Eurescom MUST project
Multimodal, multilingual information services for small mobile terminals

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

(1) France Telecom R&D
2, avenue Pierre Marzin - 22307 Lannion - France
Email : malek.boualem@rd.francetelecom.com

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

Résumé – Abstract

MUST est un projet EURESCOM[1] qui signifie "Multimodal multilingual information services for small mobile terminals". Il a démarré en février 2001 et se terminera en décembre 2002. Les partenaires sont : Eurescom (Allemagne), France Télécom, Portugal Télécom, Telenor (Norvège), Université de Nimègue, KPN et ensuite Max Plank Institute (Pays-Bas). L'objectif du projet MUST est d'étudier les possibilités de mise en œuvre de services multimodaux et multilingues sur les petits terminaux mobiles (UMTS). Un service démonstrateur a été mis en œuvre. Il permet à un touriste à Paris, muni de son terminal UMTS (simulé par un téléphone mobile et un PDA) d'accéder à des informations variées sur Paris (hôtels, restaurants, transport, itinéraires, monuments, musées, etc.). Le service intègre l'accès graphique aux points d'intérêt sur la carte de Paris, la reconnaissance vocale, le dialogue, la recherche d'information via un système de questions-réponses en langage naturel, la traduction automatique de mots-clés et la synthèse vocale.

Mobile Internet Access is expected to grow very fast in the near future, fostered by speech access to the Internet through Voice Portals and fast mobile Internet in the UMTS networks.

[1] EURESCOM, the European Institute for Research and Strategic Studies in Telecommunications, is the leading company for collaborative R&D in telecommunications in Europe. Founded in 1991, EURESCOM provides comprehensive collaborative research management services to network operators, service providers, suppliers and vendors.

Typically, mobile terminals have small screens and keyboards, which makes them difficult to use in transaction and information services that require free text input or output of complex information that is only available in textual form. Implementation of speech and language technologies in a clever way will help to solve this problem. Speech recognition may be used to overcome part of the input problem, and language technologies like Information Extraction and (multilingual) Language Generation may be used to generate condensed representations that fit the small screen. In its most basic form speech input/output can be used as an overlay for keyboard input and screen output. Yet, it is generally agreed that clever combination of speech, text and graphics in the user interface will improve the usability of mobile information services. Substantial Human Factors research is still needed to understand how speech and language technology must be used in the servers and in the mobile terminals to be able to develop high quality multimodal interfaces. The Eurescom MUST project has been launched to meet several objectives :

- To obtain a better understanding of the role that language and speech technology will play in future multimodal and multilingual services in the mobile networks accessed from small terminals and of the requirements that the technology must meet.
- To get hands-on experience by integrating existing speech and language technologies into an experimental multimodal interface to a realistic demonstrator
- To use the demonstrator to conduct human factor experiments with 'real' users to assess the value of the language and speech technologies for fast, simple and user-friendly interfaces on small mobile devices.

Keywords – Mots Clés

Multimodalité, multilinguisme, recherche d'information, questions-réponses en langage naturel, reconnaissance de la parole, synthèse de la parole, dialogue, UMTS, petits terminaux mobiles.

Multimodality, multilingualism, information retrieval, question-answering, natural language processing, speech recognition, speech synthesis, dialogue, UMTS, small mobile terminals.

1 Introduction

For Telecom Operators it is essential to invoke the widest possible use of their future UMTS services. The problems with the introduction of WAP services have proved that wide usage presupposes that at least two requirements are fulfilled: customers must have the feeling that the service offers more or better functionality than existing alternatives, and the service must have a simple and natural interface. Especially the latter requirement is difficult to fulfill with the interaction capabilities of the small lightweight mobile terminals. Terminals that combine speech and pen at the input side, and text, graphics, and speech 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 completely new technological and human factors problems of its own. Therefore, the research departments of three Telecom Operators collaborate with two academic institutes in the EURESCOM project MUST (*Multimodal, multilingual information services for small mobile terminals),* that aims to obtain knowledge about the issues involved in the implementation of a multimodal applications for small terminals.

