

# Comparing the Usability of a User Driven and a Mixed Initiative Multimodal Dialogue System for Train Timetable Information

Janienke Sturm<sup>1</sup>, Ilse Bakx<sup>2</sup>, Bert Cranen<sup>1</sup>, Jacques Terken<sup>2</sup>

<sup>1</sup>Department of Language & Speech  
University of Nijmegen, The Netherlands

<sup>2</sup>Department of User Centered Engineering  
University of Eindhoven, The Netherlands

{Janienke.Sturm,B.Cranen}@let.kun.nl, {I.H.M.Bakx,J.M.B.Terken}@tue.nl

## Abstract

The aim of the study presented in this paper was to compare the usability of a user driven and a mixed initiative user interface of a multimodal system for train timetable information. The evaluation shows that the effectiveness of the two interfaces does not differ significantly. However, as a result of the absence of spoken prompts and the obligatory use of buttons to provide values, the efficiency of the user driven interface is much higher than the efficiency of the mixed initiative interface. Although the user satisfaction was not significantly higher for the user driven interface, by far most people preferred the user driven interface to the mixed initiative interface.

## 1. Introduction

The aim of the study described in this paper is to compare the usability of two multimodal train timetable information systems. Both systems present themselves to the user by means of the same graphical interface, but differ in the way speech is used to provide input to the system. One system can be considered user driven whereas the other is in fact a mixed initiative spoken dialogue system enhanced with a GUI.

The research was carried out within the framework of the MATIS project (Multimodal Access to Transaction and Information Services), which aimed at evaluating the usability of multimodal interaction for form-filling applications on small, mobile terminals [1]. To this end, a prototype multimodal form-filling interface was built that provides train timetable information. This interface adopts a mixed initiative dialogue strategy in which both the user and the system could take the initiative. The mixed initiative dialogue has been designed in such a way that it serves both novice and experienced users: the system initiates a spoken dialogue by asking questions, in order to help novice users in completing the form. The dialogue also helps solving errors and provides clarification in order to prevent users from getting stuck. More experienced users, on the other hand, may take over the initiative by interrupting the spoken questions by pressing buttons on a graphical display of the fill-in form. In this way, experienced users can use the system in a more efficient way. An experiment was carried out in order to establish the effect of extended use on the usability of the system. The results of this experiment showed that as their hands-on experience with the system grew users indeed started to use the multimodal system in a more efficient way by interrupting system questions and by using buttons rather than speech to provide values [2].

It may well be, however, that the form-filling application

is so easy to use that even novice users do not really need the spoken system guidance [3]. In that case, an interface that is completely user driven and that does not apply a spoken dialogue may be just as effective and efficient for novice users as a mixed initiative system. To investigate this issue, in the current study the mixed initiative interface (MIMI) was compared to a completely user driven version of the same interface (Tap&Talk). This user driven interface was implemented as a tap-and-talk interface: the system does not ask the user any questions, instead the user must indicate which field he or she wants to fill in by pressing buttons on a graphical representation of the form on the screen. The present paper describes and compares the usability of the two interfaces.

## 2. Methods

### 2.1. Interfaces

As mentioned earlier, the two interfaces differ in the way the interaction is controlled. Whereas in the MIMI interface a spoken dialogue (which may be interrupted by the user) guides the user through the interaction, in the Tap&Talk interface the interaction is completely controlled by the user who uses the buttons on the screen to indicate which field he or she wants to fill in. Both interfaces show the same graphical representation of the fill-in form on the screen (Figure 1).

The screenshot shows a graphical user interface for a train timetable form. It features several input fields and buttons. At the top, there are 'From' and 'To' fields with microphone icons. The 'From' field contains 'arnhem' and has a dropdown menu open showing a list of stations: 'arnhem', 'arnhem presikhaaf', 'arnhem velperpoort', 'haarlem', and 'haarlem spaarnwoude'. Below these are buttons for 'today', 'tomorrow', and 'other day'. There are also date fields containing 'woensdag' and '26-februari-2003'. At the bottom, there are 'at' and '19:30' fields, and buttons for 'departure', 'arrival', and 'search'. Circled numbers 1, 2, and 3 are placed near the 'today', 'aarlem spaarnwoude', and 'at' fields respectively.

