Multimodal Interaction – Will Users Tap and Speak Simultaneously?

JOHN RUGELBAK AND KARI HAMNES

Well-designed multimodal interfaces can solve existing user interface problems, particularly for small handheld devices that do not allow mouse or keyboard input. For such devices, the combination of pen and speech input has proved to be efficient and effective, since the two modalities are complementary. With multimodal interaction it is easier to avoid and correct recognition errors, and dialogue completion times can be shortened.

John Rugelbak (55) received his Siv.Ing. degree (MSc) in electrical engineering in 1973 and has been a Telenor R&D employee since 1974. He has been working in the field of subjective and objective user experiments since 1980. Since 1995 he has been a member of Telenor's Speech technology group, working mainly with speech dialogues and user experiments. For the last two years he has been working with multimodal systems. His main research interests are design and usability evaluation of speech and multimodal interfaces.

john.rugelbak@telenor.com

Kari Hamnes (39) is an HCI (Human-Computer Interaction) researcher in the Future Media Group at Telenor R&D. She obtained her MSc in Electrical Engineering and Computer Science from the Norwegian University of Science and Technology (NTNU, formerly NTH) in 1986. She has studied Human-Computer Interaction at University College London (1992–95), focusing on the use of usability guidelines in product development. Her research interests include multi-modal user interfaces, mobile user interfaces and the role of usability in product development processes.

kari.hamnes@telenor.com

The next generation of multimodal interfaces must not only offer increased functionality and efficiency for expert users but must also be user friendly, natural and intuitive for naïve users. But what is natural and intuitive interaction between humans and machines? When humans communicate with each other, we use co-ordinated speech, gestures, body language and facial expressions, and we combine different input senses such as vision and hearing. Communication between humans is by nature multimodal, and it is natural to use different modalities simultaneously and with low effort. Since this is a natural way for humans to communicate, it has been considered to be natural to communicate with machines in the same way.

However, this is not necessarily the case, and a main goal for the EURESCOM project MUST – "Multimodal and Multilingual Services for small Mobile Terminals" [1] has been to obtain knowledge about user behaviour with an application that supports simultaneous coordinated pen and speech interaction.

Simultaneous coordinated multimodal interaction is a term used by the World Wide Web Consortium [2] for the most advanced and powerful form of multimodal interaction, where all available input channels are active simultaneously, and their actions are interpreted in context. For a pen-speech enabled application this means that it is possible for the user to tap while he talks, and that the different modality actions are then interpreted together.

This paper reports the Norwegian part of an expert evaluation that was run prior to the MUST user studies.

The MUST Tourist Guide – A Sample Application Using Simultaneous Pen and Speech Input

In the MUST project Telenor cooperated with researchers from France Telecom, Portugal Telecom, Max Planc Institute and the University of Nijmegen. An important part of the project was to run experiments with the purpose of investigating natural multimodal user interaction. We therefore needed:

- A platform that supported simultaneous co-ordinated multimodal input; and
- A test application/service where the user could complete tasks using different modalities one by one, or simultaneously.

Most experiments that have been run and reported previously have been based on Wizard of Oz platforms. In this project, it was decided to implement a working platform and a demonstrator/application where small experiments could be run with relatively low effort. The MUST PDA based platform is described in detail in [3a], [3b], [3c]. It was decided to implement an electronic map based tourist guide for Paris. The test application and user interface are described in the following.

Maps and Points Of Interest (POIs)

The tourist guide is organized as regional Paris maps (Figure 1), centred around different Points of Interest in Paris, such as Notre Dame, Hotel de Ville, the Eiffel Tower, Sacre Coeur, etc.

From an overview map of Paris (Figure 3) showing all available Points of Interest, the user can navigate to a regional map by tapping the POI or by saying the POI name. To move from one regional map to another, the user can either go via the overview map by tapping a button on the

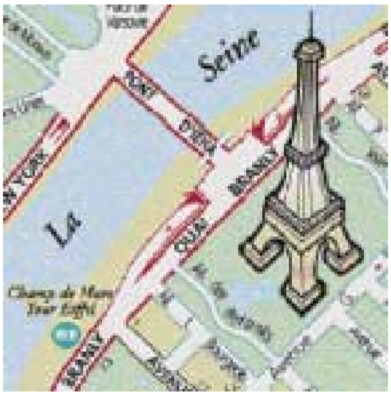


Figure 1 Regional map of Paris

toolbar, or he can use “speech shortcuts” and for example say “Show hotels near Notre Dame”.

