Implementing and evaluating a multimodal and multilingual tourist guide

Luis Almeida* (1), Ingunn Amdal (2), Nuno Beires (1), Malek Boualem (3), Lou Boves (4), Els den Os (5), Pascal Filoche (3), Rui Gomes (1), Jan Eikeset Knudsen (2), Knut Kvale (2), John Rugelbak (2), Claude Tallec (3), Narada Warakagoda (2)
* Authors in alphabetic order

(1) Portugal Telecom Inovação,
(2) Telenor R&D,
(3) France Télécom R&D,
(4) University of Nijmegen,
(5) Max Planck Institute for Psycholinguistics

E-Mail: els.denos@mpi.nl

Abstract

This paper presents the EURESCOM[1] project MUST, (MUltimodal, multilingual information Services for small mobile Terminals). The project started in February 2001 and will last till the end of 2002. Based on existing technologies and platforms a multimodal demonstrator (the MUST tourist guide to Paris) has been implemented. This demonstrator uses speech and pen (pointing) for input, and speech, text, and graphics for output. In addition a multilingual Question/ Answering system has been integrated to handle out of domain requests. The paper focuses on the implementation of the demonstrator. The real-time demonstrator was used for evaluations performed by usability experts. The results of this evaluation are also discussed.

Introduction

For Telecom Operators and Service Providers it is essential to stimulate the widest possible use of the future UMTS networks. Wide usage presupposes that services fulfil at least two requirements: customers must have the feeling that the service offers more or better functionality than existing alternatives, and the service must have a easy and natural interface. Especially the latter requirement is difficult to fulfil with the interaction capabilities of the small lightweight mobile handsets. Terminals that combine speech and pen at the input side, and text, graphics, and audio at the output side in a small form factor, promise to offer a platform for the design of multimodal interfaces that should overcome the usability problems. However, the combination of multiple input and output modes in a single session appears to pose new technological and human factors problems of its own. The research departments of three Telecom Operators collaborate with two academic institutes in the EURESCOM project MUST (Boves & den Os, 2002)[1]. The main aims of MUST are:

1. Getting hands-on experience by integrating existing speech and language technologies into an experimental multimodal interface to a realistic real-time demonstrator in order to get a better understanding of the issues that will be important for future multimodal and multilingual services in the mobile networks accessed from small terminals.
2. Use this demonstrator to conduct human factor experiments with naive non-professional users to evaluate the multimodal interaction.

Multimodal interaction has been studied for several years, see e.g. (Oviatt, 1999 and Oviatt et al, 2000). Most papers on user studies report experiments that were carried out with Wizard-of-Oz systems and professional users who manipulated objects on large terminal screens (Kehler et al., 1998, Martin et al., 1998, and Wahlster et al., 2001). For the Telecom Operators these studies

[1] Updated information from the MUST-project can be found at http://www.eurescom.de/public/projects/P1100-series/p1104/default.asp

are of interest in so far that they indicate some of the general principles of multimodal interaction. However, Telcos can only start to consider developing multimodal services if these can be built on standard architectures and off-the-shelf components, that work in real-time and that can be accessed from small mobile terminals by non-professional users. Therefore, the MUST project is focused on a user study with a real-time demonstrator of what could become a real service.
In addition, a large part of the existing literature is based on experiments that address issues such as the preference for specific modes for error repair and comparisons of several combinations of modes (including unimodal interaction). In MUST we concentrate on gathering knowledge about behaviour of untrained users interacting with one –carefully designed- multimodal system that is virtually impossible to use without combining speech and pen for input.
In this paper we first present the functionality of the demonstrator service that served as the backbone of the MUST project. Then we describe the architecture, and the user interface. Finally, we present the results of an expert evaluation of the first operational version of the demonstrator.

