EURESCOM

European Institute for Research and Strategic Studies in Telecommunications

May 2002


Project P1104

# MUST – multimodal and multilingual services for small mobile terminals

## Project leader

Els den Os

## Project supervisor

Peter Stollenmayer

## Who should read this brochure?

This brochure addresses professionals in the telecommunications and IT industry, who are aware of the critical importance of developing customer-friendly products and services

**Editors:**
Lou Boves, Els den Os

**Contributions of:**
Lou Boves
Els den Os
Malek Boualem
Pascal Filoche
Claude Tallec
Edouard Hinard
Nuno Beires
Luis Almeida
Rui Gomes
John Rugelbak
Ingunn Amdal
Jan Knudsen
Narada Warakagoda
Knut Kvale
Paal Loekstad

## Participants in the MUST project P1104:

France Télécom
Telenor
Portugal Telecom
Max Planck Institute Nijmegen
University of Nijmegen

maps by Planfax

# MUST – multimodal and multilingual services for small mobile terminals

# Table of contents

# Executive Summary

Future automatic services will very likely come with interfaces that support conversational interaction. Eventually the interface will become transparent, and interaction with the devices and services will be as easy and natural as talking to a friend. In face-to-face communication we use all our senses. We do not only talk to each other; we also see facial expressions, hand gestures, sketches and words scribbled with a pen, etc.

Face-to-face interaction is multimodal.

Today the development of multimodal interaction has just started. Terminals are emerging, which support the combination of speech and pen for input, and audio, graphics and text for output. At the same time protocols are being developed, which can support simultaneous upstream transmission of speech and pen information, and downstream transmission of audio, graphics and text.

Multimodal interaction will help to overcome the usability problems inevitably related to the lightweight mobile terminals of the future, which will come without a full-size keyboard, and with only a small screen. Clever use of automatic speech recognition and speech synthesis, in combination with pen input and graphical displays will create a new generation of interfaces. We are only at the start of this development, where we must learn how speech, pen, text and graphics can be combined effectively.

This brochure introduces multimodal interaction with small mobile terminals in cellular telecommunication environments.

After a short introduction of the concept of multimodal interaction, we discuss the most important usability issues of future terminals too small to have a full-sized keyboard and screen. Then, we sketch the development that we expect to see in the next ten years, by depicting the infrastructure and possible services in the years 2002, 2004, 2006 and 2010. We continue with a more in-depth analysis of the technical issues involved in multimodal interaction, and explain how these will affect the development of user-friendly interfaces. Rather than trying to add multimodality to an existing interface, a multimodal interface should be designed in its own right, starting from the functionality of the terminal and the service. There is a clear need for the development of standards, not only for the technology, but also for the interaction procedures. We discuss and explain an architecture for multimodal services that should enable the development of suitable standards.

Conversational interfaces are almost by definition language centric. Therefore, it is extremely important that customers can use a language in which they are really fluent. In most cases that is their mother tongue. Since many services will be offered on a roaming basis, future multimodal services will also have to be multilingual. This brochure introduces the commercial advantages and the technological implications of multilinguality.

The brochure concludes with a short discussion of the development of multimodal telecommunication services. We believe that these services can only be developed by alliances of a large number of players, including telecommunication network operators and the manufacturers of terminal and transmission infrastructure, who must work with content providers, designers and engineers who have experience in the integration of the enabling technologies.

# Introduction

Life would be much easier, and the opportunities for new services would be enormous, if we were able to interact with automatic devices in the same way as with intelligent fellow human beings. Therefore, all major computer and software companies (including IBM, Microsoft, and Hewlett-Packard, to name just the biggest ones) invest substantial resources in R&D that is aimed towards making human-computer interaction more humanlike and consequently more user friendly.

When we talk to a friend or an assistant we do not only convey information through the formal meaning of words. There may be essential information carried by the tone of voice and facial expression, and we use our hands and perhaps also a pencil or a pen to gesture, point, sketch or write. Thus, face-to-face communication is multimodal, in that we use all human senses –or communication modes- to convey and process information. Because of the advantages of multimodal interaction it seems only natural to assume that human-computer interaction would also become more efficient and easier, and eventually more 'natural' if it could be made multimodal.

In this brochure we will show that there is yet another reason for developing multimodal human-computer interfaces. In the increasingly more mobile society of the 21st century we will often find ourselves in a situation where we do not have a full-size keyboard and screen available. In these situations we will have to make do with communication channels that are less dependent on bulky equipment. Speech and pen are obvious examples. In this brochure we will provide information on how multimodal interaction can enlarge the functionality of small mobile devices.

Thinking about human-computer interaction in terms of a conversation with a friend or assistant is fundamentally different from the desktop metaphor in Windows interfaces. One difference that catches the eye immediately is the role of language. The more interactions are in the form of some form of dialogue or conversation, the more it becomes important that we are able to use a language that we really feel comfortable with. For the large majority of the population – and, therefore, of the future customers – that language will be the mother tongue. Therefore, many future 'conversational' services will be multilingual. In this brochure we will touch upon multilinguality in the context of multimodal services.

## The EURESCOM "MUST" Project P1104

EURESCOM has acknowledged the importance of multimodal services that will be feasible with the terminals and the networks of the near future. These will be services that use speech, pen and graphics on the screen of telephone handsets or palmtop computers. It has, therefore, launched the project P1104 "MUST". "MUST" is an acronym for MUltimodal and multilingual Services for small mobile Terminals. The single most important goal of MUST is to clarify some of the fundamental usability issues that are involved in multimodal and multilingual services when there is no keyboard for input, as in the example shown in Figure 1.