Buttons

Figure 2 shows buttons that are present in the PDA’s tool bar. The first two can be used to select a Facility group, e.g. hotels or restaurants. Button number three is used to go back to previous map and the fourth to go to the overview map. The fifth button is used to end the present interaction, and the sixth to request help.

Facilities

On each regional map, the user can use voice or a Tool bar button to display different facilities. Since a main goal for the project was to make a demonstrator to run experiments, rather than to implement a service, the number of POIs, facilities, etc. was limited – but sufficient to run scenario-based experiments. For the first version, only hotels and restaurants were implemented. When voice is used to display facilities, it is possible to select subsets of each facility group, for example by saying “three star hotels” or “cheap Italian restaurants”.

Selecting Objects on the Map

POIs can be selected and made the topic of the dialogue by tapping the pen or by saying the name. The user does not have knowledge about any other objects, and these objects can therefore only be selected and made active by tapping the pen on the object.

Requesting Information About an Object

To request information about an object, the user must use voice. Address, telephone number, which type of food is served at a restaurant, opening hours or a detailed description of a POI are examples of information that can be requested.

“Tap while Talk” Functionality

An important functionality of the tourist guide is that the modalities voice and pen can be used one by one in a serial way, or simultaneously, which can be more efficient. If the user for example wants to select a hotel in the Notre Dame area and request the double room rate, he can either use modalities one by one and go through a four-step procedure:

1 Select the Notre Dame regional map
2 Display a group of facilities (hotels)
3 Select a hotel
4 Request info

or he can use pen and speech simultaneously and tap on an object while he talks:

1 “Show hotels here”
2 “Get double room rate for this”

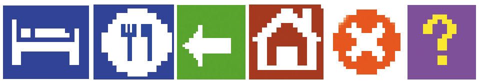


Figure 2 Toolbar buttons

For this experiment we have defined “simultaneous” as “when pen is tapped within the time window one second before “speech detected” till one second after “end of speech”. The two actions are then integrated into one combined action and regarded as one dialogue turn.

Dialogue Strategy and System Output

The overall dialogue strategy is user controlled, in accordance with what is normal for graphic user interfaces. As a consequence of this, the speech recogniser must always be open for speech input. The system’s response to the user is mostly graphics (maps displaying POIs, hotels or restaurants) or text (requested information about objects on the map). Synthetic speech is used to give additional information “we found four such restaurants”, “we found no such hotels”, or to give the user error messages such as “I didn’t understand”.

A Study of Simultaneous Pen and Speech Interaction

Aim of Experiment

The aim of the experiment reported in this paper was twofold:

- To explore the “naturalness” of simultaneous pen and speech interaction; and
- To evaluate a sample application in order to improve its usability prior to subsequent user studies.

The first part of the aim was intended to help identify the main research questions for further study. The second part of the aim was intended to maximise the effect of a larger planned study with potential end-users (novices). By eliminating potential usability problems in the sample application, the study would be able to focus better on issues related to the pen and speech based interaction styles.

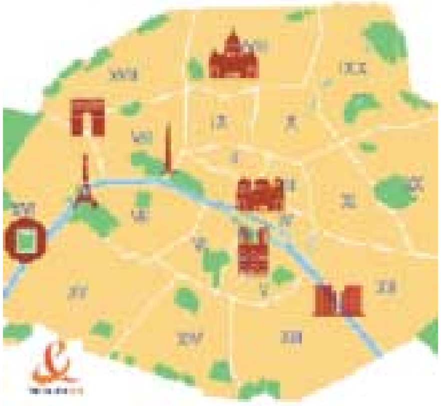


Figure 3 Overview map of Paris

Goal	Action
1 Check opening hours and entrance fee for Eiffel Tower	1 Tap: on *Eiffel Tower* 2 Say: “what are the *opening hours*?” 3 Say: “what is the *entrance fee*?”

