User Behaviour in Multimodal Interaction

Elisabeth den Os

Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 AH Nijmegen
Els.denOs@mpi.nl

Lou Boves

Radboud University Nijmegen
P.O. Box 9103, 6500 HD Nijmegen
L.Boves@let.ru.nl

Abstract

In this paper we investigate whether the behaviour of non-expert subjects who interact with a fully implemented multimodal application is comparable to the behaviours observed with Wizard-of-Oz systems. To that end we have implemented in the COMIC project[1] an application for bathroom design that accepts speech and pen input, and provides output in the form of stylized drawing and text on a screen, combined with a naturalistic talking head. In addition, we draw on data obtained in the MUST[2] project, also with a fully implemented system. We find comparable behaviour in some respects, but also different behaviours, which can be linked to differences in the functionality of the systems and the complexity and nature of the tasks.

1 Introduction

1.1 State-of-the-Art in Multimodal Interaction research

During the last decade a body of literature on multimodal interaction has accumulated, coming from different fields. One large body of research comes from the field of Immersive Environments, where the focus is on issues such as telepresence, manipulation of virtual objects, visualization of large amounts of data, etc. (Park et al, 2000). Part of this research addresses interactions between humans and virtual agents. Almost invariably, experiments in this field require that subjects wear head mounted display equipment, which restricts their movements to the confines of a closed space (e.g., Bailenson et al, 2004). With few exceptions (e.g., Thórisson, 2003) the interaction between participants and the artificial agent does not involve speech or language. At the other end of the spectrum there is research focused on users interacting with information and transaction services through a Personal Digital Assistant (PDA) or a feature-rich telephone handset in the UMTS network, with speech and pen as the primary input devices (Kvale et al., 2002; Sturm et al., 2003). The research focused on interactive map applications conducted by Oviatt also belongs in this category (Oviatt, 2003).

1.2 Multimodality or Interaction metaphor

It is often assumed that multimodal interfaces per se will facilitate universal access to services, simply because they offer users the freedom to select the mode that suits their needs for expressing their desires best. Specifically, the facts that multimodal systems can adapt to their users and that users in their turn can switch from one input mode to another have been cited as a major advantage (Xiao, et al., 2003; Sturm & Boves, 2005). However, our research in speech-centric multimodal interaction in architectural design applications in the FP5 project COMIC (den Os & Boves, 2004) has shown that the problem of universal access is far too complex than that it could be solved by adding, for example, speech to keyboard and mouse or pen input. More often than not, the crucial problem is not that users cannot express their desires by means of a graphical direct manipulation interface, but rather that they do not know what desires can be satisfied by a given application or service. And if a user does not know what (s)he can express, adding speech, gaze or gesture to keyboard input is not very helpful. Therefore, we believe that the distinction between the interaction metaphors of direct manipulation and conversational agent who guides users through a task is as least as important as the difference between unimodal and multimodal interaction (den Os et al., submitted).

---

[1] http://www.hcrc.ed.ac.uk/comic
[2] http://www.eurescom.de/public/projects/P1100-series/p1104

1.3 Tasks and Challenges

Most of the research on multimodal interaction is aimed at improving our understanding of the cognitive demands on the human user or the technical requirements of the systems. Thus, it is not surprising that there is experiments differ in the tasks and the communicative settings. This makes it difficult to appreciate the degree to which results can be generalized. In our research we do not only focus on fully implemented systems, but also on complete and ecologically realistic tasks. This seems to be different from most of the previous research, where participants must perform a series of related, but independent challenges (e.g., Oviatt et al., 2004). It also differs from the tasks in immersive environment research where participants are left to themselves to decide how to interact with a real or simulated system (e.g., Thórisson, 2002).

In all our experiments participants were required to complete tasks that were composed of a sequence of subtasks. In the Paris Tourist Guide experiments (Kvale et al., 2003) participants had first to select a specific Point of Interest, after which they were requested to obtain further information related to that PoI. In some tasks this again required selecting a specific item (for example a hotel) and subsequently searching information about that item. After each input of the user the system changed state, according to its interpretation of the input. If a recognition error occurred, the user was confronted with an unexpected graphical and acoustic display. This forces participants to take three actions: first they must notice that the display contains wrong data, then they must erase the wrong information and finally replace it with the right data. Depending on the interaction design, the latter two actions can be combined in one 'overwrite data' action. Unsurprisingly, participants not always knew how to accomplish this, let alone how it could be done in the most efficient manner. The behaviour that is elicited by this kind of error detection and repair may well be different from the behaviour observed in experiments where a wizard would generate an error message such as "I did not quite get that. Please repeat." The need to detect and repair recognition errors (or mistakes made by the participant) in order to be able to complete a scenario may lead to behaviour that is also different from what is observed in experiments where participants are left free to 'play' with a multimodal system.

