

Multimodal Interaction for Mobile Environments

Horst Rössler, Jürgen Siene, Wiesława Wajda, Jens Hoffmann, Michaela Kostrzewa

Private Network Department

Alcatel SEL AG Research and Innovation

70435 Stuttgart Germany

+49 711 821 32293

{Horst.Roessler, Juergen.Siene, Wiesława.Wajda, Jens.Hoffmann, Michaela.Kostrzewa}@alcatel.de

1 Abstract

This paper describes an approach for offering web-based multimodal services in mobile environments. After the definition of requirements a first step for an implementation is presented, that will be used to extend traditional graphical user interfaces with multimodal elements, such as speech and hand-writing recognition. Moreover requirements for a new dialog-centric Multimodal Markup Language (MML) has been proposed, where graphic, speech, handwriting, anthropomorphic avatars, touch-sensitive input and other technologies cooperate together.

1.1 Keywords

Multimodality, Mobile Environment, Distributed Speech Recognition, Synchronisation, XML, SMADA, MAP

2 Introduction

Multimodal interaction has become one of the driving factors for user interface technologies, since it allows to combine the advantages of traditional graphical interfaces that are used in the computer environment with speech driven dialog emerging from the telephony world.

Graphical interfaces allow the parallel presentation of data to the user, which enables a fast access to relevant information by using point and click interfaces. Input from a keyboard could easily interpreted without major system based misinterpretations.

On the other hand in mobile terminals the input devices are rather cumbersome: mobile phones e.g. most often have only 12 keys to input all alphanumerical letters and a limited display size. In this context speech recognition helps the user to overcome the limitation of the other input modalities. Furthermore speech input allows the parallel filling of form fields: A sentence like "I want to fly from Stuttgart to Berlin next Friday morning" will - if correctly recognised and interpreted- fasten the use of such a flight reservation application. Furthermore speech can be used also in situation where eyes and/or hands are busy, e.g. while driving a car.

Unfortunately speech recognition is currently not ready to implement such complex tasks into a mobile device. Here a distributed approach, where only the extraction of relevant features of the speech signal and the transmission to a network based server is executed in the terminal, will be necessary. To standardise such distributed technology ETSI has formed the AURORA project, where mobile phone vendors and speech recognition companies work together to settle an unique and efficient feature extraction algorithm and protocols, that will be used to distribute a speech recognition system over a network.

Especially for internet based applications the concepts of dialog presentation for graphical and vocal interfaces requires a new approach to combine interface description languages like HTML and VoiceXML. Speech and GUI may run parallel or allow supplementary operations. This needs dedicated synchronisation mechanisms between the different modalities. Furthermore a dedicated context and situation management can help to strengthen the input/output modality that fits best for the current situation.

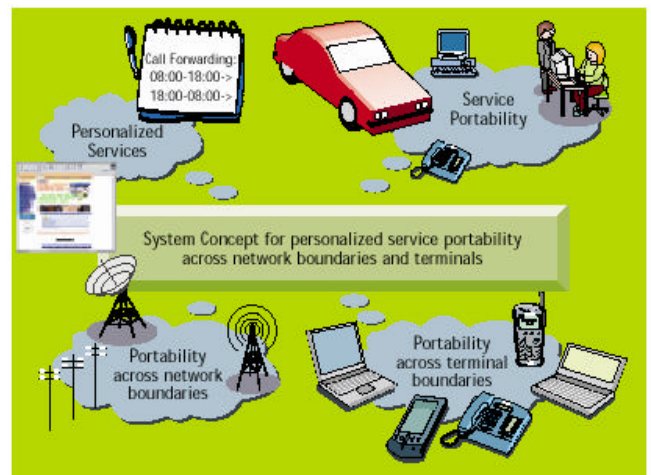


Fig. 1: Requirements for Mobility Services

Alcatel SEL is engaged in 2 public funded projects that will show multimodal interaction. While the IST project SMADA (Speech-driven Multimodal Automatic Directory Assistance) focuses on the enabling technologies for a

multimodal application in a mobile environment, the German research project MAP (Multimedia work place of the future) works on a generic description for multimodal user assistance to support secure delegation mechanisms on a mobile agent platform.