Example 1 Example of pre-defined action sequence in Cognitive Walkthrough

This paper focuses mainly on the first part of this aim.

Subjects

In order to gain maximum effect of this study, we chose to use usability/user interface experts as subjects, as these experts would be able to offer well-founded comments with respect to both the naturalness of the interaction style, and the potential usability problems of the sample application.

The study included seven expert subjects, four males and three females. While all subjects had some or extensive experience with pen based interfaces, six were also familiar with speech interfaces. All subjects had been working several years within the field HCI/Usability; five subjects had some or extensive experience designing graphic interfaces.

Cognitive Walkthrough

The experiment was based on the Cognitive Walkthrough (CW) method. CW focuses on ease of learning by exploration and evaluates each step necessary to perform a task. The technique is based on a simplified 4-step model of learning by exploration [4]. Extensive practitioner’s guides to cognitive walkthrough are provided in [5] and [6]. The technique itself will not be described in detail in this paper.

The experts performed the analysis stage of CW by walking through a set of predefined action sequences and recording problems related to interaction style and usability issues for each step of the sequence.

Experimental Procedure

The experimental procedure consisted of five steps:

- Introduction to experiment
- Exploratory phase
- Cognitive Walkthrough introduction
- Cognitive Walkthrough analysis
- Debrief interview

In the Exploratory phase, the subject was first asked to freely explore the prototype and comment on the interaction style or on apparent usability problems. In the second part of the exploration, the subject was asked to perform tasks of the type “Display the local maps for the Montmartre and Hotel de Ville”, “Display hotels near Notre Dame”, or “Find Cuban restaurants near Notre Dame”.

Having read a brief introduction to the Cognitive Walkthrough technique, the subject and the experimenter jointly performed a CW analysis on an example pre-defined action sequence in order to demonstrate how the technique works. The subject then performed a Cognitive Walkthrough analysis for three pre-defined action sequences, and identified problems in the design.

The semi-structured debrief interview focused on a number of pre-defined issues related to naturalness of interaction and usability of the MUST prototype.

Data Collection and Analysis

The Exploratory Phase and the subsequent initial comments, the Cognitive Walkthrough and the semi-structured debrief interview were recorded on audio and video.

The subjects talked aloud during the experiment to elaborate on problems, reasoning and possible design solutions. In addition, they recorded key words on the cognitive walkthrough form provided. A sample sequence is shown in Example 1.

During all sessions, the experimenters made notes of observations and comments from the subject.

The data analysis focused on three main issues, directly related to the two-fold aim of the study:

- Subjects’ observations about the pen and speech interaction styles;
- Subjects’ identification and reasoning about potential usability problems related to the interaction style and the specific MUST application;
- The experimenters’ observation of the subjects’ pen and speech interaction style.

The analysis was qualitative and relied on reviewing audio and video materials along with the subjects’ and the experimenters’ written comments and observations.

The audio/video was used for support in recording the problems and reasoning on the two main issues, and some of the contents was transcribed for exemplification of subjects’ statements.

The videos were reviewed in detail for the Exploratory Phase (the tasks that include simultaneous speech and pen actions) with respect to timing issues.

Observations

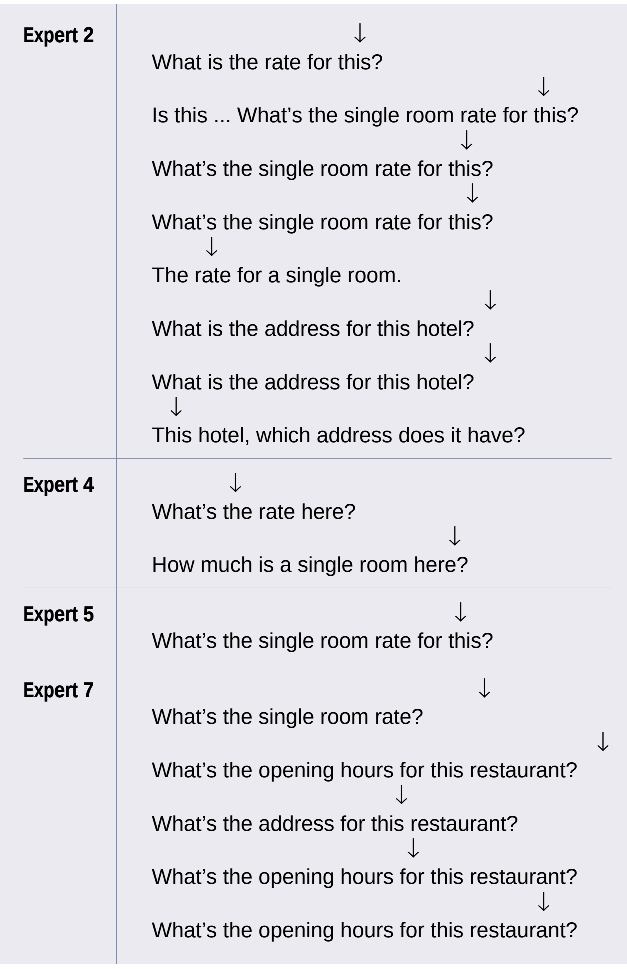
Observations on Interaction Styles During Exploratory Phase

