EURESCOM

Technical Information

Multimodal multilingual information services for small mobile terminals (MUST)

Multimodal Services – a MUST for UMTS

Editors: Lou Boves & Els den Os

Suggested readers

Strategists and business development managers of mobile operators

Abstract

The emerging mobile Internet capable terminals promise to offer a large number of new and appealing services. The small size of these terminals introduces usability problems for these services, and, therefore new interaction modes need to be developed. One of these is speech centric multimodal interaction. Speech recognition, pen, text are used at the input side and speech synthesis, graphics, text, and moving pictures are used at the output site. This document discusses the main issues in speech centric multimodal interfaces. For one thing, the pivotal role of 'language' creates the need for making most future services multilingual. An overview of present activities in the development of multimodal interfaces and some expectations for the near, mid-term and long-term future are given. The technical features are explained. To make new Mobile Internet services successful, a number of players will have to join forces: content providers, network and terminal suppliers, multimodal technology suppliers, and application developers. Since the growing use of mobile Internet is of interest for (mobile) network operators, they might want to join this chain, and even direct the developments. Only when the different players collaborate, standardized user-friendly interfaces can be developed that guarantee the acceptance and use by the large public.

EDIN 0227-1104

Project P1104

For full publication

January 2002

EURESCOM PARTICIPANTS in Project P1104 are:

- KPN (Until August 2001)
- Telenor
- Portugal Telecom
- France Télécom

[Project title] Multimodal multilingual information services for small mobile terminals (MUST)

[Document title] Expectations of mobile Internet services, terminals and technology

Editors: Lou Boves & Els den Os, KPN until August 15; since August 15th as external consultants

Project leader: Els den Os, KPN until August 15th; since August 15th as external consultant from Max Planck Institute

Project supervisor: Peter Stollenmayer, EURESCOM

EURESCOM published project result; EDIN 0227-1104

Disclaimer

This document contains material, which is the copyright of certain EURESCOM PARTICIPANTS, and may not be reproduced or copied without permission.

All PARTICIPANTS have agreed to full publication of this document.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the PARTICIPANTS nor EURESCOM warrant that the information contained in the report is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

Preface

Mobile communications have been one of the main drivers of telecommunication business during the last years. The first wave, mobile telephony, has grown so fast that it is nearly in its saturation now. The second wave, mobile Internet and mobile commerce, will probably be the driving force of the next generations of mobile communication systems. One of the most critical issues is the development of terminals, which are able to deliver the services to the customers in a most user-friendly way.

Typically, mobile terminals have small screens and keyboards, which makes them difficult to use for transaction and information services that require input or output of complex information. A clever combination of speech, text and graphics will improve the usability of mobile information services dramatically. This clever combination of different means for input and output of information is called “multimodality”. We are convinced that multimodality is the way to solve the contradiction of the complexity of the information on the one hand and the limitations of small terminals on the other hand. To facilitate the development of such terminals EURESCOM is undertaking the MUST project.

The main objectives of the MUST project are:

- Obtaining 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.
- Getting hands-on experience by integrating existing speech and language technologies into an experimental multimodal interface to a realistic demonstrator
- Using 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.

The MUST project is producing the following Deliverables:

- This first Deliverable is setting the scene by introducing the most important issues concerning multimodal services related to small mobile terminals. It identifies the different players, their roles and options in this field, gives an overview of the state-of-the-art in emerging multimodal services and research aimed at better understanding multimodal interaction per se, and presents a vision of future multimodal mobile interactions. Also some technical issues related to multimodal interactions, the requirements for future terminals and the architectures required supporting multimodal mobile services are addressed.
- A second Deliverable will provide the specifications of the experimental multimodal demonstrator and report about its implementation. This Deliverable (EDIN 0228-1104) will become available to EURESCOM shareholders and members in August 2002.
- A third Deliverable will report of the user evaluation, which will be performed with the experimental multimodal demonstrator and summarise the main findings. It will discuss the user acceptance and the possibilities of using multimodal services to provide user-friendly mobile interfaces of the near future. This Deliverable (EDIN 0229-1104) will become available to EURESCOM shareholders and members in December 2002

Executive Summary

Many persons in industry, government boards, and research environments have high expectations of the mobile Internet, once ‘always on’ has become possible. Ambient Intelligence will be a next step in the ‘total information experience’. Whatever will happen in this field, it will always be the case that some device must act as the intermediary between the information (system) and user. While it is not known today what these interaction ‘devices’ will look like, it is certain that they will be small. It is the general belief that new user interfaces and novel interaction metaphors need to be developed in order to deal with these small devices. Ideas and visions about this future interaction have already been articulated and research and development in this field have started. The common idea for these interfaces is that they must be multimodal, i.e. combining touch, vision, and speech in input and output, in order to approximate the rich communication channels in human-human interaction. It is striking that in most expectations for future interfaces speech plays an important role. Since speech technology has become mature technology (at least compared to e.g. gesture recognition or face recognition), we concentrate in this document on speech centric multimodal interfaces, where speech recognition is combined with pen and text input, and speech synthesis with text and graphics output.

We believe that in the emerging field of multimodal interaction (mobile) network operators will have a place, next to content providers, terminal and network suppliers, and multimodal technology providers. However, this place is yet to be determined; and it need not be the same for all operators. In our opinion only collaborations between these players will result in more and more user-friendly services. One of the reasons for this belief is that we see an urgent need for the development of standardised user interfaces, which guarantee better acceptation and use of services by the large public.

A growing proportion of local subscribers will have a native language different from the dominant language in the country. Customers want to be able to access services in their native language. We expect that future mobile services will be required to have the same type of language selection as present desktop applications. Therefore, telecom operators and service providers will find it increasingly important to offer their services in a large number of languages.

This document presents the issues that are related to speech centric multimodal interaction in such a way that network operators obtain the information that is needed to determine their position in the field. It might be important to push the development and research activities in order to be able to be in the frontline in offering user-friendly mobile Internet services.

To provide the information that network operators need to determine their strategy we discuss the major technical and business aspects of multimodality in more detail. We also present our expectations of the development of multimodal interfaces for the near, mid-term, and long-term future. Finally, we provide an overview of present activities in industry and academic environments.

List of Authors

Lou Boves, University of Nijmegen

Els den Os, Max Planck Institute

Bram Vromans, KPN

Luis Almeida, Portugal Telecom

Nuno Beires, Portugal Telecom

Knut Kvale, Telenor

Narada Warakagoda, Telenor

Ingunn Amdal, Telenor

Malek Boualem, France Télécom

Pascal Filoche, France Télécom

Table of Contents

List of Figures and/or List of Tables

Abbreviations

AAA	Adaptive Agent Architecture
ADSL	Asymmetric Digital Subscriber Line
AHR	Automatic Handwriting recognition
ASIC	Application-Specific Integrated Circuit
ASR	Automatic Speech Recognition
ATIS	Airline Travel Information System
DARPA	Defence Advanced Research Project Agency
DSR	Distributed Speech Recognition
EDGE	Enhanced Data for GSM Evolution
GPRS	Generalised Packet Radio System
GPS	Global Positioning System
HCI	Human-Computer Interaction
HLT	Human Language Technology
ISTAG	Information Society Technologies Advisory Group
OAA	Open Agent Architecture
OGI	Oregon Graduate Institute
PAN	Personal Area Network
PDA	Personal Digital Assistant
PSTN	Public Switched Telephone Network
SALT	Speech Application Language Tags
SLS	Spoken Language System
SMS	Short Message Service
SRI	Stanford Research Institute
TTS	Text-to-Speech
UMTS	Universal Mobile Telecommunications System
VoIP	Voice over IP
W3C	World Wide Web Consortium
WAP	Wireless Application Protocol

Definitions

Bluetooth

Bluetooth is a specification for wireless communication of data and voice based on short-range radio links. It was originally developed by Ericsson.

Browser

A *browser* is an application program that provides a way to look at and interact with all the information on the World Wide Web. The word "browser" seems to have originated prior to the Web as a generic term for user interfaces that let you browse (navigate through and read) text files online. A Web browser is a client program that uses the Hypertext Transfer Protocol (HTTP) to make requests of Web servers throughout the Internet on behalf of the browser user.

Earcon

An *earcon* is the auditory equivalent if an icon. Thus, it can be described as a short, easily recognised sound, to which a specific meaning is attached, either by convention, or by virtue of a decision imposed by the service. Earcons should be distinguished from the jingles and tunes (sometimes called 'audio logos') that are used to identify radio of tv stations, or specific programmes of a station. In general, the duration of an earcon is much shorter than an average jingle.

Fission

In the context of multimodal interaction *fission* refers to a kind of decomposition of complex chunks of information into multiple chunks that may or may not appear to be simpler. The resulting chunks may be presented through different output channels (sequentially or simultaneously) or to a single channel (in which case sequential presentation is the only option). The 'decomposition' of the information may result in redundant representations. For example, the information chunk 'Tour Eiffel' can be rendered on a screen by displaying a picture of the structure accompanied by a text label. The picture is rendered through the graphical channel, the textual label through the text channel. The name may also be spoken, adding a third channel with essentially the same information.

Fusion

In the context of multimodal interaction *fusion* refers to the combination of several chunks of information, to form new chunks, possibly of a higher order. The information chunks to be fused may or may not originate from distinct input channels or from distinct contexts. For example, the sequence of events "mouse-down, mouse-up" that occurs in the palette of a graphics editor comprises two information chunks that originate from the same input channel and from the same context (i.e., the palette agent). They are combined within the context of the palette agent to form a higher information chunk (i.e., the selection of a geometric class). Pointing with a pen at a file object on the screen and the spoken word 'open' are an example of information chunks originating from different input channels. After fusion the interpretation is 'open file'.

Piconet

Piconet is a collection of devices connected via Bluetooth technology (or possibly a similar wireless protocol for short distance communication) in an ad hoc fashion, and sharing the same physical channel.

Scattered terminal

A *scattered terminal* is a set of devices interacting via Bluetooth radio links (or a similar short-range wireless connection) and integrating a connection to the Internet in order to send requests to a server. The server sees the scattered terminal as a unique entity capable of rendering information on various physical devices.

Session

A *session* is a series of interactions between two communication end points that occur during the span of a single connection. Typically, one end point requests a connection with another specified end point and if that end point replies agreeing to the connection, the end points take turns exchanging commands and data ("talking to each other"). The session begins when the connection is established at both ends and terminates when the connection is ended.

1 Introduction

Today the Telecommunication companies are developing into full-service companies with a wide range of services. The success of these services depends to a large extent on a interfaces that make the use of the service 'intuitive', and on strategies that allow the automatic system that implements the service to act as "value-added mediator" between the user and her goals and intentions.

A service supporting interaction through more than one input/output modality is expected to provide a more friendly and richer set of interaction metaphors to its users. Superior ease-of-use is essential for customers with limited perceptual or motor capabilities, a part of the population that does not only include people with physical or perceptual disabilities, but also large proportions of the elderly population. Intelligent multimodal interfaces are also extremely important for all other customers, perhaps with the exception of young children, who are able and willing to discover the functions and adapt to the limitations of all kinds of devices. The use of these multimodal interfaces can improve speed, accuracy and naturalness of human computer interactions that will be of key importance for the success of a service in the present open Telecom market. Multimodal interaction may well turn out to be an essential factor in the development of attractive entertainment services and games played over the network.

In this document we present an overview of the state-of-the-art in the emerging speech centric multimodal telecommunication services for mobile environments that require small and lightweight terminals. These services use a multimodal interaction paradigm, combining speech, vision, and touch. We approach the subject both from a business perspective and from a technical and ergonomic point of view. The services that we are targeting are all Internet services. This means that we assume some kind of 'always-on' connection, probably in an IP-based network; we also assume that the services will provide the same rich content that is available through Internet, and interaction and browsing facilities that are matched to the richness and complexity of that information.

First we explain what is commonly understood by multimodal interaction in human-machine environments (chapter 2). In chapter 3 we deal with a number of technical and business issues involved in the development and deployment of multimodal services. Starting from the present situation we extrapolate to developments in the medium- and long-term future. In this section we will also discuss possible roles of network operators, and define other players in the field with whom the network operators must come to agreements. In chapter 4 we present an overview of the most important R&D activities in the field of multimodal services, both in industrial and academic laboratories.

In chapters 5 and 6 we discuss technical aspects of speech centric multimodal interaction. We deal with terminals (chapter 5), and architectures (chapter 6). Overall conclusions are presented in chapter 7.

1.1 Objective of this document

The objective of this document is to introduce the most important issues in multimodal services with small mobile terminals. This document consists of two parts. In the first part (chapters 1 – 3 and the Conclusions) we explain the importance of the development of new interaction modes for the success of mobile services of the future on a fairly high level of abstraction. We describe the players who are already active in this field, and we argue that they should closely collaborate to develop acceptable solutions for the interaction modes of the future. Our main aim in the first part is to define the option that the network operators have to determine their own position in the emerging field.

The second part of this document (chapters 4 – 6) provides an overview of research activities in the field of multimodality and explains some technical issues related to multimodal human-machine interaction in general, and in interaction through small terminals in particular. This part of the presentation should help strategists and business managers to better understand the opportunities, but also the problems, that come with multimodal mobile services. We address the most important requirements for future terminals, and the architectures that are required to support multimodal mobile services.

2 Multimodal Interaction

As is the case with many recent buzzwords the precise meaning of the term *multimodal* is far from clear. In this section we aim to explain the meaning of the concept *multimodality*, and of the most important related concepts. Intuitively, the meaning of 'multi' is evident (but in the case of multimodality it is not). 'Modal' is derived from the noun 'mode'. Modes are essentially the perceptual channels that a normal human person has available. Unfortunately, perception is so intimately intertwined with production that some degree of confusion between receptive and productive actions in human communication is virtually unavoidable.

The difficulty to provide a clear, unambiguous and generally accepted definition of multimodal interaction is aggravated by the fact that technology providers in the field have an interest in stretching the definition, so as to be able to bring their products under the flag of the buzzword. Without being overly strict we propose to limit the coverage of the definition at least somewhat.

2.1 Modes

Conventionally, five modes or senses are distinguished. Three modes are especially important for the services that most persons who talk and write about multimodal services have in mind:

1. audio

   in the form of speech and other 'meaningful' sounds, like music and earcons (the auditory equivalent of icons)

2. vision

   in the form of text and graphics, but also in the form of moving pictures

3. touch

   with small terminals (that often have no full-fledged keyboards) in the form of a pen to click areas on the screen, and push buttons on the terminal. With conventional desktop terminals the keyboard and mouse, and –of course- the touch screen also serve as touch-based input devices. It should be noted that *touch* in the way it is described here refers much more to production than to perception. This is in line with the common usage of touch in human-machine communication, where turning knobs, typing command and clicking icons are quite general operations, whereas it is much more difficult to imagine situations where the 'feel' of actions performed by the machine (or by an interlocutor) is relevant.

The remaining two senses viz. taste and smell, will not play a major role in the telecommunication services of the next decade. Figure 1shows a schematic illustration of the input and output modes that will be present in the multimodal applications that are currently under development in several laboratories and companies.

The literature on 'modes' is almost exclusively oriented towards perception. That makes it a bit difficult to apply the terminology to human-computer interaction, where not only the reception, but also the generation of information and signals is at stake. All three modes that are relevant in HCI can be used both to convey and to receive information. A customer can speak and listen, at the same time. And in face-to-face communication she can see the response of her interlocutor, while she is speaking and simultaneously making meaningful gestures. The fact that the modes can be used to receive and to send signals only adds to the difficulty to define the meaning of 'multi' in multimodality.

2.1.1 Modes and Media

Along with the term multimodal we also encounter the term *multimedia*. In this document we reserve the term 'media' for the different ways in which information can be presented in a certain mode or combination of modes (e.g., video combines the auditory and visual modes). Thus, speech and music are two examples of media in the auditory mode. Text and graphics are examples of media in the visual mode. None of these lists is complete, but that should not deter from their value as examples. For the purpose of this document it is a good analogy to associate ‘modes’ with technology, and ‘media’ with services, even if the distinction is not very sharp. Some authors use the term 'Communication channels' as a synonym for the word 'media', but in this document we will avoid this confusion. The distinction between modes and media that is made in this document is similar to the circumscription given in the paper by Peter Wyard and Gavin Churcher [1].

Figure 1 Pictorial illustration of multimodal interaction

Recently, a somewhat different usage of terms emerged in the context of the discussions about multimodal interaction in the World Wide Web Consortium (W3C). In that community the term ‘modes’ exclusively refers to input devices, while the term media is reserved for output channels. In reading the scientific and commercial literature one must be aware of the different uses of the terms.

2.2 The ‘multi’ issue

Now it is time to come to the ‘multi’ part of the buzzword *multimodal*. A service can be multimodal in several ways. The only aspect that is beyond discussion is that for a service to be multimodal it must involve at least two modes. However, modes can be combined in many different ways. For example, the combination of speech (audio) for input from the customer to the service with text/graphics (vision) for the response of the service to the customer combines two modes, and thus a service that uses speech for input and text/graphics for output can be considered as a multimodal service. Indeed, some technology providers who are active in the field of multimodal services promote speech-in and data-out services as multimodal. However, a service can also use a combination of speech and pen for input, and/or a combination of audio and text or graphics at the output side. Services that combine multiple input modes (and/or multiple output modes) are multimodal in a more fundamental way. It may help to avoid confusion if the term multimodal interaction were limited to situations where multiple input and/or output modes are combined. However, this interpretation of the definition of *multimodal* would call for a new definition of speech-in – data-out services (which then might perhaps be dubbed ‘mixed mode’).

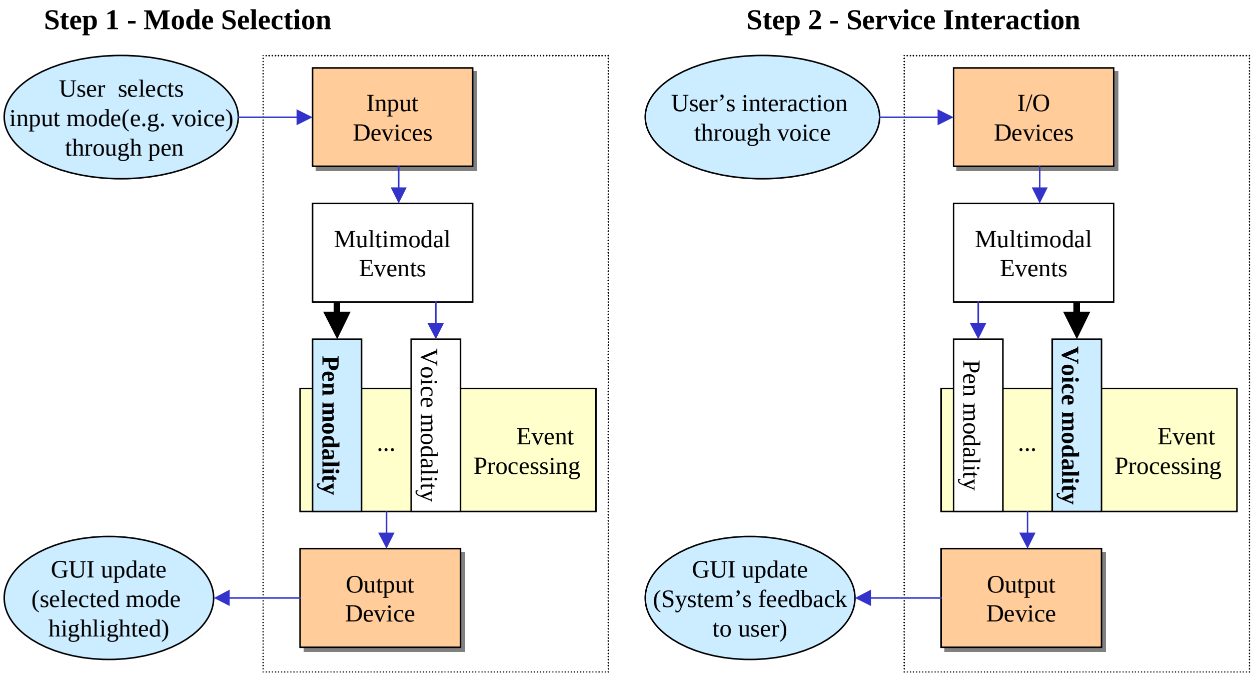

Figure 2 Schematic representation of sequential multimodal interaction

In this document we focus on services that are ‘doubly multimodal’, i.e., services that can combine speech and pen for input and speech and text/graphics for output. In other words, input is multimodal in that it combines audio and touch; output is multimodal in that it combines audio and vision. Moreover, we limit our attention to services that combine the input and output modes in a single session[1]. This excludes services that essentially require two independent sessions to complete a task, e.g., one session in which a query is formulated (for example by means of a spoken language dialogue) followed by a session in which the answer is sent in the form of an SMS message.