3 Mobile Environments

Frequently, communication is used to obtain “tailored” information rapidly and wherever the user is located. All this alters the user’s need for services, which is user-friendly ways to access them over different communication networks, efficient methods for managing and subscribing to services, or the way in which information is delivered. Communication networks are changing dramatically to support these needs. Following on from the classical networks carrying voice and low speed data, pure IP based data networks are being deployed to meet the need for high speed data transport.

In this sense, mobility might be seen in different circumstances: the mobility of the device, the mobility to access a service from different networks, the mobility of the service itself and the personalization of a given service.

Therefore the goal future advanced communication and information services should

- Provide a personal environment that “travels” with the user.
- Provide overall (especially mobile) access to the information relevant for the users actual tasks, by Internet, Intranets, etc..

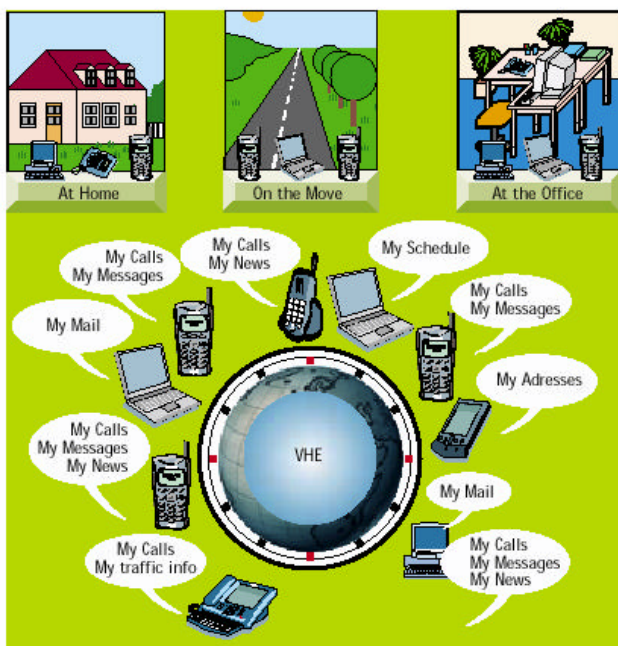


Fig. 2: Mobility using Virtual Environments

- Provide data in a format that is best suited to the user’s needs at any time.
- Provide an user interface according to the user’s profile, user’s location and situation, accessed network and the current context.

In order to present the information to the user in the best way, a module has to be included, that is aware of the user’s preferences (doesn't like to use speech recognition), the device capabilities (no graphic display) and the current situation of the user (sitting in a meeting, please do a silent alert). The MAP project is focussing on this task by implementing a context manager, which has access to the users device and location (using e.g. GPS) and offering interfaces to a dialog manager, that adapts a virtual description of the user interface to the actual needs of the user.

4 Multimodal Requirements

There are different possibilities, how multimodal applications could be implemented. The most simple approach allows only one modality at a time. Such sequential multimodal input does not need any synchronisation between the different modalities.

The next step allows different modalities at the same time, but needs an active synchronisation by the user, like the tap 'n talk interface that is used in Microsoft's MIPAD [4]. Such kind of systems may be used for form-filling in services like travel-reservation on a mobile device like PDAs. For both types of interaction the user is requested to use the touch interface, even if he wants to use only speech input, due to his preferences or the situation, in which the user is (e.g. while driving a car).

The need for multimodal interaction, where each modality can be used seamlessly, arises. In such systems a multimodal browser supervises the sequence of dialog moves that are necessary to complete a specific task. Such a browser will also integrate rendering modules for each modality envisaged in an specific environment and allow the synchronisation between them. In such environments multimodal mixed-initiative dialogues become possible.

The highest level of multimodal interaction is required, if there is a need for semantic synchronisation between the modalities. For example to point at an item on the screen and ask - using speech recognition - "Give me more information about <this> object". Here, very complex time alignments between the different input modalities are required.

In order to allow application developers to create applications independent from the target end user device, an XML based description language offers a possible solution. With such a language the application can be described by defining dialog steps, which are needed to

interact with a server. For each addressed modality the semantic of the interaction has to be defined, e.g. what are the required values that can be filled in each field of a form which short-cuts and commands could be used in the current document.

This multimodal mark-up language should allow extensibility for new input and output modalities.

Furthermore the logic of the dialog has to be described, i.e. in which ways the different modalities can be synchronised.