Although the users were told that pen and speech could be used simultaneously – and that this was our focus for the experiment – three of the seven subjects never used pen and speech simultaneously. They used pen and speech as clearly separated actions, and the most typical behaviour was to use pen to select a facility group, a single facility or a POI, and then to use speech to request information.

The typical behaviour of the remaining four was to first use modalities one by one to explore the system for a while, and then try to use both simultaneously.

Observations on Timing During Exploratory Phase

Figure 4 illustrates the timing between pen and speech during the explorative phase. The approximate timing of when the pen is used is indicated with an arrow (↓).



```
Expert 2                          ↓
          What is the rate for this?
                                              ↓
          Is this ... What's the single room rate for this?
                                     ↓
          What's the single room rate for this?
                                      ↓
          What's the single room rate for this?
                 ↓
          The rate for a single room.
                                        ↓
          What is the address for this hotel?
                                        ↓
          What is the address for this hotel?
             ↓
          This hotel, which address does it have?
Expert 4           ↓
          What's the rate here?
                                    ↓
          How much is a single room here?
Expert 5                             ↓
          What's the single room rate for this?
Expert 7                                ↓
          What's the single room rate?
                                                     ↓
          What's the opening hours for this restaurant?
                              ↓
          What's the address for this restaurant?
                                ↓
          What's the opening hours for this restaurant?
                                              ↓
          What's the opening hours for this restaurant?
```

Figure 4 Timing during Exploratory Phase

From Figure 4, we see that

- The users used deictic words like "here", "there" or "this" when they combined pen and speech. 14 out of 16 utterances contained deictic terms and for 12 of these 14, the deictic term was the last or second last word of the utterance.
- The users tended to tap near the end of the sentence (13 out of 16 utterances), but the timing seemed to be even stronger correlated to the use of deictic terms than to the sentence end. Since the deictic term in most cases occurred close to the sentence end, the users almost always tapped close to a deictic expression as well as to the sentence end. However, when the deictic term occurred early in an utterance, the user also tapped early.

Observations on Timing During Cognitive Walkthrough

For all dialogue turns that include both pen and speech, the timing between pen and speech is shown in Figure 5. Note that all sentences in this figure are predefined scripts that the users should follow, but it is up to each user to decide when to tap.

We see that there are some individual differences. It seems that users 1 and 5 prefer to tap a little earlier than the others. (Expert 1 also showed a typical tap-then-talk behaviour in the explorative phase.) However, all experts tend to tap at the end or shortly after the sentence. The largest variance is found for Task 3.1.2. This task involves the only sentence that does not contain deictic words. All other sentences have deictic words at the end. Since the users were scripted to use words like here and there, we have reason to believe that through this we influenced the users' interaction style.