In the past the telecommunication industry has paid little attention to multimodality in services. The interest in multimodal services seems to stem mainly from the design of complex control panels, such as in the process industry (or in the flight deck of large airplanes). Moreover, there is recent activity inspired by the desire to provide access to the Internet from mobile terminals. In that context W3C has published a document entitled “Multimodal Requirements For Voice Mark-up Languages” [2]. In this document a classification of multimodal interactions is proposed. This classification mainly addresses the ways in which multiple input and output modes can be combined in a single session.

- Sequential, multimodal input (or output)

  In sequential operation the input modes cannot be used simultaneously. Instead, an explicit switch between the available modes is required, so that at any given moment only a single, designated input mode is active. An example of sequential multimodal input is the ‘Tap-and-Talk’ strategy that is being pioneered by Microsoft Research [6]. In Tap-and-Talk[2] the user has

[1] For the definition of the term Session the reader is referred to the list of Definitions in the first part of this document.

[2] There are probably several tap-and-talk solutions other than the one from Microsoft. In principle, the procedure is quite generic. The ‘push to talk’ strategy that has been (and still is) widely used in the US DARPA SLS community can also be considered as belonging in this category, despite the fact that the DARPA HLT/SLS programmes are not usually referred to as multimodal research. However, the first generations of ATIS systems all combined a push to talk speech input channel with a textual/graphical output on a screen. These systems can be considered as early examples of multimodal ‘speech in, data out’ services.

to tap the screen (perhaps a specific icon on the screen) to alert the system that she is going to talk.

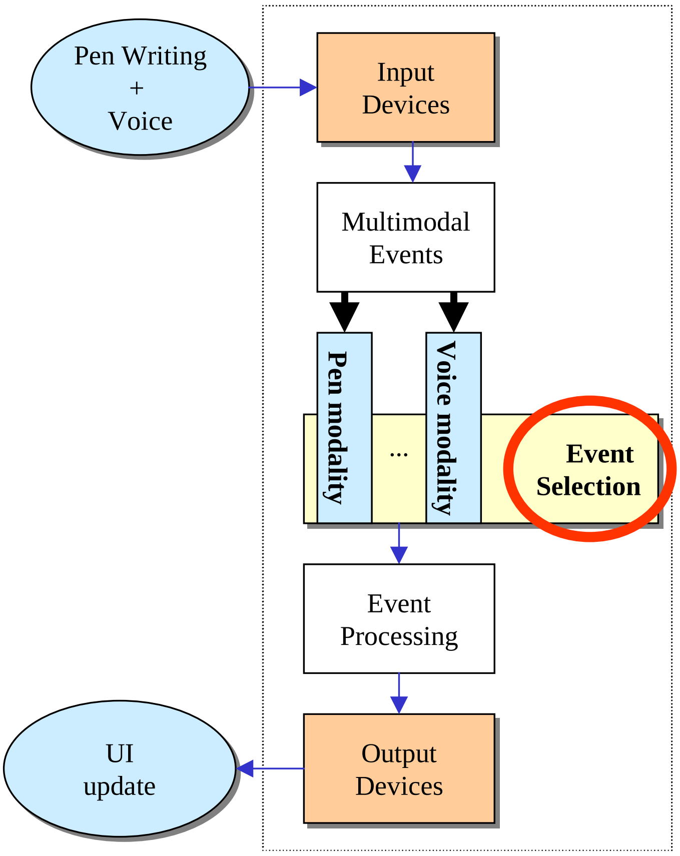

Figure 3 Schematic representation of simultaneous uncoordinated multimodal interaction

Thus, speech input is activated by a tap action. During speech input pen input is not active. Sequential multimodal interaction is graphically explained in Figure 2.

- Uncoordinated, simultaneous multimodal input (or output)

  In uncoordinated simultaneous operation all available input modes are active, but in each turn only one is selected for processing. For example, when a customer can use both ASR and a telephone keypad to enter information, and the task is to enter the account number, this number can be spoken or typed. If DTMF input is detected, speech that may be present at the same time will be ignored. Simultaneous uncoordinated interaction is explained in Figure 3.

- Co-ordinated, simultaneous multimodal input (or output)

  In this operation mode all available input devices are active simultaneously, and their actions are interpreted in context. An example of co-ordinated simultaneous input is where a user clicks an icon on the screen, and 'at the same time' asks a question about that object. It appears that the timing relation between speaking and gesturing in human communication is quite variable. Therefore, the definition of *'at the same time'* turns out to be one of the major difficulties in the design and implementation of multimodal interfaces. Simultaneous co-ordinated interaction is graphically explained in Figure 4.

As usual in specifications by W3C a distinction is made between 'must address' and 'nice to address' issues. The "must address issues" in the W3C document that are relevant for this discussion are Sequential and Uncoordinated simultaneous multimodal interaction. Co-ordinated-simultaneous multimodal interaction is in the 'nice to address' category.

The last official version of the W3C standard for voice mark-up languages, viz. VoiceXML 2.0, does not yet address multimodal interaction. In an attempt to speed up the development of a standard development environment a group of companies including Microsoft, Intel, Philips, Speech Works, CISCO, and Comverse have established the SALT Forum, to specify a set of Speech Application Language Tags (SALT), i.e.,

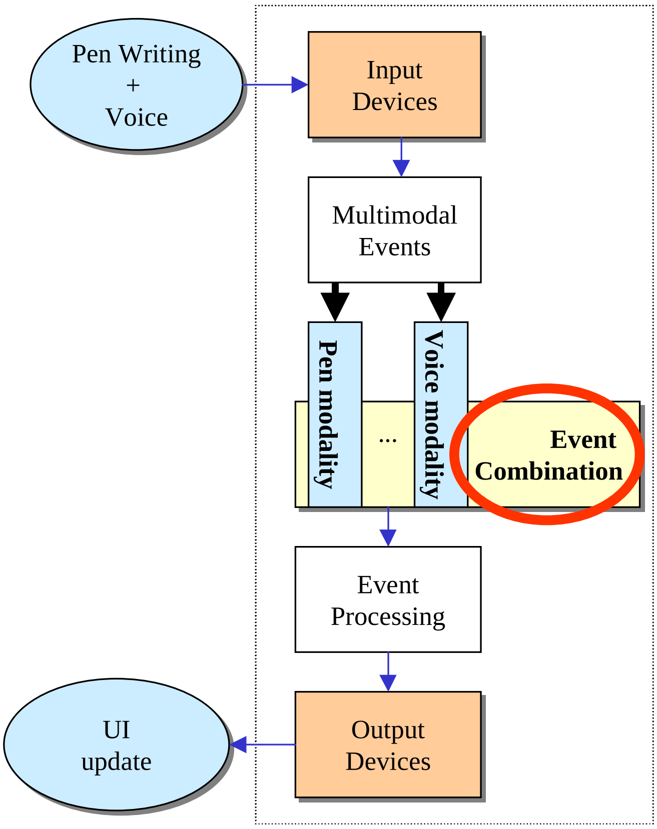

Figure 4 Graphical illustration of co-ordinated simultaneous multimodal interaction.

> ‘a lightweight set of extensions to existing mark-up languages, in particular HTML and XHTML, that enable multimodal and telephony access to information, applications and web services from PCs, telephones, tablets PCs and wireless personal digital assistants (PDAs).’ [3]

It is reasonable to expect that the results of the activities of the SALT Forum will eventually be integrated into standards supported by the W3C.

2.3 A multimodal architecture

Due to the presence of multiple simultaneous input and output modes the overall architecture of a multimodal system is much more complicated than what one is used to observe in more conventional monomodal systems [11]. The simultaneous presence of pen and speech input requires some kind of arbitration or interpretation of the two input modes. Information from these channels may be complementary (as will probably be the case in co-ordinated simultaneous operation) or it may be in competition (as in uncoordinated simultaneous operation). This arbitration and interpretation is the responsibility of the Fusion module (cf. Fig. 1). Due to the relatively novel and fluent status of R&D in multimodal interaction, Fusion is rather a research issue than a concept that is well understood. Some authors want to distinguish between Fusion on different levels. For example, when the signals coming from speech and handwriting recognition both represent the same words, fusion can already be attempted at an early processing stage. When speech and pen input contribute more independent pieces of information or evidence, fusion is more appropriate on a higher level of semantic and pragmatic abstraction. The former kind of fusion is dubbed ‘early fusion’ by some authors, who call the latter ‘late fusion’.

Fission in the architecture of Figure 5 refers to the distribution of information over the parallel (or sequentially used) output modes. The state-of-the-art in Fission is also research rather than established best practice. Consequently, proper use of available output modalities is more of an art than a science in user interface design.

Fig. 1 shows a global –and highly schematic- picture of the architecture of a multimodal system that has speech recognition, pen input (but also some kind of alpha-numeric keyboard) and GPS input, combined with text, graphics and speech (recorded or synthetic) for output.

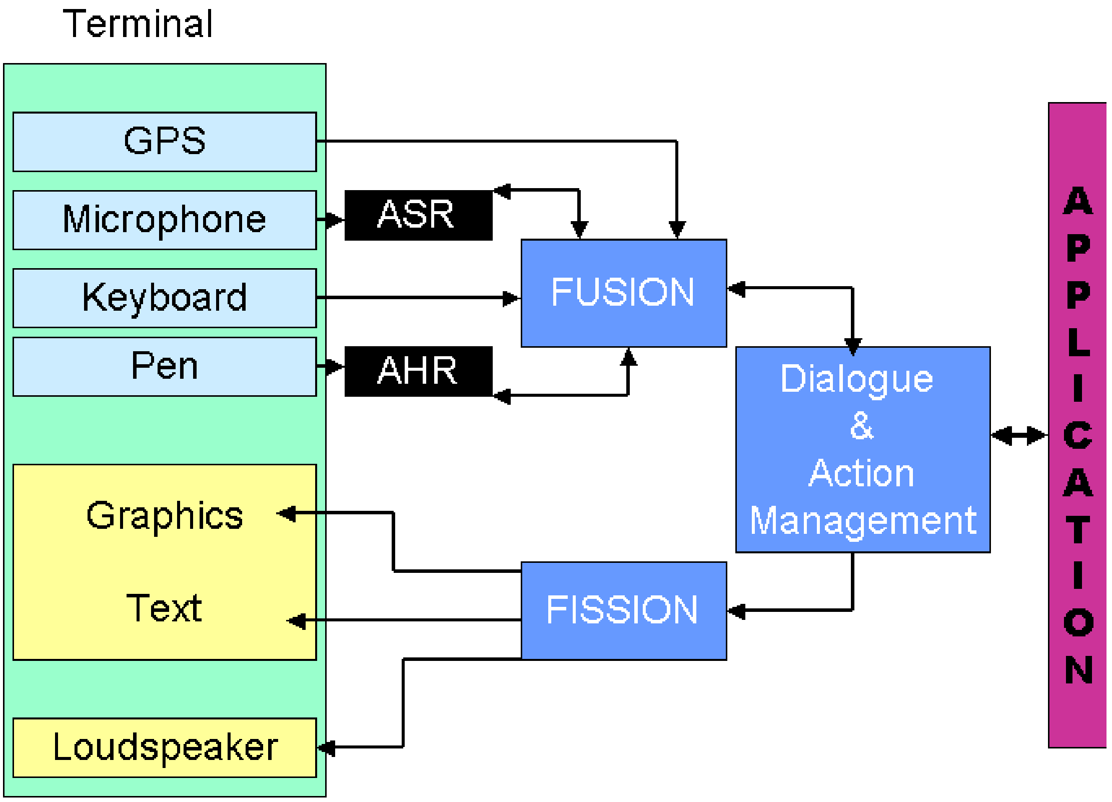

Figure 5 Schematic diagram of the modular architecture of a multimodal service

2.4 Why is multimodal Interaction important?

By default, all human communication behaviour is multimodal. People are known to gesture even during telephone conversations, despite the fact that they know that the interlocutor cannot see these gestures. Essentially all monomodal communication settings are imposed upon the users by the restrictions imposed by the technology (as for example with the plain old telephone). Not all information is easy to process and digest if speech is the only communication mode. When one deals with a travel agent over the phone, to arrange even a relatively simple journey, one almost always uses paper and pencil at the side to make notes of times, prices, etc. Interestingly, the travel agent takes recourse to the same means to simplify the task. Although it is perhaps less clear, Windows based interfaces for applications on desktop PCs also suffer from the constraints imposed upon the interaction by the fact that often only the screen, the mouse and the keyboard are available. Referring to objects that are not visible on the screen is difficult. In the very least it involves a sequence of actions to produce an icon or a text string referring to that object on the screen, so that it can be (double) clicked. Speech and natural language, on the other hand, make referring to invisible objects extremely easy, and natural – for that matter.

Thus, many information and transaction services are expected to become easier to use, and therefore more attractive for the general public, if the constraints imposed by the limited availability of I/O modes could be removed. The need for multimodality is only aggravated by the shrinking size of future mobile terminals. Maybe it is reasonable to expect the screen size to become less of a constraint as foldable or rollable screens will become widely available at low cost, but there is no obvious lightweight substitute for full-size keyboards in the pipeline that can be used for palmtop operation[3] (except for high quality, large vocabulary, robust ASR).

At the time of this writing the first generation of multimodal services are just emerging. We expect that the development of multimodal services will follow several different lines in parallel. In the following subsections we sketch two different types of services that are targeted to develop into fully multimodal services. The first class of services develops from the 'plain old telephone' (PSTN) networks; the other class will grow on the basis of Internet services. Both developments can be subsumed under a single keyword, viz. 'make services accessible from al terminals, and through all modes that a customer chooses to use'. The telecom network operators (i.e., the 'old' EURESCOM shareholders) are heavily involved in these developments. Therefore, these companies should be aware of the technical, and even more so of the usability issues that are involved in these developments. Despite the fact that all services under discussion need a telecommunication network infrastructure, the Telcos are no longer the only players in this field. Internet Service Providers, Application Service Providers, Content Providers, but also technology providers and system/application integrators have entered the field. Due to the presence of Internet-related players, and the fact that the next generation mobile network architectures (GPRS, EDGE, UMTS) support IP protocols, the emphasis in ongoing developments seems to be on device-independent access to Internet services, rather than on enhancing voice-based services. However, the future may well show that the developments converge so rapidly and so completely that the distinction looses much of its appeal.

The field of multimodal services is characterised by a tension between what is currently possible and what visionary scientist expect to see emerging in the medium- and long-term future. The expectations are generally characterised by a vision of 'ambient intelligence', i.e., a situation in which users can communicate with other persons and with a networked distributed information and support system as if they were talking to expert human assistants. The interface between the user and the ambient intelligent environment must be completely transparent [16]. The interaction with the networked environment will definitely multimodal, in the same way as human-human communication is multimodal.

2.4.1 Voice portals

'Voice Portals' is a cover term for a range of services that provide access to a more or less coherent set of information and transaction services from the voice telephone network. For the purpose of this document there is no need to distinguish between fixed and cellular networks. The emphasis is on the implication of the use of small terminals that have the keys integrated in the mouthpiece (so that it is cumbersome to key in DTMF sequences while listening to spoken prompts). At the same time the screen size makes it difficult (if not impossible) to display the information retrieved by the service in a textual or graphical form that is easy to read (so that it does not need to be remembered after hearing it just once). Although a number of voice portals based on proprietary technology are still in operation (and perhaps quite successfully) there is a clear trend towards a tighter integration of voice access and 'conventional' Internet services. This development is fostered by the increasing acceptation of VoiceXML as the standard for mark-up in information and transaction services that are accessed via the telephone. In the future it should become possible to generate VoiceXML compliant versions of the large majority of the XML based Internet services. On the one hand this should allow almost any Internet service to become accessible from a wide range of terminals, in-

[3] Detachable and foldable keyboards are available for some palmtop computers and PDAs. However, these appliances can only provide laptop-like functionality if the device can be placed on a sufficiently solid table, at a height and distance that is suitable for typing.

cluding voice telephones. However, it is evident that a large proportion of Internet services is inherently not suitable for voice-only access. Many services would be easier to use if there were at least some form of graphical display to support navigation or to capture and keep complex factual information. Thus, we see a growing need for multimodal terminals to overcome the usability constraints of voice-only access to Internet based information and transactions.

2.4.2 Mobile internet

Parallel to the development to provide access to Internet services through VoiceXML platforms we see an independent growth of mobile Internet services. Many PDAs are advertised as fully functional palm computers, which will be coupled to the Internet in the same way as desktop terminals, as soon as GPRS, EDGE and, eventually, UMTS networks will become operational. WAP and I-mode have started a development in telephone handsets towards Internet terminals. In addition, we will see a range of communication devices designed for use in cars, where hands-free – eyes-free operation must be possible during driving, in alternation with the use of manual input and graphical output devices that can be used when the car is parked. However, all these terminals are expected to suffer from the lack of full-fledged keyboards that can be used to enter text of some length, such as an SMS or an e-mail message. Easy access to conventional Internet services from terminals that have substantially less I/O facilities than desktop terminals is only possible if the limitations of the devices that must be small and light can be overcome by offering alternative I/O facilities, properly integrated in the user interface to the terminal and to the service.

2.4.3 The role of the network operators

Voice portals and mobile Internet services are at the heart of the business of the telecom network operators. The importance of these services will only increase as the integration of now disparate telecommunication services continues. In the last decade quite a large number of advanced integrated services have been subjected to market trials, conducted by network operators and infrastructure providers. Most of those new services (e.g. personal call assistants) have not survived the first rounds of market trials, because the customers found the offerings too complex and too difficult to use comfortably. There are two tightly intertwined issues that play a role in this context. First, prospective customers find it very difficult to form a correct mental image of the new services. Such conceptual problems are inherent in every new type of service: customers must learn why they are useful, and how they can be used. Second, the capabilities of the terminals with which these new services had to be operated did not allow to create the type of user interfaces that could support new customers in building appropriate mental images of the services. It is expected that the developments towards full multimodal interaction will alleviate both problems. As mobile telecommunication services acquire a look and feel that resembles what one is used to have on the desktop in one’s professional work, it will be easier to build an appropriate mental image. At the same time the multimodal terminals will make it easier to design the type of user interfaces that can actively support the creation of correct mental models (in addition to making the interaction more comfortable in many other respects).