Error detection and correction in multimodal interaction also differs from what has been observed in spoken dialogue systems. Especially in system driven interactions users are confronted with explicit requests to confirm the system's recognition result. If an error has occurred, users may not know that the can say "No, that is wrong. I meant YYY", so that they can correct the error in a single turn. Yet, few users will wonder how they can undo an error when they are confronted with the confirmation request. If the error is displayed on a screen, users must know how to erase the wrong output.

1.4 Wizard-of-Oz experiments

Previous research in speech centric multimodal interaction was conducted to demonstrate proofs-of-concept, rather than to develop systems that could be taken out of the laboratory anywhere in the near future (Kehler et al., 1998). At the same time, few fully operational multimodal systems were available. Consequently, most of the research was conducted in Wizard-of-Oz settings. While WoZ experiments have made absolutely essential contributions to the progress of multimodal interaction research and the emerging theories and models in the field, they also incur risks that have not always been given due attention.

The development of our bathroom design system started with a number of WoZ experiments in which we investigated how uninformed users enter blueprints into a computer system (Rossignol et al., 2004). It appeared that the complexity and variation of the pen and speech input gestures that produced by the participants was far beyond the capabilities of our recognizers. Also, interactions between users and wizard appeared to rely in a fundamental manner on the intelligence and communicative competence of both interlocutors. We have encountered three major hurdles in going from a WoZ simulation to a fully implemented system, viz. recognition performance, the amount of artificial intelligence that is needed to manage a dialogue when users interact with an error-prone system, and the way in which turn taking is implemented. These issues have a large impact on the flow of the communication between the user and the system. It is well known that recognition errors tend to elicit counterproductive reactions, such as over-articulation. If the system lacks the intelligence to handle the exceptional situations caused by the errors appropriately, the risk that users behave in an unexpected manner increases. In a WoZ setting it is easy to overlook difficulties that impede the completion of a full task, because the Wizard cannot imagine just how unintelligent a system can be. Miscommunications are aggravated by the way turn taking is implemented. With the

exception of Thórisson (2002) all operational multimodal systems seem to rely on what is essentially half-duplex interaction. This may be reasonable for the speech channel, although conversational human-human dialogues show substantial amounts of overlapping speech (ten Bosch et al., 2004), but not for the information exchanged in the visual channel. It is difficult to explain to users why it is not possible to draw or write while the system's avatar is speaking, or –for that matter- showing that he is working to interpret previous input.

The impact of the way in which turn taking is implemented is especially clear when it comes to combining speech and pen input. The system's responses are based on its interpretation of the input, and this depends on how data from the two input channels are combined. Information in the channels can be redundant (if someone writes '3 m' and at the same time says 'three meters'), complementary (in the case of "put that there") or unrelated (e.g., when selecting a radio button for date, while saying a city name in a timetable information system). In the first two situations a decision must be made about the timing relation between the channels: must speech and pen input be interpreted in combination, or is the time distance between the two such that the channels must be interpreted independently. Of course, the time window within which acts are considered as 'simultaneous' is arbitrary. In the Paris Tourist Guide we considered acts as simultaneous if they occurred within 1 second of each other. However, it seems that Xiao et al.(2003) consider acts only as simultaneous if they overlap at least partially.

1.4.1 Test procedures

An important consequence of testing with a fully implemented system is that users must stay within the limitations of the system set by the limitations of the input recognizers, the dialogue manager and the output rendering. Since there is no way for the system to convey those limitations, there is hardly an alternative for scenario-based testing, where participants are 'guided' towards behaviour that is understandable for the system by the scenario (den Os & Boves, 2004) and perhaps also by short instructions at the start of an experimental session (Kvale et al. 2003). At the positive side, scenario-based testing provides a natural means for measuring task completion rate.

1.5 Aims of this Paper

In the previous subsections we have identified a number of issues in multimodal interaction research that may affect the extent to which the results and conclusions of one experiment may be generalized to other tasks and setting. These issues include the importance of the Conversational Agent metaphor rather than the addition of input and/or output modes, the possible impact of 'unrealistic' system intelligence in Wizard-of-Oz settings, and the fact that realistic applications will almost always require successive steps that build upon each other. In the research reported in this paper we address these issues in more detail. We report on two experiments in which users had to interact with fully implemented multimodal systems, viz. the MUST Paris Tourist Guide and the COMIC bathroom design system. In both experiments the participants were confronted for the first time with multimodal systems. In the experiment with the bathroom design system also the task was unfamiliar. Our interpretation of the findings is guided by related results obtained in a series of experiments with several versions of a time table information system. Thus, our appreciations are based on experiences with three generic applications: architectural design, map based interaction and form filling interaction. On the basis of the results we will discuss the issues raised above, and try to formulate conclusions and recommendations for the design of multimodal interaction.