Figure 1: Screen shot of the fill-in form

In both systems speech must be used to fill in the station names, times and dates other than today or tomorrow. Other values can be filled in by direct manipulation, using the buttons on the screen. Providing input via the screen can be done in the following three ways (cf. Figure 1):

1. Radio buttons – can be used to select mutually exclusive values, such as “today / tomorrow”.
2. Drop-down lists – can be used to select recognition alternatives or alternative stations in case the station names were recognized incorrectly.
3. Microphone buttons – can be used to start the recording for a specific field. Once the button has been pressed, the user can fill in the field using speech. In the MIMI interface a short, spoken instruction (e.g. “say the departure station”) is played if a microphone button is pushed, whereas in the Tap&Talk interface the user can immediately start speaking.

Finally, pressing the *Search* button forces the system to query the database. All fields must be filled to do this, but the values do not have to be verified yet. Thus, pressing the Search button is a form of implicit verification. (For detailed information about the graphical input facilities and how they can be used in combination, the reader is referred to [4]).

The display also provides the user with information about the fields that need to be filled, the status of the dialogue (whether the system is recording speech, recognizing speech, etc.) and the recognition results.

Although originally conceived for small devices such as palmtops and mobile phones, for practical reasons both interfaces were implemented as a Java-applet on a desktop computer with a touch screen and no keyboard, for practical reasons. To interact with the system an ordinary telephone was used, with a headset in order to keep both hands free for interaction by means of the touch screen.

## 2.2. Subjects & tasks

Seventeen subjects took part in the evaluation (eight male and nine female, between 20 and 71 years of age, with mixed educational backgrounds). They were paid for participating. Except for one subject, all had experience with computers. Half of the subjects were regular train travelers (at least once a week); the other half consisted of occasional travelers (between once a month and twice a year). The subjects declared that they mainly used the Internet to get train timetable information. Eight subjects had used a different spoken dialogue system before; six subjects had used the commercial version of the speech-only dialogue system for train timetable information ever before.

All subjects tested both interfaces. They were divided into two groups; the first group tested the MIMI interface first and then the Tap&Talk system (8 subjects), the other group tested the two systems the other way round (9 subjects).

After a short introduction all subjects completed one practice scenario and three test scenarios with each of the two systems. The scenarios were presented graphically in order to avoid influencing the manner in which subjects express themselves (see Figure 2). We used different scenarios for each of the two interfaces, in order to avoid any learning effect. To ensure that the test would provide information about how users deal with speech recognition errors, some scenarios concerned station names that are highly confusable for the automatic speech recognizer.

All sessions were conducted in the usability lab of the UCE department of TU Eindhoven, which is furnished as a living room.

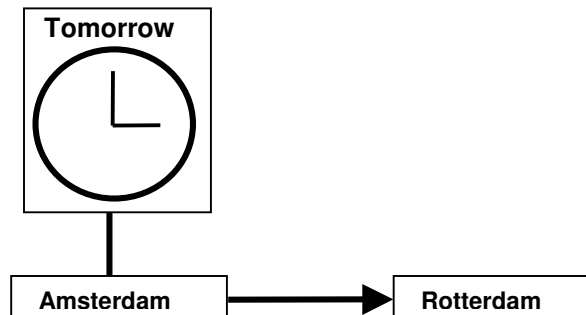


Figure 2: Example of a test scenario

## 2.3. Data capture & evaluation measures

Speech and clicking actions of all dialogues were automatically logged (including time stamps). Additionally, audio and video recordings were made of all dialogues.

