How do non-expert users exploit simultaneous inputs in multimodal interaction?

Knut Kvale, John Rugelbak and Ingunn Amdal[1]
Telenor R&D,
Norway
knut.kvale@telenor.com , john.rugelbak@telenor.com , ingunn.amdal@tele.ntnu.no

Abstract

This paper evaluates scenario based user tests of speech-centric multimodal interaction on a small mobile terminal. Non-expert users solved tasks in the tourist guide domain using a functional multimodal PDA-based application. The tasks required pen and speech input, but the users were free to choose either sequential or simultaneous pen and speech input at each step in the dialogue.

Multimodal interfaces are still a novelty to most users so we had to explain this functionality to the test users. The format of the introduction had a noticeable effect on user behaviour. Users who had seen a video demonstration used simultaneous pen and speech input more often than the users who had had a text only introduction even if the same information was present in both formats. 9 of 14 subjects who had seen the video demo, applied simultaneous pen and speech input instantly. We therefore claim that people will use simultaneous multimodal input when they have been properly introduced to this functionality. However, simultaneous use of pen and speech may impose an extra cognitive load, at least until people get familiar with this kind of interface.

The users considered the multimodal interaction attractive and expressed that they enjoyed the freedom of choosing input mode at each step in the dialogue.

Key words: Multimodal interaction, small mobile terminals, user behaviour

1. Introduction

Multimodal human-computer interfaces give the user the opportunity to choose the most natural interaction pattern depending on context and application. Multiple input and output modalities can be combined in several ways. Here we distinguish between combining the multimodal inputs sequentially or simultaneously. Systems allowing either mode have several parallel input channels active at the same time.

In a sequential multimodal system only one of the input channels is interpreted (e.g. the first input at each dialogue stage). In a simultaneous multimodal system all inputs within a given time window are interpreted jointly depending on the fusion of the partial information from the different input channels.

In interaction between humans simultaneous input is natural, but it is by far the most complicated scenario to implement for human-computer interaction, especially in mobile terminals. Before huge investments are spent on implementing simultaneous pen and speech functionality in future mobile terminals and networks, we need to study how people actually interact with a mobile terminal. In this paper we report experiments on a system that allows simultaneous multimodal input, to study how users exploit simultaneous pen and speech

[1] Ingunn Amdal is currently at The Norwegian University of Science and Technology

functionality. The mobile terminals used in our study had limited size and processing power, and we therefore restricted the functionality to speech centric multimodal interfaces with two input modes: Speech (audio) and touch, and two output modes: Audio and graphics/text.
In this work we cooperated with researchers at France Télécom, Portugal Telecom, Max Planck Institute for Psycholinguistics, and the University of Nijmegen in the EURESCOM - project called MUST – "Multimodal and Multilingual Services for Small Mobile Terminals". We designed and implemented a test-platform for speech-centric multimodal interaction with small mobile terminals, offering the possibility of simultaneous pen and speech input [1,2,3,4].
This paper reports the Norwegian part of the MUST experiments and is organized as follows: Section 2 provides an overview of the experimental set-up. Section 3 provides the results from the user tests. The data are discussed in section 4 and conclusions are provided in section 5.

2. Experimental set-up

2.1. The platform

The Norwegian part of the MUST project multimodal test platform is based on the Telenor R&D voice platform [5]. The Automatic Speech Recognition (ASR) used Philips SpeechPearl® 2000 for Norwegian with a fixed 65 word open grammar covering 10 concepts. For Norwegian Text-to-Speech Synthesis (TTS) we used Telenor R&D's Talsmann®.
The GUI Client (i.e. the small mobile terminal) runs on a Compaq iPAQ Pocket PC running Microsoft CE 3.0/2002. The modules communicate asynchronously by message passing through a Hub that distributes the messages according to a set of rules in accordance with the service logic. More technical details of the platform are provided in [1,2,3,4,5].
The experiment was carried out in a quiet room to reduce the effects of background noise on the ASR performance. The PDA was placed on a table in a cradle to facilitate video recording of the interactions.