2 The Experimental System

2.1 The bathroom design system

The COMIC system for bathroom design is targeted as a tool for customers in a bathroom shop to get an impression of the available options before they take time from an expert sales person. Thus, it is essential that the system can be used by persons who have no experience with bathroom design without prior instructions. The complete process comprises four phases. In the first phase the user is requested to enter the shape and size of the room, and to indicate the position of the door(s) and window(s). From each door the system needs to know how it opens, and for the window(s) the system needs to know the height of the window sill, the width and the height. In the second phase the system proposes several proven solutions for the selection and placement of sanitary ware, based on the physical possibilities obtained from the information collected in the first phase. In the third phase users can browse through a

collection of tiles and decorations. The users can see the result of their selection immediately in a 3D representation of the room, viewed from a fixed angle. If the user is not satisfied, it is possible to return the previous phase.

The physical design of the system is shown in Figure 1. The user sees a terminal screen and an electronic tablet. The pen can be used to draw, to write and as a point-and-click device. In addition to the pen users can input information by means of speech. The terminal screen shows a naturalistic talking head, which acts as a helpful agent who guides the user through the interaction. The avatar interacts with the user by means of synthetic speech and by facial expressions and head gestures (thinking, listening) that should provide back channel information to control turn taking. The language of the system is (British) English.

In our experiment we have only tested the interaction in phase 1, i.e., the phase in which the data about the room must be entered. Ideally, phase 1 would start with the agent inviting the user to enter the required data, using pen, speech or both, depending on the user's preferences. Since it is quite unlikely that uninformed users know exactly what the system requires, the system must be able to ask follow-up questions, for example where the door is, and how it opens. In the task at hand the system has no means of knowing that the data are complete and correct. The fact that a user has not yet entered data on a window does not mean that the input is not complete, because bathrooms do not need to have windows.

To obtain a system with which most participants could complete the task, it was necessary to guide them so that they stayed within the limitations imposed by the functionality of the recognizers and the dialog manager (Rossignol et al., 2004). We were forced to use a strongly system driven interaction style, in which the system prompts the user to enter individual information elements.

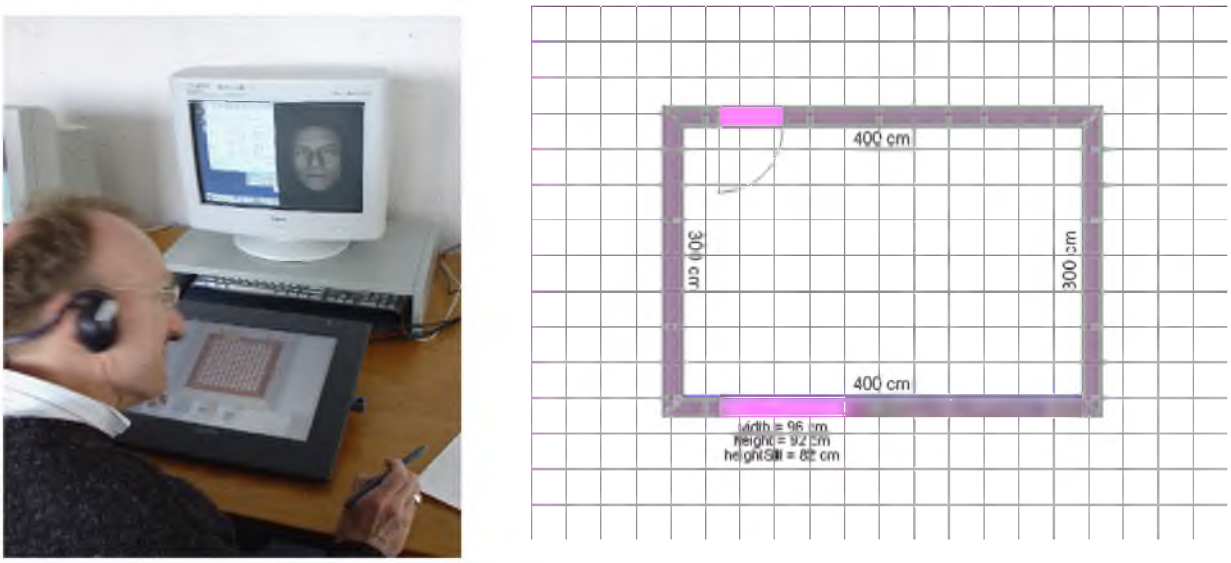

Figure 1: Physical appearance of the Conversational Agent application for bathroom design. Left: user interaction; right: tablet screen showing the eventual recognition result.