The usability of the two interfaces is evaluated in terms of effectiveness, efficiency and user satisfaction. Effectiveness is defined as the number of dialogues that were completed successfully (the dialogue success rate). Efficiency is defined as task completion time (i.e. the time span between the start of the first user utterance and the moment at which the query is sent to the information database). User satisfaction is measured by means of 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*” (see Table 3). The subjects expressed their agreement or disagreement with these statements on a five-point Likert-scale (1 = I strongly disagree, 3 = I agree nor disagree, 5 = I strongly agree). Furthermore, subjects indicated their preference for one of the two interfaces concerning a number of aspects.

## 3. Results & discussion

In total, 51 dialogues were recorded with the MIMI interface and 49 with the Tap&Talk interface<sup>1</sup>. In this section results are given for effectiveness, efficiency and user satisfaction.

### 3.1. Effectiveness

The effectiveness of the interfaces is measured in terms of the number of successfully completed dialogues (i.e. dialogues in which the user obtained the travel advice he or she requested).

Table 1: Effectiveness per scenario

| Scenario | MIMI      | Tap&Talk  |
|----------|-----------|-----------|
| 1        | 16 (94%)  | 15 (94%)  |
| 2        | 17 (100%) | 16 (100%) |
| 3        | 16 (94%)  | 12 (71%)  |
| Total    | 49 (96%)  | 43 (88%)  |

<sup>1</sup> For the Tap&Talk interface the figures do not add up to 51 (17 times 3), because due to technical errors one subject could not complete the first scenario and another subject could not complete the third scenario.

Table 1 shows that overall the effectiveness of both interfaces is rather high and slightly higher for the MIMI system (96%) than for the Tap&Talk interface (88%). A Wilcoxon test showed that this difference is not significant ( $z = .63$ ; N.S.). All unsuccessful dialogues were caused by persistent recognition errors, after which the subject hung up.

Table 1 shows that only for Scenario 3 the Tap&Talk interface is less effective than the MIMI interface. As this was the scenario with the most confusable station names, this seems to indicate that solving recognition errors is easier with the MIMI interface than with the Tap&Talk interface. This is probably caused by an extra feature in the MIMI interface that facilitates error correction: if a user has explicitly denied a certain station name, this station name will not be recognized in subsequent attempts to fill in a value for this field. In the Tap&Talk interface, as there is no explicit verification in this interface, the same station name may be recognized over and over again.

No significant effect was found of the order in which the two systems were tested.

### 3.2. Efficiency

The efficiency of the interfaces is measured as time to completion (of the successfully completed dialogues). For each scenario the mean duration of a dialogue was calculated in seconds measured from the start of the first user utterance to the query to the information database. Using time to completion as a performance measure in a user-driven interface is not trivial: people may take some extra time to figure out what the next step should be, when they are not being rushed by a spoken dialogue. We tried to diminish this effect by telling subjects that they were paying for this service. Nevertheless, some care should be taken in comparing the efficiency figures for the two interfaces.

Table 2: Efficiency per scenario (in seconds)

| Scenario | MIMI | Tap&Talk |
|----------|------|----------|
| 1        | 69.7 | 56.1     |
| 2        | 42.6 | 31.7     |
| 3        | 71.5 | 56.4     |
| Average  | 60.9 | 47.1     |

Table 2 shows that on average dialogues are completed faster using the Tap&Talk interface than using the MIMI interface (the difference is on average 13.8 seconds). A three-factor mixed ANOVA, in which the missing values (unsuccessful dialogues) were replaced by the mean of their

respective condition (system \* scenario \* group), revealed that this difference is significant ( $F(1,15) = 8.13$ ;  $p < .05$ ).

The main cause for the difference in efficiency is the duration of the spoken prompts in the MIMI interface (recall that there are no spoken prompts in the Tap&Talk interface). The mean total duration of the system prompts in the successfully completed dialogues using the MIMI interface is 10.3 seconds.

