Multimodal Interaction for Next Generation Networks

Jürgen Sienel, Dieter Kopp, Horst Rössler,
Private Network Department
Alcatel SEL AG Research and Innovation
70435 Stuttgart Germany
+49 711 821 32293
{ Juergen.Sienel, Dieter.Kopp, Horst.Roessler }@alcatel.de

1 Abstract

This paper highlights the advantages of advanced interaction capabilities like multimodal services as the essential components for the future fixed and mobile networks. But telecom operators due to their need to look for new models generating revenues from data services are interested in convincing applications, which need more but only enhanced interface capabilities in end-user devices.

Due to emerging mobile Internet capable terminals a large number of new and appealing services will be available. The small size of these terminals introduces usability problems for these services, and, therefore new interaction modes need to be deployed. One of these is speech centric multimodal interaction. Speech recognition, pen and text are used at the input side and speech synthesis, graphics, text and moving pictures are used at the output side. Providing end users with multi-modal access to availability and location of other end users improves flexibility and efficiency of human-machine communication, and support the user in person-to-person communication

Concepts integrating presence and context information with multimodal interaction may influence the use of such services. In order to achieve this, the network has to provide basic interaction functions and dialog control, wherever the device itself is not capable to process all the information or several devices are used in parallel.

1.1 Keywords

Multimodality, Mobile Environment, Distributed Architectures.

2 Introduction

Telecommunication business faces currently a move from just selling connection time to enhancing the network by offering new exciting services, where both public and private networks are concerned. The customers will see both the convergence of mobile and fixed network applications and services, but also the integration of services known from the internet into telephony networks. Furthermore end-users are able to connect different types of devices to the networks ranging from desktop PCs, laptops, PDAs and all sets of mobile and fixed network phones and new set of devices become connected like game consoles, television sets and telematic devices running in a car or a truck. All these devices have different capabilities in presenting information to the user and interacting with services. Especially in mobile environments the interaction component is essential for using services at any time and anywhere, since the interaction possibilities depend not only on the device, but also on the situation of the user.

Multimodal interaction i.e. the combination of speech, graphics, pen and other 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 computer environments with speech driven dialogs emerging from the telephony world.

Especially for internet based applications the concepts of dialog presentation for graphical and vocal interfaces require a new approach to combine interface description languages like Hyper Text Mark-up Language (HTML) and VoiceXML. Speech and Graphical User Interface (GUI) may run in parallel or allow supplementary operations. This needs dedicated synchronisation mechanisms between the different modalities as proposed by an interaction manager. Furthermore a dedicated context and capabilities management can help to strengthen the input/output modality that fits best for the current situation.

Multimodal interaction has significant advantages:

- user can select at any time the preferred modality of interaction;
- can be extended to selection of the preferred device (multi-device);
- user is not tied to a particular channel's presentation flow;
- improves human-machine interaction by supporting selection of supplementary operations;
- interaction becomes a personal and optimised experience;

- multimodal output is an example of multi-media where the different modalities are closely synchronized.

3 Future Multimodal Networks

The increasing demand coming from large network operators in converging their different infrastructure for fixed and mobile communication networks, results in a couple of interesting questions that have to be resolved in near to midterm future. This consists in the use of different terminals and access modes for communication and access to information services. In fact, operators are interested in deploying services available in mobile networks like Short Message Service (SMS) to their fixed networks customers. Similar services will be introduced from a computer based environment like management of large directories. All this will introduce a new way of communication, since a customer does not need anymore to remember a large set of numbers (fixed, cell-phone, email, etc.) but can contact the partner by using the name, letting the network decide which is the appropriate method for establishing the communication based on the context and presence information held in the network. With respect to information services this requires the possibility to adapt the content to the capabilities of the terminal that is used in the current situation or even more the use of several devices interacting for a service, like the presentation of a video on a television set, while the mobile phone is used to as voice channel and as remote control.

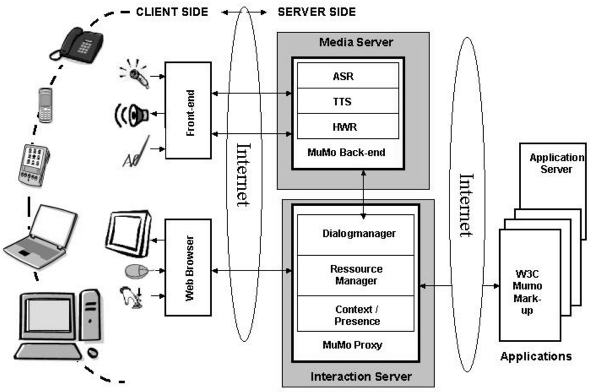

Figure 1 – Multimodal Architecture

The overwhelming number of information presented today to individuals will continue to increase, from which the largest part is irrelevant in the actual situation to the user. Currently 50% of received emails contain spam, at Google you have to visit several links before you have a relevant one for the information you look for.

The need currently arises to personalise the information that arrives and provide a virtual assistance function knowing about relevant context information to guide the user through the information space. Thus requires the functions that enhance the interaction between the user and the service, using speech and graphical information enhanced with multimedia content where possible, desired and cost effective. The need of application servers, media processing units, intelligent switching and routing and presence & context servers will be the coming push in enhancing the network technology of communication providers.