2.2 The Paris Tourist Guide System

The Paris Tourist Guide is meant as a tool for preparing visits to Paris (Almeida et al., 2002). The application starts with an overview map of the city centre, from which users can select a Point of Interest. Associated with each PoI is information about opening hours, hotels and restaurants, etc. The interaction strategy is user controlled. This implies that the speech recogniser must always be open to capture input. Obviously, this complicates signal processing and speech recognition. Thus, the application is designed as a "Tap While Talk" system. When the pen is used during or shortly before or after a speech utterance, the two input actions are interpreted in combination. An example is a tap at Notre Dame on the start screen, while saying "Show hotels here". When a user taps one second before the start or after the end of speech, the pen and speech actions are considered to be sequential and independent. The output

information is mainly presented in the form of text (e.g. ”the entrance fee is 3 €”) and graphics (maps and pictures of hotels and restaurants). The text output appears in a text box on the screen (cf. Fig. 2).

To help the user keep track of the system status, the system will always respond to an input. In most cases the response is graphical. For example, when a PoI has been selected, the system will show the corresponding detailed map. If the system detects an ambiguity (e.g. if audio input was detected, but ASR was not able to recognise the input with sufficiently high confidence), it provides an audio prompt saying that it did not understand the utterance.
The Dialogue/Context Manager is designed such that the interaction starts without a focus for the dialogue. Thus, the first action that a user must take is to select a PoI on the overview map. The selected object ten becomes the focus of the dialogue: all deictic pronouns, requests etc. now refer to the selected object. Selection can be made in three ways: by speaking, by pointing, or both simultaneously. Irrespective of the selection mode, the application responds by showing the map that contains the PoI. A selected object is marked by a red frame. Additional active objects on a map are indicated by green frames. When the user has selected a PoI, facilities such as hotels and restaurants can be shown on the maps. This can be done by means of speech (e.g., by asking ‘What hotels are there in this neighbourhood?’), or by tapping on the ‘facility’ button in the menu at the bottom of the screen. The speech modality supports information filtering, i.e. the user can say "Show me the Italian restaurants" to see only the restaurants of this particular type.

Figure 2 shows the buttons that are present in the toolbar of the second version of the GUI. The first four buttons are related to the service: the first button opens a menu with facilities to chose, the second button ends the present interaction, the third button provides context dependent help and the fourth button brings the user back to the overview map. The next two buttons refer to the set-up and closing down of the telephone interaction. The last button refers to the soft keyboard that was not used in this application.

Speech input allows what we call shortcuts. For example, at the top navigation level (where the overview map with POIs is on the screen) the user can ask questions such as ‘What hotels are there near the Notre Dame?’. That request will result in the detailed map of the Notre Dame, with the locations of hotels indicated as selectable objects. However, until one of the hotels is selected, the Notre Dame will be considered as the topic of the dialogue.

Figure 2: Screen Layout of the MUST tourist guide

3 Experimental tasks

In both experiments participants had to complete one or more scenarios. The tasks are summarized in the following sections.

Thus, some operations in this application can be performed by speaking or by tapping on buttons or on an object on the map (e.g. selecting a PoI on the overview map, highlighting a PoI, displaying facilities), but others can only be performed by speech, (e.g. asking for opening hours, entrance fees, or general information about a PoI). We deliberately designed the interface to force the users to use both pen and speech input to be able to study multimodal interaction patterns.

3.1 The bathroom scenario

Participants had to imagine that they were in the process of re-designing their bathroom, and that they were visiting a large bathroom store in which two systems were available that could help them in the design process. They were asked to copy the blueprint, which represents a rectangular room of 2.5 by 3 meters. Exactly in the middle of one wall of 3 m is a door that opens to the inside and which is 85 cm wide. In the middle of the opposite wall is a window that is 100 cm wide, 75 cm high, and with a window sill 120 cm from the floor. Since we wanted to investigate a situation in which users were confronted for the first time with bathroom design systems, we did not offer any practice time.

To complete the scenario a minimum of 14 (complex) information items had to be entered, 4 walls with 4 measures, one window of which the position, the height and width and the height of the window sill had to be entered, the position of the door and the way it opens. For the measures the system wanted to know both the numbers and the units. Therefore, it could take two turns to enter a complete measure. The walls and the positions of the door and window could only be entered by pen. For the measures, users had the choice between speech and handwriting, or a combination of speaking and writing. 20 subjects took part in the experiment, 10 males and 10 females. All were students or staff of Nijmegen University (den Os & Boves, 2004).