This brochure summarises the results of the first year of the MUST project. We take a closer look at where multimodal interaction is today and where it will go in the near and not so near future. Multimodality and multilinguality indeed hold great promises. However, to earn revenues, substantial investments are still needed, not only in the hardware infrastructure, but also in the development of knowledge about the design of user interfaces that will enable to harness all the advantages of conversational interaction.

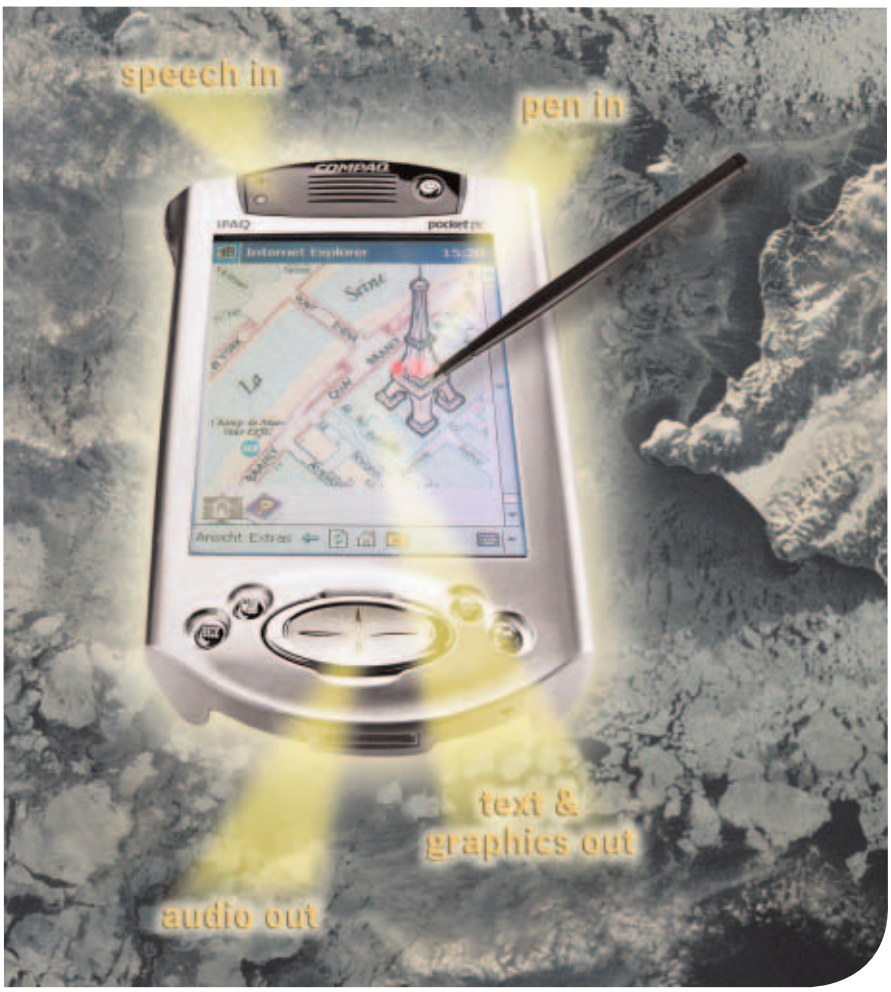



**Figure 1:**
**Multimodal interaction with a small mobile terminal**

# New interaction paradigms

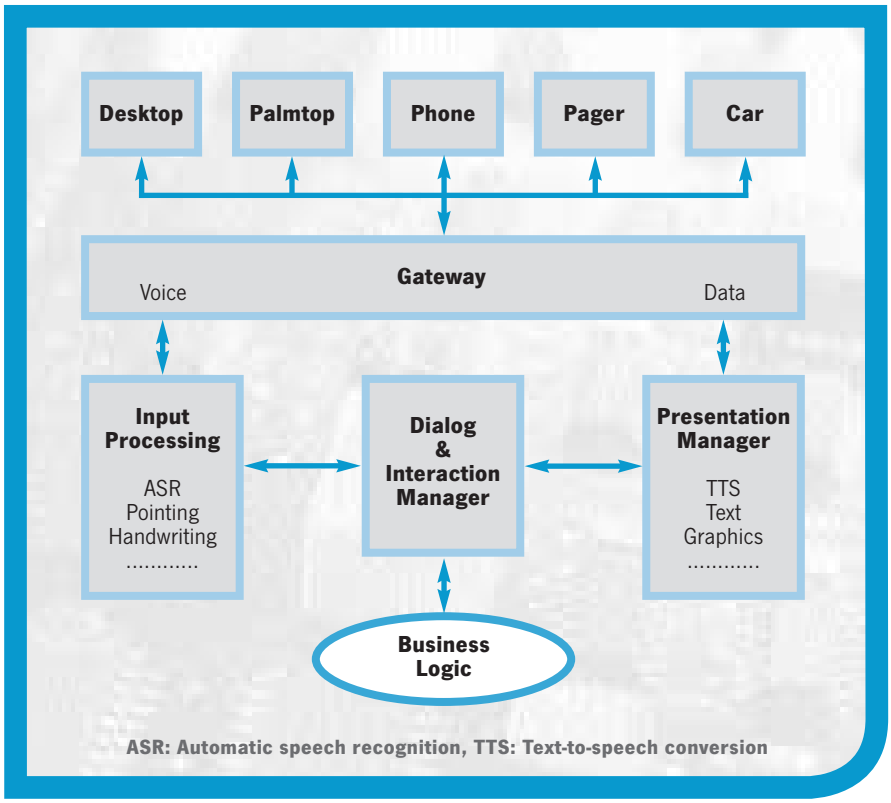
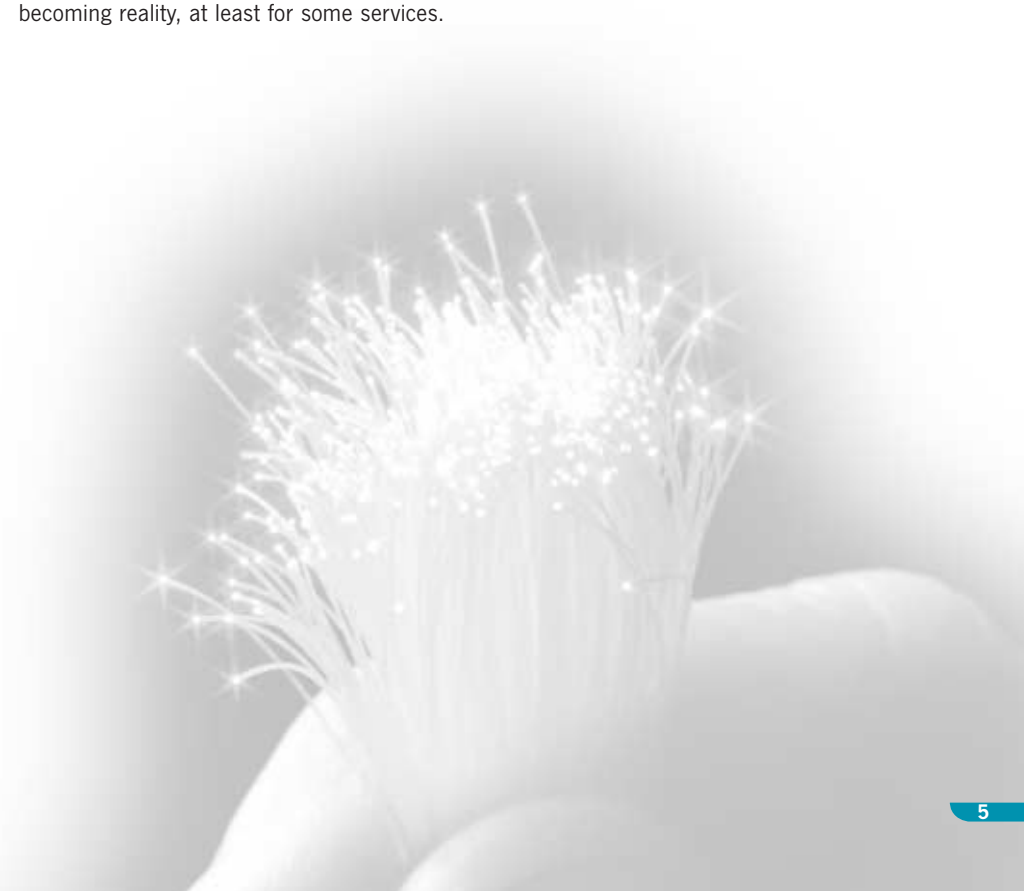