1 The functionality of the demonstrator

Multimodal interaction comes in several forms that imply different functionalities for the user. In MUST we decided to investigate the most powerful approach, i.e. simultaneous coordinated multimodal interaction[2]. We want to provide Telecom Operators with information on what this type of interaction implies in terms of implementation effort and on how users will appreciate this new way of interaction.
Only some of the services that one might want to develop for the mobile Internet networks lend itself naturally to the use of simultaneous coordinated interaction combining speech and text input. A necessary requirement for such a service is the need to talk about objects that can be identified by pointing at them on the screen. One family of services where pointing and speaking can be complementary is when a user is required to talk about objects on a map. This probably explains why multimodal map services have been so popular in the research community (Oviatt, et al, 2000; Martin et al., 1998). Tourist guides that are organised around detailed maps of small sections of a city are an example of this family of services. Therefore, we decided to model the MUST demonstrator service after this metaphor. Paris was selected as the object city.
Thus, the MUST Guide to Paris is organized in the form of small sections of the town around "Points of Interests" (POI's), such as the Eiffel tower, the Arc de Triumph, etc. These POI's are the major entry point for navigation. The maps show not only the street plan, but also pictorial representations of major buildings, monuments, etc. When the user selects one of the POI's, a detailed map of the surroundings of that object is displayed on the screen of the terminal (cf. Fig. 2). Many map sections will contain additional objects that might be of interest to the visitor. By pointing at these objects on the screen they become the topic of the conversation, and the user can ask questions about these objects, for example "What is this building?", and "What are the opening hours?". The user can also ask more general questions about the section of the city that is displayed, such as "What restaurants are in this neighbourhood?' The latter question will add icons for restaurants to the display, that can be turned into the topic of conversation by pointing and asking questions, for example about the type of food that is offered, the price range, and opening hours. The information returned by the system is rendered in the form of text, graphics (maps, and pictures of hotels and restaurants), and text-to-speech synthesis.
For mobile network operators a substantial part of access to services comes from roaming customers. It is well-known that most people prefer to use their native language, especially when using speech recognisers, that are known to degrade in performance for non-native speech. Therefore, information services offered in the mobile networks must be multilingual, so as to allow every customer to use the preferred language. The MUST demonstrator is developed for Norwegian, Portuguese, French and English.

Users will be allowed to ask questions about POI's for which the answers are not in the database of the service, perhaps because only a small

[2] Simultaneous coordinated multimodal interaction is the term used by W3C http://www.w3.org for the most complicated multimodal interaction, where all available input devices are active simultaneously, and their actions are interpreted in context.

proportion of the users is expected to be interested in this information (e.g., 'Who is the architect of this building?' and 'What other buildings has he designed in Paris?'). For the answers to these questions access will be provided to a multilingual Question/Answering (Q/A) system, developed by France Télécom R&D, that will try to find the answers on the Internet (Boualem and Filoche, n.y.).

2 The architecture of the demonstrator

The overall architecture of the MUST demonstrator is shown in Figure 1. The server side of the architecture combines a number of specialised modules, that exchange information among each other. The server is accessed by the user through a thin client that runs on the mobile terminal. The application server is based on the Portugal Telecom Inovação (Azevedo and Beires, 2001) and Telenor R&D (Knudsen et al., 2000) voice servers, which were originally designed for voice-only services, i.e. there are two versions of the demonstrator that only differ in the voice platforms used. The voice servers provide an interface to ISDN and PSTN telephony and advanced voice resources such as Automatic Speech Recognition (ASR) and Text-to-Speech Synthesis (TTS). The ASR applied is Philips *SpeechPearl2000*, that supports all the languages in the project (English, French, Portuguese and Norwegian). ASR-features such as confidence scores and N-best lists are supported. The TTS engine is used to generate real-time speech output. Different TTS-engines are used for the different languages in MUST. Telenor and France Télécom use home-built TTS engines, while Portugal Telecom uses *RealSpeak* from L&H.

The multilingual question-answering (Q/A) system uses a combination of syntactic/semantic parsing and statistical natural Language Processing techniques to search the Web for potentially relevant documents. The search is based on a question expressed in natural language, and the system subsequently tries to extract a short answer from the documents. The size (in terms of number of characters) of the answer cannot be predicted in advance, but it is expected that most answers are short enough to fit into the text box that is used for presenting information that is already available in the database. If an answer is too long, it will be provided by Text to Speech.