3.2 The Paris Tourist Guide

In the experiment with the Tourist Guide the scenarios provided predefined tasks that should be followed by the subjects (e.g. "You consider to visit the Eiffel tower, you want to be early, therefore you want to know the opening hours"). Some of the subtasks could only be accomplished by speaking (e.g., asking for detailed information about a restaurant) while other subtasks could be completed by both pen and speech. Subjects were also told that it was possible to accomplish several actions in a single turn by combining pen and speech. Two versions of the instruction were developed, one a text only explanation, and another one that combined text with examples in the form of video clips. The test scenarios explored all possible interaction ways. A short practice scenario was performed first. Then three test scenarios followed, each divided into two parts, and each part consisted of three tasks (Kvale et al., 2003). User tests took place in Norway, France, and Portugal. The backgrounds of the subjects who participated in the tests were somewhat different for the three countries. In Norway they were employees of Business Units of Telenor (excluding the Research lab), in France they were employees of the of the R&D lab of France Télécom, and in Portugal they were friends and relatives of employees of the R&D lab of Portugal Telecom. In total 46 subjects participated in the user tests: 15 in Norway, 16 in Portugal, and 15 in France. All subjects completed the test (although not all objective measurement data are available from all subjects).

4 Results and Discussion

4.1 The bathroom design system

Table 1 shows the mean durations of complete interactions, as well as mean task completion measure for the bathroom design system. We defined six yes/no decisions: walls present?, wall measures OK?, door present?, door features OK?, window present?, window features OK?. A task completion measure of 6 means that the interaction ended with all items fully correct. If a user was not able to get anything right, zero points were given. Users were also asked to indicate *estimated* durations of an interaction. All participants needed more than the minimum of 3.8 minutes in which a trained user can complete the task, if no recognition errors occur. The mean number of turns is 21.7 (with a range between 15 and 31). The minimum number of turns (if no corrections are necessary) is 14. The extra turns are the result of the corrections that were needed to get the final result. A large proportion of the errors in pen input were due to the fact that subjects were given a fixed amount of time to complete their input. The time window of 8 seconds, set on the basis of a large number of interactions during the development of the system, turned out to be too short for a number of participants. Few participants understood that gestures produced after the end-of-turn detected by the system were not processed. This problem was aggravated by the fact that the electronic ink left in one turn remained on the tablet screen during next turn. This confused users who were requested to re-enter a measure that was not correctly recognized. Often, they just tried to make parts of the handwriting more clear, while the system expected them to re-enter the complete string. In many cases where the system did not get the user's input completely correct, the system's response appeared to confuse the user. This confusion then triggered user behaviour that was beyond the system's capabilities. Unsurprisingly, this unexpected and unmodelled behaviour caused additional interaction errors. These effects are not likely to occur in WoZ settings, or experiments in which participants perform a number of unrelated tasks.

Table 1: Objective duration and standard deviation (min.), estimated duration (min.), and overall task completion (min=0, max=6). N=20.

Objective duration	6.5 (1.8)
Estimated duration	8.2 (3.6)
Task completion	5.1 (0.8)

From the difference between duration and estimated duration it appears that users tend to overestimate the duration of an interaction session. The mean values for task completion show that not all users were able to input all items correctly.

4.2 Mode Switches in the COMIC System

The way the COMIC system was designed left users no freedom in the way they entered the walls and the position of the door and window: these data had to be provided by means of pen drawing. However, users had the choice between speech and pen for entering the measures (four for the walls, and three for the window). Pen and speech recognition were optimized for the task at hand, but that did not prevent recognition errors. Thus, most of the users had to correct one or more recognition errors. The most relevant data are summarized in Table 2.

In Table 2, it can be seen that 7 participants preferred the use of the pen, 12 preferred speech, and one participant did not show a clear preference for the input mode. It is also apparent that 10 participants exclusively used one of the two modes. From the 7 'pen users' 2 switched to speech; from the 12 'speech users' 7 also used the pen. Most of the mode switches were triggered by recognition errors, but it also happened that participants switched input mode for no apparent reason. One person (pp 2) used pen and speech simultaneously three times; it is not clear why he did so. The data also show that some participants (17, 18, 19) essentially stick with speech, despite the large number of recognition problems. The same holds for the pen preference of participants 6. The large number of pen gestures produced by participant 7 is not due to recognition errors. The first time that he tried to enter a measure, he did not manage to complete the unit (centimeter). The system recognized the number (300), and prompted for the unit in the next turn, where 'cm' was correctly recognized. This person adapted immediately to the system: for all subsequent measures he first wrote the number, and waited for the system to prompt for the unit.

Table 2: For 20 users we present gender, usage of pen and speech turns (min number of turns = 7), and mode switches.