Up to now only a small number of experiments have been performed to assess the major Human Factors issues involved in multimodal services. The results of these studies are not always easy to interpret. Yet, it has been convincingly demonstrated that in order to be potentially successful, mobile multimodal services will have to be designed in their own right, observing a completely new set of best practice guidelines. Due to the fact that all bits and pieces that will eventually make up mobile multimodal services are just emerging, it is not evident what the ‘natural’ role of the network operators is. We are convinced that these services can only become successful if they address real communication and information needs of a large part of the population, and if they are sufficiently easy to operate. The European telecom network operators are eminently positioned to gain a strong position in the emerging field.

In chapter 3 we will go into more detail about the ways in which the multimodal services of the future can be designed and developed. We will argue that the potential and prospective players can take several different positions in this activity. We will provide basic information on the technical, usability and marketing aspects of that development, which should help the shareholders to define their position.

2.5 The role of Multilinguality

There is a tendency in the literature on multimodal interaction to emphasise the challenges and the promises of the combination of input modalities, and at the same time to de-emphasise issues related to fission, i.e., to the ways in which the output of a service is rendered. In GUI design it is now customary to separate the business logic from the presentation layer. This allows to design applications in an essentially language independent manner. Localisation can be restricted to the presentation layer, where linguistic representations of the information can be generated and subsequently rendered on the screen. This is probably a somewhat optimistic view of the world, as a full separation of the business logic from a truly language-independent presentation layer is far from trivial, as was shown in the EURESCOM project P923 BabelWEB [12].

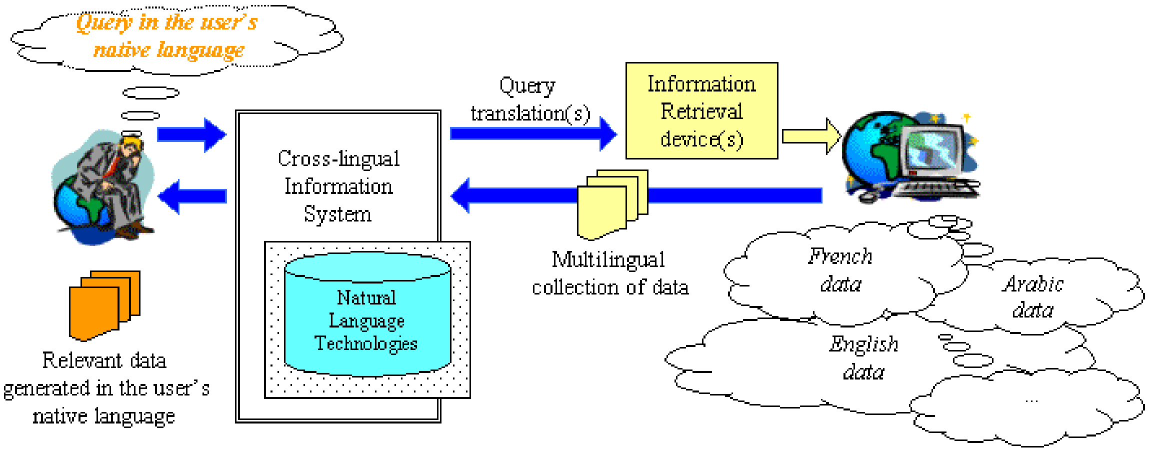
Query in the user's native language
Query translation(s)
Information Retrieval device(s)
Cross-lingual Information System
Natural Language Technologies
Multilingual collection of data
French data
Arabic data
English data
...
Relevant data generated in the user's native language

Figure 6 Typical multilingual Information service

Nevertheless, users of future mobile services are likely to consider their mobile terminal as a universal access device for all the services they subscribe to. And they will expect these services to ‘behave’ in the same manner, irrespective of the place where they happen to be at a given point in time. This is somewhat like the interface of application programmes on a laptop computer, which do not adapt the language in the interface to the country where the user happens to be. Therefore, we expect that future mobile services will be required to have the same type of language selection in the interface as present day desktop/laptop applications. More specifically (as appeared from the overview of multilingual web services in BabelWeb) customers want to be able to access services in their native language. The success of services clearly depends on the ability to address prospective users in their native language. This finding is reflected in the fact that there is a rapidly increasing number of web services that are offered in many different languages. English is becoming somewhat of a fallback option, which is available in case the native language of a user is not supported. Since the majority of the future mobile services will rely on some kind of Internet access, this implies that the services for which we are developing user interfaces need to be offered in several, perhaps many, different languages. Figure 5 shows a conceptual representation of a typical multilingual service based on Internet access.

2.5.1 What is ‘multilinguality’?

Services and applications can be multilingual in different ways. In desktop applications programmes such as WinWord are usually configured for a specific language, at least as far as the user interface is concerned. Modern text processors automatically determine the language of the text that is being edited, and dynamically adapt spelling and grammar checking to that language. It is reasonable to expect that small mobile devices will be personalised, at least in the sense that the owner must set a language preference (much in the same way as with the present generation of GSM phones). The architecture in Figure 5 should allow language negotiation between a terminal and the service logic. To accomplish this, ASR and AHR at the input should consist of an array of language specific modules, each of which must have the same functional specifications. It goes without saying that the modules also must have the same API, independent of the language. At the output side the Fission module must contain a complementary array of text generation modules and TTS systems.

It is unlikely that all network operators and all service providers in the multilingual Europe of the next decades will be able to offer ASR, AHR, text generation and TTS for all languages of Europe, in addition to Japanese, Chinese and Korean (to name just the presently most important non-European languages). For example, the number of Norwegians travelling in Portugal may not justify the addition of Norwegian speech processing to all information services offered through the network of Portugal Telecom. In these cases the customer must have the possibility to explicitly select the language (s)he is most comfortable with for the interaction in the user interface. It is necessary to make a distinction between the language choice in the interface on the one hand, and the language(s) used in the ‘application’ module in the architecture of Figure 5 on the other. Some applications (for example timetable information, flight reservation, etc.) are essentially language independent: it is easy to see how destinations, dates, fares, etc. can be mapped on symbols that are no longer dependent on a specific language. However, if an application needs to access free text databases in order to search for information requested by the user, the situation becomes very different. In many cases the information may only exist in one single language, which may very well be different from the preferred interface language. An obvious example is detailed information about a small town in a region of –again- Portugal that has not enjoyed interest of large crowds of tourists. And even for the more popular destinations it may not be possible to provide all information in a large number of languages. In these situations some kind of cross-lingual operation is called for. Therefore, information retrieval systems must be developed that enable querying and displaying results in the user language, even if the information that the user is seeking is written in another language. Highly reliable -and domain specific- automatic translation mechanisms are likely to be used, although the large number of languages to be processed calls for generic (domain independent) translation procedures. The need to be able to provide cross-lingual services will certainly lead to the emergence of a clear role for multilingual content providers in the pattern of ISP economy. The details of the implementation of multilinguality are likely to differ between services and service providers.

2.5.2 Multilinguality and Multimodality

As multimodal interaction matures, and the set of input and output modes provided by ‘standard’ terminals grows, the role of Fusion and Fission as sketched in Fig. 1 will become more important and more evident. It will appear that the semantic interpretation of the input signals will grow more abstract, and become less tightly bound to linguistic meaning representations. In the generation of the output a similar trend is already visible, in projects that aim to develop techniques for rendering of the information to a wide range of devices with widely different screen size and resolution.

At the same time we are witnessing the emergence of a type of content providers who offer information in a number of different languages. Publishers of tourist guides that are available in many languages are just one example. The report of the BabelWeb project contains many additional examples and references [12]. Not just content providers, ISPs and ASPs are moving towards European (of global) presence, crossing language borders; a similar trend has been visible among the European network operators. Despite the backlash caused by the present economic difficulties it is widely believed that this trend towards international presence is irreversible. Therefore, there can be little doubt that network operators must be able to offer services in many different languages in parallel.

2.5.3 The role of the network operators

As was the case with multimodal services, there is no ‘natural’ role in the field of multilinguality for the network operators. The presence of new international (or multi-national) operators such as Vodafone, Hutchison Wampoa, and to some extent also NTT DoCoMo on the European market will force the national operators to compete with roaming services that present the same user interface, irrespective of the country where the customer happens to be. However, it is reasonable to expect that these services will be offered by some kind of ‘joint venture’ between network operators, content providers, technology integrators, and perhaps also terminal manufacturers.

In the emerging multicultural society network operators and service providers will see a growing proportion of local subscribers whose native language is different from the dominant language in the country. First generation (but perhaps also part of the second generation) immigrants may prefer to interact with complex services and applications in their native language. Already today almost 50% of the inhabitants of a city like Amsterdam do not have Dutch as their first language. Being able to cater to the needs of this growing customer base will certainly strengthen the competitive strength of all service providers. Therefore, the European telecom service providers will find it increasingly important to offer their own local services in a large number of languages.

One reasonable scenario (but certainly not the only one) would be that network operators offer the infrastructure that is needed to handle multilingual input and (speech) output. In this scenario the local/national network operator must provide the capability to negotiate the language used in the interface. This negotiation may have to be repeated for each application, since it need not be the case that all applications that run in a network support the same set of languages. The functionalities of the handset/terminal may also interfere, for example when codes are used to trigger the display of fixed messages that are rendered linguistically.

In general terms all players in the field of multimodal services must think about the optimal distribution and allocation of the building blocks in the architecture depicted in Fig. 1. In the scenario sketched in the previous paragraph ASR, AHR and TTS engines would be located in the switches, and therefore owned and operated by the network operators. It is reasonable to expect that Fusion and Fission, and even more so Dialogue and Action Management will be developed and owned by content and service providers (in close collaboration with technology providers and system integrators). However, if it appears that future PDAs will become powerful enough to do ASR, AHR and TTS in the terminal, the distribution of the work between the client and the servers in the network will change, and affect the share of the network operator. In this highly uncertain and rapidly evolving situation it will be essential to adhere to formal and emerging standards. Fortunately, there is a clear trend in the industry to foster the development of standards that support interoperability of building blocks, making them independent of the hardware and software platform that they run on. Companies that intend to develop and/or offer multimodal services should use their presence in standards organisations such as W3C, ETSI, ITU, etc. to closely monitor the development of such standards.

3 Issues in present and future Multimodal Services

3.1 Introduction

Presently there are only very limited number of commercially available multimodal services in operation, mainly built on proprietary technology (see also chapter 4). However, many techniques and theories are getting into shape that eventually will make it easier to develop and implement multimodal systems and services for small devices. Therefore, we are convinced that in the short- and medium-term future multimodal services will emerge, irrespective of the business strategies of the European network operators. The main aim of this chapter is to show the present and expected developments of multimodal services (how far are we, and what are the problems). Until now the European network operators are not so much involved in these development. Since they have a vital interest in the growth of the use of the Mobile Internet, they should be aware of what is going on, so that they are able to determine their position in the field.

In this chapter we will discuss present and future multimodal services. We concentrate on services in the cellular networks. Invariably, these services are accessed with small, lightweight mobile terminals. Most of the discussion is limited to multimodal services in which the customer can use multiple input channels, such as speech, keyboard, mouse or pen, etc., to enter information[4] required by the service. At the same time, the response of the service can use multiple output media, such as graphics and text displays, speech and non-speech audio, or vibration. It should be clear that the input and output modes that can be used depend both on the service design and on the hardware and software capabilities of the terminals. The look-and-feel of the user interface is further determined by the interaction protocol, i.e., by the decision to use sequential, uncoordinated-simultaneous, or co-ordinated-simultaneous interaction (cf. section 2.2).

It is generally expected that virtually all new services that will become available thanks to the combination of the new generation of cellular networks and mobile multimodal terminals will be Internet services. During the last decade a substantial amount of experience with the design and the marketing of Internet services has accumulated. However, the class of services that is going to be offered through mobile terminals will differ from the conventional Internet services on the desktop in three important respects:

1. Unlike desktop workstations and laptop computers that come with a standard set of input and output devices, multimodal mobile terminals are only just emerging, and there is still a long way to go to reach some kind of standardisation.
2. It is expected that the trend in mobile terminals towards smaller size and lower weight will continue. Because of the size and weight reduction of the mobile terminals, their input-output capabilities will differ significantly from those of desktop terminals. As a consequence, the use of mobile terminals will require the development of a completely new brand of interaction metaphors, with all attendant implications for Human Factors issues and the overall design of services.
3. Desktop Internet services could develop on the basis of the original packet switched IP protocol, since all user interaction could be supported by local processing of essentially digital I/O signals and transfer of data packets over the network. Mobile multimodal services must combine packet and streaming data types from the very beginning. And they must do so in the absence of well-established protocols for the transfer of a combination of packet and streaming data.

[4] GPS can also be considered as an input device for an increasing number of services that adapt their behaviour to the location of the user. In this document we will not explicitly discuss GPS, despite its commercial importance. However, users will hardly ever explicitly manipulate GPS input in the interface for a service. Therefore, GPS is not a major issue in the usability of the interface. On the other hand it should be clear that the interpretation of speech and pen input in location based services will depend to a large extent on the information provided by GPS.

Therefore, while designers of desktop Internet services can rely on the availability of a well-understood set of I/O functions in fairly standard terminals, the designers of mobile Internet services must work under conditions that are less well defined, and that will keep changing for a considerable time to come. Up to now only a small number of experiments with multimodal interaction have been conducted. Yet, all results point in the same direction: the interface for the services needs to be developed from scratch. Porting these services from a desktop workstation to a terminal that is much less capable will fail almost by necessity.

3.2 Mobile Internet Terminals

In this section we briefly sketch the most likely development in the field of mobile Internet terminals. In doing so, we will indicate the most important consequences for the volume of the services that can be offered for those terminals.

It is, of course, possible to aim at the use of laptop computers as the terminals for mobile Internet services. There will undoubtedly be a market for that class of services. However, that market is likely to be rather small. The trend in the development of laptops is towards retaining the same complement of I/O devices as desktop PCs have, rather than towards decreasing weight and size. This is quite natural, because in actual practice laptops mainly function as ‘portable desktop workstations’. Therefore, we do not expect that the market for mobile Internet services with full-fledged laptop computers will attain a size that is sufficient to recover the investments in 3G mobile networks.

At the other extreme of the scale we have the present generation of GSM phones, which are already essentially multimodal in that they come with audio and keyboard input, and with audio, graphics, and text output. Despite the fact that the 12-key keypad is far from optimal for entering text, SMS has become one of the most successful GSM services. This shows that –at least under some conditions- inadequate input (and output) devices can be overcome. However, it is unlikely that customers will also be willing to use the limited input capabilities of the 12-key keypad to access the type of Internet information and transaction services that are successful with desktop terminals. The lack of success of WAP is here to support that expectation. Therefore, it is quite likely that the present generation of GSM handsets can support only some highly specific niche services, if at all.

In between desktop PCs and GSM handsets another class of terminals is emerging, viz. the organisers and PDAs. These devices are also called palmtop computers. The latter name suggests that they are closer to desktop PCs than to GSM handsets, and in many respects that suggestion is true. However, it appears that many of the applications that are being developed for PDAs will be much more useful if some kind of on-line link with a network of powerful workstations and mainframe computers can be established. At the same time it appears that applications for PDAs that require substantial amounts of (text) input are hindered by the absence of a full-fledged keyboard. This explains why PDA manufacturers are working towards the addition of telecommunication interfaces as a standard feature, and at the same time towards an enhanced set of I/O capabilities such as ASR and automatic recognition of handwriting to solve the Human Factors problems in services where the user must enter substantial amounts of information.

Future mobile multimodal services will be dependent on the functionality of the handsets and terminals. For this reason it is essential that network operators define their position relative to the terminal manufacturers.

3.3 Enhancing voice or data services?

The development of GPRS and UMTS terminals must guarantee downward compatibility with the present generation of GSM services. In chapter 2 we have already introduced the two major development lines, viz. voice versus Internet services. At its inception the GSM network was designed for digital voice services. However, the success of SMS has shown that voice is definitely not the only service that caters to the requirements and needs of the customers. The GPRS, and even more so the future UMTS networks seem to be conceived as mobile equivalents of the landline Internet network, probably with conventional Internet-type services in mind, rather than voice services. At the same time it is reasonable to assume that GPRS and UMTS terminals will have some features in common with the present generations of both GSM handsets and PDAs.

The present design of PDAs may not be suitable for voice telephony communication. On the other hand, they are probably better suited for SMS messages and Internet access than the present GSM terminals. Whether GSM handsets or PDAs make for the most promising starting point seems to depend on the business strategy of the network operator. If GPRS and UMTS services are going to be marketed as value added to a voice network, then the GSM handsets are probably the best starting point. However, if the GPRS and UMTS services are going to be marketed as mobile versions of existing Internet services, then the PDAs are the starting point of choice. However, irrespective of the starting point (voice or Internet), the Human Factors issues in the use of multimodal mobile services will be similar, if not identical.

3.4 The players in the field of multimodal services

Presently, we identify five major stakeholders in the field of mobile multimodal services, viz.

1. Network operators, although it seems that so far these companies have not taken a very active role in the developments
2. Providers of network infrastructure, which have not been particularly active in the field either
3. Terminal manufactures, a group of companies that partly overlaps with network infrastructure providers
4. Content providers, i.e., the companies who own the information that end users might want to buy
5. Multimodal technology suppliers, i.e. companies that provide solutions and technology components for building services (like handwriting recognition, speech recognition, fusion, fission, dialogue control, etc). This category of companies has been most visible in the development of multimodal services.

What is still missing in this picture is the application developer that designs and creates the services. This situation seems to provide room for a new type of stakeholder, in the form of companies who invest in the development of services, who define the functionality of new services and design appropriate user interfaces on the basis of whatever is made available by the terminal manufacturers and supported by the network operators. But the void can also be filled by the conventional stakeholders, who can join forces in several different ways. The players in the field of multimodal and multilingual services of the near future are also shown in Figure 7. It should be noted that the order in which the players appear in that figure does not intend to suggest a conceptual organisation of the relations that will emerge.

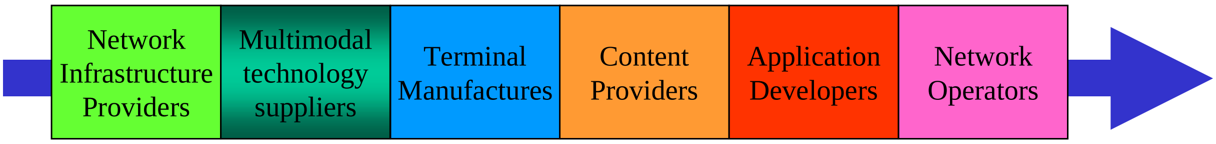

Figure 7 Overview of the players in the field of future multimodal multilingual services. The order in which the players are shown does not suggest 'natural' relations between the players.