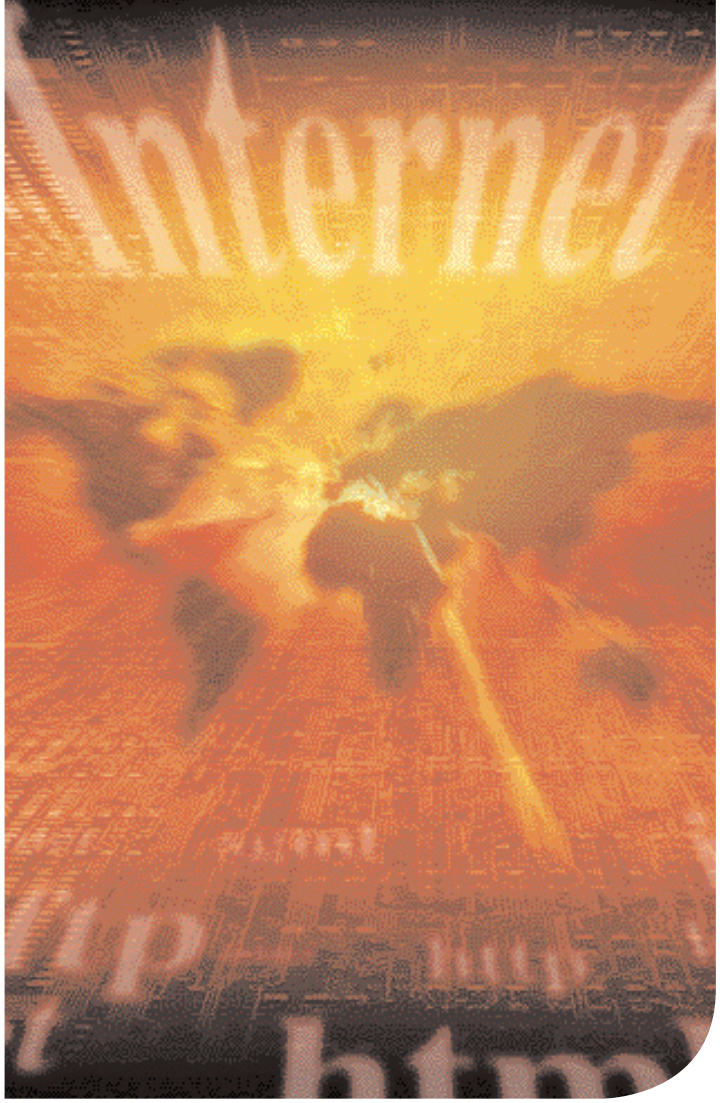



**Figure 2:**
**Device independent access to services**

We are at the beginning of the next revolution in telecommunications and computing. This revolution may change the scene more dramatically than ever before, because it will affect the end users in the most inescapable way possible: it will change the design, appearance and functionality of the handsets. In a few years from now, telephone handsets will have the same functionality as electronic organisers. People from the computer industry will say that the PDAs will acquire the functionality of a telephone. Whichever way one may look at it, soon we will have terminals that provide both speech and Internet communication over the cellular GPRS and UMTS networks.