2.2. The application

For some tasks it seems natural and almost necessary to use both pen and speech simultaneously. A classical example is "Put That There" in Bolt's concept demonstrator [6]. In a map navigation task we may naturally ask, "What's the distance from here to there" and tap on map locations as we say "here" and "there".
In the MUST project we chose a "Tourist guide to Paris" application [1,2,3]. This service required use of pen and speech actions to accomplish the tasks, but the users were free to interact either sequentially, i.e. to tap with the pen first and then talk, or simultaneously, defined as a pen action in the time window from one second before start of speech to one second after end of speech. The tasks could be completed in many different ways, and it was not obvious which approach was most efficient, i.e. the users did not necessarily benefit much from using simultaneous input. In this way we wanted to explore whether users prefer simultaneous to sequential input mode.

2.3. The user interface

The graphical part of the user interface consists of two types of maps: An overview map showing all Points Of Interest (POI), such as the Eiffel Tower, and detailed maps with a POI in the centre, see Figure 1.

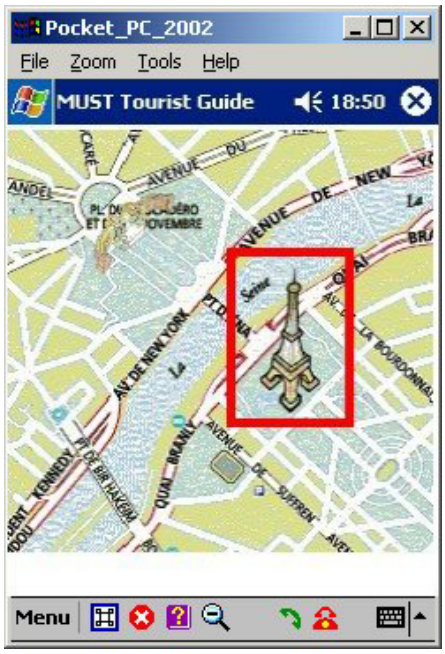

Figure 1: The PDA-screen layout of the MUST tourist guide showing a detailed map with the Eiffel Tower as the selected POI.

The application initially shows the overview map without a focus for the dialogue. The user must first select a POI by pen, speech, or both pen and speech. Using speech it is possible to go directly from one detailed map to another. When the user has selected a POI, several facilities such as hotels and restaurants can be shown as objects on the detailed maps. This can be accomplished by asking a question such as 'Which hotels are in this neighbourhood?', or by a tap on the 'facility' button at the bottom of the screen. To select a specific facility (hotel, restaurant etc) the user must use the pen. The selected object (i.e. POI or facility) becomes the focus of the dialogue and all subsequent requests refer to this object. Information about selected objects can be obtained by asking e.g.: 'What's the address?'. It is also possible to ask questions that filter out special attributes, e.g. 'Show me the Italian restaurants?', to select only the restaurants of a particular type.

2.4. The test subjects

We recruited 13 men and 8 women from the business units of Telenor (i.e. excluding the Research lab) through internal advertisements that briefly explained the project and service. All subjects used a PC in their daily work and described themselves as positive to trying new technology. 6 subjects used PDA regularly, 8 used PDA sometimes, and 7 had never used a PDA. When asked about level of education, with alternatives "low", "medium" and "high", 9 answered medium and 12 high education. Thus, the subjects were not "naïve users" representing the mass market. They were, however, "non-expert users" having no previous experience with multimodal interfaces.

2.5. Explaining the service to non-expert users

The subjects were briefly introduced to the MUST-project and the purpose of the test. Then the experimental procedure was explained including the scenarios and what to do when speech recognition errors occurred. Since we wanted to keep the test conditions equal for all

subjects the supervisor was not allowed to interfere during scenarios (between scenarios some help was allowed). If the application stopped due to technical problems the supervisor restarted the system. The subjects were told that they could solve recognition problems by repeating what they had said or by using other words.
To study the effects of different introduction formats the users were divided into three groups, each of seven subjects. The groups were introduced to the service in one of the following ways:

- "Text_only": The subjects got a textual introduction on the PDA.
- "Video": The last part of the text introduction, explaining how to combine pen and speech, see below, was replaced by a short video showing how to operate the service using simultaneous speech and pen input in exactly the same wording as the text version.
- "Video_short": The instructions in the video were changed to a less verbose speaking style (for example just saying "Four star hotels" instead of "Show four star hotels here").