User	# Pen	# Speech	Mode Switches	User	# Pen	#Speech	Mode Switches
1 (F)	7	0		11 (F)	0	8	
2 (M)	7	4 3*P+S	1*P→S 1* S→P	12 (F)	0	9	
3 (F)	9	0		13 (F)	1	8	
4 (F)	10	0		14 (F)	2	7	1*S→P 1*P→S
5 (M)	8	3	1*P→S 1*S→P	15 (F)	2	8	1*P→S
6 (M)	11	0		16(M)	3	7	2*P→S 1*S→P
7 (M)	17	0		17(M)	1	14	1*S→P 1*P→S
8 (M)	0	7		18(M)	1	19	1*S→P 1*P→S
9 (M)	0	7		19 (F)	3	18	1*S→P
10(M)	0	8		20 (F)	5	7	1*P→S 1*S→P

Thus, our data confirm previous findings that the general population consists of two or perhaps three different groups with respect to their preference for input mode, when they are given the choice: one with persons who prefer to write, another who prefers to speak, and a probably smaller group of people who do not have a clear preference. Our data do not give a clear picture of the mode switching pattern. First, some subjects do not need to switch, because they did not encounter errors. Others persisted using one mode, despite many errors. Some of the subjects who switched stayed in the other mode, while others returned to the first mode. For the persons who do switch there is no clear pattern as to the point where the switch occurred. Thus, we are led to conclude that persons who interact with a multimodal system and who are given the freedom to use pen or speech (or both simultaneously) do indeed use both input modes, be it in person specific manner.

4.3 The Tourist Guide System

All interactions with the Tourist Guide system were videotaped, all spoken utterances were recorded, and all subject actions that can be perceived by the system as well as the system actions were logged. All turns were annotated by hand, using the system logs and the video tapes. This annotation was mainly meant to classify the turns according to the performance of the speech recognizer.

Statistical analysis of the data proved to be very difficult, mainly for two reasons. Firstly, raw data are available only at the level of individual turns. Because all user actions depend on the state of the system, which is unpredictable if users have several options and if errors occur, it is not obvious how these data can be converted into scores that are meaningful at the level of a scenario, or at the level of an individual subject. We ended up counting numbers of turns per scenario with and without ASR errors, but this representation of the data is still too coarse. For example, it does not allow one to distinguish between ASR errors caused by background noise being interpreted as speech that could not be understood and errors where meaningful input was misrecognized. Secondly, there was a wide range of variation in the interaction between subjects and systems. This variation cannot be controlled, because users are free to choose their approach in this type of user initiative applications. At the same time, the between-subject variation in all conditions (actually even in all scenarios) was so large that it dominated the data.

4.3.1 Distribution of Turn Types

Fig. 3 summarizes the distribution of the turn types (pen, speech, or pen+speech simultaneously) in the two conditions (Text and Video) and the three countries. From the figure it is clear that the proportion of turns in which pen and speech are used simultaneously is rather small. Thus, when left to their own, people do not discover the most efficient and effective way to interact with a multimodal system spontaneously, even if they have received a short instruction in which the optimal strategy was explained. It is also apparent that speech-only turns are by far the most frequent, but that is due to the fact that some actions in the scenarios could only be performed by speech. Thus, it is not possible to draw conclusions about preferences from the data that we have gathered. This is the more so because, as in the experiment with the bathroom design system, users may discover an effective strategy early in the session, after which they stick to the strategy that has proven to work, even if it is not the most efficient one.

Another interesting result that can be seen in Fig. 3 is that there is a substantial difference between the proportions of simultaneous turns in the Text and Video conditions in Norway, but not in the other two countries. Although we tried to make the experimental conditions in the three countries as similar as possible, we still think that the difference between the Norwegian subjects on the one hand and the French and Portuguese on the other is due to differences in the way in which the service was introduced to the subjects. In fact, we believe that the video introduction will be more effective in inducing simultaneous use of speech and pen, but the effect is not very strong, so that it can be overridden by interactions between subjects and experiment leaders.

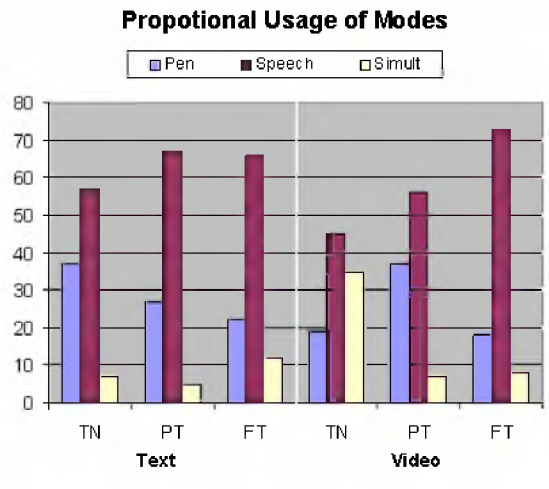

Figure 3: Distribution of turn types in the three countries

Even medium-scale experiments in multimodal interaction generate so many data that labelling by hand is prohibitive. Therefore, there is an urgent need for automatic procedures to annotate multimodal interaction corpora.