Moreover, customers will expect to be able to access services irrespective of time and geographic location, using the terminal and interaction modes that are most appropriate under the given conditions. In the office that is likely to be a desktop computer, but on the move it will be a smaller lightweight device. In the car the terminal will be connected to the car-PC, much in the same way as a laptop is plugged into the docking station in the office. The situation sketched in Figure 2 is already becoming reality, at least for some services.





## Usability issues

Ubiquitous access may be highly desirable, but it will not come easy. This is because the lightweight, energy-efficient small terminals of the future will come without a real keyboard, and with only a small screen with a limited number of pixels. Although we may eventually see roll-able or foldable screens and keyboards that can be projected on any surface, we will have to make do with limited input and output capabilities of the tiny terminals that have become so popular. Because of the functional limitations of those terminals, it will be necessary to develop completely new human-computer interaction paradigms. The most promising metaphor for this new way of interaction is the conversation. Therefore, spoken natural language will be pivotal. At the same time it is evident that speech is not the best medium for rendering complex factual information, so that there is a need to combine speech with other interaction modalities. Although this combination is straightforward for humans, it appears to be surprisingly difficult to let computers do a similar job.

## Technical issues

Access to services with different terminals over different networks creates technical problems that may interfere with the design of natural and transparent interaction procedures. To a large extent the problems are due to the need to transfer multiple parallel data streams over narrow band cellular networks. At the moment several efforts are under way to develop protocols that allow connecting a wide range of devices to automatic services. Some of those developments aim at proprietary technology, while others are committed to open standards, especially in the framework of the World Wide Web Consortium (W3C). The lack of widely accepted protocols complicates the development of promising services. Fortunately, transmitting parallel data streams will become easier when broadband cellular networks based on the Internet Protocol (IP) will become available. We expect that this will facilitate the development of sufficiently powerful open standards protocols for multimodal communication in the near future.

# A look into the future

The telecommunication networks of the future will all be based on the Internet Protocol (IP). From the service development point of view IP networks have a number of advantages. This is especially true for multimodal services, which will require simultaneous transmission of multiple parallel data stream, such as voice, text, graphics and information about the geographical location of the user. Location information can, for example, be provided by the Global Position System (GPS) or its future European counterpart Galileo.

Reliable, low-cost and ubiquitous access to mobile and fixed IP networks will enable the rollout of many new services. Although it is impossible to present the 'definitive' roadmap for the development of those services, we think that it is useful to present a sequence of scenarios that suggest increasingly complex and appealing services. In doing so we focus on services for people on the move.

## Imminent introduction of multimodal services

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, by offering services based on automatic speech recognition (ASR) and text-to-speech conversion (TTS). VoiceXML will serve as the bridge between the voice network and Internet based services. The alternative development will try to make Internet services originally designed for PC use accessible from palmtop computers, which by default do not have a full-size keyboard. In both cases the challenge will be to overcome the limited input/output capabilities of the terminal devices. Combining speech and pen for input, and speech, text and graphics for output is expected to provide the solution. Because of the lack of standardised terminals that allow combining these modes, we will see mainly proprietary developments, promoted by technology providers who aim to reap the fruits of their position as first and fast movers. The extent to which input and output modes can be combined to form a flexible interface will be limited, due to the restricted functionality of the terminals and the transmission protocols. We expect that the ongoing R&D in multimodal interaction will not result in services with substantial new functionality in the year 2002. Rather, we expect to see several field trials of services with a focus on ease of use, thanks to the combination of speech, pen, and graphics. The services will allow customers to find a nearby restaurant serving a specific type of food, in a given price range, where cell information will be used to determine the location. It will also be possible to consult movie and theatre program guides and to make seat reservations, or to book flights, make reservations for car rentals and more.

## Full multimodal interaction will be available in 2004

The multimodal service field trailed in 2002 will be further developed, expanded and commercialised. Customers will be able to access up-to-date dynamic tourist information for the major cities in Europe with a small mobile handset, in their preferred language, and using the interaction modalities that they find most pleasant. Speech and pen can be freely combined for input, while output will be a combination of graphics, text and speech. This will make it possible to circle an object on the screen and at the same time ask questions about it, even open questions such as 'What is this thing?' or 'How can I get there?' The system's response can be a combination of graphical output, text displayed on the screen and speech.

Full multimodal interaction will enhance a range of professional services in 2004, allowing business people to maintain full synchronisation between their PDA and the information in the company Intranet. New appointments and orders can be entered with the PDA using speech recognition. The entered data is immediately and automatically transferred to the central agenda of the company. In addition, a wide range of entertainment services will be available. For example, a customer who has a day off in Paris can be alerted about a special exhibition in the Centre Pompidou. Customers can subscribe to information push services, especially in the fields of sports and music entertainment. A tourist 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 a match is going on at the same time, he can set his handset to receive interesting events, such as goals, penalties or near misses.