We have implemented a multimodal application using an iPAQ terminal. The service concept that we chose as an example to focus the development work and to show the potential of additional functionality is an interactive tourist guide to Paris. Obviously these services cannot be deployed without the underlying networks. The users perceive the network performance and quality when they interact with the services and applications. The deployment of multimodal services due to its characteristics puts additional requirements on the underlying networks. Network capabilities required to enable for instance the simultaneous transmission and reception of voice and data within the same service session should be available as well as coordination and synchronization. The feedback of the users interaction will be very important to ensure the appetency of users to adhere and subscribe to multimodal services in the near future. The introduction of such services will cause impact on business cases for mobile services.

2 Functionalities of the MUST tourist guide

The MUST tourist guide for Paris combines speech and pen at the input side, and text, graphics, and speech at the output side. Basically, the service is the equivalent of a printed tourist guide that provides information about a small section of the city, and that uses a detailed map of that section as a navigation and orientation aid. The tourist guide 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. 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 iPAQ (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 general questions about the section of the city that is displayed, such as "What restaurants are in this neighborhood?' 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, such as the type of food that is offered, the price range, opening hours, etc. Simultaneous coordinated interaction allows the pointing and speech gestures to overlap in time. 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.

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 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 Telecom R&D, that will try to find the answers on the Internet. This system is based on natural language processing technologies. Access to the Q/A system should provide a graceful failure solving solution in the case of out-of-domain questions (although it is evident that there also remain unresolved issues in the field of automatic speech recognition and natural language understanding).

3 Architecture of the MUST tourist guide

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 so-called "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 pre-defined time window.

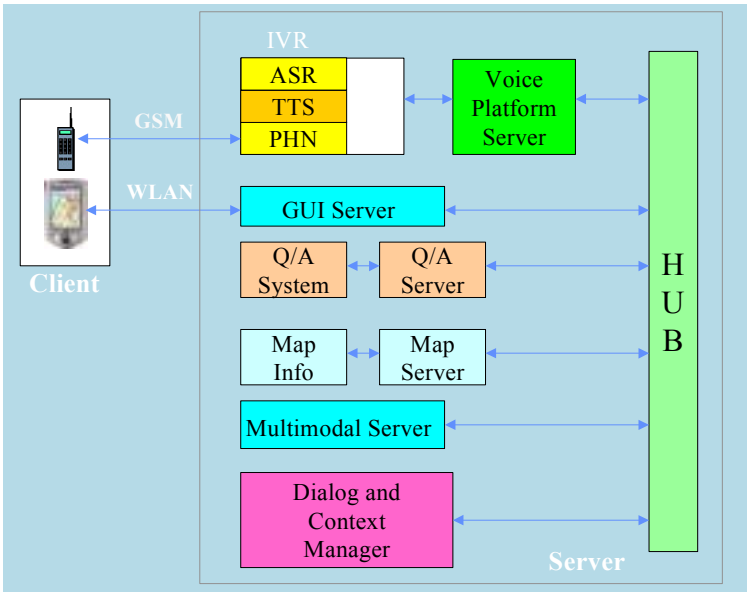


Figure 1 : Overall architecture of the MUST tourist guide to Paris

The overall architecture of the MUST-demonstrator shown in Figure 1 consists of a complex application server and a relatively simple (thin) client. The application server is based on the Portugal Telecom Inovação and/or Telenor R&D voice platforms with interface to ISDN-telephony, SS7, Analogue line and advanced voice resources like 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-engines are different for the different languages.

The Q/A is a question-answering system that searches the Web for potentially relevant documents, based on a question expressed in natural language, and that subsequently tries to extract an answer from the documents. The size (in terms of number of characters) of the answer fetched on the Internet can not be predicted in advance, and due to the reduced dimensions of the screen device, it has been decided to output the answer through speech. A Text-to-Speech engine will be used to generate real-time voice output .

