Supporting the Construction of a User Model in Speech-only Interfaces by Adding Multi-modality

Jacques Terken, Saskia te Riele

IPO, Center for User - System Interaction,
Technische Universiteit Eindhoven, The Netherlands
j.m.b.terken@tue.nl

Abstract
Comparing to graphical user interfaces, speech-only interfaces face several problems: robustness, making clear what functionality is available, and making clear how the functionality may be accessed. We explore a potential solution for these problems by presenting a visual representation of the domain of discourse and of the state of the dialogue. We describe an experiment in which uni-modal and multi-modal interfaces are compared in terms of effectiveness, efficiency and satisfaction. The results of the experiment show a strong learning effect. Subjects who start using the multi-modal interface subsequently have a strong advantage when switching to the uni-modal (speech-only) interface, compared to subjects who start by using the uni-modal interface, switching to the multi-modal interface later on. The results are discussed in terms of the need to establish an appropriate user model as early as possible. We discuss implications of this interpretation for interaction design.

1. Introduction

Language and speech technology have been advocated as key technologies for user interfaces in the 21st century. In the first place, spoken language interfaces allow direct access to functionality rather than forcing the user to control individual functions in a stepwise manner, for instance by means of menus. Secondly, the use of natural language provides by definition for natural interaction. However, there are several reasons to doubt that speech interfaces will indeed replace other kinds of interfaces. Two important ones are the following. It has been proven difficult for users to find out what functionality is available. Also, the robustness of the language and speech technology is quite low, so that many utterances produced by the user are not adequately dealt with. Unfortunately, these two factors interact. The less adequate the understanding of the user of how the system works, the more likely he is to produce utterances that cannot be adequately handled.

In order to explore solutions for this situation, several research groups are exploring multi-modal extensions. Most of this research focuses on combining multiple input modalities. In the current paper we focus on system output, and investigate whether speech-based interaction can be supported by offering a visual representation of the domain of discourse. We started from a speech-only interface for a train table information system. The domain of discourse can be represented as a form that has to be filled during the dialogue, showing the parameters by which to query the train table database (see Fig. 1). In the multi-modal version, the form to be filled was shown on a PC screen, and the values for the different fields in the form were filled in as the dialogue progressed. In the following sections we describe an experiment aiming to find out whether the visual presentation of the form helped to improve the interaction in terms of the traditional usability factors (Effectiveness, Efficiency and Satisfaction).

2. Methods

2.1. Materials

In order to evaluate the contribution of the visual presentation to the success of the interaction, the performance for the multi-modal interface was compared with that for the uni-modal interface, that had been developed in the framework of a research programme on language and speech technology [1]. The current version of the multi-modal interface extends the speech-only interface with the visual representation of the form to be filled during the dialogue (cf. Fig. 1). As the dialogue proceeds, the system fills the fields with the values provided by the user, so that the user at each moment in the dialogue has an overview of the current state of the dialogue.

Several design decisions need to be mentioned concerning the multi-modal interface, some of which were inherited from the speech-only interface.

- The system would always verify the information presented by the user explicitly, by asking a yes-no question (“Did you say ...?”). In earlier versions an implicit verification strategy was used, but this strategy turned out to be confusing for the listener in case of

![Figure 1: Graphical representation of domain of discourse presented on the screen.](image-url)
speech recognition errors, because the expression containing the error is not part of the linguistic focus.

- The language generation component of the speech-only interface was left unchanged, so that utterances were rather verbose. This may be less appropriate for a multi-modal interface.
- In the multi-modal version of the system, the values for the form fields that were recognized from the utterances were shown on the screen immediately, even though they might be wrong. Again, this may be less appropriate for a multi-modal interface, as written text might leave the user confused about whether the system still allows the user to correct erroneous information.
- In the speech-only interface, processing the user utterance and preparing the next system utterance took about four seconds, for the most part consumed by the language generation and speech synthesis. In the multi-modal interface, the recognized information was shown on the screen right away, so that it appeared on the screen approximately two seconds before the user would hear the utterance of the system verifying that piece of information. This might lead the user to react immediately, but since no barge-in was allowed the system would not react to such immediate reactions by the user.

We realize that these properties may affect the quality of the multi-modal interface. In particular, we would prefer to tune the verbal interaction so that it would be more appropriate in a multi-modal context. For the current experiment we decided to keep the two systems identical in these respects so that differences in results for the two systems would reflect the true effect of the visual presentation rather than of corollary changes.