The last part of text introduction that was replaced with video:

> **How to combine pen and speech**
> Pen and speech can either be used:
> - after each other, or
> - at the same time (simultaneously)
>
> You can for example say:"show four star hotels here" while you at the same time tap on a Point of Interest.
> Or you can say: "what is the single room rate" while you tap on a hotel on the map

The "Video_short" introduction was recorded after a preliminary analysis of the two other versions. The sentence error rate was higher for the "Video" group than the "Text_only" group and our impression from the preliminary analysis was that the "Video" group used a slightly more verbose speaking style. We therefore made an additional video introduction using a less verbose speaking style, called "Video_short". The analysis in this paper is based on all three versions. Note that all participants got the same information; we only changed the format of some part of the information. Thus, all participants should know the multimodal functionality of the terminal.

2.6. Three scenarios

After a short introduction and training scenario, the subjects conducted three larger scenarios. The scenarios had identical structure and consisted of 2 parts each containing three tasks. Part A of each scenario starts at the overview map and the three tasks are: Display a group of facilities located near a certain POI; get information about two of the displayed facilities and get additional information about one of the facilities. Part B of each scenario continues from part A (i.e. at a detailed map) and the three tasks are: Display a group of facilities near another POI not shown on the detailed map; get info about one of the facilities; and get information about the POI.

3. Results

In the following data analysis of dialog turns we have:

- Ignored data from out of scenario actions (especially one person did not always adhere to the scenarios).
- Ignored the effects of ASR errors.
- Ignored dialogue turns that were used to handle/correct situations caused by ASR-errors

We will characterize users who never or accidentally applied pen and speech simultaneously once or twice as being "sequential users", and users who applied pen and speech simultaneously typically 15 times or more as being "simultaneous users".

3.1. Impact of introduction

The introduction format had a clear impact on the number of users using simultaneous input, as shown in Table 1.

Table 1: Number of users using simultaneous versus sequential input style as a function of introductions.

Introduction	Simultaneous users	Sequential users
Text_only	1	6
Video	5	2
Video_short	4	3
Total	10	11

The introduction format also had a clear influence on the total number of simultaneous pen and speech inputs. The distribution of dialogue turn types is shown in Figure 2. Only 5% of all turns in the text group are simultaneous inputs (all due to one user), while the corresponding figures are 28% and 31% for the two video groups. We observed no major differences between the video groups regarding interaction style.

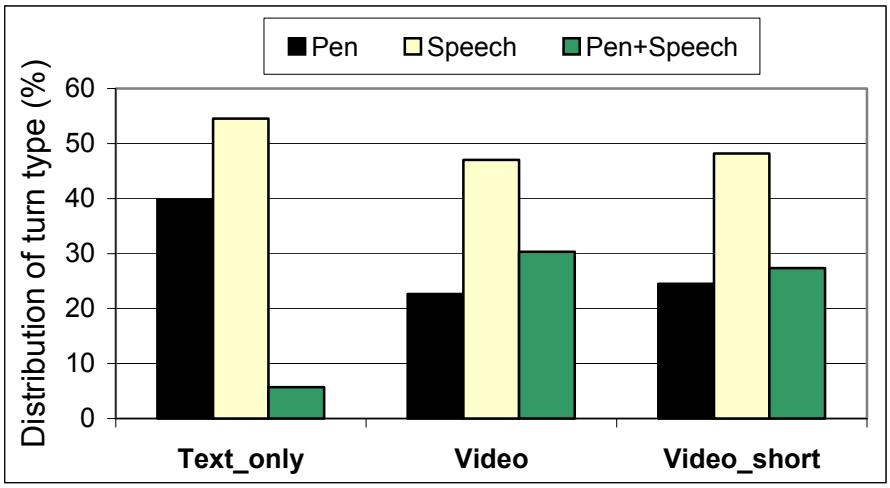

Figure 2: Distribution of dialogue turn types for different introductions.

The average sentence error rate for the group with text introduction was 18%, the "video_short" group had 21 % and the "video" group had 25% sentence error rate.
The average utterance duration for the text, video and video_short groups were 1.8, 2.0 and 2.0 seconds respectively. The variation in utterance duration was large: The fastest speaker on average used 1.2 seconds per utterance whereas the slowest speaker used 2.8 seconds.