Results from Cognitive Walkthrough Analysis

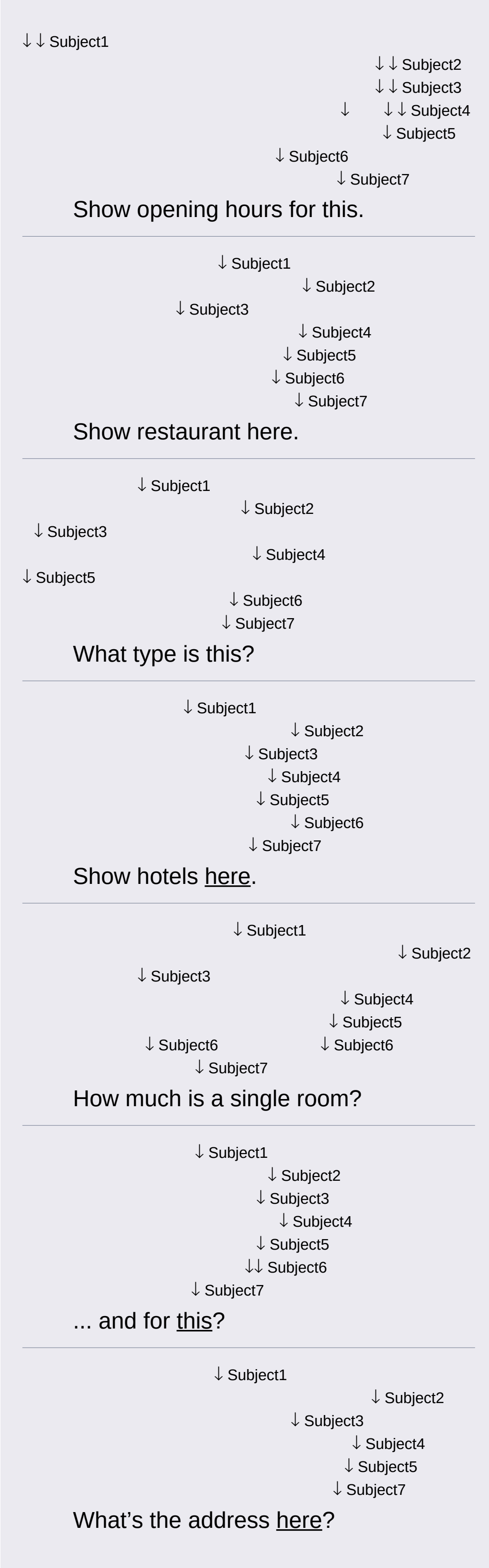
The problems identified by the experts in the Cognitive Walkthrough were classified as belonging to one of 19 main design issues. These issues were further categorized according to the main aims of the evaluation, namely whether they related to the pen and speech interaction styles or to the usability issues specific to the MUST application. This paper will only discuss the six interaction style issues

- Domain knowledge
- Prompting
- Training/Instruction
- "Tap" to select
- Timing
- Speech as shortcut

Domain Knowledge

Several experts commented that the interaction style (particularly the speech part) would be more intuitive and work better in a domain in which the user is familiar. The MUST Tourist

Figure 5 Timing during Cognitive Walkthrough



Guide requires domain knowledge in the form of detailed knowledge of Paris and buildings in Paris, as well as knowledge about what type of information it is possible to get. Knowledge of the domain would also help the user by forming expectations with respect to vocabulary. Yellow Pages (YP) is an example of a "domain" which most users know. They know that YP contains various classes of professions, and contact information for the professionals or businesses, and the users have a fairly good idea of the vocabulary they can use.

Prompting

Several experts commented that prompting could be one strategy for encouraging users to explore the multimodal interaction style. The users could be given hints about the available functionality during the dialogue, for example that it is possible to use speech- or combined speech and pen shortcuts.

Training/Instruction

All experts agreed that without some initial training and instruction, users would probably not use a multimodal interaction style. Indeed, initial training/instructions is a requirement to even understand that the MUST Tourist Guide is multimodal. The screen does not indicate that it is possible to use speech. It is not intuitive that it is possible to use speech at all, and in particular to use pen and speech simultaneously, or to use shortcuts related to objects that are not visible.

"Tap" to select

Several experts commented that PDA users would be more inclined to select objects (e.g. POIs) by tapping, as opposed to selecting objects using speech, due to previous learning. One expert commented that she would probably tap, tap, tap – until there are no more choices, and then try to speak.

Another comment was that when one has a limited domain, and does not exactly know which alternatives are available, a PC or PDA user is used to tapping or using the mouse again and again, to narrow the "search space".