The task for the subjects in the experiment consisted of completing six scenarios in which they had to find out departure times for pre-specified trajectories. For instance, in one scenario subjects would have to find out how late the train departed given that they needed to go from Eindhoven to Nijmegen on Sunday morning, arriving at 11:30. The train departed given that they needed to go from Eindhoven to Nijmegen on Sunday morning, arriving at 11:30. The scenarios included mostly station names that had a high probability of correct recognition, in order to avoid a situation where the appreciation would be strongly influenced by the performance of the speech recognition component.

### 2.2. Subjects

Subjects’ ages (8 males, 8 females) ranged between 21 and 51 years, four of them being older than 35 years.

### 2.3. Procedure

The comparison was done in a mixed between- and within-subjects design. All subjects used both systems, so that they might express a preference judgement. In order to avoid order effects two conditions were created. One group of subjects (Group I, n=8) first completed three scenarios with the speech-only interface and then three scenarios with the multi-modal interface. In the other group (Group II, n=8) the order was reversed. Following each scenario, the subjects hung up the phone and started a new phone call, so that for each subject we obtained six phone calls, made up of two series of three. After having finished a series of three, subjects filled in a questionnaire asking for their subjective judgements concerning a number of usability aspects. In addition, after having finished both series, they were asked for preference judgements.

### 3. Results and Discussion

There was no overall effect on effectiveness. Both interfaces gave rise to the same amount of successfully completed dialogues (42/48 for the uni-modal interface, 40/48 for the multi-modal interface). However, we observed strong order effects. There were less successful dialogues in the first sequence of three dialogues than in the second series, 36/48 and 46/48, respectively, equally divided across both systems. This suggests that subjects build some expertise about how to establish a successful interaction with the system during the first series of calls, regardless of what system they use during the first series, which they exploit during the second series.

With respect to efficiency, operationalized by time to completion, again there was no effect of system. If we count only the true part of the dialogue between user and system, i.e., the duration from the end of the opening question of the system until the end of the last user utterance before the system starts to present the result of the query, and excluding the failed dialogues, for both interfaces the time to complete a dialogue was approximately the same: 58 seconds for the uni-modal interface, 59.5 seconds for the multi-modal interface. Again, there was a clear effect of order: dialogues in the second series were completed faster than dialogues in the first series (see Figure 2). This effect was found to be significant in a mixed factorial Analysis of variance with series as repeated measures, collapsing across dialogues within series for individual users ($F_{1,14}=12.58, p=.003$).

### Table 1: Average durations for components of turn cycle. Group I order: first uni-modal, then multi-modal. Group II order: first multi-modal, then uni-modal.

<table>
<thead>
<tr>
<th></th>
<th>User pause</th>
<th>User utterance</th>
<th>System pause</th>
<th>System utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1 Series 1 (uni-modal)</td>
<td>1.60 s</td>
<td>0.99 s</td>
<td>4.50 s</td>
<td>1.86 s</td>
</tr>
<tr>
<td>Group 1 Series 2 (multi-modal)</td>
<td>0.99 s</td>
<td>1.26 s</td>
<td>5.10 s</td>
<td>1.98 s</td>
</tr>
<tr>
<td>Group 2 Series 1 (multi-modal)</td>
<td>1.29 s</td>
<td>1.12 s</td>
<td>4.84 s</td>
<td>1.84 s</td>
</tr>
<tr>
<td>Group 2 Series 2 (uni-modal)</td>
<td>0.70 s</td>
<td>0.93 s</td>
<td>4.24 s</td>
<td>1.64 s</td>
</tr>
</tbody>
</table>
Also, there was a considerable effect of Groups. Group II (multi-modal interface first) completed the dialogues in the first series faster than group I (uni-modal interface first) (see Figure 1). This advantage of Group II over Group I was then preserved when Group II worked with the uni-modal interface. For Group I, series 2, with the multi-modal interface, also went faster than series 1, with the uni-modal interface. Yet, their time to completion for series 2 was slower than for Group II, which used the uni-modal interface in the series 2. However, the effect of group turned out not to be significant (F1,14=.80, p=.386), due to large within-group variation.

For further analysis the dialogue was decomposed into turns, and pause and utterance durations were considered separately. Means for the different conditions are shown in Table 1. As can be seen, the system has a major contribution to the duration of the dialogue. It needs a lot of time to process the utterance of the user and generate the next utterance. Also, the utterances of the system are longer than the utterances by the user.

Pauses preceding the replies by the user are longer in the first series of dialogues than in the second series (F1,14=3.62, p=.078). To the extent that this difference is indicative of a true difference between uni-modal and multi-modal systems, it may be due to a tendency for users to produce longer utterances when interacting with a multi-modal system: as can be seen in Table 2, the number of words per utterance is higher in the multi-modal conditions than in the uni-modal conditions.