5 General Discussion

Although a direct comparison is not possible, it is interesting to note that the bathroom design system, implemented as a Conversational Agent was not evaluated more positively than the Paris Tourist Guide implemented as a Direct Manipulation interface, despite the fact that the agent's help yielded a higher task completion rate in the bathroom design system than the same subjects obtained with a DM interface (den Os & Boves, 2004). This can be explained by the fact that the users of the Tourist Guide did not need assistance in completing the task proper. If they needed help, it was only in using the interface, not the underlying application. In the bathroom design application users did need help with the task per se. Thus, it is

not possible to draw general conclusions about which interaction metaphor should be preferred: the best one can say is that conversational agents have an advantage if users are likely to need assistance in completing their tasks.

All our experiments involved medium-sized groups of subjects. Yet, we have always seen a wide range of different behaviours, as soon as we relinquished tight control over the order in which subjects performed the subtasks needed to reach their final goal. Because subjects had to react to the responses of the systems, they were fully dependent on the system's recognition performance and the system's intelligence in conducting a dialogue and especially in detecting and handling mistakes (whether of the user or the system itself). The unpredictable occurrence of recognition errors increases the number of possible state sequences in a complete task substantially, with the inevitable effect that the number of different behaviours of the users is also increased. The fact that this range of behaviours is already apparent in medium-sized strongly suggests that we will see a range that is at least as wide in the general population of prospective customers. Thus, it is evident that interaction designers must anticipate a large degree of variation in user behaviour. At the positive side it can be concluded that iterative experiments during the design process can be safely conducted with groups of between 20 and 40 participants.

In both experiments we observed that most subjects appear to find an effective strategy for completing their tasks early on in a session, and that most subjects then stick to that strategy, even if it is not the most efficient one. Thus, it is probably not justified to say that a person who starts using speech and who then sticks to the speech mode has a clear preference for that mode. It can also be that this person just sticks to what appeared effective in the first turn. Our data do not allow general conclusions about mode switching. Virtually all switches are caused by recognition errors, but some subjects do not even switch on persistent errors. Here too, the fact that users seem to stick to an interaction strategy that has appeared at least somewhat effective may play a role. Furthermore, since our experiments were focused on natural interaction rather than –for example- multimodal error correction, our fully implemented systems were optimized to minimize the number of errors. Therefore, our subjects were seldom confronted with situations in which they had to make more than two (or exceptionally three) attempts to enter some information element. Yet, it appeared that our users were quite unforgiving of recognition errors. Thus, we must conclude that also multimodal interaction design should focus on avoiding errors, rather than on deploying multimodality to facilitate error correction. And because user behaviour is so variable and difficult to predict, designers should try to provide alternative input modes whenever that is reasonable.

The latter being said, it is still the case that the importance of multimodality appears to depend strongly on the nature of the subtasks that make up a complete task. In both systems that we tested subjects were not completely free to choose an input mode in all dialogue states; rather, some of the subtasks could only be completed by either pen or speech. In all cases the mode that was enforced was so obviously superior, that the obligation to use that mode did not cause problems. This is not to say that interaction designers do not need to worry about the way in which interaction modes are made available. But it is also true that the risk that users are frustrated by the fact that they cannot accomplish a 'pen goal' with speech is small. However, in the Tourist Guide application there were a few goals that could only be accomplished by speaking. Here, participants commented that they might want to avoid speech in environments where they can be overheard. Thus suggests that applications for the general public should be designed in such a manner that they can be used without speaking, even if the ability to use speech would enable shortcuts.

In the process of designing the systems we have experimented with we have seen that seemingly simple functions may be rather difficult to implement and that they may have far-reaching ramifications. The way lengths were handled in the bathroom design system is a good example. One would expect a system to be able to infer the unit (metre or centimetre) from the magnitude of the number. In actual practice, the interpretation of numbers appeared to be context dependent, and implementing the intelligence for full disambiguation would have required a project of its own. Therefore, we had no other choice than have the system prompt for information that in many cases seemed redundant and confused users.