Timing

The experts commented on the timing issue of simultaneous coordinated input. In general, they appreciated the functionality and indeed felt that it was quite natural after having used it for a little while. However, they felt that users would be unsure about when they would have to tap in relation to what they said. Many of the experts said that it felt more natural to tap towards the end of the sentence. Several experts said they would feel it as an unwanted restriction, if they had to tap exactly during speech. A user-friendly system should therefore be flexible regarding when the user is allowed to tap: The system should allow the user to tap during speech, as well as shortly before or after.

Speech as Shortcut

It was mentioned that PDA users would be more likely to tap, in general, but that speech/multimodal interaction could have a potential as shortcuts to specific data. It is however not intuitive that one can request information about objects that are not visible.

Discussion

Naturalness of Simultaneous Pen and Speech Interaction

Since only seven experts participated in this evaluation, results should be interpreted with due caution. The most noteworthy observations will be discussed here.

During the Exploratory Phase of the evaluation, most experts started to use the two input modalities one by one, and some of them never tried to use them simultaneously. After a while, four of the seven experts started to use pen and speech simultaneously.

Timing between speech and pointing has been studied in other experiments, e.g. [7] and [8].
In the expert evaluation we observed that the experts typically tapped at the end or shortly after the utterance. This was especially the case when the utterance ended with deictic expressions like ‘here’ or ‘there’. If no deictic expressions were present, tapping often occurred somewhat earlier. Timing relations between speech and pointing will be investigated in more detail in the user evaluation experiment that is now being designed.

The results from the Exploratory Phase indicate that frequent PC and PDA users are so accustomed to use a single modality (pen or mouse) to select objects or navigate through menus to narrow down the search space, that even if they are told that it is possible to use speech and pen simultaneously, they will have to go through a learning process to get accustomed to the new simultaneous coordinated multimodal interaction style. But once they have discovered and experienced it, the learning curve appears to be quite steep.

It was not intuitive and obvious that the interface was multimodal, and in particular that the two modalities could be used simultaneously. This indicates that for the naïve user evaluation we should pay much attention to the introduction phase where we explain the service and the interface to the user.

During the expert evaluation many usability issues were revealed. They can be divided into interaction style issues and issues that are specific for the MUST tourist guide. The MUST guide specific issues were mainly related to buttons, feedback, prompts, the way selected objects were highlighted, and the location of the POIs on the screen. Most of the problems can be solved rather easily. The comments from the experts gave helpful advice to improve the graphic interface and button-design for the second version of the demonstrator that will be used for the user evaluation experiments.

Almost all experts agreed that without some initial training and instruction, the users would probably not intuitively use a simultaneous multimodal interaction style. They also believed that the users would probably be able to use such an interaction style with small cognitive effort, once they are aware of the systems capabilities. This is also supported by our observations of the experts’ behaviour during the Exploratory Phase.

With the present lack of multimodal applications for the general public, there is a need to introduce the capabilities of simultaneous coordinated interaction explicitly before customers start using the new products. According to the experts a short video or animation would be suitable for this purpose. The introduction that is given to the users before they start to use the tourist guide will be the main parameter in this experiment.

In the introduction to the explorative phase, the experts were explicitly instructed to use the two modalities both one by one and simultaneously. Still only four experts used simultaneous interaction, and only 16 out of approximately 250 to 300 dialogue turns were “simultaneous” (contained both pen and speech). The far most typical interaction style was to use modalities one by one. There are several possible explanations for this, such as the fact that users are accustomed to operating graphic interfaces in a serial manner. Another possible explanation is the cognitive load associated with pointing and speaking simultaneously. During inter-human dialogue, speech and pointing actions are occurring simultaneously, obviously without effort. This also includes the use of available aids such as pencils and pointers. In [9] these pointing actions are denoted “Natural Pointing”. Simultaneous multimodal systems have made it possible to simulate gestures and pointing found in inter-human communication. However, for these systems, the user must also touch a small object on a screen (“Tactile Pointing”). If the user speaks and uses Tactile Pointing simultaneously, it is likely that there will be a resource competition between talking and pointing, and that this cognitive load is sufficiently large to influence the user’s choice of interaction style (use modalities simultaneously or one by one).