Alcatel Research & Innovation is active on the implementation of a distributed multimodal demonstration system for mobile networks, which allows the use of speech recogniser and TTS in the network by accessing from different en-user equipment like PDA, PC or cell phone. Figure 1 shows the architecture. The idea is that the interaction server acts like a proxy between the graphical browser, the voice resources and the application server on which the multimodal application resides. It is responsible for distribution of data to the handling resources, the activation and synchronisation of the speech resources and the presentation based on the context information and the device capabilities. It integrates the dialog manager for keeping track of the current dialog steps and the integration of the receiving events from the media processors. The media server will process the data received from the terminal like ink, or speech and realises the speech output component which sends the data to the end-user device e.g. by RTP. On the client a standard web-browser running applets or java-script is used to provide the synchronisation mechanisms between the interaction server and the client. Additional a media client has been implemented to connect the audio devices to the RTP channel at the media server.

4 Instant Messenger as multimodal application prototype

Based on the architecture described above, an Instant Messaging application (IM) has been implemented comprising three different interaction modes: pure graphic, voice and mixed multimodal communication.

The different modes should show the adaptability of the interaction layer to different output modes and to adapt to the user preferences in specific situations. People prefer the mode “graphic” in public areas, e.g. train or airport, whereas “voice” may be used in car.

The figure 2 shows essential functions of the IM application. Basically the IM provides functionality to the user which enables him to send messages to his buddies. Whenever the status of the buddies, e.g. being online or offline, has been changed or a new message from those buddies has been sent this status has to be updated (shown) on the client side.

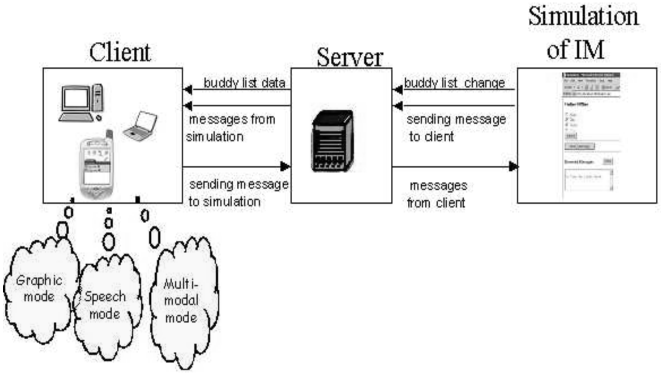

Figure 2 – IM Architecture

The system begins with a start-page for welcoming the user which says: “Welcome to Instant Messenger!”. Dialogue is spoken and written underlined as a link.

Afterwards the user can use the link and gets to a multimodal page where he can choose between the three modes “graphic”, “multimodality” and “speech”. Because this is a multimodal page the user can select his mode by clicking on the appropriate button or saying the mode and gets so to the menu of the chosen mode. Some screenshots of the multimodal instant messenger are shown in figure 3.

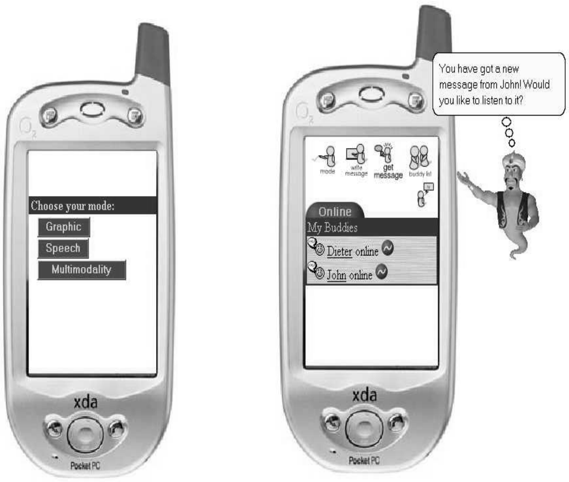

Figure 3 - Screenshots from IM Prototype

5 Multimodal Standardisation Requirements

Alcatel’s position on important standards is mostly covered in the W3C Multimodal Interaction Activity (MMI) The group is extending the web voice and graphical user interface to allow multiple modes of interaction, offering users choosing their voice, or an input device such as a key pad, keyboard, mouse, and stylus. For output, users will be able to listen to spoken prompts and audio, and to view information on graphical displays. The Working Group is developing markup specifications for synchronization across multiple modalities and devices with a wide range of capabilities. With respect to the architecture we are promoting in this paper all major topics are tackled by this initiative. From our perspective most important is the definition of the modality interface component and the work of the system and environment subgroup. Although EMMA provides an very powerful method to promote events throughout different components of the framework, synchronization issues have to be carefully looked at, especially if in future access from several devices in parallel should become possible. Lastly, from realization point it has to be considered , that a lightweight protocol, which can be realized rather fast, may be with restriction s in terms of flexibility, is helpful to deploy first services in a short time frame. There performance issues have to be considered, to open the new multi-modal services to wide range of customers.

6 Conclusions

We presented in this paper a view to a distributed multimodal architecture based on a proxy concept, that is able present web based application with a speech driven multimodal user interface, which could be adapted to the preferences of the user. As a sample application we realized a demonstration of a multimodal instant messaging application running on a PDA like device.

In future work we are considering the use of multiple devices and an increased use of capability and context information which is currently used for adaptation of the presentation. The work of the W3C MMI working group provides meaningful input for continuation of our work.

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. Lou Boves & Els den Os, Multimodal Services – a MUST for UMTS, EURESCOM, http://www.eurescom.de/~pub/deliverables/documents/P1100-series/P1104/p1104-d1.pdf
3. Niklfeld G., Pucher M., Finan R., Eckhart W. 2002 Mobile multi-modal data services for GPRS phones and beyond. http://userver.ftw.at/~niklfeld/pub/niklfeld_icmi02.pdf
4. W3C. Multimodal requirements for voice markup languages W3C working draft 10 July 2000;http://www.w3.org/TR/multimodal-reqs, 2000.
5. W3C Modality Component to Host Environment DOM Requirements and Capabilities Assessment, Public Note, 10 May 2004