The GALAXY Communicator Software Infrastructure, a public domain reference version of DARPA Communicator maintained by MITRE (http://fofoca.mitre.org), has been chosen as the underlying inter-module communication framework of the system. It also provides the HUB in Figure 1, through which nearly all the inter-module messages are passed. The main features of this framework are modularity, distributed nature, seamless integration of the modules, and flexibility in terms of inter-module data exchange (synchronous and asynchronous communication through HUB and directly between modules). GALAXY allows to 'glue' existing components (e.g., ASR, TTS, etc.) together in different ways by providing extensive facilities for passing messages between the components through the central HUB. A component can easily invoke a functionality that is being provided by other components without knowing which component provides it or where it is running.

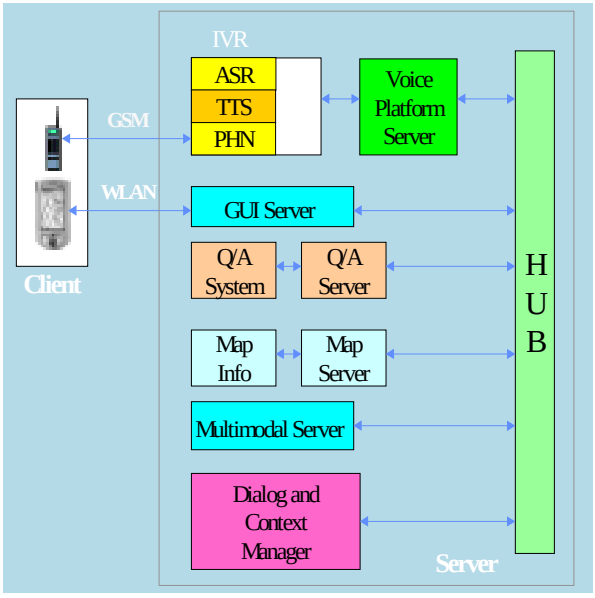

Figure 1. Schematic architecture of the MUST tourist guide to Paris

The processing in the HUB can be controlled using a script or it can act as a facilitator in an agent based system. In MUST the HUB messaging control is script based. The modules are written in Java and C/C++ under Linux and Windows NT.

In order to keep the format of the messages exchanged between the modules simple and flexible, it has been decided to use an XML based

mark-up language named MxML - MUST XML Mark up Language. MxML is used to represent most of the multimodal content that is exchanged between the modules. Parameters required for set-up, synchronization, and disconnection of modules use key pair (name - value) attributes in Galaxy messages.

The client part of the demonstrator is implemented on a COMPAQ iPAQ Pocket PC running Microsoft CE with WLAN connection. The speech part is handled by a mobile phone. The user will not notice this "two part" solution, since the phone will be hidden and the interface will be transparent. Only the headset (microphone and earphones) with a wireless connection will be visible for the user.

The spoken utterances are forwarded to the speech recogniser by the telephony module. The text and pen inputs are transferred from the GUI Client via the TCP/IP connection to the GUI Server. The inputs from the speech recogniser and the GUI Server are integrated in the Multimodal Server (late fusion) and passed to the Dialogue/Context Manager (DM). The DM interprets the result and acts accordingly, for example by contacting the Map Server and fetching the information to be presented for the user. The information is then sent to the GUI Server and Voice Server via the Multimodal Server that performs the fission. Fission consists of the extraction of data addressed to the output modalities (speech and graphics in this case).

MUST set out to investigate implementation issues related to coordinated simultaneous multimodal input, i.e. *all* parallel inputs must be interpreted in combination, depending on the fusion of the information from all channels. In our implementation we opted for the "late fusion" approach, where recogniser outputs are combined at a semantic interpretation level. The temporal relationship between different input channels is obtained by considering all input contents within a reasonable time window. The length of this time window has a default value of 1 second and is a variable parameter that can be adjusted dynamically according to the dialog context.