Conclusions and Future Work

The main goal for this experiment has been to identify research issues to be studied further in a planned user experiment within the MUST project. Seven experts in the fields HCI and Usability participated in an experiment supporting simultaneous pen and speech input.

The main conclusions and topics for further study were that:

- This is a new way of interacting with machines, and the users will need an introduction to understand or be aware of this new functionality (that it is possible to both tap and speak, and particularly that it is possible to do both simultaneously). An animated instruction (video) showing "how to do it" may be more effective than text.

- It is not intuitive or natural for new users to tap and speak simultaneously. They are used to operating PCs, PDAs etc. in a sequential way, and the typical behaviour will probably be to tap, tap, tap etc. and then speak. Even when they are aware of the simultaneous functionality, they may choose to use the interface sequentially, because of the larger cognitive load. But users seem to "learn" quickly and the cognitive load will be smaller when users become expert users. Since speech centric multimodal interfaces are new, there is little research data on the mental effort the user spends in processing multimodal input, and we see this as an interesting research question.

- When users tap and speak simultaneously, and the utterance contains a deictic word, there seems to be a strong timing relation between pen and the deictic word. If deictic words are actively used in the introduction and system prompts, it may be possible to influence the users' pen timing and interaction style, since the users will probably mimic words used by the system.

The experts agreed that multimodal pen and speech systems have a great potential, and that users can and will use such interfaces. To what extent users will tap and speak simultaneously will depend on at least three issues:

1. Whether they will continue to use the serial interaction style they are used to when they operate graphical interfaces (PCs, PDAs).

2. Whether the cognitive load associated with using two modalities simultaneously is sufficiently low. If not, the users may prefer to use modalities one by one.

3. The application and how much there is to gain by using pen and speech simultaneously. If a user finds simultaneous interaction style more efficient – or maybe a must, he probably can and will use pen and speech simultaneously, even if he normally prefers to use modalities one by one.

References

1 *EURESCOM Project p1104*. 2003, June 18 [online] – URL: http://www.eurescom.de/public/projects/p1100-series/p1104

2 *Multimodal Requirements for Voice Markup Languages*. W3C Working Draft, 10 July 2000. 2003, June 18 [online] – URL: http://www.w3.org/TR/multimodal-reqs

3a Almeida, L et al. The MUST guide to Paris : Implementation and expert evaluation of a multimodal tourist guide to Paris. *Proc. ISCA tutorial and research workshop on Multi-modal dialogue in Mobile environments (IDS2002)*, Kloster Irsee, Germany, 2002.

3b Almeida, L et al. Implementing and evaluating a multimodal tourist guide. *Proc. International CLASS workshop on natural, intelligent and effective interaction in Multimodal dialog system*, Copenhagen, Denmark, 28–29 June 2002.

3c Almeida, L et al. User friendly multimodal services, a MUST for UMTS. *EURESCOM Summit*, Heidelberg, 21–24 October 2002.

4 Polson, P G, Lewis, C. Theory-based design for easily learned interfaces. *Human-Computer Interaction*, 5 (2 & 3), 191–220, 1990.

5 Wharton, C et al. The cognitive walkthrough method: A practitioner's guide. In: Nielsen, J, Mack, R L (eds.). *Usability inspection methods*. New York, NY, John Wiley, 1994.

6 Lewis, C, Wharton, C. Cognitive walkthroughs. In: Helander, M, Landauer, T K, Prabhu, P (eds.). *Handbook of Human-Computer Interaction*. New York, Elsevier, 1997.

7 Martin, J-C, Braffort, A, Gherbi, R. 2000. Measurement of cooperations between pointing gestures and constrained speech during human-computer interaction. *3rd International Conference on Methods and Techniques in Behavioral Research "Measuring Behavior 2000"*, Nijmegen, The Netherlands, 15–18 August 2000.

8 Kehler, A et al. 1998. On Representing Salience and Reference in Multimodal Human-Computer Interaction. *Proceedings of the AAAI'98 workshop on Representations for Multi-modal Human-Computer Interaction*, Madison, Wisconsin, 26–27 July 1998.

9 Schmauks, D. Natural and simulated pointing. *Proceedings of the 3rd European ACL*, Copenhagen, Denmark, 1987, 179–185.