The functionality of these services can, at least in principle, also be offered with interfaces based on monomodal interaction. However, the ability to combine speech, pen, and graphics will make a decisive difference in terms of usability. Clever multimodal interfaces will greatly facilitate the navigation in large content databases. We expect to see the first examples of conversational interaction, to overcome the need for the customer to know and understand the structure of the content databases. As soon as it is possible to ask directly "what is shown in the Opéra Bastille", it is no longer necessary to know how to navigate the complex menu structure to reach the requested node.

## Conversational interaction in 2006

In the year 2006, the terminals will have bigger screens, and the bandwidth of the networks will have increased significantly. Terminals will come with an integrated video camera. Moreover, enabling technologies, such as automatic speech and gesture recognition, speech synthesis, generation of convincing avatars and artificial intelligence in the back-end application, will have improved substantially. This will enable basic problem solving services. For example, a customer may use a service for planning a holiday, suggesting options that fit the family situation, financial and time constraints, and entertainment preferences. This can be done through multimodal communication, much like the situation where the customer and the travel agent discuss face-to-face over a brochure. Professional services will include interactive design of interior decorations by architects who are in their customer's house.

## Long term developments: 2010 and beyond

Governments and commercial companies have developed scenarios that may seem quite futuristic today, but are based on realistic expectations about the development in enabling technologies. These technologies will enable services that rely on pervasive networked computing. Customers will carry their Personal Area Network (PAN), which uses the integrated telephone/PDA as a kind of server. PANs are constantly connected to sensors in the environment and to Local and Wide Area Networks that run a large number of advanced intelligent services. Customers will communicate with other persons and with automatic services in a fully transparent way. The total will have the look and feel of an Ambient Intelligence Landscape. Eventually, customers will be able to act as if they were constantly supported by a team of intelligent assistants, some of whom may indeed still be humans.

# Usability issues in multimodal interaction

If two or more input and output modes are available, there are several different ways to combine them. The simplest way to turn a service multimodal is to use one mode for input and another mode for output. Today there are several companies promoting speech-in, data-out services. After entering such a service the terminal screen shows a form or some other graphical representation like a schematic city map. Customers use automatic speech recognition to enter queries and commands, and the response is shown on the screen. Although speech-in, data-out interaction can in principle be considered as multimodal, we will focus on 'true' multimodal services, i.e. services that combine speech and pen for input, and speech, graphics and text for output. An interface can enforce a kind of interaction protocol, for example: First use the pen to indicate a field and to activate speech recognition, then speak out the information that should go into the field. Or the interface can allow the user full freedom in the way the input modes are used. For obvious technical limitations total freedom on the user side will not be feasible for quite some time to come. Consequently, it is necessary to develop interaction procedures that unobtrusively guide users in such a way that it is always evident what they can and cannot do.

## Different ways of combining pen and speech

Different ways to combine modes correspond to different interaction styles. As always there is a trade-off between 'expressive power', i.e. the freedom of a user to express what she wants to accomplish and complexity of the software needed to enable it. It may seem that total freedom for the user to express herself would be ideal, but maybe it is not. Even human-human interaction is more efficient and more pleasant, if people adhere to culturally accepted behaviour, related to politeness, physical proximity, gestures and eye contact. For human-computer interaction, 'culturally accepted' interaction styles remain to be developed. Several bodies, all related to the World Wide Web Consortium (W3C), are working to develop some kind of standards for multimodal interaction. One such group intends to develop multimodal extensions for VoiceXML. Another group called SALT Forum, with Microsoft as one of the leading partners, is working on standards for multimodal interaction trying to go beyond the inevitable limitations of existing platforms such as VoiceXML.

## Standardised interaction styles

Standards for multimodal interaction serve multiple goals. They make it possible to develop building blocks for multimodal services. But at the same time, and for the moment perhaps even more importantly, these standards help to develop consistent conceptual models of the ways in which speech and pen can be combined to enter information and to control services. So far, most experience has been accumulated with interfaces modelled after the form-filling paradigm. For completing forms, pen and speech can be used sequentially: First the pen is used to 'put the cursor in a field' and then the user can speak out the information to go into that field. The field selection can act as a trigger for the speech recogniser to listen to the audio input. The requirement to trigger the ASR by selecting a field simplifies one of the most difficult parts of the speech recognition process, namely to determine that there is speech to recognise. Command actions in form filling can be accomplished in two ways: Either via clicking a button, such as the 'Submit' button on a form, or via just saying 'submit'. If two input modes are simultaneously available for the same function, it is natural to interpret only the command that arrives first. Provisions must, of course, be made for situations in which conflicting commands arrive simultaneously from different input modes.

In some cases it is natural to convey part of the information in the form of speech, and another part by pointing. This happens, for example, in a conversation with a travel agent, where it is natural to point at some picture in a brochure, while at the same time saying something like 'and what about this?' In these situations neither mode can be interpreted in isolation; the information carried by the modes must be combined to construct a meaningful dialogue act. For people this combination is effortless, but we do not quite know how to accomplish this feat in an automatic system. In the technical literature this form of combining speech and pen inputs is called 'simultaneously and co-ordinated', because the inputs arrive more or less simultaneously, and they must be co-ordinated to understand what the user intended to accomplish.