We expect that the network infrastructure manufactures will not make large contributions to the development of new services. They must be involved, of course, because their products must support the requirements of the services, for example in the form of suitable protocols for the combination of packet and streaming data.

It is clear that the network operators cannot develop new services on their own. They will need suitable terminals on the one hand, and appealing content on the other. The same is true for the other key players. Therefore, it appears that the development of appealing new services requires some kind of contribution of each of the key players. In this context it may be illuminating to point out that the success of I-mode as introduced by NTT-Docomo was the result of close collaboration between content providers, terminal manufacturers, network suppliers, and the operator NTT-Docomo, under the guidance of the latter.

The network operators are probably guaranteed to have one part of the cake, because they operate the basic infrastructure that is needed to offer the services. Just how big this share is, is less clear. If the network and service infrastructure is designed along the lines of the present fixed network, the network operators may be able to provide the billing part of the service to generate extra revenues. However, if billing is based on the value of the information that is provided, the billing service may well be part of the share of the content provider. The role of the network operators as providers of the communication infrastructure that is needed for the services defines a position very close to the manufacturers of the network equipment.

The network operators are not limited to the provisioning of the infrastructure that is needed to support new services. On the contrary, they can take a more active role, and work with terminal manufacturers and content providers on the development of new services. In doing so, the network operators will be able to influence and guide the development of the terminals (and probably also the communication protocols that are supported by the terminals) and the formatting of the information in the hands of the content providers, so as to make it maximally suitable for access by mans of small mobile terminals and limited bandwidth networks. A more active role can take several different forms, depending on who takes the lead, and the degree of financial risk. Two different options are sketched below.

3.4.1 Contributing partner

In every consortium partners have specific and different roles. One such role is that of the initiator, i.e., the party who has a vision and who assembles the consortium to make that vision become reality. Other partners join because they believe that this vision may become valuable, and because they think they can eventually profit from contributing to its realisation. In the field of future mobile multimodal services such a consortium will probably be formed to develop a range or a family of services, all based on a single family of terminals, and similar types of information.

Probably the most important asset that the network operators bring to a consortium that aims at the development of a new range of services is their subscribers, and their knowledge about the behaviour and the needs of their subscribers.

If the focus of a consortium is either on the development of future terminals, or on the organisation of potentially appealing content, the role of the network operators is almost by nature subsidiary. That opens up possibilities for pre-competitive collaboration between several operators.

In our opinion the experience with the development and market introduction of WAP has shown that the network operators should contribute much more to a consortium that aims at the development of a new family of services than just technological expertise. On the contrary, the operators should also harness their knowledge of the usability and the marketing aspects of those services.

Participation in a consortium that develops the equipment, protocols, information databases and user interfaces for new services can yield essential benefits, the most important of which is probably a shorter time to market than competitors who must begin to understand the details of the new equipment, and the new interaction paradigms after they are released. It is well known that the type of complex services that we are dealing with here involve many details that same unimportant from the outside, but that appear to be decisive in many practical circumstances. By actively contributing to the development network operators can build up the knowledge and the understanding that is necessary to avoid the pitfalls on the way to commercial deployment. It has been rumoured that one of the major reasons for Philips' failure to reach its targeted market share in GSM handsets is its absence in the market for network infrastructure. The companies that develop the infrastructure and have an advantage in the development of the terminals that go with the infrastructure that is almost impossible to make up for outsiders.

3.4.2 Leading partner

In the preceding section the leading partner in a consortium was characterised as the one who has the vision, and pushes towards the realisation of that vision. This partner can very well be a network operator, provided that it considers the development of new classes of services as part of its core business. Network operators who have decided to focus on their role as providers of the communication infrastructure are not likely to take on this role.

It is questionable whether the difference between the role of participant and the leader is very large for the European network operators. In today’s open markets, where volume production is essential for the terminal manufacturers, it is unlikely that new equipment can be developed for a single operator.

3.4.3 Why should network operators work on multimodality

Despite the rapidly growing number of articles in the press about multimodal telecommunication services, the actual number of mobile multimodal Internet services is quite small. There are several reasons for this state-of-affairs, which are of crucial importance for network operators. Multimodal services do not come in an off-the-shelf fashion. On the contrary, the few services that do exist are all tailor made, relying on in-depth knowledge of the component technologies, platforms (both hardware and software) for the integration of the components and for the coupling of the I/O channels with the back-end application. For a network operator it would be virtually impossible to develop mobile multimodal services without comprehensive knowledge of the technologies and platforms. But the situation is not different for the other players in the ballpark. Because of the lack of 'standard' multimodal services network operators should not count on the possibility to buy turnkey solutions from external suppliers.

Network operators must be aware of the importance of in-house expert knowledge and experience with multimodal services, because it is virtually impossible to separate the specification of the functionality of such a service from the capabilities of the interface. Adding seemingly simple functions may very well make the difference between a situation in which co-ordinated simultaneous multimodal input is mandatory, and a situation in which one can suffice with unco-ordinated simultaneous, or even sequential multimodal input, which are much easier to implement in a standard fashion. Only an expert will see these implications in time, and only experts are able to take appropriate actions.

Last but not least, there is close interaction between the functionality of networks and terminals and the functionality and interface of telecommunication services. The development of successful new services in the mobile Internet will be crucially dependent on the collaboration between network equipment and terminal manufacturers on the one hand, network operators on the other, and service providers as third parties. The network operators need substantial in-house expertise to play their part in that development.

It is expected that future multimodal services will be accessible from a wide range of different terminals, each with their own specific functionality, and their own transmission protocols. For a long time to come these services will rely on both the switched and the packet based networks. Figure 8 shows how a single service may be accessed from several different environments, with different terminals, that use streaming and data transmission protocols.

3.4.4 Importance of Multimodality for terminal Manufacturers

Currently, all partners in MUST represent network operators. It is difficult (and perhaps unwise) for representatives of network operators to speak on behalf of network equipment and terminal manufacturers. Therefore, we will confine ourselves here to a very concise summary of what we think are the most relevant aspects of multimodal services that require active involvement of the manufacturers.

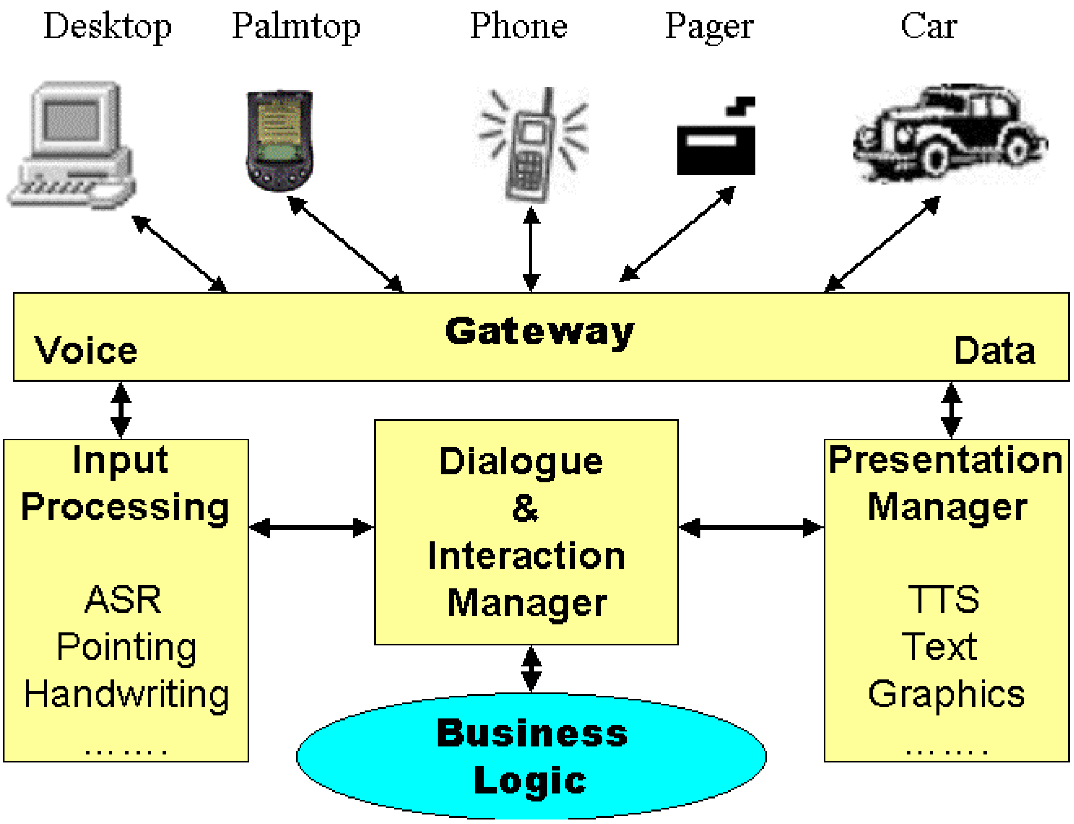

Figure 8 Multimodal multilingual services will be accessed from many different terminals, environments and networks.

3.4.4.1 Hardware/software specs of terminals

Multimodal services are fully dependent on the availability of multimodal terminals. First, the basic I/O modalities must be provided in the terminal. Therefore, decisions must be made as to the availability and functionality of pen input and ASR for input and audio, text and graphics for output. The functionality of the terminal will determine the interaction strategies that can be used in a service. For example, if a terminal provides speech and pen input, but it does not properly time stamp the pen input, co-ordinated simultaneous use of speech and pen is not possible.

The terminal manufacturer also influences the service interfaces by decisions about the browser[5] that comes with the handset.

Issues that must be addressed specifically include Distributed Speech Recognition (cf. section 5.3.1) and speech synthesis. In more general terms, terminal manufacturers must make decisions about the priorities for the implementation of functions in ASIC (Application-Specific Integrated Circuit) or as software.

3.4.4.2 Protocol stacks to support realistic services

In the recent past we have encountered several occasions where it appeared impossible to carry out experiments with multimodal services in operational networks and with commercial terminals because of the lack of proper protocols for the transmission of parallel information streams, representing different modalities. For example, it is generally known that WAP services might be made much easier to use if it were possible to enter text through ASR. However, the WAP protocol stack

[5] For the definition of the term Browser the reader is referred to the list of Definitions in the first part of this document.

does not provide the possibility to combine speech and keyboard data in a single session. Thus, the only way to use ASR in a WAP service would be to implement the complete ASR in the terminal, and provide the software to overlay ASR upon keyboard entry, also in the terminal. Obviously, this is not feasible.

Network equipment and terminal providers have a large stake in the developments that are now coming off the ground in standardisation bodies such as W3C, precisely because of their interest in the relation between the functionalities of the terminals, the protocols and the networks.

3.4.4.3 Knowledge of future applications

Given the tight interrelation between the terminal, the network and the service, knowledge of future services and knowledge of the contents on which those services will rely is becoming increasingly important. For the time being the equipment manufacturers seem to be at a larger distance from the service providers than the network operators. Closer relations may prove to be extremely important for the manufacturers, to strengthen the basis on which they must make decisions for the development of future terminals (and network infrastructure).

3.5 Scenarios

In the previous section we introduced the major commercial and technical issues related to mobile multimodal services. In this chapter we present a couple of scenarios for concrete services that should make our vision of the present and future of mobile multimodal interaction more concrete.

These scenarios describe speech centric multimodal services from a functional and interface point of view. The technical implications for hardware and software are indicated on a high level of abstraction. More detailed information on the technical features are provided in chapters 5 and 6.

3.5.1 2002:

In the year 2002 we will see two types of mobile services develop in parallel. One type will build upon the facilities of the conventional voice network; ASR will be combined with limited support from graphical feedback. The alternative development will build upon Internet services; the lack of suitable keyboards will be remedied by the introduction of ASR technology. Due to the limited processing power of the terminals the services will rely heavily on distributed processing. Additional scenarios can be found in the marketing material of companies that are offering multimodal services. References to those companies can be found in section 4.1 below.

3.5.1.1 Voice-based services

Perhaps the single most important drawback of ASR based voice-only services is that customers find it difficult to maintain a consistent mental model of the status and progress of the dialogue, and to remember detailed factual information that they have only heard, not seen. Voice-In/Graphics-Out interaction can kill two birds with one stone: by displaying a kind of form to be completed on the screen and by showing the detailed information that was sought after on the screen when it has been found, both problems can be solved. In order to facilitate the completion of the form ASR can be used. However, since the terminal does not have sufficient processing power to implement the flexible vocabulary ASR, a coded version of the speech must be transferred to a powerful computer in the network. Since the terminal must maintain a simultaneous data link for the transportation of the graphical output, some kind of proprietary technology is needed to combine speech and data in a single session. Several companies are offering technical solutions that implement what can be considered as voice browsing based on GPRS or EDGE. In fact, any mobile packet network with sufficient bandwidth would do the job.

Based on this technology it is possible to offer restaurant location services for users who have only a standard GPRS phone. If the customer calls the service –perhaps through a portal that offers a menu that contains a range of related (entertainment) services- she will first be asked whether the restaurant should be in the direct neighbourhood of where the call is made. The location of the caller can be determined sufficiently accurately by means of the cell information. If the question is answered in the positive, a list of nearby restaurants is retrieved. If the list is too long, the service will ask the caller a couple of additional questions, such as regarding the type of food, price, atmosphere, etc. Each response will reduce the number of options, and if the remaining number is small enough to be browsed on the screen of a mobile phone, the service will advance to showing the selected restaurants. The service may display pictures of a selected set of restaurants, to assist the caller in making a decision. When the caller has made her choice, the service will call the restaurant, and attempt to make a reservation. If that attempt succeeds, the service will confirm the success by means of an SMS message. The customer can then establish a video call with her friend, using the CMOS mini-camera built into the handset, to make final arrangements for the dinner.

3.5.1.2 Internet-based services

In the year 2002 Internet-based services in Western Europe will also build on the GPRS network. The most important difference with voice-based services will be the type of terminal. To users who have a PDA with a GPRS radio connection built in a wide range of information and transaction services can be offered, adapted from conventional Internet services to account for smaller bandwidth and less powerful displays. However, the PDAs have significantly better display facilities than GPRS phones. In addition, they will also allow pen input, in addition to ASR. Multimodal input will be limited to sequential interaction, pretty much like the tap-and-talk strategy explained in section 4.2.4. With a PDA-like terminal customers will be able to make flight and hotel reservations world- wide. Destinations can be entered by voice, while dates and times can be selected through drop down menus. Entertainment services such as horoscope, movie and theatre information, etc. are also easy to access, again with a combination of ASR and pen clicks. Thanks to the presence of colour displays it will be possible to show short sequences of pictures, for example of the interior decoration of a restaurant, or the posters of the movies showing in a theatre.

3.5.2 2004:

In the year 2004 the bandwidth of the mobile networks in densely populated areas has increased to such an extent that QoS is comparable to basic rate ISDN (in other words, equivalent to 128 kbit/s). Terminals have evolved to a point where it has become difficult to tell the difference between a PDA and an advanced phone: all terminals have a colour display of at least 25 cm^2 that is touch sensitive to allow point-and-click actions for input. Also, all terminals come with a wireless earphone-microphone set, and a CMOS mini-camera. Protocols for simultaneous audio and data transmission have been standardised; the same holds for the XML mark-up that is needed to support services running on the small mobile terminals.

A wide range of professional services is available, allowing business persons to maintain full synchronisation between their PDA and the information in the company Intranet. Adding new appointments to the agenda, sending standard replies to e-mails and entering new orders will be done by means of a mix of ASR and pen input. Interaction with the services is still somewhat structured, to take account of the limited capabilities of the ASR and NLP modules, which invariably run on computers in the network, to be able to dedicate processing power and battery power to processes that must be performed in the terminal. Most services will offer simultaneous, but uncoordinated use of ASR and pen for input. A small number of advanced services will support simultaneous coordinated input.

The information, transaction and entertainment services that were emerging in 2002 have been further developed and expanded. Customers can access detailed tourist information for the major cities in Europe through their handset, in the language of their preference. Cell information, perhaps combined with GPS input, will be used to keep track of the position of the customer, who can enable services that push information that may be relevant. For example, a customer who has a day off in Paris can be alerted about a special exhibition in the Centre Pompidou. When she is close to the exhibition site, her attention may be drawn to special offers in a nearby shop.

Customers have a choice of a wide range of information push services, especially in the fields of sports and music entertainment. A user who subscribes to football information can go and sit in the

Jardin du Luxembourg, and browse through the major events in the matches of the previous evening. If matches are going on at the same time, he can set his handset to receive interesting events, such as goals, penalties, near misses, etc. of one or more of the matches. Activating the service and selecting the matches can be done with a combination of ASR and point-and-click.

3.5.3 2006:

The most important developments since the year 2004 have yielded broader bandwidth transmission and larger screens, with higher resolution. This enables the addition of a presentation agent on the screen for a range of services. The agents will be able to show easily recognised facial expressions, thus making the interaction with the user easier and more comfortable. For example, the presentation agent can indicate uncertainty about the user's intention, if the spoken instruction was not completely understood, or too ambiguous.

Speech recognition will be combined with natural language processing, and the simultaneous interpretation of gestures and body posture. Larger displays will come with two synchronised mini-cameras, which will allow a basic type of gaze tracking. This will alert the service that the user may have missed information that was only temporarily visible on the screen, because the user has not fixated the location of the screen during the presence of the information. The service will then offer the missed information in other –unobtrusive- ways.

The combination of a presentation agent, large vocabulary continuous speech recognition, natural language processing and gesture recognition will enable problem-solving services. For example, the user may use a service that helps her to plan a holiday, by suggesting options that fit the family situation, financial and time constraints, and the cultural preferences of the user. Professional services will include interactive design of interior decorations by architects who are in their customer's house.

3.5.4 2010: ISTAG-begin. Agents, problem-solving,

For the long-term and medium-term vision we decided to re-use the scenarios developed on behalf of the Information Society Technologies Advisory Group (ISTAG), the body that supports the European Commission in shaping the Information technology line in their Framework Programmes [4]. These scenarios were developed just before the downturn of the global economy and the waning confidence in the future of the telecommunication industry. Because of their visionary optimism these scenarios may seem far too futuristic for our purposes, which aim at realistic expectations for the next five years. Still, the ISTAG scenarios were useful as a source of inspiration, and as a source of indications of the directions in which services and the underlying technology are likely to develop. For the convenience of the readers of this document short versions of the scenarios are given in Appendix A.

One common aspect of the four futuristic scenarios developed by the ISTAG is the reliance on tightly coupled networks, from Personal Area Networks (PAN) at the lowest level, to the Wide Area Networks at the top level. The PAN networks are assumed to employ data transmission technology such as Bluetooth, while the higher level networks will employ more conventional technologies, such as ADSL in the fixed networks, and GPRS, and eventually UMTS or its successors in mobile connections. In this 'network of networks' the processing power of the terminals is mainly dedicated to capturing and displaying signals, and to store personal identity and preference information. In other words, the terminals proper mainly have the function of sensors and actuators, combined with a personal information store. The expectation that future terminals and future networks will heavily rely on some kind of distributed processing, and that the processing power and memory in the terminals will be dedicated to signal capturing and display was confirmed in consultations with the industry. Therefore, we have based our short-term scenarios on the assumption of limited local processing power, with the attendant higher demand for data transport between the terminals and the network computers that are responsible for heavy-duty computation.