This description language should be independent of the implementation of the presentation on a specific device.

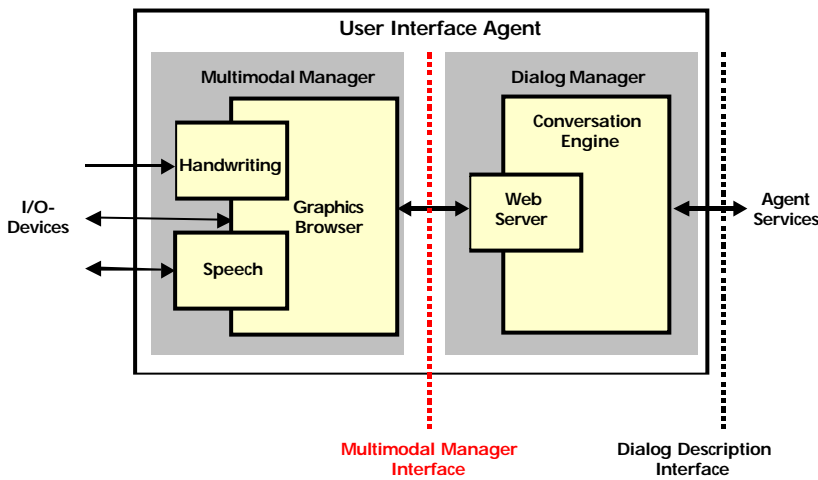


Fig 3: Architecture of Multimodal Demonstrators

Chapter 6 will address a first approach to combine graphic and speech based elements in a single language.

5 Demonstration Systems

Alcatel has evaluated two technologies to integrate different modalities into a browser environment, and built prototypes to show the functionality of each approach.

The first system integrates in addition to a standard graphic interface speech synthesis and handwriting recognition as applets on a standard browser platform. The system can run on any computer with a touch sensitive display, audio-interface and a connection to the Internet.

Processing of multimodal elements such as speech recognition, speech synthesis and handwriting recognition needs a lot of computing power if high recognition rates are mandatory. Therefore a client/server architecture has been realised. The input is pre-processed on the terminal, a reduced feature set is transmitted over network connections and finally processed on the server.

The system provides several independent servers such as a handwriting recognition server, a text-to-speech server and in the near future a speech recognition server. The

Java front-end performs processing and synchronisation of multimodal inputs, data pre-processing and the communication protocol between front-end (client) and server.

The applet provides specific procedures for different input modalities. The applet procedures may be invoked from any HTML-tag and/or JavaScript-tag.

The second system implements a simple multimodal browser, by extending the Microsoft Internet Explorer web-browser control with the integration of speech recognition, TTS and handwriting.

For controlling the different modalities, standard tags of the HTML document can be additionally analysed and will initiate the processes like speech recognition or TTS. While TTS and hand-writing work on any webpage, speech recognition needs a specific document, that contains the actual valid grammar. This will be loaded from the web-server and fed into the speech recognition engine. The synchronisation is implemented by an external control module which captures the results of the different modalities and interacts with the actual document accordingly.

At the current step the tap 'n talk like synchronisation method is implemented in this system, while a dialog based approach for more speech centric users will be available soon.



Fig. 4: Java Based Multimodal Demonstrator

This architecture could easily be extended to integrate new synchronisation principles and will help in user tests to clarify demands that arise for the development of a new multimodal mark-up language.

Results of this evaluation will be used to implement a multimodal demonstration system for SMADA, where also distributed speech recognition will be integrated.



Fig. 5: Multimodal Browser

6 The Multimodal Mark-up Language (MML)

Although we have described a way to introduce simple multimodality, a more generic approach is needed to fulfil the above requirements to allow seamless multimodality over different devices in a mobile environment.

Therefore a new description language for dialogs is currently developed, where graphic, speech, handwriting and anthropomorphic avatars co-operate together.

The new multimodal language combines advantages of HTML and VoiceXML considering that graphic and speech, will be the most important modalities to be combined.

HTML offers authors the possibility to publish documents on graphical devices. It proposes possibilities to structure documents in sections or to establish direct relationship with other documents by linking them together. Interactive dialog elements like text edits, check or select fields and buttons are defined.