### Form filling

Today almost all knowledge about the ease of use of the different ways of combining modes is based on laboratory experiments. These experiments suggest that sequential interaction is perceived as quite 'natural' for tasks that can be completed by filling out a form. Users find it easy and natural to learn that they must first indicate a field in the form, and then speak, if only because they have learned to do this in many applications on their desktop PC. One service that is suitable for this type of interaction is Directory Assistance. Figure 3 gives an impression of how the screens in such a service could look like. The left part of the figure shows the form to be filled; the right part shows how the output can be presented.


**Figure 3: Example of tap-and-talk style interaction.**


**Figure 4: Example of a screen that requires co-ordinated simultaneous interaction.**

### Map services

However, laboratory experiments with services that cannot 'naturally' be presented as a form filling task suggest that some applications will need the full complexity of simultaneous co-ordinated interaction. Tourist information is a good example to demonstrate the need for this type of interaction. With a screen as shown in Figure 4, users will find it difficult to learn that they must always first point at an object before they can start asking a question about it.

## Design strategies

Multimodal interaction is only in its infancy. There is little experience with the design of effective multimodal services, especially services intended to be used with small portable terminals. However, laboratory experiments have provided ample evidence that it is not a good idea to try to 'upgrade' existing monomodal services by adding additional interaction modes, simply because a terminal and a network appear to support the combination. Rather, it will always be necessary to design the interaction strategy with the functionality of the service and the requirements of the customer as starting points. Several research labs are now working to develop such user centred design strategies.

# An architecture for multimodal interaction

Though we will not present in-depth technical information here, it is useful to disclose some aspects of the complexity of multimodal interaction. As an example, we use the most complex, the simultaneous co-ordinated mode. Figure 5 gives a sketch of the system architecture needed for a service that allows customers to freely combine several input modes. Because the user may provide part of the information in the form of speech and another parts by using the pen, provisions must be made to combine and merge information of parallel channels. For example, when a user points at the tower on the screen in Figure 4 and at the same time says 'What is this?' the meaning of the word 'this' must be decoded as 'the building at the geographical co-ordinates corresponding to this point on the map' and this must be combined with the remaining part of the information in the input. This would result in the question 'What is the building that is located at the co-ordinates (x, y)?' This combination, or fusion of information items, is accomplished in the module that is labelled 'Fusion' in Figure 5. Subsequently, the question must be converted into some action that is perceived as helpful and meaningful by the user. That is accomplished by the 'dialogue and action management' module, which will consult an application database to find out that the co ordinates (x, y) correspond to the location of the Eiffel Tower. The dialogue and action management module must then decide what information to pass on to the user. It could just provide the name of the building, but it might also decide to add additional information, such as the year when it was completed or its function. The module labelled 'Fission' must decide how to render the information. For example, the name of the building could be displayed in a small text box on the user's screen, while information about the age and the functions of the building might be read using a text-to-speech converter.