3 The user interface of the demonstrator

One important feature for the user interface is the "Tap While Talk" functionality. When the pen is used shortly before, during or shortly after speech, the two input actions are integrated into one combined action. An example is the utterance "Show hotels here", while tapping at Notre Dame. When the time between tapping and speech is longer than a pre-set threshold, the actions are considered as sequential and independent.

The overall interaction strategy is user controlled, in accordance with what is usual in graphical user interfaces. This implies that the speech recogniser must always be open to capture input. Obviously, this complicates signal processing and speech recognition. However, it is difficult to imagine an alternative for a continuously active ASR without changing the interaction strategy. Users can revert to sequential operation by leaving enough time between speech and pen actions.

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

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 Point of Interest (POI) has been selected, the system will respond by showing the corresponding 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 a prompt saying that it did not understand the utterance.

The graphical part of the user interface consists of two types of maps: an overview map showing all POIs, and detailed maps with a POI in the centre. 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. The selected object automatically becomes the focus of the dialogue: all deictic pronouns, requests etc. now refer to the selected object. Selection can be accomplished in three ways: by speaking, by pointing, or by both simultaneously. Irrespective of the selection mode, the application responds by showing the section map that contains the POI. A selected object is marked by a red frame surrounding it, as a graphical response to the selection action. All additional selectable objects on a map are indicated by green frames. When

the user has selected a POI, several facilities such as hotels and restaurants can be shown as objects on the maps. This can be accomplished by means of speech (by asking a question such as 'What hotels are there in this neighbourhood?'), or by tapping on one of the 'facility' buttons that appear at the bottom of the screen, just below each section map.

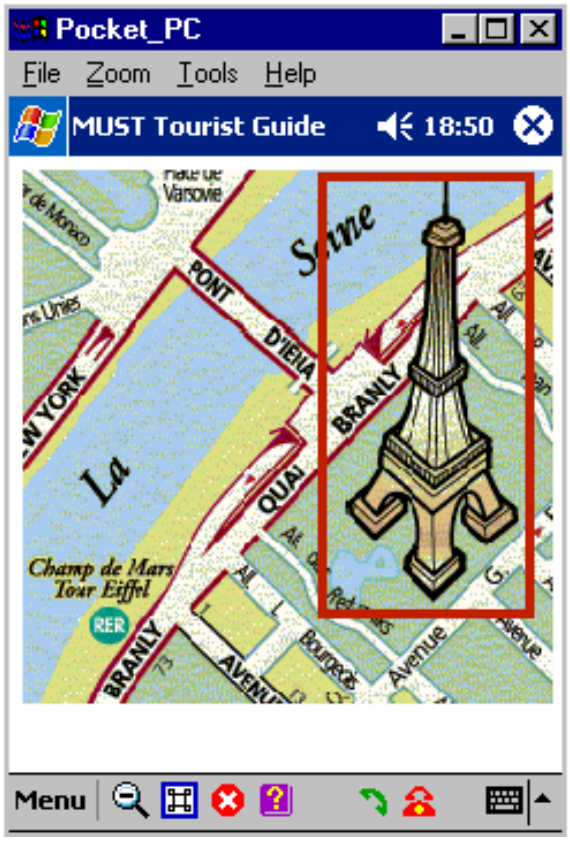

Figure 2. Screen Layout of the MUST tourist guide

Fig. 2 shows the buttons that were present in the toolbar of the first version of the GUI. Two buttons are related to the functionality of the service (hotels and restaurants), and three buttons are related to navigation: a help button, a home button, and a back button. The back button will make the application go back to the previous state of the dialogue as a kind of error recovery mechanism to deal with recognition failures. 'Help' was context independent in the first version of the demonstrator; the only help that was provided was a short statement saying that speech and pen can be one by one or combined to interact with the 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.

4 Expert review

The MUST application was investigated by Norwegian and Portuguese experts in human-machine interaction. Since only twelve experts participated in this evaluation, results should be interpreted with due caution. There were great similarities between the remarks and observations of the Portuguese and Norwegian experts. 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 five of the twelve experts started to use pen and speech simultaneously.