Another common aspect of the scenarios is the presence of very large numbers of non-personalised sensors and actuators in the immediate environment. The signals captured by these sensors are merged with the signals provided by the personalised terminal. The actuators will be used to adapt the environment to the needs of the customers. In our short-time scenarios we will assume that the range of sensors and actuators will be limited to cell information, GPS and Bluetooth. The infrastructure for cell information and GPS is already available around the globe. What remains to be done is to integrate GPS transceivers in the handsets. The precision of cell information will limit the range of services that can be offered. GPS offers higher precision, probably at higher cost, especially if the demands are set very high. Bluetooth has been designed with a totally different goal in mind than GPS. This too will affect the type of services that can be built on top of that technology.

Also, all scenarios assume that the interface between the customer, her personal device, and the sensors in the environment is invisible and completely intuitive. For one thing, this means that high quality, large vocabulary automatic speech recognition (ASR) is taken for granted, but also that it is absolutely essential. In addition, most of the scenarios seem to assume that automatic recognition and interpretation of gestures and body posture is also possible. For speech and gesture recognition the assumption seems to be that the processing is essentially done in the network. For the short-term scenarios we will assume that a restricted form of ASR is available, and –equally importantly-that customers are familiar with its functionality. Moreover, we will assume that all terminals offer some kind of pen input, sufficient to select icons and to enter short texts by means of a soft keyboard.

4 Research on Multimodal Interaction

This chapter aims to give a brief overview of the state-of-the-art in multimodal (telecommunication) services. Due to the lack of large scale operational multimodal services the focus will be on field trials and R&D in the laboratories of the network operators, terminal manufacturers, and –interestingly- software companies.

Virtually all research in multimodal interaction is inspired by the need to provide ‘intuitive’ interfaces. Another adjective that has been uses is ‘perceptual’ interfaces, loosely defined as interfaces that have human-like interaction skills (including human-like social skills, e.g. in terms of knowledge about how to behave in a dialogue).

Intuitive or perceptual interfaces do not only make high demands on performance and capabilities of the basic input-output modes. Perhaps even more important is that they –more or less tacitly- assume advanced artificial intelligence. All visions about perceptual interfaces are based on the assumption that we must move towards some kind of declarative, instead of procedural interaction. It should become possible that a customer tells the ‘network’ *what* she want to done, and leave it to the intelligence in the service to figure out *how* the task can be accomplished. This shows again that substantial artificial intelligence is inevitable. Artificial Intelligence may very well turn out to be the single most important problem in the development of advanced information and transaction services.

Another thing that virtually all visions about future user-friendly interaction have in common is the central role of natural language, and especially speech. In many cases speech will indeed be the most natural, attractive and efficient way to express declarative commands. For this reason a large part of the research into multimodal interaction takes speech as the pivotal interaction mode. Despite the body of research results that is now becoming available, it is still difficult to formulate conclusions about the advantages of speech centric multimodal interfaces over speechless interaction that can safely be generalised beyond the limits of the particular tasks in particular experiments, using particular choices for the component technology and the interaction (dialogue management) strategy. The only general conclusion that seems to hold is that multimodal interaction is never considered as less pleasant or rewarding than a monomodal interface. In addition, most of the theoretical advantages and disadvantages of speech, pen, text and graphics are substantiated in very general ways. For example, speech makes it very easy and natural to refer to objects that are not visible (on a screen), but it is definitely not the preferred medium to render detailed factual information, such as telephone numbers or train schedules. Pen input is ideal for making selections from short lists that can be displayed on the screen, but people find it very hard to use a soft keyboard.

Because of the uncertainties about the general applicability of the results obtained in individual experiments we will not give general conclusions or recommendations relating to speech centric multimodal interaction in this document. However, after reading this chapter one should have a fairly comprehensive picture of what is going on at the moment, and can be expected in the near-to-intermediate future.

4.1 Operational services

Operational multimodal services are rare for two closely related reasons: suitable terminals and the necessary protocols are missing. The last years have witnessed the advent of multimedia PCs that are in principle able to support truly multimodal services. However, few such services seem to be in commercial operation. In addition, MUST wants to concentrate on mobile Internet services. Today, virtually no mobile terminals and transmission protocols that can support multimodal access to the Internet are available.

Recently a couple of services have been announced that are multimodal in the sense that they combine speech input with text/graphics output in a single session. This kind of services has been baptised ‘speech in, data out’, or SIDO services[6]. Four offerings that have attracted special attention

[6] Note that we have said that we do not consider ‘speech in, data out’ services as truly multimodal services (cf. section 2.2).

are summarised here. It may well be that there are other 'products' on the market that have escaped our attention. It is at once interesting and alarming that two of the three technology providers behind the offerings do not refer to VoiceXML and its future multimodal extensions.

4.1.1 Vodafone AirTel Móvil

In August and September 2001 AirTel Móvil has conducted a trial with multimodal interaction in the Madrid area, using the GPRS network. The service was based on technology provided by Auvo, a company that develops software for all-IP (Internet Protocol) wireless environments. Users had the option to send and receive information using a combination of voice, text or graphics – all through a single IP connection. During the course of the trial, participants had wireless access to a variety of different applications including: Address Book, Virtual Village (chat like application) and Horoscope. In the case of horoscopes, for example, participants will be able to use voice commands, handset keypad or stylus and touch screen to access horoscopes for the range of zodiac signs. With a touch-screen stylus, users were able to activate the horoscope icon graphic on the screen of their wireless device and hear their daily horoscope information. The trial was intended to provide invaluable insight into the human behaviour surrounding the use of the new multimodal wireless application technology. At the time of this writing the results of the trial are not yet (publicly) available.

The AirTel Móvil trial used Compaq iPAQs as terminal. Auvo has not been very verbose about their technology. However, it is reasonable to assume that they have implemented a VoIP protocol.

For additional information refer to http://www.auvo.com

4.1.2 VerbalNet™

VerbalNet is a Voice-in Data-out product offered by Verbaltek Inc., a Santa Clara, California, company devoted to the development of advanced technologies for speech recognition on mobile platforms, such as PDAs and mobile phones. The product is based on a proprietary implementation of Distributed Speech Recognition (cf. section 5.3.1) that enables Verbaltek to develop services using a purely data oriented transmission protocol. As far as one can see, Verbaltek has opted for the use of the IP protocol.

Verbaltek also provides tools that help in reformatting screen layouts to adapt the rendering to the limitations of the small screens of PDAs.

For additional information refer to http://www.verbaltek.com/

4.1.3 PocketPresence

PocketPresence AB is a Swedish company that offers Voice over IP (VoIP) software, presently only for the Compaq iPAQ running Windows CE. This solution effectively bypasses the lack of protocols for mixing information streams. Speech recognition must be done in the network. Assuming sufficient bandwidth and QOS, VoIP should be able to develop into a viable means to combine speech with other I/O modalities, such as pen input and text/graphics output. In this sense the PocketPresence solution offers the promise of a development towards an open standard.

For additional information refer to http://www.pocketpresence.com/default2.asp

4.1.4 Lobby7

LOBBY7, Inc., and SpeechWorks International, Inc. recently released the demonstration of a multimodal application prototype using Mapquest functionality. This application is developed with support from the DARPA Communicator project.

Showcasing MapQuest functionality, using LOBBY7's multimodal application server and SpeechWorks' speech recognition and text-to-speech technologies, the prototype features a full integration of speech, touch and graphics on a PDA device. The user can use his voice to tell the device where he is going. The user can also simultaneously tap the screen and say, "Show me how to get here." He will also see a map on the screen of the device and can receive command driven audio directions to ensure ease of use. All of these modes of interaction happen simultaneously as a response to user demand, environment, and command. The user experience for the prototype was based on years' of user design expertise from SpeechWorks, which has implemented its technology in such large-scale customer facing applications as AOL by Phone, United Airlines and CIBC bank.

The prototype also introduces a new kind of user interface. The interface features a visual layout that not only allows the user to easily see what he needs to say in order to activate the device with his voice, but also works with a minimal amount of display space. The combination allows the user to easily see the information he needs, while being able to input information naturally using speech, making the interface optimal for use regardless of the user's environment (i.e. if he's in a restaurant, on a subway, in a car, etc.).

See http://www.Lobby7.com among other things for a short demonstration of multimodal interaction, as well as several scenarios for services that Lobby7 expects to be able to deploy in the very near future. The example of the most advanced services qualifies for the epithet 'truly multimodal', since it implements what looks like simultaneous co-ordinated input of speech recognition and pointing.

4.2 R&D projects using small terminals with limited I/O capabilities

The majority of the R&D in the field of mobile multimodal Internet access on which public information is available appear to use existing PDAs or palmtop computers to simulate future handsets. Sometimes a GSM card is added to the device, in other cases the experiments rely on some kind of Wireless LAN for the coupling of the device to the network. Increasingly, communication is based on GPRS or (simulated) UMTS networks. At the same time developments are under way to connect mobile terminals to local and global networks through Bluetooth or similar short-range (Piconet) wireless networks. Invariably, the services that are being developed and tested rely on a distributed architecture. The exact distribution of the processing seems to differ between the projects.

Here we make reference to a small number of projects to provide an admittedly sketchy picture of the field. However, attempting to give a comprehensive and reasoned overview of recent and current activities in the field would far exceed the resources available in MUST, if only because almost all major conference in the fields of speech, language, image processing and human-computer interaction nowadays feature special sessions on multimodal mobile services. The major aim of the account given below is to give an impression of what is happening in the world just across the border of the terrain that used to be owned by the network operators.

4.2.1 The Archipel project of France Télécom R&D

The Archipel project is aimed at the development platform for the implementation of user-friendly multimodal applications that allow a wide range of client devices (PDA, WAP phones, Notepads, etc.) to provide access to Web services. The project relies on the concept of Piconet (based on Bluetooth technology). The platform is intended to act as an interface between Web servers and scattered terminals, allowing the servers' response to be adapted to the actual configuration of the terminal. Information is sent to the client via the appropriate device (e.g. PDA, Wap phone, graphical notepad…).

The Archipel platform allows terminals to communicate with the service by means of speech recognition and speech synthesis, in addition to keyboard and mouse input and screen output. Seamless migration of conventional web services to multimodal interaction is supported by the 'Archipelizer' tools. Simultaneous multi-publishing allows the nomadic user to consult or deposit information via multiple devices.

The Archipel project intends to deliver a service prototype in Q4 of 2001. Field trials are planned for Q3 of 2002.

An example of a service that can be provided on the Archipel platform would allow a user to consult an all-purpose news multimedia server with his PDA. Thanks to Archipel, images are rendered on the PDA's graphical notepad, while audio commentaries are sent to the earplugs of the user's mobile phone or digital audio player.

4.2.2 SMADA

The IST project SMADA (Speech-driven Multimodal Access to Directory Assistance) is carried out by a consortium of telecommunication network operators, a terminal manufacturer and a number of universities [10]. The project anticipated the shift of speech-based interaction from offices and private homes towards mobile Internet access for information services, such as searching for telephone numbers. Therefore, two work packages in the project address issues in mobile multimodal interaction. The Compaq iPAQ was chosen as the terminal platform for the experiments. Preliminary results of usability tests show the importance of the theoretical distinctions proposed by W3C: users become confused when the interface switches between sequential and uncoordinated simultaneous use of speech recognition and pen input [7]. However, it also appeared that users have no problem in switching between pen and speech input as long as the graphical interface clearly prompts pen input (e.g. for initiating commands by clicking a button) and when they are able to select the input mode themselves (for entering names).

4.2.3 MATIS

The Department of Language & Speech of the University of Nijmegen collaborates with two other Dutch universities (Eindhoven and Delft) in the project MATIS, that will continue under the new name CRIMI as of January 1, 2002. The project is part of a medium size R&D programme, funded by the Dutch Ministry of Economic Affairs, aimed at strengthening the position of the Dutch industry in the field of Human-Machine Interfaces. In MATIS graphical output and pen input are added to what was originally conceived as a unimodal (speech-only) train timetable information service. The single most important usability problem of the unimodal service was the difficulty for the callers to maintain a correct mental picture of the progress of the dialogue. Especially the navigation through the timetable after a possible connection was found proved to be extremely difficult. Adding graphical feedback, and offering pen input for commands and selections appears to improve the usability of the service [15]. The approach taken in MATIS (adding graphical output and pen input to what used to be a speech-only service) is reminiscent of work that is under way in Telenor [8].

4.2.4 Microsoft

Microsoft Research are working on a multimodal interaction device that is known as MIPAD. In the vision of Microsoft a device like MIPAD will become the major vehicle for interacting with intelligent home appliances (that are expected to come without a keyboard and display of their own, much like modern radio an tv sets, which are designed for operation through a remote control) on the one hand and with remote applications in cellular networks on the other. In Microsoft's vision the device will combine speech and pen for input, and audio and a small graphical display for output. Fusion of the input modes is based on the 'Tap and Talk' strategy: the user must tap a designated icon on the screen before he can start speaking. This obviates the need for a continuously active speech detector.

The latest series of experiments conducted with MIPAD use the Compaq iPAQ to emulate the terminal [6]. It goes without saying that the applications are based on a client-server architecture, in which the iPAQ is used as a relatively thin client.

So far Microsoft has conducted experiments with Personal Information Management services using MIPAD. Preliminary user studies show a clear preference for multimodal interaction over pen only input for tasks like setting up a new appointment and writing e-mail messages.

4.2.5 IBM

IBM’s Watson Research Center are working on what they call a ’personal speech assistant’ (PSA). The most recent version is built by adding a substantial amount of processing and network power to a Palm III [5]. Although IBM also aims at a client-server architecture, they seem to have a much thicker client in mind than Microsoft. IBM’s PSA can do small-to-medium vocabulary isolated word ASR in the terminal. It also handles most of the dialogue management in the client.

So far, IBM seems to have experimented mainly with the PSA for voice dialling in the car. Other applications that are referred to include Personal Information Management, but also a multilingual phrase book for foreign travellers (actually a kind of speech-to-speech translation).

4.2.6 SpeechWorks/Compaq/Auvo/Lobby7

The US Defence Advanced Research Projects Agency (DARPA) has expressed its interest in moving speech centric applications to mobile environments. In that context SpeechWorks have won a contract to develop mobile multimodal interfaces, in collaboration with Compaq, Auvo and Lobby7. In this consortium Compaq provides the hardware platform for the client and the server. The client is based on the iPAQ. Auvo is a company that advertises itself as providing a break-through platform and architecture required for wireless implementation of the integrated, multi-media interfaces. Auvo intends to provide to wireless carriers and Internet content companies the technology, software and services solutions they need to bring the ultimate Internet experience to wireless consumers around the world: the seamless integration of voice, text and graphics. Lobby7 is a company closely linked to MIT, dedicated to development of multimodal mobile services that integrate speech, text and graphics.

SpeechWork’s does not provide information about the services that they have in mind. However, combining bits and pieces it seems that the consortium intends to build upon VoiceXML.

4.2.7 The Leitprojekte of the German BMWI

The German Bundesministerium für Wirtschaft und Technologie funds six so called Leitprojekte in the field of 'Human-Technology Interaction'. All six projects are related to multimodal interaction, in one way or another. The most obvious links between MUST and these Leitprojekte are through the projects MAP and SmartKom. The easiest way to access the list of Leitprojekte is probably via the web site of the MAP21 project.

4.2.7.1 MAP21

MAP is a German acronym for multimedia workplace of the future. The project, co-ordinated by Alcatel, addresses several issues, among others 'Support for Mobility' and 'Agent Technology'. The project started recently; therefore, there are no results that can be used immediately in MUST.

For additional information, refer to http://www.map21.de/map/index.phtml

4.2.7.2 SmartKom

The SmartKom project, co-ordinated by DFKI in Saarbrücken, consists of four sub-projects, each of which addresses another type of application domain. For MUST SmartKom Mobile, a mobile communication assistant, is the most interesting sub-project [9]. The following is a literal citation of the text on the English version of the SmartKom website:

> SmartKom-Mobile serves as a mobile platform for numerous information services, such as, for instance, access to the Internet.
>
> Access to the Internet is provided via a GSM connection and by means of an integrated GPS independent movement may be followed on a digital map. A loudspeaker and display are provided for the delivery of information.

> In order to facilitate dialogues in spoken language, SmartKom-Mobile has a built-in microphone [….]. A camera has also been integrated [….] to record mimics. Handwriting or explanatory gestures co-ordinated with language may also be entered via use of a pen.
>
> SmartKom-Mobile has been designed as a mobile platform to provide numerous information services, which integrate pre-existing functionalities such as an address book or telephone but which is also capable of including new future technologies. In particular, the possibilities offered by information and communication networks are to be accessible allowing the user to combine interesting and useful information at all times in all places. This includes the use of Internet services which are already capable of providing innumerable services, from weather reports to hotel reservations.
>
> http://www.smartkom.org/start_en.html

Although SmartKom as a whole is well under way, the sub-project on mobile applications has not yet produced a publicly accessible demonstrator at the time of this writing.

4.2.8 European Media Lab (EML)

EML is working on several multimodal applications. One of their projects in this field is called Deep Map. It aims at building a web-based application that enables customers to plan a trip to Heidelberg. A closely related project is Talking Map that aims at providing intuitive interfaces for personal mobile systems. The mock-ups used in the Talking Map project are based on the Compaq iPAQ.

Additional information on the projects of EML can be found in their latest Annual Report at http://www.eml.villa-bosch.de/english/news/news.html

4.3 Projects aiming at kiosk-type applications

Most of the ’older’ R&D in the field of multimodal HCI have targeted kiosk-like applications. After all, mobile devices with appropriate I/O facilities and compute power are only now beginning to appear. Most of the experiments that have been carried out so far were based on applications that were originally conceived with a full-blown desktop terminal in mind, but that needed to be restyled in order to allow operation without a keyboard. In public kiosks keyboards are just too volatile to be effective. The research has focused on speech as an alternative for soft keyboards or for complex menus navigated on touch screens. Research has also addressed issues such as the use of menu/list selections to overcome ASR errors.

A fairly comprehensive overview of the issues, projects and experiments conducted by several groups in the USA and preliminary results can be found in Oviatt et al. [11]. Presently, there is no comprehensive overview of the work in the field that has been done in Europe or in Pacific Rim countries. MUST does not provide the resources to make such an overview. Therefore, we limit ourselves to a list of large projects that we are aware of (ongoing and completed). We are aware of many additional smaller projects, but to the best of our knowledge none of those is making contributions to the field that go significantly beyond what is mentioned here. Moreover, a recent project in the USA, based on the originally monomodal DARPA Communicator platform, viz. the Listen-Communicate-Show systems under development by Lockheed Martin deserve mentioning [14].

- The BMWI Leitprojekte mentioned before also address kiosk-like applications. Actually, the only presently available demonstrator in SmartKom is a kiosk application that provides information about Heidelberg.
- LingWear is the name for a set of demonstrators developed at the University of Karlsruhe and Carnegie Mellon University [13]. LingWear demonstrators are available for tourist information (on Heidelberg and Karlsruhe) and medical information. The most striking feature of LingWear is its combination of multimodal interaction and multilinguality: the information returned by the system can be cast in another language than that of the query. http://www.is.cs.cmu.edu/js/lingwear.html provides additional information on the LingWear project.