In order to find out whether differences in length and number of words of user utterances are associated with different interaction strategies in the two systems, an analysis was conducted on the type of dialogue acts in the different conditions. The reasoning is that the uni-modal speech-only system would give rise to more system-driven dialogue and elicit on average shorter utterances as replies to the system prompt, whereas the multi-modal system, showing the whole form, would encourage the users to supply relevant information even when not prompted to do so. A set of ad-hoc dialogue acts was constructed, containing the following dialogue acts: (a) Supply-info; (b) Confirm; (c) Rejection; (d) Mixed-initiative; (e) Confirm+mixed-initiative; (f) Implicit confirmation+mixed-initiative; (g) Rejection+mixed-initiative; (h) Implicit rejection+mixed-initiative; (i) Silence; (j) Hesitation; (k) Unclear situation.

Proportions for the different types of speech acts are shown in Figure 3. As can be seen, the distribution is by and large the same for the different systems. On closer inspection, there are subtle differences. The group that starts with the multi-modal system (group 2 series 2) has fewer simple “supply information” acts and confirmations. Instead, they use more acts involving mixed initiative (categories d, e and f). However, by the time they get to series 2, everything is back to normal. It is as if they are prompted by the visualization of the domain model to supply more information than asked for, but return to the standard strategy of system driven dialogue upon finding out that the mixed-initiative causes problems for the speech recognizer.

In this respect, there is a noteworthy observation from the sessions: in the multi-modal condition, sometimes users reacted as soon as the system filled one of the fields with information extracted from the user’s utterance. However, in the current set-up the system would not react to this user action, but instead would simply go on to prepare the associated utterance. This suggests that, whereas from a designer’s point of view displaying information on the screen and producing an utterance were considered as components of a single multi-modal dialogue act [2], users would treat them as separate actions if they did not occur at the same time.

Table 2: Average number of turns cycles, average duration per turn cycle in seconds, and average number of words per user turn, in series 1 and 2 for groups I and II. Group I order: first uni-modal, then multi-modal. Group II order: first multi-modal, then uni-modal.

<table>
<thead>
<tr>
<th>Series 1</th>
<th>Series 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group I: First uni-modal, then multi-modal interface</td>
<td>Group II: First Multi-modal, then Unimodal interface</td>
</tr>
<tr>
<td>Number of user turns</td>
<td>Duration of user turn</td>
</tr>
<tr>
<td>9.0</td>
<td>2.5 s</td>
</tr>
<tr>
<td>7.7</td>
<td>2.4 s</td>
</tr>
</tbody>
</table>
Finally, the results for the questionnaires showed that almost all subjects preferred the multi-modal interface, even though the objective differences between the uni-modal and multi-modal interfaces described above may seem rather modest.

4. Conclusion

The results suggest that users who start with the multi-modal interface have an advantage over users who start with the speech-only uni-modal interface, and that this advantage is preserved later-on, when they switch to the uni-modal interface. We take this finding as evidence that the multi-modal interface helps these users to develop an adequate model of the system. Once they have internalized this model, they can do without, so that they will also profit from the appropriate understanding of the system when the external model is no longer available. The users who start with the uni-modal interface do not so easily develop an adequate understanding of the system, even though in this group a learning effect is observed as well.

From these findings we tentatively conclude that, even for such relatively simple systems as a train table information system, adding multi-modality may help, albeit not in the way expected. The main effect seems to reside in helping the users to develop an adequate model of the underlying structure of the system. Once this model has been internalized, the presence of the model in the external world is no longer a precondition for adequate use of the system.

This is in agreement with findings by Kamm et al. [3], who compared different ways of instructing users about the possibilities of a telephone interface and about how to use it. They compared “no instruction”, and two different instruction methods, one being the standard written instruction, the other being a set of scenarios that were described on a webpage and which the user could go through actually using the system. As might be expected and in line with informal observations of most people, the written instruction was hardly more effective than the “no instruction” condition. The scenario-based instruction in which users actually performed a number of exercises was very effective. Cast in the current terminology, one might say that the scenario-based instruction provides the novice user with an appropriate model of the system. Once the user has internalized this model, he is no longer dependent on an actual model in the external world.

For more complex systems, however, one may expect that the user will remain dependent on a model that is present in the external world. By adding multi-modality to a speech-only interface we add affordance to the system, thereby prompting appropriate actions from the user and overcoming one of the main drawbacks of speech-only interfaces. Also, the external model may serve as a set of retrieval cues for the user’s knowledge of the system. Finally, without external assistance it may become quite difficult to keep track of the current state of the system.

5. Acknowledgements

The research reported here was supported by a grant to the MATIS project from Senter/EZ in the framework of IOP-MMI (Innovatiegerichte Onderzoeksprogramma Mens–Machine Interactie).

6. References