Timing between speech and pointing has been studied in other experiments (Martin et al. 1998; Kehler et al., 1998). 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 utterances 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 graphical 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 use a simultaneous multimodal interaction style. They also believed that the users will 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 explorative 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. This issue will be studied during the user experiments that will be carried out in September. The introduction that is given to the users before they start to use the tourist guide will be the main parameter in this experiment. Then we will also gain more information on how naïve users benefit from adding the simultaneous coordinated actions in a multimodal tourist guide. In our demonstrator it is not necessary for the user to input several modalities simultaneously. The choice of sequential/simultaneous mode is controlled by the user.
Another issue pointed out by the experts is the importance of a well-designed help mechanism in speech-centric user initiative information services. In these services it is difficult for the system to convey information about its capabilities and limitations (Walker and Passonneau, 2001).

5 Conclusion and further work

The aim of MUST is to provide Telecom Operators with useful information on multimodal services. We have built a stable, real-time multimodal demonstrator using standard components without too much effort.
The first version was evaluated by human-factor experts. One of the main conclusions was that naïve users will need instructions before being able to benefit from a simultaneous coordinated multimodal interaction. Once aware of the systems capabilities they should be able to use the system with small cognitive effort. This will be studied more in forthcoming user experiments. Another issue we will study in this experiment is the timing of the input, especially when deictic expressions are used.

References

Azevedo, J., Beires N. (2001) *InoVox - MultiService Platform Datasheet*, Portugal Telecom Inovação.

Boualem, M. and Filoche, P. (n.y.) Question-Answering System in Natural Language on Internet and Intranets, *YET2 marketplace*, http://www.yet2.com/

Boves, L., and Den Os, E. (Eds.) (2002) *Multimodal services – a MUST for UMTS*. http://www.eurescom.de/public/projectresults/P1100-series/P1104-D1.asp

Cheyer, A. and Julia, L. (1998) Multimodal Maps: An agent-based approach. In: H. Bunt, Beun, Borghuis (Eds) *Multimodal Human-computer communication*, Springer Verlag, pp. 111-121.

EURESCOM (2002) *Multimodal and Multilingual Services for Small Mobile Terminals.* Heidelberg, EURESCOM Brochure Series.

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.

Knudsen, J.E., Johansen, F.T. and Rugelbak, J. (2000) *Tabulib 1.4 Reference Manual*, Telenor R&D scientific document N-36/2000.

Martin, J.-C. Julia, L. and Cheyer, A. (1998) A theoretical framework for multimodal user studies, *CMC-'98*, pp. 104-110.

Nielsen, J. and Mack, R.L. (eds), (1994) *Usability Inspection Methods*, Jon Wiley & Sons, Inc

Oviatt, S. (1999) Ten Myths of Multimodal Interaction, *Communications of the ACM*. Vol. 42, No. 11, pp. 74-81.

Oviatt, S. et al. (2000) Designing the user interface for multimodal speech and gesture applications: state-of-the-art systems and research directions for 2000 and beyond. In: J. Carroll (ed) *Human-computer interaction in the new millennium.* Boston: Addison-Wesley Press.

Oviatt, S. & Cohen, P. (2000) Multimodal Interfaces That Process What Comes Naturally, *Communications of the ACM*, Vol. 43, No. 3, pp. 45-53.

Oviatt, S. L., DeAngeli, A. & Kuhn, K. (1997) Integration and synchronization of input modes during multimodal human-computer interaction, *Proc. Conf. on Human Factors in Computing Systems: CHI '97*, New York, ACM Press, 415-422.

Wahlster, W., Reithinger, N., and Blocher, A. (2001) SmartKom: Multimodal Communication with a Life-Like Character, *EUROSPEECH-2001*, Aalborg, Denmark, pp 1547–1550.

Walker, M. A., and Passonneau, R. (2001) DATE: A Dialog Act Tagging Scheme for Evaluation of Spoken Dialog Systems. *Human Language Technology Conference*. San Diego, March 2001.

Wyard, P. and Churcher, G. (1999) The MUeSLI multimodal 3D retail system, *Proc. ESCA Workshop on Interactive Dialogue in Multimodal Systems*, Kloster Irsee, pp. 17-20.