Other explanations for the observed difference in efficiency can be found in the way people interact with the two interfaces. In the MIMI interface the spoken dialogue encouraged users to provide values by speech rather than by pressing buttons (which is more efficient), whereas in the Tap&Talk interface using buttons was the only option to provide values for a number of fields (such as departure/arrival). Furthermore, in the MIMI interface values were verified by means of a spoken verification question, which leads to a large number of yes/no utterances, whereas in the Tap&Talk interface values were only verified visually. As a result, the total number of user utterances per dialogue was smaller in the Tap&Talk interface than in the MIMI interface: The Tap&Talk interactions contained on average 4.1 spoken utterances, whereas dialogues with the MIMI interface contained 5.7 spoken utterances. Not only the number of utterances, but also the duration of the utterances was shorter in the Tap&Talk interactions (1.3 seconds per utterance vs. 1.8 seconds in the MIMI dialogues). In the Tap&Talk interface subjects could only provide one value per utterance, whereas in the MIMI dialogues subjects tended to provide more information in one utterance (e.g. a combination of departure station and arrival station).

For both interfaces the time needed to complete scenario 2 was significantly shorter than the time needed to complete scenarios 1 and 3 ( $F(2,30) = 17.50$ ;  $p < .05$ ). This is probably due to the fact that the station names used in this scenario were relatively easy to recognize for the automatic speech recognizer and therefore caused less misrecognitions (which is in accordance with the high effectiveness values for scenario 2 shown in Table 1).

No significant effect could be found of the order in which the two groups tested the two interfaces ( $F(1,15) = 3.16$ ; N.S.).

### 3.3. User satisfaction

The results of the user satisfaction questionnaire are shown in Table 3. Most statements concern both systems, except for statements 15-19; those only concern the MIMI interface. For the negative statements 4, 9 and 13 both the statement and the scores have been inverted, so that high scores always denote the positive end of the scale.

Table 3: Results of user satisfaction questionnaire (1 = "I completely disagree" - 5 = "I completely agree")

| Statement  | MIMI | Tap&Talk |
|--|------|----------|
| 1. I consider the system easy to use                   | 3.4  | 3.9      |
| 2. I always understood what was expected from me       | 3.8  | 4.7      |
| 3. I found it easy to correct errors                   | 3.3  | 4.0      |
| 4. I thought the system was NOT slow                   | 1.0  | 1.9      |
| 5. I thought the travel advice was clear               | 4.5  | 4.7      |
| 6. The combination of speech and graphics was useful   | 3.8  | 3.8      |
| 7. The system reacted adequately to the combined input | 3.4  | 4.1      |
| 8. Visualizing the filling form was useful             | 4.4  | 4.5      |
| 9. I was NOT distracted by the display                 | 3.1  | 3.4      |

| Statement   | MIMI | Tap&Talk |
|---|------|----------|
| 10. Visualizing the travel advice was useful                                  | 4.8  | 4.8      |
| 11. After a while I started using the system differently                      | 3.5  | 2.5      |
| 12. I used the touch screen more often as I got more experienced              | 2.9  | 2.8      |
| 13. I did NOT feel uncomfortable when I had to speak to the system            | 2.8  | 2.8      |
| 14. I would use this application if it were on my PDA or mobile phone         | 3.7  | 3.9      |
| 15. Speech and graphics were well tuned to one another regarding the contents | 3.7  | -        |
| 16. Speech and graphics were well tuned to one another regarding the timing   | 3.1  | -        |
| 17. The length of the spoken utterances was appropriate                       | 3.8  | -        |
| 18. Giving the travel advice in spoken form was useful                        | 2.5  | -        |
| 19. Being able to interrupt the system speech is necessary                    | 3.4  | -        |