- Catch2004 is an IST project started in January 2000. It aims at multilingual and multimodal access to information, from a range of terminals including kiosks and WAP phones. There is not yet much public information about the status and progress of he project available. The information that exists can be accessed through http://www.catch2004.org/
- MASK was an ESPRIT project carried out by MORS (co-ordinator, F), SNCF (F), LIMSI-CNRS (F), and UCL (UK). The aim of the Multimodal-Multimedia Automated Service Kiosk (MASK) project was to pave the way for advanced public service applications by user interfaces employing multimodal, multi-media input and output. A prototype information kiosk was developed and tested in the Gare St. Lazare in Paris. The kiosk should improve the effectiveness of information and transaction services by enabling interaction through the co-ordinated use of multimodal inputs (speech and touch) and multimedia output (sound, video, text, and graphics). Speech input is managed by a spoken language system, which aims to provide a natural interface between the user and the computer through the use of simple and natural dialogs.

  Additional information can be found at http://www.limsi.fr/Recherche/TLP/mask.html

4.3.1 The MUeSLI programme of BT

BT Adastral Park has been investigating basic issues in MUltimodal Spoken Language Interfaces in their MUeSLI project [1]. The research started with focus groups that discussed storyboards of possible system interactions. In the next phase focus groups were conducted with mock-ups of the target systems. In the third phase the mock-ups were replaced by Wizard-of-Oz based interactive experiments, resulting in a fourth phase in which fully operational systems were used. The task that the user must complete related to the selection of fabrics for the remodelling of a living room. The working environment is similar to a desktop Internet workstation or to a kiosk.

4.4 Conclusions

The survey of current R&D projects on speech centric multimodal interaction show a substantial degree of commonality. All close-to-operational services are aiming at some kind of speech-in, data-out type of interaction, that combines the strong features of speech (mainly the possibility to directly refer to items that are not visible on a screen) and text/graphics (the ability for the user to digest the information at her own will and in her own pace, combined with the possibility to store the information for future re-use). In addition, all close-to-operational services employ terminals such as the Compaq iPAQ or Palm III, for which VoIP protocols have been developed.

The more advanced research projects focus on the problems raised by fusion of simultaneous input channels and the division of output information over the available channels (fission).

5 Future Terminals

It goes without saying that the terminals play a decisive role in the development of (future) multimodal services. Uncertainty about the time lines of the development of fully functional UMTS terminals and the details of their functional specifications has been cited as one of the causes of the uncertainty surrounding the development and roll-out of UMTS networks and services. After all, services can only be offered successfully if both infrastructure and terminals are sufficiently widely available. This reminds one of the explanation of the advantage that NTT DoCoMo could create for itself in I-Mode, and may now again be creating with FOMA.

Multimodal services make specific requirements to the terminals. They also make requirements to the network architecture, and to the architecture of service platforms. In this chapter the emphasis will be on the terminals. However, advanced multimodal services always will live in a distributed environment. It may not always be evident what the best division will be of the functionality of an end-to-end service between the terminal and the 'network'. It is clear that the terminal must host the sensors for the input and the screen and loudspeaker for the output. It is much less clear what part of the processing of the input and output signals must –and will- be done in the terminal, and what will be delegated to other computers in the network. It is also clear that there will be some trade-off between the processing in the terminal and the amount of information that must be transmitted to the terminal. In this document the focus is on local processing in the terminal. The discussion of distributed processing will be limited to Distributed Speech Processing (DSR) that is especially important in a study of speech-centric multimodal services.

In this document we will not explicitly address the requirements for the network infrastructure. However, it must be pointed out that speech-centric multimodal services are likely to make specific requirements, if only in the form of communication protocols that support simultaneous transmission of streaming data (such as speech and moving pictures) and data for which a conventional packet protocol is adequate (such as point-and-click actions). Since the research in the MUST project is focused on usability, rather than on the technical infrastructure, we will limit the discussion of networks and protocols to a short section.

This chapter, then, starts with a discussion of the state-of-the-art and short-term expectations of input and output modes in mobile terminals, followed by a more general section on signal processing in the terminal (with a sub-section on DSR). Next, we briefly discuss protocols and networks, and the chapter concludes with an overview of Operating Systems and Application Development software that runs on the terminals.

5.1 Sensors

All mobile multimodal terminals will come with a basic set of sensors, including a microphone to capture speech and other audio signals, a pen, to capture selections from menus displayed on the screen and to capture structured gestures (handwriting, and perhaps also a limited number of drawing patterns), and a position sensor (probably related to GPS). It is unlikely that future terminals will have no keyboard whatsoever, not even the 12-key numeric keypad of a telephone handset. In the late nineties Alcatel offered a GSM telephone that had no electromechanical keyboard, but a soft keyboard instead. Although this design provided much more space for the graphics screen, the product never became successful, probably because of a mix of causes, among which the price of the device and the ergonomic problems with simple dial operations were the most important. Even if the price of terminals with a relatively large screen can be brought down, we expect that customers will be reluctant to give up the numeric keypad, as long as the major function of a mobile terminal is person-to-person communication by voice or text. However, we do not expect that future mobile terminals will come with a fully functional, ergonomically attractive keyboard.

The next generation of mobile terminals may come with the option to attach a video camera. We expect that these cameras will only be used to support person-to-person communication. Until the year 2006, the farthest point in the future we dare to speculate about, the technology to automatically interpret video images captured by the camera on a mobile terminal will not be generally available. Consequently, we do not expect the possibility to use advanced techniques such as detection of emotions in facial expressions or body posture in mobile services that will be operational and accessible for the general public in 2006.

5.1.1 Microphone

In our vision there will never be mobile multimodal terminals without a built-in microphone, at least not until the moment when person-to-person communication will disappear completely. For the type of speech centric multimodal services that we have in mind the quality of the microphone will be crucial. The quality may not so much be determined by the bandwidth of the sensor, but rather by its noise cancelling properties. So far it has appeared that high performance automatic speech recognition is feasible with speech limited to the conventional telephone band with an upper cut-off frequency of approximately 4 kHz. Improving signal-to-noise ratio is much more effective in reducing the error rate than widening the transmission band.

We do not envisage developments in microphone technology that will provide excellent signal-to-noise ratios for a sensor mounted in a terminal that is used at arm length distance. Rather, we expect that future terminals will come with a monaural microphone input for a close talking microphone. By default the microphone will be physically attached to some kind of earpiece. The transmission between the handset and the microphone earpiece set can be wired or wireless (e.g., using Bluetooth), without any major effect on the functionality. The noise cancellation properties of microphones built into the case of a mobile terminal may be enhanced by making the sensor extremely directional. This is feasible, because it is safe to assume that the direction from which the speech input comes is always perpendicular to the plane formed by the screen. Future terminals will have sufficient compute power to support a beam-forming microphone, consisting of directional microphones at the edges of the case.

5.1.2 Position Sensor

Since location based services will be among the most important ones, we expect that future terminals will all come with some kind of fairly accurate position sensor. Several technologies may be used; among which access to the satellite based GPS system probably provides the best accuracy. In 2006 positioning may also be possible by means of interaction with local beacons, that get in touch with the terminals through a Bluetooth connection. Obviously, such a beacon based positioning system will only be available in densely populated areas, where specific service offerings can be attached to individual beacons. As GSM, GPRS and UMTS cells become smaller in circumference, useful positioning systems can be based on the cell information.

5.1.3 Pen Input

Already today terminals are available that can capture most conceivable types of pen input, perhaps at the cost of the need to use a special stylus that comes with the terminal. Inputs that can be captured include transmission of the co-ordinates at which the screen is touched, and a sequence of co-ordinates that represent a stroke or line segment drawn on the screen. Isolated co-ordinate input and continuous sequences of co-ordinates are likely to come with proper time stamps. It is not completely clear whether time stamp information for strokes is standardised. At least two different options are available: time stamps only for the first and last co-ordinate in a gesture during which the pen remains in contact with the screen, or time stamps for each individual co-ordinate pair. Evidently, only the latter protocol allows one to recover the exact dynamics of the gesture.

Research is under way in several academic and commercial laboratories to develop automatic handwriting recognition. High performance recognition of letters drawn in isolation is already available, but it is not clear whether consumers are willing to use that option to enter more than an occasional single word. In most cases recognition of cursive script relies on the dynamics of the strokes, in addition to the characteristics of the resulting ‘electronic ink’ pattern. This makes automatic recognition different from what is usually done in banks and letter/parcel sorting application, where only the static ink patterns are available for recognition. Soft keyboards, which only need sufficient resolution of the co-ordinates where the screen is touched, come as a standard option with all PDAs. But most users seem to find soft keyboards inconvenient when substantial amounts of text must be entered. Automatic word completion that can be added at low cost in terms of CPU power, only works if the completed word can be displayed very close to the keyboard on the screen; else, users fail to notice the automatic completion on the screen.

In principle pen input also includes pen pressure on the writing surface. It is not clear whether the terminals that we will see until 2006 will provide pressure sensing.

5.1.4 Bio-sensors

Bio-sensors form a special class of sensors, if only because they have a special and highly dedicated function. Bio-sensors may include systems to capture fingerprint and hand palm geometry information, and –in more advanced systems- also retina pattern information. Capturing this information has a single goal, viz. to identify the user, or to verify the claimed identity. Information to verify the identity of the user is probably essential for a range of services that involve financial transactions or access to privacy sensitive data.

In the period until 2006 we expect to see only the use of fingerprint information to enhance security. In addition, we expect to see an increasing use of speaker verification technology, if only because the latter does not require additional input sensors. In fact, we expect to see a resurgence of R&D in speaker verification, aimed at providing an additional level of privacy protection in the use of mobile terminals, which will tend to be highly personalised, yet easy to lose or to steal.

5.1.5 Additional input devices

In the laboratory additional techniques for capturing intentional and unintentional information conveyed by the user are under development. One quite promising input channel is gaze tracking. Gaze tracking delivers the screen co-ordinates on which the eyes are focused. However, with the technology that exists today accurate tracking is only possible if the sensors can be fixed to the subject’s head, so that the distance between the sensors and the eyes is relatively constant. The only convenient way of accomplishing this would be to integrate the sensors in a special purpose pair of spectacles. Although these devices have been successfully demonstrated, we do not expect to see them on the mass market by the end of the decade. For the time being it is not evident that the information provided by a gaze tracking system can be used to improve the quality of the interaction with a sufficiently wide range of services to warrant the development costs. However, without gaze tracking the need to design the graphics on the screen in such a way that relevant information is guaranteed to appear in the focal area should not be underestimated.

It has already been said that we do expect to see optional detachable video cameras for a range of mobile terminals. However, we have also said that we do not expect automatic recognition or interpretation of the images captured by these cameras. The performance of automatic image processing in the period until 2006 will be crucially dependent on image quality, and especially on homogenous and constant illumination conditions. Neither of these can be guaranteed with the small mobile cameras in mass production in 2006.

5.2 Output Actuators

All terminals that are on the market today offer simultaneous audio and graphics/text output. We do not expect to see additional output channels in the period until 2006. It is possible that the screens will become larger and the resolution higher in the next five years. However, we do not expect to see foldable or roll-able large screens in mass production before the end of the decade. We are aware of the emergence of head-mounted displays, but it is difficult to predict the success of these appliances.

As more complex graphical information must be displayed (such as moving video and high quality audio) in environments where bandwidth will be at a premium, high demands will be made on signal coding at the source, and decoding at the client side where the display must be regenerated. We expect to see most of the local processing power in mobile terminals dedicated to decoding and rendering tasks.

5.3 Local Signal Processing

The ways in which the signals produced by the input sensors can be used in the service depends on the processing that can be applied. In principle, computationally intensive processing is possible, provided that the signals can be properly transmitted to powerful compute servers in the network.

We expect that the processing of the input signals in the terminal proper will be limited. Rather, we expect that the precious processing power in the terminal will be dedicated to processes that cannot be delegated to another computer in the network. Especially in the types of handsets that evolve from the present generation of cell phones local processing power is more likely to be dedicated to decoding of audio and picture information for local display. The situation may happen to be somewhat different in devices that evolve from the present generation of PDAs, where programme code may be stored in some kind of local background memory, and activated on demand.

In the period until 2006 we expect to see standards emerge for local signal processing in the terminal. For the audio signal this processing may take either of two forms: (1) signal encoding that is necessary for transmission over a packet switched network, or (2) computation of a standard set of parameters for input to a speech/speaker recognition system. Both routes are compatible with the IP protocol at the higher levels of the network architecture.

The processing power in the terminals will be sufficient to do flexible vocabulary isolated word recognition or word spotting locally. Sensory Inc., an American company that builds and markets software for embedded speech applications (www.Sensoryinc.com), has announced ASR and TTS for the StrongARM processor in the Compaq iPAQ. Lernout & Hauspie is also known to have worked on software for embedded speech recognition and synthesis. In November 2001 e.Digital (www.edig.com) announced the MPX100, a voice controlled MP3 player that can store some 100 audio tracks, and lets the user select songs by saying the name of the player or the song (based on speech recognition technology of Lucent). This kind of application is similar to ‘voice browsing’, in that limited size vocabularies can be built on the fly, containing the words that are relevant in the condition of the moment (the links on the screen, or the titles of the songs in the local memory). Vocabulary updates can be accomplished in the server, which can send the newly relevant words together with the new screen text, or the new songs. At the time of this writing no information about the performance of embedded speech applications on a ‘standard’ PDA is available. It is to be expected, however, that the first generations of speech-driven services that use embedded speech software will be tailor made for specific applications. Embedded implementations of speaker verification are also possible without overtaxing the CPU and memory of a PDA.

In all existing devices substantial local processing is applied to pen input. If soft keyboards are available, all processing to identify the selected characters is done locally, but this kind of processing does not make large requirements. Many existing terminals also provide the software and compute power for the recognition of (isolated) hand written characters. It remains to be seen whether future terminals will provide sufficient memory and CPU power to extend the processing of pen input to a wider range of gestures, such as circling areas on the screen, sketching simple forms that can be interpreted at a later stage, etc.

If processing of pen input in a network computer is necessary, properly standardised protocols must be developed and implemented for the representation and transmission of the signals. These protocols must include decisions about the amount of information that is encoded and transferred, such as time stamps attached to integral strokes or to all individual co-ordinate pairs that make up a stroke.

From a purely technological point of view it is already now possible to integrate a GPS processor in a cell-phone or PDA. It remains to be seen what services need the additional precision of GPS, as compared to the location information that can be derived from the identity of the cell in which a nomadic user is operating. The introduction of ever-smaller cells (eventually culminating in Piconets) may make high precision GPS somewhat redundant, at least for services that are targeted at densely populated areas.

We expect that terminals that come with a fingerprint sensor will also provide the software and CPU power for fingerprint recognition.

5.3.1 DSR

The obvious alternative for power hungry computations in the terminal is to distribute the work, and to allocate heavy processing to computers in the network. This type of distributed processing will become the norm, rather than the exception, in the future 'ambient intelligence landscape'. One of the tasks in multimodal interaction for which distributed processing was first considered is automatic speech recognition. The basic idea behind Distributed Speech Recognition (DSR) is to distribute the work for ASR between the terminal (client) and a network server in such a way that the capabilities of both parties and the network that connects them are optimally exploited. DSR is directly related to the issue of thin vs. thick clients. At the same time it makes an attempt to address the problems caused by the radio links in the 2G and 2.5G mobile networks.

There is a substantial body of evidence that shows that the standard GSM codecs (Full Rate, Enhanced Full Rate, Half Rate) have only a minor effect on ASR performance, as long as all bits in all frames are received intact. However, under all realistic conditions frames get damaged because of interference, fading, hand over at cell boundaries, etc. These bit errors appear to have a substantial impact on recognition performance. DSR is one of the ways in which the impact of radio transmission errors can be diminished.

The idea underlying DSR is that conventional ASR software only uses information about the spectral envelope of the speech signal. In each GSM codec this information is computed in the handset. This suggests the possibility of transmitting only the spectral information, for instance in the form of cepstral coefficients. These coefficients can be handled as data. Since they require less bits that the full speech signal, the data frames can be protected better in the channel coder. This should result in a less error prone input to the ASR software.

The ETSI project AURORA is working on the definition of standard front ends for ASR, based on Mel Frequency Cepstral Coefficients. WI007 has provided a standard for MFCCs in clean conditions. Presently, WI008 is approaching a proposal for a standard set of cepstral parameters that can be used under noisy conditions.

For MUST is probably more important that DSR allows one to keep the client (terminal) thin, at least for what concerns ASR. This frees resources in the terminal for tasks that cannot easily be relocated to the server (i.e., a workstation in the network, behind the switch).

Moreover, DSR offers the possibility to enable voice-in/data-out services, by setting up a data oriented session (probably using the IP protocol). The technology under development by Verbaltek is just one example of this approach.

5.4 Protocols and networks for multimodal services

Definition of suitable transmission protocols is closely linked with the type of services that are envisaged. Questions that arise are: 'Must ASR input be combined with speech output?' so that a data uplink must be combined with a speech downlink in the same call. 'Must ASR input be combined with other input modes, like point-and-click?' so that two input data streams must be merged, probably with proper timing information. 'Is the interaction at the network side limited to a computer, or should fallback to a live agent be considered?' so that it may not be sufficient to transmit only the cepstral coefficients, since they do not allow to reconstruct intelligible speech.

From the discussion in section 4.1 it appears that multimodal services are likely to use simultaneous input and output in several modes. Form the point of view of protocols and network infrastructure the distinction between co-ordinated and non-coordinated simultaneous operation may not be very large. In both types of combination provision must be made for simultaneous transmission of streaming and packet data. Presently, there is no agreement in the industry about the best ways in which this combination can be accomplished. Three trends to solve protocol issues in actual services or to side step them in experiments and proprietary products seem to be emerging:

1. use of DSR can enable speech-in/data-out services by using only a single data connection. However, for the time being this solution requires proprietary technology, because suitable standards for DSR are not yet available.

2. VoIP can be used for more advanced services (but also for speech-in/data-out services). This approach raises concerns about available bandwidth. Moreover, VoIP remains to be standardised.
3. use Wireless LAN, Bluetooth (or some other Piconet protocol) to connect a palmtop PC that emulates the terminal to a network of workstations.

We expect to see all three solutions competing on the market in the next five years. It may well be that all survive in their own market niches.

5.4.1 Networks supporting multimodal services

The type of services addressed in the MUST project is best characterised as mobile Internet services. Therefore, the networks must be able to support the IP protocol. This will be the case for the GPRS, EDGE and UMTS networks that are under construction. All these networks support always-on connections, which is essential for future mobile services to have the same functionality as desktop Internet applications.

One additional issue that may turn out to be of crucial importance is the effective bandwidth that the network can offer. Multimodal services require immediate response of the system to each meaningful action of the user. Moreover, many multimodal services will entail the need to transfer substantial amounts of graphical information from the server to the user's terminal. The bandwidth must be available that is needed for smooth transmission.

5.5 Local Software

Today, small appliances such as PDAs, mobile phones, tv set-up boxes, etc have an embedded microprocessor and they are typically focused on a more restricted set of applications than desktop PCs or laptops. However, this situation is changing. Many companies are developing technology to enable the development and deployment of new, and much more varied applications for small mobile devices. The convergence between the PDA and the mobile phone and the existence of environments for the development of applications will leverage the appearance of new richer applications, combining several modalities for input and/or output.