Furthermore, it is possible to include other pictures sound-clips and other elements in the document. This often needs additional modules to be integrated in the browser in order to “view” them.

VoiceXML is designed as a description language for voice based dialogs, that could be accessed over telephone networks. It offers synthesised speech and digitised audio data as output modalities and speech recognition as well as DTMF key input as input.

Using VoiceXML allows application developers to bring the advantages of web-based development and content delivery to interactive voice response systems, without the need to know the used speech recognition engine.

Next, VoiceXML defines different possibilities, do handle errors like “no speech input detected” or “rejections” (the system has recognised a word, but is unsure which one).

Finally, VoiceXML is able to handle the filling of different fields with one utterance, a feature that is needed in mixed-initiative dialogs. In multimodal systems mixed-initiative could be even more favourable, than in speech only applications, since the user can “see” if all information has been recognised correctly.

The multimodal mark-up language is an XML language. This allows a clear separation in content, structure/ format and interaction. Furthermore, a fourth element has to be included – the cross modality synchronisation.

The multimodal mark-up language will consider the way, how speech output could be annotated: to vary the speaker and/or the voice will bring a kind for formatting to the content presented through the vocal interface. A text that will be presented with a bold font on the graphical device, might have higher volume in the audio channel.

In some cases the same presentation can be used in different modalities – like text, that can be shown on a display or read by a TTS system. For other types like figures or pictures the representation might differ between the modality or the device.

A synchronisation mechanism is needed to combine different output, e.g. an speaking avatar that is presenting information on a graphic display, has to be synchronised with the output of the TTS system.

In the same way the recognised phrase has to be split to the fields in the graphical display.

To fulfil the above requirements extra attributes are used to provide additional information for dialog elements:

- **modalityOut** specifies the modalities for the element and how the different modalities will present the element,
- **modalityIn** specifies the use of the input modalities and there combination, including grammars for speech recognition or hand-writing.
- **production** controls the format of output elements e.g. for synthesised speech, the emphasis of words (stressing or accenting), pauses, volume, pitch and speaking rate,
- **timing** defines the synchronisation information e.g. text spoken by the computer to extend given textual information or to replace graphical information, e.g. picture description,

- **bargeIn** which enables or disables interruption of voice announcements,
- **initiative** specifies the control flow, e.g. the sequence in which the vocal part of the dialog interaction is presented and the error recovery .

Currently, we are developing a language environment for MML, which will be a bases for device and media independent applications.

7 Conclusions

A number of techniques and theories exists in the field of multimodal architectures, but yet there are no commercially available system. The need for clearly defined architectures that support building multimodal applications which can be used on different types of terminals, require the development of common standards and solutions for synchronisation mechanisms. The proposed architecture may overcome some of the issues and might be helpful for further investigation on user interaction technologies.

8 Acknowledgements

This work was carried out within the European research project SMADA, which is supported by the European Commission, under the Action Line Human Language Technology in the 5th Framework IST Program and the German lead project MAP, supported by the German Ministry of Economics and Technology.

REFERENCES

1. Oviatt, S.L., DeAngeli, A. and Kuhn, K. Integration and synchronization of input modes during multimodal human-computer interaction. In Proceedings of Conference on Human Factors in Computing Systems CHI '97 (March 22-27, Atlanta, GA). ACM Press, New York, 1997, pp. 415-422.
2. Vildan Bilici, Emiel Kraemer, Saskia te Riele and Raymond Veldhuis Ipo, Preferred Modalities In Dialogue Systems
3. Oviatt, S. and Cohen, P., Multimodal interfaces that process what comes naturally, Communications of the ACM, 43(3):45-53, 2000.
4. X. Huang et al., "MiPad: A Next Generation PDA Prototype", ICSLP, Beijing, China 2000
5. Wahlster W., Reithinger N., Blocher A., "SmartKom: Multimodal Communication with a Life-Like Character", Eurospeech Aalborg, Denmark, 2001-09-13
6. W3C, "Multimodal requirements for voice markup languages W3C working draft 10 july 2000", <http://www.w3.org/TR/multimodal-reqs>, 2000
7. Peters, M., Rombaut, P. "Mobile or mobility? Evolution of mobility services ", in Alcatel Telecom Review 2/2000
8. www.map21.de
9. smada.research.kpn.com