Table 4: User preferences

| Question  | Preferred system |          |               |
|---|------------------|----------|---------------|
|   | MIMI             | Tap&Talk | No preference |
| Which system did you consider the easiest to use?                         | 4 (24%)          | 13 (76%) | -             |
| With which system did you know best which information you had to provide? | 2 (12%)          | 14 (82%) | 1 (6%)        |
| With which system was correcting errors easiest?                          | -                | 13 (76%) | 4 (24%)       |
| Which system did you consider the most fun to use?                        | 6 (35%)          | 11 (65%) | -             |
| With which system was understanding the travel advice easiest?            | 4 (24%)          | 5 (29%)  | 8 (47%)       |
| Which system would you prefer to use in the future?                       | 5 (29%)          | 12 (71%) | -             |

In general, user satisfaction was rather high. Most statements were judged about equal for both interfaces, but in some cases there were minor differences mostly in favor of the Tap&Talk interface. Subjects understood significantly better what was expected from them using the Tap&Talk interface than using the MIMI interface ( $z = 3.13$ ;  $p < .05$ ) and correcting errors was considered easier in the Tap&Talk interface, although the effectiveness data suggest that correcting errors was easier using the MIMI interface. Also, subjects judged both systems as slow, but the MIMI system was judged significantly slower than the Tap&Talk system ( $z = 2.56$ ;  $p < .05$ ), which is in accordance with the efficiency results described in the previous section. According to the scores for statement 11, the learning effect for the MIMI interface was stronger than for the Tap&Talk interface. In general, subjects appreciated the visualization of the filling form (st. 8), and they thought the travel advice was clear (st. 5). Also, they judged the visualization of the travel advice useful (st. 10). Finally, most subjects would use this type of application if it were on their mobile device (st. 14).

Table 4 shows the user preferences for the two interfaces. On most aspects subjects clearly preferred the Tap&Talk interface, as could be expected given the satisfaction data in Table 3. Once again, subjects indicated that correcting errors was easier using the Tap&Talk system than using the MIMI system, which is in contradiction with the effectiveness data. The only aspect on which there was no clear preference for one of the two interfaces concerned the travel advice; apparently, users were not helped by the spoken version of the travel advice in the MIMI interface. 71% of the subjects indicated that they would choose to use the Tap&Talk interface in the future.

#### 4. Conclusions

The aim of the research presented in this paper was to compare the usability of a user driven version and a mixed initiative version of a multimodal interface for train timetable information.

The results show that the effectiveness of the two interfaces does not differ significantly. However, due to the absence of spoken prompts and the obligatory use of buttons to provide values, the efficiency of the successful dialogues with the user driven interface is much higher than the efficiency of the mixed initiative interface. Apparently, also for novice users the user driven interface is much more efficient than the mixed initiative interface without a real loss of effectiveness. However, based on the results of a previous experiment, we expect that the learning curve will be much steeper for the mixed initiative interface than for the user driven interface, so that the differences will reduce once users get more experienced [2].

Finally, although the overall user satisfaction was not significantly higher for the Tap&Talk interface, by far most people preferred the Tap&Talk interface to the MIMI interface.

#### 5. Acknowledgement

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

#### 6. References

- [1] <http://www.ipo.tue.nl/projects/matis/>
- [2] Sturm, J., Bakx, I., Cranen, B., Terken, T. (2002a). 'The effect of prolonged use on multimodal interaction'. *Proceedings ISCA Workshop on Multimodal Interaction in Mobile Environments*, Kloster Irsee, Germany.
- [3] Shneiderman, B. (1997). "Direct Manipulation for Comprehensible, Predictable and Controllable User Interfaces". *Proc. of International Conference on Intelligent User Interfaces*, Orlando, Florida.
- [4] Sturm, J., Bakx, I., Cranen, B., Terken, T., Wang, F. (2002b). 'Usability evaluation of a Dutch multimodal system for train timetable information'. *Proceedings LREC2002*, Gran Canaria de Las Palmas, Spain.