Several companies such as Microsoft, Symbian and Sun MicroSystems are investing in the development of OS and other software components that enable the development of multimodal terminals.

5.5.1 Symbian Platform

Symbian is a company owned by mobile computing and telecommunication device manufacturers like Erickson, Nokia, Motorola, Panasonic and Psion. The main goal is the development of core software for wireless devices, according to the architecture depicted in Figure 9. Symbian OS is a software platform that is used by manufacturers as the basis of wireless information devices, for the development of applications integrated with wireless telephony and data. It is a platform for deployment of applications (programs and content) developed in a wide range of languages and media. Symbian provides a set of software development and customisation kits to enable development of wireless information devices and applications that run on them. It is based on EPOC technology (initially developed by PSION) that provides an operating system, customisable user interfaces and colour support fit-for-purpose application suites, Internet connectivity, software development kits and PC connectivity software.

Symbian provides a set of reference designs, named Device Family Reference Designs (DFRDs), on which mobile computing and cellular phone manufacturers can base their devices.

The last version of the platform (version 6.1) is shipped with two complete communicator reference designs, Quartz and Crystal. Quartz is a reference for the "pocket-sized" tablet Communicators. Crystal is a reference for keyboard-based wireless information devices.

Both Quartz and Crystal reference designs are based upon Symbian's Generic Technology components. These components provide a set of services - the multi-tasking EPOC kernel, data management, communications, graphics, multimedia, security, application engines, messaging engine, browser engines for WAP and HTML, Java™ runtime environment, support for data synchronisation and world-wide locales - used by both reference designs.

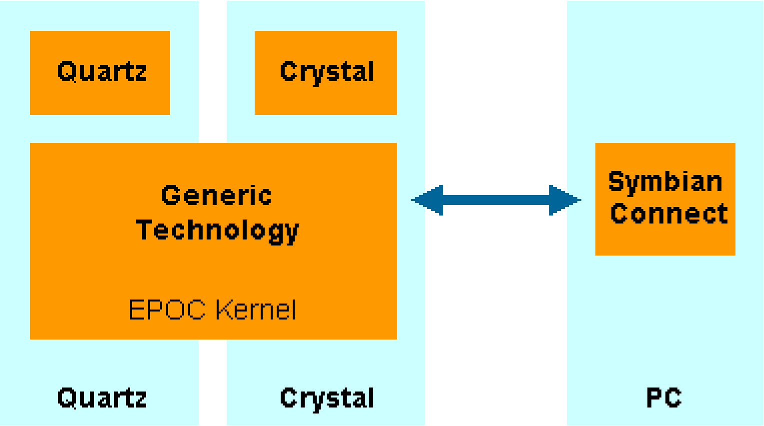

Figure 9- Symbian high-level architecture

The use of generic components allows more portability of the applications. An application developed for a certain reference design can be easily adapted to work in other reference design because those parts that use the same generic technology APIs do not need to be changed.

The Nokia 9210 Communicator and Ericsson R380x Smartphone are examples of 2.5G devices already powered by Symbian platform. The success of the Nokia 9210 has made Symbian OS the platform with the highest market share in the second half of 2001.

More information about Symbian can be found at http://www.symbian.com

5.5.2 Sun's J2ME

The technology developed by Sun for wireless devices is Java-based and its principal platform is Java 2 Micro Edition (J2ME). The J2ME architecture was designed to be modular and scalable to support the growing requirements in terms of flexibility and customisable deployment. This is achieved with a model organised in three layers of software built upon the operating system of the device (cf. Figure 10).

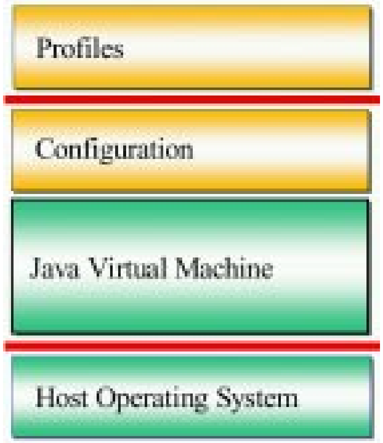

Figure 10 J2ME High level architecture

The first layer implements a Java Virtual Machine according to the device operating system characteristics. The next layer is the Configuration Layer and is less visible to the users but very important to the profile implementers. It is the layer where the minimum set of Java virtual machine features and Java Class Libraries is defined that are available to a particular “category” of devices that represent a particular horizontal market segment. This layer represents the “minimum common denominator” of the Java platform and libraries that users can assume to be available on all devices. The current J2ME architecture supports two configurations: Connected Device Configuration (CDC) and Connected Limited Device Configuration (CLDC). The CDC technology uses the classic virtual machine (VM), a full feature VM with all the functionalities available on a Desktop system. The devices for this configuration must have at least a few memory megabytes available. The CLDC is used in devices with constrained memory environments such as wireless devices. This configuration defines Java platforms targeted for devices with memory range of 160 KB to 520 KB. The CLDC technology includes the K Virtual Machine (KVM) and the core class libraries.

The top layer named Profile Layer is more visible to the users and application providers. It is the layer responsible by the definition of the minimum set of Application Programming Interfaces (APIs) available on a vertical market segment. Profiles are implemented on top of a particular configuration and applications are designed for a particular profile and are portable to any device that “supports” that profile. A device can support several profiles viz.:

- The Mobile Information Device Profile –MIDP is the profile designed for cell phones, two-way pagers and PDAs. This profile is based upon CLDC configuration.
- The Foundation Profile serves two purposes. First, provides a profile of the Java 2 Platform suitable for devices that need support for rich network enabled Java environment, but do not require a graphical user interface. Second, provides a base profile for inclusion by other profiles that need to build on the functionality it provides by adding graphical user interfaces or other functionality.
- The Personal Profile provides the J2ME environment for those devices with a need for a high degree of Internet connectivity and web fidelity. This Profile is intended to provide the next generation of Sun's PersonalJavaTM environment, and as such has the explicit requirement of providing compatibility with applications developed for versions 1.1.x and 1.2.x of the PersonalJava Application Environment Specification.

More information about J2ME can be found at http://java.sun.com/j2me

5.5.2.1 Implementations

Motorola announced that the Java technology will be supported in virtually all of its wireless products by 2002, allowing developers to create a wide range of Internet-based applications for commerce, entertainment, and communications.

Motorola at the 2000 JavaOne Conference has demonstrated a smart phone prototype which combines an easy-to-use touch screen with organiser, messaging and Internet browsing functions. The Java technology-enabled prototype is a GSM dual band (GSM 900/1800) smart phone that incorporates a touch screen user interface with dual Internet browsers (HTML 3.2 and WAP 1.1). Motorola provides already J2ME mobile phone device - Motorola i85s - for Motorola's iDEN technology. This technology combines the capabilities of a digital wireless phone, a two-way radio, a text pager, always-on Internet access, two-way e-mail and wireless modem functionality into a palm-size handset. By mid 2002, Motorola anticipates that all its handsets will have J2ME capability as standard.

NTT DoCoMo has launched a Java enabled service called i appli that provides access to rich enhanced content applications viz. multi-player games and mobile e-commerce. The terminal (new 503i cellular phone from Fujitsu and Matsushita/Panasonic) is enabled with the Java 2 Platform, Micro Edition™ (J2ME™) technology.

A J2ME toolkit is available for the Palm product family that enables the deployment of Java applications based on MIDP profile.

5.5.3 Other environments to run third-party applications

There are several other companies providing run-time environments that can be installed on handheld devices. These environments allow the expansion of the set of applications that is shipped with the device.

5.5.3.1 Kaffe

Kaffe provides an independent implementation of the Java Virtual Machine and class libraries. The libraries include classes that handle graphics, file management, input/output, and networking.

According to the information available, Kaffe is mostly compliant with JDK 1.1, except for a few missing parts and parts of it are already JDK 1.2 (Java 2) compatible.

Kaffe has been ported to more than 30 Operating Systems, running on about 8 different processors.

At moment the distribution of Kaffe requires the compilation of the source code in the target platform. http://www.kaffe.org

There is one Linux distribution for handheld devices, named Pocket Linux, which is built upon the Kaffe java virtual machine. More information about pocketlinux distribution is available at http://www.pocketlinux.com

5.5.3.2 Jeode Platform

Insignia, company dedicated to the development of Java virtual machines, provides a Java runtime environment named Jeode Platform that enables Java functionality on information appliances and embedded devices such as PDAs, smart phones, set-up boxes, etc. This product is compliant with Sun’s PersonalJava 1.2 and EmbeddedJava 1.0.3 specifications, which is roughly equivalent to JDK version 1.1.8.

Target operating systems currently supported include Windows CE 2.12 and 3.0, Windows NT4, VxWorks, Lynux, uITRON, Nucleus, BSDi UNIX, pSOS and others. CPUs currently supported include ARM, MIPs, x86, Super H-3, Super H-4 and Power PC.

More information can be found at http://www.insignia.com

.

6 Architectures for Multimodal Services

In this chapter we introduce two issues that are important in the design and implementation of multimodal services, irrespective of the type of terminal that they intend to use. These issues are (1) the software architecture of multimodal services, and (2) strategies for the control of multimodal dialogues. The issues discussed here apply both to services with small mobile terminals and to kiosk services.

First, we take up the issue of software architectures. Speech-centric multimodal services are still in a development stage. Despite activities launched by the W3C, and more recently by the SALTForum, there are no 'standard' application development platforms. Consequently, developers of multimodal services will face the need for substantial software development. In section 4.1 several product prototypes have been mentioned. It is too early to predict, which of these offerings will set the standard. As said before, we expect to see several approaches to persist for some time into the future, perhaps dependent on the specific type of application, or on the specific vendor. In this chapter we will focus on architectures that can be extended towards 'simultaneous co-ordinated' multimodal interaction (cf. section 2.2).

Because of the need to combine (fuse) the information from several input channels, and/or to distribute information over output channels (fission), the general architecture of a multimodal application is more complex than what is usually found in a monomodal application. Fig. 1 (in section 2.3) shows a somewhat simplified picture of the most important modules in a multimodal application, where it is assumed that the terminal offers speech, keyboard, pen and 'GPS' input, while output can consist of audio (first and foremost speech, natural or synthetic), of text and of graphics. We have abstracted away form the –potentially complex-- internal architecture of the application.

What Fig. 1 does not (and cannot) show is the software architecture that is used to implement the service. Nor does it show the options that are available for the control of the dialogue between the service and the customers. Options for software architectures will be dealt with in section 6.1; options for dialogue control are discussed in section 6.2.

6.1 Architectures for multimodal services

The development of multimodal applications requires a software architecture that enables flexible dynamic addition or substitution of modules and components, permitting rapid application development and future evolution. This evolution can be for example the addition of a new input modality such as gesture recognition to an application that supports only pen and voice as input modalities. Moreover, the technologies used in multimodal applications such as speech recognition or speech synthesis are progressing continuously, and it is advisable to use a software architecture that enables the addition or substitution of modules without the need to change the rest of the application.

At a very high level of abstraction two software architectures seem to be available for building multimodal services, viz. script-based and agent-based systems. Both architectures are highly modular and support the use of many hardware configurations, ranging from a standalone PC to several networked PCs and workstations. These architectures also offer substantial flexibility in terms of the programming languages to write the modules. The major difference is in the way in which the modules communicate with each other and in the way the system is controlled.

Script-based architectures are best considered as finite-state rule-based systems. It is fair to say that the script defines system control. Essentially, for every state/condition the system can find itself in, and for every action of the user, a response must have been foreseen, which is formulated in the script. Existing script-based systems tend towards system driven interaction strategies, although user initiative can be accommodated by processing also unsolicited information provided by the user. Perhaps not surprisingly, script-based interaction strategies have been most popular in the spoken dialogue community. The best-known example of a script-based architecture that is suitable for the development of multimodal services is the DARPA Communicator. It should be mentioned that the Communicator architecture can handle virtually empty scripts, which would turn it into an agent-based architecture.

The heart of agent-based systems is often called a 'Facilitator', rather than a control module. This suggests a degree of flexibility that goes beyond the capabilities of a finite state machine. Of course, agent-based architectures are not anarchistic; in addition to the Facilitator module each realistic system will have some kind of Dialogue and Action Management module, which is responsible for the management and co-ordination of the system and its interaction with the outside world. Individual modules in an agent-based architecture still have specialised functions (ASR knows how to recognise speech, TTS knows how to convert text to audible speech, etc.), but this functionality is wrapped in a layer that contains the intelligence that is needed to know under what –perhaps very general- conditions the individual skills can contribute to the successful completion of the task. Agent architectures have been around for quite some time, although they used to be ignored by the speech community, probably because of the failure of ASR systems using blackboard architectures that were investigated in the seventies and early eighties. Well-known agent architecture include SRI's Open Agent Architecture (OAA) and OGI's Adaptive Agent Architecture (AAA).

6.1.1 Combining modes

One of the most important questions to be answered in the design of a multimodal service is addressed in W3C's request for proposals for "Multimodal Requirements For Voice Markup Languages", where a distinction is made between sequential and simultaneous use of the multiple modes that are available for input and/or output. Moreover, if simultaneous use is opted for, the distinction between co-ordinated and uncoordinated usage becomes relevant.

The founding of the SALTForum in the Fall of 2001 testifies of the difficulties to enhance the original VoiceXML architecture to support multimodal interaction. For the near future we expect to see quite a large number of proprietary, and to a large extent application specific solutions for multimodal interfaces. At the same time we expect a strong movement towards some kind of de facto standardisation, much in the same way as GUI interfaces have become standardised. Standardisation will –at the very least- provide a common set of interface building blocks for all services that are offered on the same family of devices from a single manufacturer. Yet, it is quite possible that we will see fierce competition between interface styles designed on top of different operating systems, such as Symbian, Windows CE, etc.

At this moment it is not completely clear whether there is any interference between the choice of the overall architecture of a multimodal application and the ease with which (un)co-ordinated usage of multiple input and/or output modes can be implemented. It may well be that co-ordinated simultaneous usage is much easier to implement in an agent architecture than in a script-based one. The fact that the co-ordinated simultaneous system proposed by Oviatt and her co-workers is based on the AAA may be an important piece of evidence in this respect.

6.2 Dialogue Design

The decisions about dialogue/interaction strategy in the design of a service have a deep impact on the ways in which the customer can (and sometimes must) interact, i.e., on the look-and-feel of the service, and on the ways in which the input and output modes can be used and combined. In this section a short summary is given of Dialogue Management Issues involved in multimodal human-machine interaction. The information is based on a survey of the literature, in addition to discussions with scientists who are actively involved in some of the bigger multimodal projects that are currently under way.

6.2.1 Dialogue Management

The term 'dialogue management' refers to the strategies that determine how the flow of the interaction is determined. The flow can be quite complex, especially when the number of information items that must be discussed grows larger. For the purpose of this explanation the concept of 'information item' can loosely be defined as the kind of information that needs to be entered in the slots of some form. In the case of a conference registration information items might include 'title', 'first name', 'surname', 'middle initials', 'credit card number', etc. In a travel service 'information items' might be 'departure city', 'date', 'destination', etc. From these examples it is already clear that some ’items’ can be molecular, rather than atomic: a date is a ’molecule’ made up of three ’atoms’, viz. day, month, and year.

It is well known that human-human dialogues (especially when speech is the medium) contain many turns that are not devoted to the transmission of factual information (information items), but rather to the process of keeping the information exchange going (hopefully in the right direction). A large proportion of the turns can be devoted to confirmation actions, i.e., attempts of one (or both) of the partners in a dialogue that aim to confirm that the factual information is correctly understood. For the purpose of this exposé we will call dialogue terms that convey factual information 'primary'; dialogue acts devoted to the maintenance of the process will be indicated as 'meta acts'.

Recently it has been pointed out that human-computer dialogues need a third type of turn, which is relatively rare (but certainly not impossible) in human-human interaction. This type of turn is concerned with attempts of the computer to explain its capabilities and its command of the domain of the conversation. For example, a travel information service might need to explain, at some point in a dialogue, that it can only provide information on domestic (and not on international) travel; or that is can only make reservations for customers who have previously authorised the service to bill against a credit card. We will use the term 'about-task' for this type of dialogue acts [17]. In a way, about-task acts are reminiscent of system initiated (context sensitive) *help* procedures. One can, of course, easily envisage the complementary situation where a user explicitly invites a system to explain its capabilities.

The term 'dialogue management' refers to the way in which it is determined what the type of the next turn in the interaction will be (primary, meta, or about-task). In the case of primary acts decisions must be made regarding the information item(s) that are addressed in the next turn.

The basic issues of dialogue management in multimodal interaction are not essentially different from unimodal interaction: at the three extremes of the triangle we have

1. **user driven control**

   The choice of the type of dialogue act, and the information item to be addressed is always determined by the user. The simplest example is where a user determines the order in which the fields in a form on a screen are filled (e.g. by positioning the cursor in the field).

   It should be noted that user driven control does not imply complete freedom for the user. For example, some fields in a form may only accept a predefined set of valid values. However, this does not detract from the essential freedom of the user to determine the order in which the fields are filled.

2. **system driven control**

   The system determines the type of dialogue act, as well as the information item(s) to be addressed in the next turn. The most apparent example of a system driven service are the conventional IVR services, where the customer is prompted by the system to make specific selections in each turn of the interaction. It is tacitly assumed that at each dialogue node there is a closed set of valid responses from which the user must make a selection.

   Almost invariably the closed set of response alternatives includes a 'backup' option, that the customer can use to alert the system that the dialogue is no longer on the right track, and that it must back up to a previous state. This (essential) possibility for the human partner to change the straight course of the dialogue is compatible with system control, since the backup option is one of the valid responses in the set.

3. **mixed initiative control**

   In mixed initiative dialogues both the human and the computer have the freedom to propose the type of the next dialogue turn, as well as the information item that is addressed. This is the default in human-human dialogues, where a question can be replied to with another question, if only the request to explain the meaning of the first question.

Most spoken dialogue systems seem to adhere to a system driven interaction style. Desktop applications, on the other hand, are typically user driven. It remains to be seen whether multimodal interactions will profit most from the desktop or from the telephone metaphor. For the time being we expect that multimodal interaction will tend towards some kind of user driven interaction.

7 Conclusions

Multimodal human machine interaction is quickly becoming an issue of vital importance. It is an emerging field, where almost all aspects remain undecided and uncertain. The aspects include the terminals, the network protocols, the service architectures, and the roles of different types of companies in the development and deployment of the services.

Multimodal terminals will evolve along two lines. First, cell phones will get larger displays, colour displays, and more features, such as pen input, video cameras, and built-in Java engines that offer the possibility to implement a large range of services. At the same time PDAs will evolve to include the full range of short-range (Bluetooth, Piconet) and long distance (GPRS, UMTS) communication ports. Despite the different routes, all types of future mobile terminals will be small and lightweight, and all future services will be essentially Internet based. The combination of increasing complexity of the applications and the lack of desktop input/output devices will require the development of a completely different interaction style for the future mobile terminals and the attendant services. Strategists expect to see a completely transparent communication with an ‘ambient intelligence landscape’, where sensors and compute power are pervasive and distributed. In this ambient intelligence landscape communication with the systems will rely heavily on a combination of speech, natural language, gestures and graphical displays. Because of the importance of speech and natural language customers will expect to be able to interact with their environment in their native language. This explains the importance of multilinguality for the future mobile services.