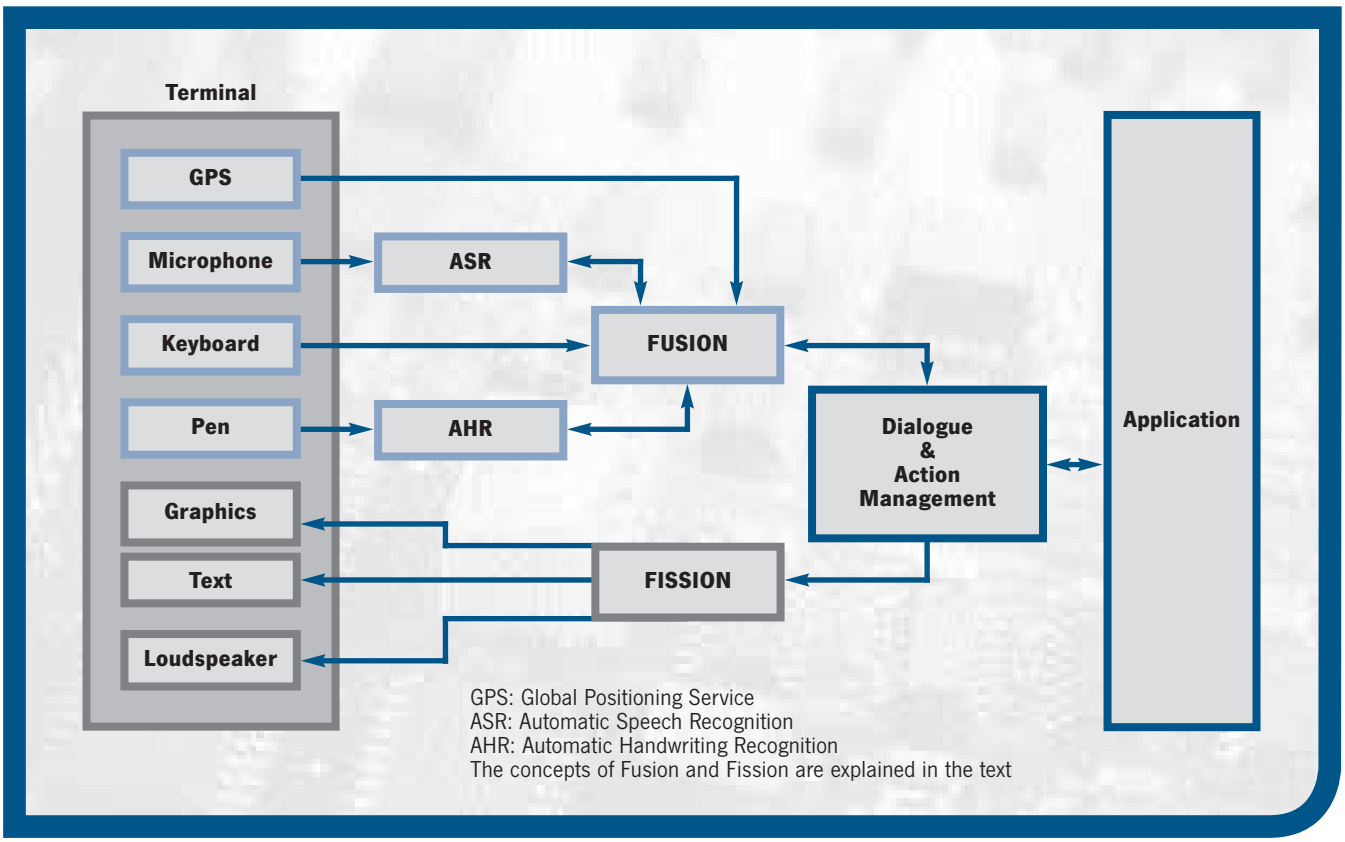



**Figure 5: Schematic diagram of the modular architecture of a multimodal service.**

## Fusion and fission

Today, fusion, fission, and 'dialogue and action management' in simultaneous co-ordinated interaction are areas of active research, rather than well-established engineering disciplines. For example, little is known about the timing relations between speaking and pointing when questions such as 'What is this?' are asked. It is evident that the timing of speech and pointing actions is variable, and that it may differ between subjects. There are lots of other hidden questions, of which even laboratory research caught only glimpses. Fortunately, most of the complexities of multimodal interaction will not come to the surface in applications limited to sequential interaction. Then 'fusion' is reduced to a kind of switch that is fully controlled by the 'dialogue and action management' module. In form filling applications 'fission' will boil down to well-understood procedures for presenting graphical and textual information.

## Standardisation and usability

The single most important challenge for the development of future applications based on simultaneous co-ordinated interaction will be to design standard interaction procedures with a good trade-off between user intuition and technical complexity. We predict that something similar to the present Windows standards will emerge for multimodal interaction styles.

An important question remaining to be answered is whether it is possible to design interaction procedures that combine sequential and simultaneous usage of input modes in one service, without confusing the customers. It is clear that such a combination of styles can only be successful if it is always obvious, which actions a user can and cannot perform. It is also clear that this will require excellent designs of the graphical information on the screen. But we expect that excellent designs per se will not be sufficient. Rather, we must join forces to create standardised interaction styles, which will enable users to apply what they learned in one service on many other similar services.

# Multilinguality

For many services it will be essential that customers can use their mother tongue. Perhaps the single most important reason is that the performance of automatic speech recognition (ASR) degrades substantially for non-native speakers. But the ability to express oneself in the preferred language adds considerably to a satisfying experience.

Fortunately, large parts of the architecture for multimodal interaction sketched in Figure 5 will treat information in a symbolic form, which is largely independent of the language in the interface. For example, the relation between (x, y) co-ordinates and the identity of a building needs very little natural language for its representation. The semantics of questions and commands must also be represented in some formal way, which abstracts from the peculiarities of a specific natural language.

At the output side customers will prefer information to be rendered in their native language, especially if complex instructions are involved. For this purpose, language technologies, such as automatic text generation or machine translation may need to be used. In mobile environments, where the user may not be able to print or to store large text files, automatic summarisation of the information may also come in handy.

# Hardware and protocols

Multimodal services will not just introduce a new type of interaction, but also define a new generation of terminals and of communication protocols. It is only natural that fast moving companies are trying to claim the newly discovered territories by promoting their proprietary technology as some kind of industry standard.

## Terminals

The functionality of future multimodal services will depend to a large extent on the input/output functionality that is implemented in the terminals. All terminals will have audio input and output that can be used for speech including speech recognition and synthesis and music. Terminals will also have a screen that is suitable for displaying video, graphics and text, and that can be used as an additional input device for a special pen, which can be just a point-and-click device, but which may also provide a link to some handwriting facilities. Recently, terminals have been introduced that come with a digital camera that makes it possible to capture and transmit still pictures. Last but not least, terminals will further shrink in size, weight and energy consumption.

It is widely believed that telephones and PDAs will converge into a single multi-functional device, and that this will happen soon. However, it may turn out that a large proportion of the customers will be happy with a device that allows to make telephone conversations and send SMS messages. At the same time business people may find it easier to carry two or three devices that are specialised for specific functions such as making phone calls, or updating the agenda or the order book than a single device that is clumsy for all or some of these functions. The emergence of successful multi-function terminals will be contingent on the success of attempts to develop efficient and effective multimodal interaction procedures.

**Distributed processing**

The multimodal functionality that meets the eye of the customer can be implemented in different ways, depending on the distribution of the computations. Advanced robust automatic speech and handwriting recognition require more CPU cycles, more memory, and may consume more energy than what is available in a lightweight handset. Small vocabulary isolated word recognition and isolated print letter recognition can be accomplished with the compute power of today's GSM handsets. Multimodal services must make a decision about the way in which computation is distributed between a terminal and number crunchers in the network. These decisions will have an impact on the architecture of the services. If speech recognition were handled completely in the terminal, the ASR module in Figure 5 would be located in the terminal. This will definitely affect the flexibility with which services can be updated, as well as the ways in which services are best developed and marketed.

There is another interesting possibility, which will only add to the complexity of the service creation chain: It may be advantageous to distribute the work for computationally intensive tasks such as automatic speech and handwriting recognition between the terminal and number crunchers in the network. For speech recognition, a standard has been proposed that implements most of the initial signal processing in the handset, and that leaves the 'intelligence' part of the recognition to a computer in the network. This standard is known under the name Distributed Speech Recognition, or DSR. Similar standards are expected to emerge for handwriting recognition.