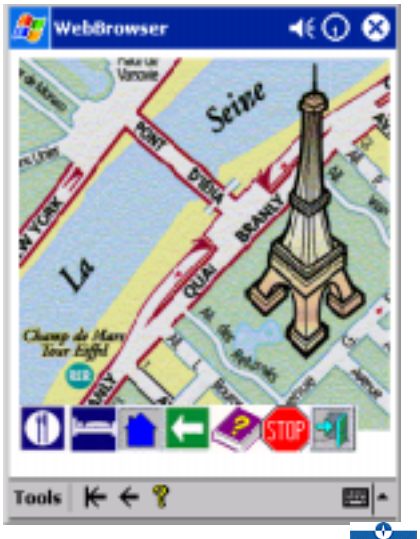
The GALAXY Communicator Software Infrastructure, a public domain reference version of DARPA Communicator maintained by MITRE (http://fofoca.mitre.org) has been chosen as the underlying software architecture which provides the HUB in Figure 1. The main features of this framework are modularity, distribution, seamless integration of the modules and flexibility in terms of inter-module data exchange (synchronous and asynchronous communication through HUB and directly between modules). The GALAXY platform allows then to 'glue' existing components (e.g., ASR, TTS, etc.) together in different ways by providing extensive facilities for passing messages between the components through a central

'Hub'. A component can very easily invoke a functionality that is being provided by other component without knowing which component provides it or where it is running. The processing in the Hub can implement a script or it can act as a facilitator in an agent based environment. In MUST the Hub messaging control is script based. The modules are written in Java and C/C++ running on Linux or 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. The MxML is used to represent most of the multimodal content that is exchanged between the modules. Other required parameters for setup, synchronization, and disconnection are based in key pair (name plus value) attributes written in the Galaxy messages.

The client part of the demonstrator is implemented on an iPAQ (Compaq) 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 Bluetooth connection will be visible for the user. The multimodal input is processed as follows. The spoken utterances are forwarded to the speech recognizer 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 recognizer 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.

4 The user interface of the MUST tourist guide



Copyright© of Paris maps: Planfax

The graphical part of the user interface consists of two types of maps: An overview map showing all POI's, and more detailed maps centered around one single POI. Different groups of facilities such as hotels and restaurants can be shown as objects on the maps. The map objects can be selected/activated through speech and by tapping the pen on an object. To initiate an action, the user must tap a button or speak a command, e.g., 'show info', 'display hotels', 'help', 'go back to previous map'. The user can use speech shortcuts and for example request actions like: " Show me the Eiffel Tower". The final paper will contain a full description of the user interface.

5 The expert evaluation

The evaluation of the MUST user interface will be carried out in two steps: First a simplified version of the demonstrator will be presented for an expert review. The results from this review will be used in the design of the final user interface. For this expert review which will be finished before the end of April, we only perform qualitative evaluation, i.e. no quantitative measurements. Secondly, usability tests with naive users will be performed on an improved version of the demonstrator. In the expert evaluation we apply a Usability Inspection method called Cognitive Walkthrough Method (CWM) (Nielsen and Molich, 1994).

References

EURESCOM P1104 MUST Deliverable 1 ”Multimodal Services – a MUST for UMTS”, January 2002 (http://www.eurescom.de/public/projectresults/P1100-series/P1104-D1.asp).
Nielsen, J. and Mack, R.L. (eds), (1994) “Usability Inspection Methods”, Jon Wiley & Sons, Inc.
W3C, “Multimodal requirements for voice markup languages”, W3C working draft July 2000.
Galaxy Communicator, MITRE (http://www.mitre.org).

List of Acronyms

ASR	Automatic Speech Recognition
IP	Internet Protocol
GPRS	Generalised Packet Radio System
PDA	Personal Digital Assistant
TTS	Text To Speech
UMTS	Universal Mobile Telecommunication System
W3C	World Wide Web Consortium
XML	Extended Markup Language