3.2. Sequential versus simultaneous mode

Figure 3 illustrates that “simultaneous users” on average only needed 30 turns to solve the three tasks, whereas “sequential users” used 45 actions on average. We observe that half of the turns for the "simultaneous users" were simultaneous and that still a substantial part of the turns were speech only turns.

The differences between “sequential users” and “simultaneous users” did not show any correlation with gender, education level, or PDA experience.

Average utterance durations were similar when using simultaneous pen and speech as when speaking only. This observation is in contrast to other experiments, e.g. Oviatt [7], who claimed that multimodal language is substantially simplified compared to “unimodal forms”, i.e. speech. The average sentence error rate was 18% for sequential users and 25% for simultaneous users.

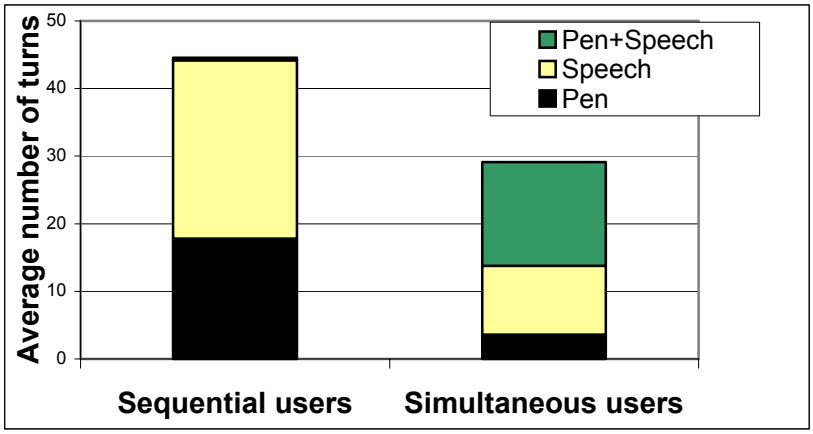

Figure 3: Average number of turns used to complete all three scenarios.

3.3. Learning effects

Since the three scenarios had exactly the same structure, it was possible to observe user behaviour over time. The total number of dialogue turns used to complete one scenario is almost constant over the three scenarios for both sequential and simultaneous users. Figure 4 shows the number of simultaneous pen and speech input turns used in scenario parts A and B over time (excluding the training scenario). We observe that simultaneous use of pen and speech does not increase over time. There is rather an opposite tendency, particularly for part B of the scenarios.

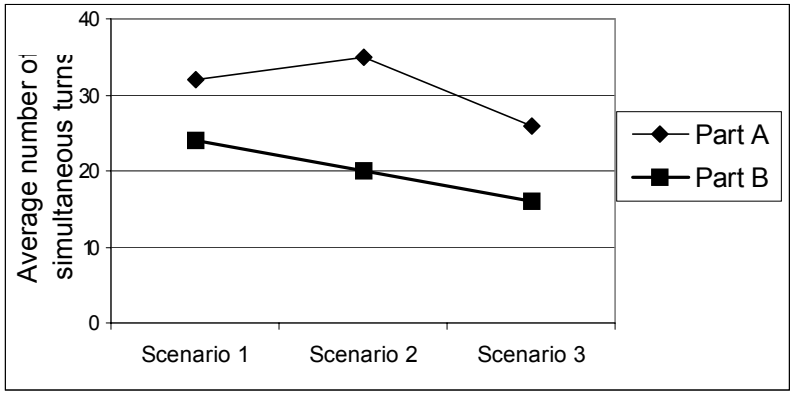

Figure 4: Average number of simultaneous inputs used in scenario part A and B as a function of scenario number.

4. Discussion

In this paper we have studied preferred user interaction styles, rather than effectiveness of a particular service. In our experimental service the interaction was user driven and the users could spend variable time between turns, e.g. studying the graphical outputs to make sure what the next step in the scenario should be. The users were instructed to carry out the scenarios as specified, and were not told to complete the scenarios as quickly as possible. The supervisor was always present and helped if the application stopped due to technical problems.

Experts on human-machine interfaces reviewed an earlier version of our service [1,2,3]. They concluded that users would need an introduction to multimodality since this is new technology and PC/PDA users are in the habit of tap or click sequentially. The experts expected that none or very few would use pen and speech simultaneously unless they were explicitly told that this was technically possible. In our experiment, the users were introduced to the new type of interaction. The main variable in this experiment was that one part of the introduction (focusing on simultaneousness) was given as text or video.