The route towards the pervasive intelligence landscape will be long and uncertain. It is beyond doubt that we will have to learn how to communicate by means of ‘devices’ that have no conventional keyboard. The concept of the ‘keyboardless society’ already appeared as one of the targets in the 5th Framework Programme of the European Union. The lack of full-size keyboards puts automatic speech recognition at the centre of the stage. (And in its wake we will see increasing attention for speech synthesis, and also for speaker verification.) However, little is known at this moment about the ergonomy of speech-centric multimodal interaction. To avoid disappointments with service trials that fail because of inadequate multimodal user interfaces, companies intending to develop multimodal services should pay close attention to the developments in the field of interface design for mobile services. It is extremely important to allow the user to always be in control, and able to select the interaction modes that are most appropriate in a given situation. This is the more so because mobile terminals will often be used in situations where the customer needs to pay attention to other things than just the interaction with the service. In-car applications, where eyes-busy situation will be frequent, are just one example. Also at home there are many circumstances in which users are combining several tasks at the same time.

Future mobile terminals and the services that they will access will be highly personalised. Moreover, customers will use their terminals and the services from many different locations. In the multilingual Europe this will inevitably mean that services must be accessible in the native language of the customer, independent of the location. Therefore, the interface to the services must be language independent. The only way in which this can be accomplished is through clear architectural separations of the language dependent and language independent parts of the service and the underlying information databases. The language independent kernel must then be made accessible through interfaces that operate in multiple languages. Over the last years the developments in speech-driven services have clearly shown the importance of standard building blocks with which new services can be built quickly and inexpensively. VoiceXML is probably the most outstanding result of that development. In the field of multimodal interaction attempts to define some kind of standard building blocks have already been initiated, in the first place in the form of multimodal extensions of VoiceXML. However, it appears that such extensions may be difficult to define. That is the reason for the founding of the SALTForum that aims at the development of standards for multimodal interaction that are not constrained by decisions made in the specification of VoiceXML. Standardisation of building blocks for multimodal interaction will be rather complicated, because of the large number of input and output parameters..

This study has identified a number of players who must join forces in order to develop and deploy attractive and successful multimodal services. In addition to the telecom network operators there are essential roles to be played by the companies that manufacture network infrastructures, companies manufacturing terminal devices, content providers, interaction technology providers (speech recognition and synthesis, handwriting recognition, etc.) and service developers. In this complex and rapidly developing field of force all prospective players must find their proper position. Different telecom network operators (and the same holds for all other players) may want to make different decisions, depending on such factors as their relations to content and technology providers. This study offers concrete data and pointers to external information sources that should be able to support strategists and decision makers in the companies to make reasoned decisions about the most desirable position and role. It should be noted that NTT DoCoMo could only reach its dominant position in I-Mode services because it was able to specify the network infrastructure, the protocols, the terminals and the services platform, and have everything built to its specifications for a large and affluent home market. The technology catered to a need for information and entertainment and the demands posed by the Japanese writing tradition, that needed this kind of technology.

Last but not least, this study provides pointers to knowledge and expertise in the field of usability research for multimodal interaction. Because the field is so young, that experience is rather rare. Yet, it is clear that this expertise is essential to be able to separate the wheat from the chaff in the many offerings of multimodal service platforms that are at the horizon. The telecom network operators should either ensure that they build and maintain this expertise in house, or get access to it through structural links with external research laboratories.

References

[1] Peter Wyard and Gavin Churcher (2000) "All Channels Open: Multimodal Human-Computer Interfaces", BT Technical Journal, Vol. 18, No. 1, January 2000.

[2] W3C, “Multimodal requirements for voice markup languages”, W3C working draft July 2000, http://www.w3.org/TR/multimodal-reqs, 2000.

[3] http://www.saltforum.org.

[4] Ducatel, K. Bogdanowics, M., Scapolo, F. Leijten, J. & Burgelman, J.-C. Scenarions for Ambient Intelligence in 2010. ftp://ftp.cordis.lu/pub/ist/docs/istagscenarios2010.pdf.

[5] Comerford, L., Frank, D., Gopalakrishnan, P., Gopinath, R., Sedivy, J. “The IBM Personal Speech Assistant”, Proc. ICASSP-201.

[6] Huang, X, Acero, A et al. ‘‘MIPAD: A Multimodal Interaction Prototype’’, Proc. ICASSP-2001.

[7] den Os, E., de Koning, N., Jongebloed, H., Boves, L. “Usability of a Speech Centric Multi-modal Directory Assistance Service”, Proc. CLASS Workshop, Verona, 13-15 December 2001.

[8] Kvale, K., Warakagoda, N.D., Knudsen, J.E. Speech-Centric Multimodal interaction with Small Mobile Terminals. Proc. NORSIG-2001, 18-20 October 2001, Trondheim.

[9] Wahlster, W., Reithinger, N., Blocher, A. SmartKom: Multimodal Communication with a Life-Like character. Proc. EUROSPEECH, Aalborg, Denmark, 2001, pp. 1547-1550.

[10] Béchet, F., den Os, E. Boves, L., Sienel, J. “Introduction to the IST-HLT project Speech-driven Multimodal Automatic Directory Assistance (SMADA)”, *Proc. ICSLP-2000*, Beijing.

[11] Oviatt, S.L., Cohen, P.R., Wu, L.,Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J. & Ferro, D. “Designing the User Interface for Multimodal Speech and Pen-based Gesture Applications: State-of-the-Art Systems and Future Research Directions”, Human Computer Interaction, 2000, vol. 15, no. 4, 263-322.

[12] Almeida, L., et al. “Multilingual Web sites: Best practice guidelines and architectures”, EURESCOM, 2001, Project P923.

[13] Fügen, C. Westphal, M., Schneider, M., Schultz, T., Waibel, A. „LingWear: A Mobile Tourist Information System“, Proc. of HLT 2001, pp. 230-234.

[14] Daniels, J.J. Bell, B. „Listen-Communicate-Show (LCS): Spoken Language Command of Agent-based Remote Information Access”, Proc. of HLT 2001, pp. 235-238.

[15] Sturm, J., Wang, F., Cranen, B “Adding extra input/output modalities to a spoken dialogue system”, Proc. 2nd ACL SIGdial Workshop on Discourse and Dialogue, Aalborg, September 2001.

[16] “Emerging Thematic Priorities for Research in Europe: Information and Communication Technologies”, Institute for Prospective Technological Studies, Sevilla, http://futures.jrc.es

[17] Walker, M., R. Passonneau “DATE : A Dialogue Act Tagging Scheme for Evaluation of Spoken Dialogue Systems”, Proc. HLT2001, pp. 66-73.

Annex A ISTAG Scenarios

In this Annex we present short versions of the four scenarios for future use of telecommunication services developed by the Information Society Technology Advisory Group (ISTAG) in order to allow the decision makers in the European Commission to understand the technologies that need to be developed. The text in this Annex is cited from reference [1].

A.1 Maria – Road Warrior

After a tiring long haul flight Maria passes through the arrivals hall of an airport in a Far Eastern country. She is travelling light, hand baggage only. When she comes to this particular country she knows that she can travel much lighter than less than a decade ago, when she had to carry a collection of different so-called personal computing devices (laptop PC, mobile phone, electronic organisers and sometimes beamers and printers). Her computing system for this trip is reduced to one highly personalised communications device, her ‘P–Com’ that she wears on her wrist. A particular feature of this trip is that the country that Maria is visiting has since the previous year embarked on an ambitious ambient intelligence infrastructure programme. Thus her visa for the trip was self-arranged and she is able to stroll through immigration without stopping because her P-Comm is dealing with the ID checks as she walks.

A rented car has been reserved for her and is waiting in an earmarked bay. The car opens as she approaches. It starts at the press of a button: she doesn’t need a key. She still has to drive the car but she is supported in her journey downtown to the conference centre-hotel by the traffic guidance system that had been launched by the city government as part of the ‘AmI-Nation’ initiative two years earlier. Downtown traffic has been a legendary nightmare in this city for many years, and draconian steps were taken to limit access to the city centre. But Maria has priority access rights into the central cordon because she has a reservation in the car park of the hotel.

Central access however comes at a premium price, in Maria’s case it is embedded in a deal negotiated between her personal agent and the transaction agents of the car-rental and hotel chains. Her firm operates centralised billing for these expenses and uses its purchasing power to gain access at attractive rates. Such preferential treatment for affluent foreigners was highly contentious at the time of the introduction of the route pricing system and the government was forced to hypothecate funds from the tolling system to the public transport infrastructure in return. In the car Maria’s teenage daughter comes through on the audio system. Amanda has detected from ‘En Casa’ system at home that her mother is in a place that supports direct voice contact. However, even with all the route guidance support Maria wants to concentrate on her driving and says that she will call back from the hotel.

Maria is directed to a parking slot in the underground garage of the newly constructed building of the *Smar-tel Chai*n. She is met in the garage by the porter – the first contact with a real human in our story so far! He helps her with her luggage to her room. Her room adopts her ‘personality’ as she enters. The room temperature, default lighting and a range of video and music choices are displayed on the video wall. She needs to make some changes to her presentation – a sales pitch that will be used as the basis for a negotiation later in the day. Using voice commands she adjusts the light levels and commands a bath. Then she calls up her daughter on the video wall, while talking she uses a traditional remote control system to browse through a set of webcast local news bulletins from back home that her daughter tells her about. They watch them together.

Later on she ‘*localise*s’ her presentation with the help of an agent that is specialised in advising on local preferences (colour schemes, the use of language). She stores the presentation on the secure server at headquarters back in Europe. In the hotel’s seminar room where the sales pitch is take place, she will be able to call down an encrypted version of the presentation and give it a post presentation decrypt life of 1.5 minutes. She goes downstairs to make her presentation…this for her is a high stress event. Not only is she performing alone for the first time, the clients concerned are well known to be tough players. Still, she doesn’t actually have to close the deal this time. As she enters the meeting she raises communications access thresholds to block out anything but red-level ‘emergency’ messages. The meeting is rough, but she feels it was a success. Coming out of the meeting she lowers the communication barriers again and picks up a number of amber level communications including one from her cardio-monitor warning her to take some rest now. The day has been long and stressing. She needs to chill out with a little meditation and medication. For Maria the meditation is a concert on the video wall and the medication….a large gin and tonic from her room’s minibar.

A.2 Dimitrios and the Digital Me (D-Me)

It is four o’clock in the afternoon. Dimitrios, a 32 year-old employee of a major food-multinational, is taking a coffee at his office’s cafeteria, together with his boss and some colleagues. He doesn’t want to be excessively bothered during this pause. Nevertheless, all the time he is receiving and dealing with incoming calls and mails.

He is proud of ‘being in communication with mankind’: as are many of his friends and some colleagues. Dimitrios is wearing, embedded in his clothes (or in his own body), a voice activated ‘gateway’ or digital avatar of himself, familiarly known as ‘D-Me’ or ‘Digital Me’. A D-Me is both a learning device, learning about Dimitrios from his interactions with his environment, and an acting device offering communication, processing and decision-making functionality. Dimitrios has partly ‘programmed’ it himself, at a very initial stage. At the time, he thought he would ‘upgrade’ this initial data periodically. But he didn’t. He feels quite confident with his D-Me and relies upon its ‘intelligent ‘ reactions.

At 4:10 p.m., following many other calls of secondary importance – answered formally but smoothly in corresponding languages by Dimitrios’ D-Me with a nice reproduction of Dimitrios’ voice and typical accent, a call from his wife is further analysed by his D-Me. In a first attempt, Dimitrios’ ‘avatar-like’ voice runs a brief conversation with his wife, with the intention of negotiating a delay while explaining his current environment. Simultaneously, Dimitrios’ D-Me has caught a message from an older person’s D-Me, located in the nearby metro station. This senior has left his home without his medicine and would feel at ease knowing where and how to access similar drugs in an easy way. He has addressed his query in natural speech to his D-Me.

Dimitrios happens to suffer from similar heart problems and uses the same drugs. Dimitrios’ D-Me processes the available data as to offer information to the senior. It ‘decides’ neither to reveal Dimitrios’ identity (privacy level), nor to offer Dimitrios’ direct help (lack of availability), but to list the closest drug shops, the alternative drugs, offer a potential contact with the self-help group. This information is shared with the senior’s D-Me, not with the senior himself as to avoid useless information overload.

Meanwhile, his wife’s call is now interpreted by his D-Me as sufficiently pressing to mobilize Dimitrios. It ‘rings’ him using a pre-arranged call tone. Dimitrios takes up the call with one of the available Displayphones of the cafeteria. Since the growing penetration of D-Me, few people still bother to run around with mobile terminals: these functions are sufficiently available in most public and private spaces and your D-Me can always point at the closest…functioning one! The ‘emergency’ is about their child’s homework. While doing his homework their 9 year-old son is meant to offer some insights on everyday life in Egypt. In a brief 3-way telephone conference, Dimitrios offers to pass over the query to the D-Me to search for an available direct contact with a child in Egypt. Ten minutes later, his son is videoconferencing at home with a girl of his own age, and recording this real-time translated conversation as part of his homework. All communicating facilities have been managed by Dimitrios’ D-Me, even while it is still registering new data and managing other queries. The Egyptian correspondent is the daughter of a local businessman, well off and quite keen on technologies. Some luck (and income…) had to participate in what might become a longer lasting new relation.

A.3 Carmen: traffic, sustainability & commerce

It is a normal weekday morning. Carmen wakes and plans her travel for the day. She wants to leave for work in half an hour and asks AmI, by means of a voice command, to find a vehicle to share with somebody on her route to work. AmI starts searching the trip database and, after checking the willingness of the driver, finds someone that will pass by in 40 minutes. The in-vehicle biosensor has recognised that this driver is a non-smoker – one of Carmen requirements for trip sharing. From that moment on, Carmen and her driver are in permanent contact if wanted (e.g. to allow the driver to alert Carmen if he/she will be late). Both wear their personal area networks (PAN) allowing seamless and intuitive contacts.

While taking her breakfast coffee Carmen lists her shopping since she will have guests for dinner tonight. She would like also to cook a cake and the e-fridge flashes the recipe. It highlights the ingredients that are missing milk and eggs. She completes the shopping on the e-fridge screen and asks for it to be delivered to the closest distribution point in her neighbourhood. This can be a shop, the postal office or a franchised nodal point for the neighbourhood where Carmen lives. All goods are smart tagged, so that Carmen can check the progress of her virtual shopping expedition, from any enabled device at home, the office or from a kiosk in the street. She can be informed during the day on her shopping, agree with what has been found, ask for alternatives, and find out where they are and when they will be delivered.

Forty minutes later Carmen goes downstairs onto the street, as her driver arrives. When Carmen gets into the car, the VAN system (Vehicle Area Network) registers her and by doing that she sanctions the payment systems to start counting. A micro-payment system will automatically transfer the amount into the e-purse of the driver when she gets out of the car. In the car, the dynamic route guidance system warns the driver of long traffic jams up ahead due to an accident. The system dynamically calculates alternatives together with trip times. One suggestion is to leave the car at a nearby ‘park and ride’ metro stop. Carmen and her driver park the car and continue the journey by metro. On leaving the car, Carmen’s payment is deducted according to duration and distance.

Out of the metro station and whilst walking a few minutes to her job, Carmen is alerted by her PAN that a Chardonnay wine that she has previously identified as a preferred choice is on promotion. She adds it to her shopping order and also sets up her homeward journey with her wearable. Carmen arrives at her job on time.

On the way home the shared car system senses a bike on a dedicated lane approaching an intersection on their route. The driver is alerted and the system anyway gives preference to bikes, so a potential accident is avoided. A persistent high-pressure belt above the city for the last ten days has given fine weather but rising atmospheric pollutants. It is rush hour and the traffic density has caused pollution levels to rise above a control threshold. The city-wide engine control systems automatically lower the maximum speeds (for all motorised vehicles) and when the car enters a specific urban ring toll will be deducted via the Automatic Debiting System (ADS).

Carmen arrives at the local distribution node (actually her neighbourhood corner shop) where she picks up her goods. The shop has already closed but the goods await Carmen in a smart delivery box. By getting them out, the system registers payment, and deletes the items from her shopping list. The list is complete. At home, her smart fridge screen will be blank. Coming home, AmI welcomes Carmen and suggests to telework the next day: a big demonstration is announced downtown.

A.4 Annette and Solomon in the Ambient for Social Learning

It is the plenary meeting of an environmental studies group in a local ‘Ambient for Social Learning’. The group ranges from 10 to 75 years old. They share a common desire to understand the environment and environmental management. It is led by a mentor whose role it is to guide and facilitate the group’s operation, but who is not necessarily very knowledgeable about environmental management. The plenary takes place in a room looking much like a hotel foyer with comfortable furniture pleasantly arranged. The meeting is open from 7.00-23.00 hours. Most participants are there for 4-6 hours. A large group arrives around 9.30 a.m. Some are scheduled to work together in real time and space and thus were requested to be present together (the ambient accesses their agendas to do the scheduling).

A member is arriving: as she enters the room and finds herself a place to work, she hears a familiar voice asking “Hello Annette, I got the assignment you did last night from home: are you satisfied with the results?” Annette answers that she was happy with her strategy for managing forests provided that she had got the climatic model right: she was less sure of this. Annette is an active and advanced student so the ambient says it might be useful if Annette spends some time today trying to pin down the problem with the model using enhanced interactive simulation and projection facilities. It then asks if Annette would give a brief presentation to the group. The ambient goes briefly through its understanding of Annette's availability and preferences for the day's work. Finally, Annette agrees on her work programme for the day.

One particularly long conversation takes place with Solomon who has just moved to the area and joined the group. The ambient establishes Solomon's identity; asks Solomon for the name of an ambient that 'knows' Solomon; gets permission from Solomon to acquire information about Solomon's background and experience in Environmental Studies. The ambient then suggests Solomon to join the meeting and to introduce himself to the group.

In these private conversations the mental states of the group are synchronised with the ambient, individual and collective work plans are agreed and in most cases checked with the mentor through the ambient. In some cases the assistance of the mentor is requested. A scheduled plenary meeting begins with those who are present. Solomon introduces himself. Annette gives a 3-D presentation of her assignment. A group member asks questions about one of Annette's decisions and alternative visualisations are projected. During the presentation the mentor is feeding observations and questions to the ambient, together with William, an expert who was asked to join the meeting. William, although several thousand miles away, joins to make a comment and answer some questions. The session ends with a discussion of how Annette's work contributes to that of the others and the proposal of schedules for the remainder of the day. The ambient suggests a schedule involving both shared and individual sessions.

During the day individuals and sub-groups locate in appropriate spaces in the ambient to pursue appropriate learning experiences at a pace that suits them. The ambient negotiates its degree of participation in these experiences with the aid of the mentor. During the day the mentor and ambient converse frequently, establishing where the mentor might most usefully spend his time, and in some cases altering the schedule. The ambient and the mentor will spend some time negotiating shared experiences with other ambients – for example mounting a single musical concert with players from two or more distant sites. They will also deal with requests for references / profiles of individuals. Time spent in the ambient ends by negotiating a homework assignment with each individual, but only after they have been informed about what the ambient expects to happen for the rest of the day and making appointments for next day or next time.