## Protocols

By definition, multimodal services require parallel transmission of several different kinds of data. Speech and perhaps the output of local ASR must be transmitted in parallel to the actions of the pen. In the near future images captured by a digital camera must also be transmitted. At the output side, speech and soon also video, text and graphics must be transmitted to the handset.

The transmission protocols form the link between the hardware functions of the terminal on the one hand and the business logic of the service on the other. It should be clear that there is a tight interaction between the hardware capabilities of the handset and the protocols it has to implement. If a handset does not do local ASR, there is no need to transmit ASR output. However, if there is local ASR functionality, then the protocol definition determines the type of recognition results that can be transmitted. In turn, this determines the information that the business logic in the network will receive from the user. Therefore, for a service developer it will not be enough to know or to influence the functionality of the handset; it will also be necessary to know or to determine the transmission protocols.

# Alliances for multimodal services

Multimodal services are very much white spaces on the map of telecommunications and computer services. To cultivate the newly discovered land, it will be necessary to forge a new type of alliance between network operators, network infrastructure manufacturers, terminal manufacturers, enabling technology suppliers, software companies and content/service providers.

The infrastructure manufacturers should develop networks that are capable of providing the bandwidth and quality of service that is needed for attractive multimodal services. They must be able to support emerging new protocols. The enabling technology providers must develop robust high quality ASR, AHR, TTS and, in the future, gesture recognition technology that can be implemented in the handsets, in the network and in the applications. The terminal manufacturers have to implement the necessary functionality in attractive devices, while the content and service providers will contribute the 'flesh' that should turn the bare bones into an appealing living creature. Due to the lack of knowledge about what will eventually emerge as successful multimodal interaction strategies, it may also be necessary to include a novel type of company in a successful alliance, which brings knowledge and expertise in the integration of enabling technology such as ASR, AHR and TTS, and interface design.

It is not so evident where the 'natural' position of the network operators is in the multimodal service chain. We can imagine at least three different business strategies the operators may choose from. One option is to only provide transmission and billing services for the other players. Alternatively, operators may move towards a position where they combine transmission services and content provisioning. Last but not least, operators might want to strengthen their role as experts in interface and interaction design, perhaps in combination with content and service provisioning. Many network operators have in-house R&D groups that have the knowledge and expertise that is necessary to develop novel interaction procedures.

## Who takes the lead?

It is interesting to see that the alliances alluded to above do not have a natural leader. In fact, we see different companies make attempts to take the lead in different markets. In Japan, NTT DoCoMo, a network operator, took the lead in the development of I-Mode. NTT DoCoMo specified the terminals, the protocols, and the application platforms. The USA and Europe are still awaiting the advent of an appealing multimodal application. In the meantime, large hardware and software companies – Hewlett-Packard and Microsoft, to name just a few – and also small high-tech companies are competing for their share of the emerging market. So far, the European network operators have not been prominently present in the field, though it is widely held that the success of UMTS may very well depend on the degree to which it will be possible to develop appealing services for small, lightweight mobile devices. Yet many of the operators do have the in-house expertise that is necessary to play a prominent and leading role in the development of multimodal services, and to influence the specification of the terminals, the networks and the software needed to implement those services.

## The role of EURESCOM

EURESCOM supports network operators and other companies in the multimodal services arena to establish and secure their position. In doing so, the crucial importance of the usability requirements of multimodal services is fully acknowledged. Projects like MUST are at the crossroads where advanced technology development meets R&D focusing on the users and services, which will attract new customers.

## Further reading

A growing number of papers on multimodal interaction are now appearing in the scientific literature. For a critical appraisal of the present state-of-the-art in the scientific field, as well as for an overview of recent offerings of commercial companies, the reader can consult the web pages of the MUST project P1104 on the EURESCOM Server.

**http://www.eurescom.de/public/projects/p1100-series/p1104**

and read the publicly available Deliverable 1

**http://www.eurescom.de/public/projectresults/p1100-series/p1104-d1.asp**

Recently, several workshops on multimodal interaction have been organised. The web site of the MUST project contains pointers to those workshops and their proceedings.

## List of Acronyms

| | |
|---|---|
| AHR | Automatic Handwriting Recognition |
| ASR | Automatic Speech Recognition |
| CPU | Central Processing Unit |
| DSR | Distributed Speech Recognition |
| IP | Internet Protocol |
| IT | Information Technologies |
| GPRS | General Packet Radio System |
| GPS | Global Positioning System |
| GSM | Global System For Mobile Communication |
| PDA | Personal Digital Assistant |
| SMS | Short Message Service |
| TTS | Text To Speech Conversion |
| UMTS | Universal Mobile Telecommunication System |
| W3C | World Wide Web Consortium |
| XML | Extended Markup Language |

EURESCOM

European Institute for Research
and Strategic Studies
in Telecommunications
Schloss-Wolfsbrunnenweg 35
69118 Heidelberg, Germany
Tel.: +49 6221 989-0
Fax: +49 6221 989 209
E-mail: info@eurescom.de
http://www.eurescom.de

EURESCOM is the leading company for collaborative R&D in telecommunications.

Founded in 1991, EURESCOM provides comprehensive collaborative research management services to network operators, service providers, suppliers and vendors who wish to collaborate on the issues facing the telecommunications industry today and tomorrow.

If you wish to join EURESCOM, please contact us.

ISO 9000
KEMA