We observed large individual differences in user behaviour. The limited size of the experiment and the large within-group variance made it difficult to discover significant differences in the data. However, there were a number of interesting tendencies in the data that are discussed below.

The introduction format had a significant effect on the number of subjects using pen and speech simultaneously and on the total number of simultaneous turns. Video is a far more powerful medium than text. 9 out of 14 subjects in the two video groups used simultaneous input, whereas only 1 of 7 in the text group did. This supports the expert's hypothesis: People are so used to operate graphical interfaces in a sequential way, that they need a proper introduction to utilize the new functionality with simultaneous multimodal inputs.

Using pen and speech simultaneously seems to introduce an extra cognitive load. We had the impression that simultaneous users were slightly more verbose, that they spoke more staccato and that there were more hesitations. Sentence error rate was higher (25% versus 18%), average utterance duration was longer (2,1 versus 1,8 seconds) and the number of simultaneous turns used in each scenario decreased over time. We expect that these effects will be less significant when the new interaction style is learned and "automated".

All users seemed to combine pen and speech input either sequentially or simultaneously with low effort, and 9 out of the 14 who saw one of the videos exploited simultaneous input immediately. Although there were individual differences regarding interaction style, number of turns and time spent to complete the scenarios, all subjects were positive to using this type of interface. To what extent users will apply simultaneous interaction style over time will probably depend on the "automation effect" and on the given application.

When the scenario testing was finished, the users were given a questionnaire. All users were positive to the new interface. There was no difference between sequential and simultaneous users. A high score was given to statements like: "I found it easy to use this service", "It felt natural to combine pen and speech", "If this type of service will be offered in the future (also for other cities and for a reasonable price), I will certainly use it".

5. Conclusions

In this paper we have studied how users exploit simultaneous pen and speech functionality solving tourist guide tasks on a PDA. The tasks to be solved in this study could be completed in many different ways, and it was not obvious which approach would be most efficient, i.e. the users did not necessarily benefit from using simultaneous compared to sequential pen and

speech input. 21 non-expert users were introduced to a multimodal interface for the first time. 9 out of the 14 users who saw a video introduction immediately exploited simultaneous pen and speech input. We therefore claim that people will use simultaneous pen and speech input when they have been properly introduced to this functionality.
Since users seem to be able to handle simultaneous pen and speech input with rather low effort, we believe they can and will actively use this interaction style when they find it necessary or efficient and effective. This will depend on application, context and users domain knowledge. Another argument for implementing interfaces capable of simultaneous inputs, is that subjects appreciated the freedom of choosing input style at each step of the dialogue.

6. Acknowledgements

We would like to thank our colleagues in the MUST project and in the Speech Technology Group at Telenor R&D for valuable discussions and cooperation.
The recording of the video_short set-up and parts of the user behaviour analysis was carried out and financed by the research program "Knowledge development for Norwegian language technology" (KUNSTI) of the Norwegian Research Council.

7. References

[1] 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 (ITRW) on Multi-Modal Dialogue in Mobile Environments, (IDS 2002), pp. 49-51, Kloster Irsee, Germany, 2002.
[2] Almeida L, et al., "User friendly multimodal services, - A MUST for UMTS". In: Proc. EURESCOM summit 2002, Heidelberg, Germany, Oct 2002.
[3] Almeida, L. et al. "Implementing and Evaluating a Multimodal Tourist Guide", Proc. International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, pp.1-7, Copenhagen, Denmark, , 2002.
[4] Kvale, K., Warakagoda, N.D. and Knudsen, J.E., "Speech centric multimodal interfaces for mobile communication systems", to be published in Telektronikk, 2003.
[5] Knudsen, J.E., Johansen, F.T. and Rugelbak, J., "Tabulib 1.4 Reference Manual", Telenor R&D N 36/2000, 2000.
[6] Bolt, R., "Put That There: Voice and Gesture at the Graphics Interface", Computer Graphics, 14(3), pp 262-270,1980.
[7] Oviatt, S., "Ten Myths of Multimodal Interaction", Communications of the ACM, 1999, Vol. 42, No. 11, pp. 74-81.