier in the MATIS system than in the speech-only system (Table 2), this was not appreciated in the subjective ratings: the scores for the MATIS and the speech-only system for statement G3 were the same. People were not positive about the speed of the dialogue in the MATIS and the speech-only system (G4): these systems were considered slow.

Table 5 shows the results of the Likert-scale statements for the statements that concern the MATIS system only. (Again, scores for the negative statement (M5) were inverted so that high values denote the positive end of the scale).

The data in Table 5 show that the MATIS system is appreciated primarily for its visualisation features: visualising the travel advice and the filling form (M6 and M7) is considered very useful. Users have less pronounced opinions about the opportunities for multimodal interaction (M4, M5, and M8). The design is judged moderately positive (M1, M3, and M10), although the time synchronisation of speech and graphics (M2) was rated relatively low, possibly due to unexpected delays that occurred in the spoken system output.

Statement	Rating
M1. Speech and graphics were well tuned to one another regarding the contents	4.0
M2. Speech and graphics were well tuned to one another regarding the timing	3.2
M3. The length of the spoken utterances was appropriate	4.2
M4. The combination of speech and graphics was useful	3.2
M5. I was not distracted by the display	3.3
M6. Visualising the travel advice was useful	4.8
M7. Visualising the filling form was useful	4.4
M8. I liked using speech besides the touch screen	3.9
M9. I used the touch screen more often as I got more experienced	3.7
M10. The system reacted adequately to the combined input	3.5

Table 5 Results of the Likert-scales for the statements concerning the MATIS system (1 = "disagree", 5 = "agree")

The Likert-scales scores for the MATIS-specific statements as well as the general statements showed no significant difference between the three groups, i.e. no effects of the order of presentation of the three systems were observed.

Question	Rating		
	MATIS	GUI	Speech-only
Which system did you consider easiest to use?	2.1	1.2	2.7
With which system did you know best which information you had to provide?	2	1.2	2.5
With which system was correcting errors easiest?	1.8	1.2	2.8
Which system did you consider the most fun to use?	1.5	2	2.5
With which system was understanding the travel advice easiest?	1.8	1.2	2.9
In two of the systems the travel advice is shown on the screen, in the other it is not. Which one do you prefer?	1.9	1.1	2.8
Which system would you prefer to use in the future?	2.2	1.1	2.7

Table 6 Preference judgments (1 = "most preferred", 3 = "least preferred")

As explained in Section 3.2, after having used all systems, subjects rank ordered the systems as to their preference on a number of aspects, assigning 1 to the most preferred system and 3 to the least preferred system. Table 6 shows the average preference judgments across participants for all questions (1 = highest preference).

As can be seen, average preferences are close to 1 ("most preferred") for GUI, close to 3 ("least preferred") for Speech-only, and MATIS is in the middle, for most questions. Thus, the preference judgments support the conclusion that the subjects liked the GUI system best. The only aspect on which MATIS outperformed the GUI was the question concerning fun: people thought the MATIS system was more fun to use than the GUI and speech-only system. Furthermore, it may be noted that, although the Likert-scales did not show clear differences between the scores for the MATIS and the speech-only system for the first four statements, Table 6 shows that, if users are forced to choose, the MATIS system is preferred to the speech-only system on all aspects.

5. Discussion and conclusions

The results of this experiment show that a multimodal system can solve a number of problems with speech-only interfaces. More in particular, Tables 2 and 3 show that the MATIS system is more effective and more efficient than the speech-only version of the system. Furthermore, people were satisfied about the system and indicated that they prefer the multimodal system to the speech-only version because of its visualisation features. However, users clearly preferred the GUI version of the application both to the multimodal and the speech-only system, as this is fastest and least error-prone.

Whereas a keyboard provides an efficient means to provide error-free input, in the MATIS system there is no way around using error-prone speech to accomplish the task. The fact that several dialogues could not be completed successfully (both in the MATIS system and in the speech-only system) indicates that the performance of the speech recogniser is too poor to use speech as the only input modality. Providing a (virtual) keyboard or a well-designed graphical menu as a fallback option to deal with speech recognition errors seems necessary and would probably make the interaction in the MATIS system more effective.

A further observation is that the users in the test population have lots of experience interacting with GUI's. Since therefore our group of subjects cannot be considered real novice users with respect to all aspects of the multimodal interface, the results of this study may be slightly biased. The spoken dialogue may be more suited for people who have no experience in using computers in general, as it guides them through the dialogue without forcing them to touch any buttons. In an attempt to create a self-explanatory device that remains close to a spoken dialogue system, the system was designed so that it produces speech-prompts whenever a user presses a button. However, for users with experience in using graphical interfaces or computers in general - i.e. people who are not afraid to touch buttons - the speech prompts may be annoying, as they tend to slow down the interaction. As a consequence, the speech output features of the multimodal system would probably have been evaluated differently had the subject group been more heterogeneous. An interface where there is no spoken dialogue at all - the tap-and-talk implementation of the interface - may be better suited for the more experienced user. To investigate to what extent a tap-and-talk interface would be disadvantageous for a novice user, a comparison of the current implementation and a tap-and-talk implementation will be made in a separate user test.

Finally, we need to consider the possibility that the amount of experience a user has with a specific interface is important as well. User studies show that with experience users learn to better integrate speech with other modalities and to switch to the most effective modality (Karat et al., 2000). Because of this learning effect the usability of the interface must be interpreted in terms of the amount of experience that people have in using the interface. In the current experiment, this type of experience has been of minor influence, because subjects only had to carry out three scenarios with each of the systems, which is probably not enough to see any effect of experience. Despite this, subjects already indicated that they agreed with the statement "I used the touch screen more often as I got more experienced". The effect of experience on the usability of the interface will be studied in a separate user test described in Sturm et al. (2002).

It must be noted that the results from this test apply primarily to form-filling interfaces and as such do not necessarily generalise to other domains. In form-filling applications, there is a way around multimodal interaction, and it may very well be that when people are offered a choice between different modalities, they will stick to what they are used to, because that ensures relative effectiveness and efficiency. In other types of applications the need for multimodal interaction may be more evident.

6. Acknowledgment

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

7. References

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

Karat, J., Horn, D., Halverson, C., Karat, C. (2000), Overcoming unusability: developing efficient strategies in speech recognition systems. *Proceedings CHI-2000*, The Hague, The Netherlands.

Kvale, K., Narada, D.W., Knudsen, J.E. (2001), Speech-centric multimodal interaction with small mobile terminals. *Proceedings NORSIG-2001*, Trondheim, Norway.

Oviatt, S.L., Cohen, P.R., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J. & Ferro, D. (2000a), Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions. *Human Computer Interaction*, 2000, vol. 15, no. 4, 263-322.

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

Sturm, J., Wang, F., Cranen, B. (2001), Adding extra input/output modalities to a spoken dialogue system. *Proceedings 2nd SIGdial Workshop on Discourse and Dialogue*, Aalborg, Denmark.

Sturm, J., Bakx, I., Cranen, B., Terken, J., Wang, F. (2002), The effect of experience on interaction with multimodal systems. *Accepted for IDS02*, Kloster Irsee, Germany.