We have seen little simultaneous use of pen and speech in the Tourist Guide service. It seems as if simultaneous use of pen and speech is less 'natural' one would expect. Perhaps this is due to the fact that most users discovered an effective unimodal interaction strategy, and were reluctant to try an alternative. The almost complete absence of simultaneous pen and speech input in the bathroom design system is easier to explain: the task simply did not invite simultaneous input. However this may be, these findings so raise questions about the added value of simultaneous use of pen and speech input in the types of applications that we have investigated.

6 Acknowledgements

We thank all members the COMIC team for their contributions to the work reported in this paper. The COMIC is funded by the EU, under contract IST-2001-32311. Special thanks are due to Stéphane Rossignol and Louis ten Bosch for their assistance in conducting the experiment. We also are indebted to the members of the team that carried out the EURESCOM project MUST, especially Luis Almeida, Malek Boualem and John Rugelbak.

7 References

Almeida, L., Amdal, I., Beires, N., Boualem M., Boves, L., den Os, E. , Filoche, P. , Gomes, R., Knudsen, J.E., Kvale, K., Rugelbak, J., Tallec, C., Warakagoda, N. (2002). The MUST guide to Paris; Implementation and expert evaluation of a multimodal tourist guide to Paris. *Proc. ISCA Workshop Multi-Modal Dialogue in Mobile Environments*, Kloster Irsee June 17 – 19.

Bailenson, J. N., Swinth, K. R., Hoyt, C. L., Persky, S., Dimov, A., and Blascovich, J. (2004). The independent and interactive effects of embodied agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in Immersive Virtual Environments. *PRESENCE: Teleoperators and Virtual Environments.*

Buisine, S., Abrilian, S. & Martin, J.-C. (2004) Evaluation of multimodal behaviour of Embodied Agents. In: Zs. Ruttkay, C. Pelachaud (Eds.), *From Brows to Trust: Evaluating Embodied Conversational Agents*, Dordrecht: Kluwer Academic Publishers, pp. 217-238.

Den Os, E. & Boves, L. (2004) Natural multimodal interaction for design applications. In: P. Cunningham & M. Cunningham (Eds.) *eAdoption and the Knowledge Economy*. Amsterdam – Berlin: IOS Press, pp. 1403 – 1410.

Den Os, E., Boves, L., Rossignol, S. & ten Bosch, L. (submitted) *Conversational Agent or Direct Manipulation in Human-System Interaction, submitted to* Speech Communication

Kehler, A., Martin, J.-C., Cheyer, A. Julia, L., Hobbs, J. and Bear, J. (1998) On representing salience and reference in multimodal human-computer interaction. *AAAI'98, Representations for multimodal human-computer interaction*, Madison, pp. 33-39.

Kvale, K., Rugelbak, J., & Amdal, I, (2003) How do non-expert users exploit simultaneous inputs in multimodal interaction?, *Proc. International Symposium on Human Factors in Telecommunication,* Berlin, 1-4. December 2003.

Oviatt, S.L. Multimodal interfaces. In: J. Jacko and A. Sears (Eds.) *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, Lawrence Erlbaum Assoc., Mahwah, NJ, 2003, chap.14, 286-304.

Oviatt, S.L., Coulston, R. & Lunsford, R. (2004) When do we interact multimodally? Cognitive load and multimodal communication patterns. *Proceedings of the International Conference on Multimodal Interfaces*, Vancouver, B.C., ACM Press, 2004.

Park, K., Kapoor A., Scharver, C., Leigh, J., Exploiting Multiple Perspectives in Tele-Immersion, *Proceedings of IPT 2000: Immersive Projection Technology Workshop,* Ames, IA, 06/19/00-06/20/00.

Rossignol, S., ten Bosch, L., Vuurpijl, L., Neumann, A., Boves, L., den Os, E., and de Ruiter., J.P. (2003) Human Factors issues in multi-modal interaction in complex design tasks. In Proceedings, HCI International 2003.

Sturm, J., Bakx, I., Cranen, B. & Terken, J. (2003) Comparing the Usability of a User Driven and a Mixed Initiative Multimodal Dialogue System for Train Timetable Information. *Proceedings of Eurospeech*, Geneva, Switzerland.

Thórisson, K. (2002) Natural turn-taking needs no manual: Computational theory and model from perception to action. In: B. Granström, D. House & I. Karlsson (Eds.) *Multimodality in Language and speech systems.* Dordrecht: Kluwer Academic, pp. 173-207.

Walker, M.A., Kamm, C. & Litman, D.J. (2000). Towards developing general models of usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue systems.*

Xiao, B., R. Lunsford, R. Coulston, M. Wesson & Oviatt, S.L. (2003) Modeling multimodal integration patterns and performance in seniors: Toward adaptive processing of individual differences, *Proceedings of the International Conference on Multimodal Interfaces*, Vancouver, B.C., ACM Press, pp. 265-272.