Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is an author's version which may differ from the publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/75072

Please be advised that this information was generated on 2020-10-29 and may be subject to change.

Towards Ambient Intelligence: Multimodal Computers that Understand Our Intentions

Els DEN OS[1], Lou BOVES[2]
[1]*Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525XD Nijmegen, Netherlands*
Tel: +31 24 3521333, Fax: +31 24 3521213, Email: Els.denOs@mpi.nl
[2]*Dept. Of Language & Speech, Univ. of Nijmegen, Erasmusplein 1 6525HT Nijmegen, NL*
Tel: +31 24 3612902, Fax: +31 24 3612907, Email: L.Boves@let.kun.nl

Abstract: This paper discusses the state-of-the-art and future developments of conversational multimodal interaction with computers, one of the corner stones of all future eCommerce and eWork applications. On the basis of past and present experience it is argued that natural, conversational interaction can only be achieved on the basis of powerful cognitive models of multimodal human-machine interaction. The paper explains how these models can be developed, how they can be used to make the design of multimodal applications more cost-effective, and how they facilitate non-experts in using complex applications. In addition the paper also identifies a number of threats that must be recognized by system and service developers. Throughout the paper we use a prototype system for the design of bathrooms as an example to illustrate the major points. We also discuss how the principles of founding multimodal interaction on sound cognitive models can be generalized to different application domains.

1. Introduction

On a rainy Saturday afternoon a tired family enters the fourth bathroom shop. In the preceding weeks they have studied brochures and catalogues of a large number of sanitary and tiles manufactures, and in the three shops they had visited earlier that day they have seen samples of the tiles and sanitary displayed in mock-up spaces that looked as remote to the bathroom in their home as the flashy pictures in the catalogues. They still have no idea how things could fit nicely in their own bathroom. In the fourth shop, which they enter without much hope to leave it wiser than they entered, they find a computer that displays an artificial but friendly person who invites them to tell it the shape and dimension of the bathroom in their home, and promises to show designs of their choice displayed exactly in the way they will look in their home. The layout of the bathroom can be sketched with a pen and a tablet, and during the process the avatar asks questions for clarification where it does not immediately understand the drawing and the spoken comments of the family members. It takes no more than a couple of minutes to enter all relevant data. Then the computer asks what style of tiles and sanitary ware the family prefers, and to help them make up their mind it shows examples, already adapted to the size and shape of the family's bathroom. The family members can ask the computer to add, remove or shift sanitary ware, to replace items with similar ones from other brands, series or styles, and change the choice and layout of the tiles. They can speak and point to items on the screen. It is not necessary to know technical terms; saying something like 'this thing', while pointing at some item on the screen is sufficient. If the computer is not sure what the family members mean, it will ask for clarification. Moreover, for each design shown on the screen it volunteers to explain why it made this choice of sanitary and tiles. It takes the family less than 15 minutes to settle for a design that fits their taste as well as their budget. The design is printed and

stored on disk, so that it can be used to discuss the details of the choice with an expert salesperson in the shop.

This scenario is just one example of how intelligent multimodal user interfaces can make the difference between window shopping in frustration and a customer who has been helped to find his way and ends up buying. With the ever-growing multitude of on-line databases there is no lack of data for customers anymore; if anything, there is overkill. However, it is by now common knowledge that raw data is not enough. What customers need to be satisfied is not data, but rather information, accessed in a way that is intuitive for the user, and presented in a way that is easy to grasp and understand. Intelligent conversational multimodal interfaces promise to offer excellent opportunities to accomplish this. However, creating such interfaces is by no means a well-understood engineering exercise.

In this paper we will give an appraisal of the state-of-the-art of multimodal interaction for *e*Work and *e*Business applications. To that end we will introduce the FP5 project COMIC, that is working at the forefront of multimodal interaction in those fields, preparing the ground for the much-touted Ambient Intelligence landscape. We will explain the role of cognitive models for building (and using) conversational multimodal interfaces. Last but not least, we will discuss the hurdles that still must be overcome before we will see large-scale deployment of such interfaces.

2. The COMIC Project

COMIC [1] is an FP5 project in Key Action 2, in the area of Long Term, High Risk Research. The project is coordinated by the Max Planck Institute of Psycholinguistics in Nijmegen; the other partners are the Max Planck Institute of Biocybernetics in Tübingen, DFKI in Saarbrücken, the Department of Computer Science of Sheffield University, the Human Communication Research Centre in Edinburgh (HCRC), the Nijmegen Institute for Cognition, Information and Neuroscience (NICI) in the Netherlands, and ViSoft GmbH in Sindelfingen (Germany). Except for ViSoft all partners have their core business in fundamental and applied research in Cognitive Science and Artificial Intelligence. The COMIC project addresses all issues involved in the kind of natural multimodal interfaces that are sketched in the introduction. Specifically, COMIC investigates the cognitive modelling that is needed for this type of applications. To test the applicability of those models, the project develops an operational demonstrator of a system that can give pre-sales consultancy for bathroom remodelling to non-expert customers.

COMIC combines software and system development with experiments in human-human and human-computer interaction in language-centric multimodal environments. The experiments are based on two types of scenario that can be tightly controlled, but that at the same time are relevant for *e*Commerce and *e*Work applications. The first scenario is the bathroom design application mentioned above. The other scenario is specifically designed to investigate human-human interaction in collaborative problem solving. Experiments focus on the role of eye contact, gaze and body posture, in combination with drawing with a pen and speaking when two persons work together to find the best solution for a common problem.

The bathroom design application has speech and pen input recognition at the input side. The user is able to make sketches on a pen tablet, where possible designs will also be shown in 2D and 3D representations (cf. Figures 1 and 2). In addition, users will be able to point at objects on the screen, such as bathtubs, basins, faucets, etc., and ask the system to shown alternative designs. The system will explain advantages and disadvantages of specific designs. In doing so, it will attempt to take into account a dynamically evolving model of the preferences, likes and dislikes of the user. In addition to the tablet screen, where designs and drawings can be shown, the system will also feature a second screen that

displays a highly realistic talking head. To enhance the naturalness of the interaction this 'avatar' will be able to express the moods and attitudes that a customer expects from an expert sales consultant (except for the fact that the automatic system will always stay polite and will never show irritation). A schematic image of the eventual layout of the application during the phase when the shape and dimensions of the room is being entered is shown in Figure 1. The avatar guides the user through the application by explaining what it is expecting and by asking questions if the input is ambiguous. The user can simultaneously draw or write and speak.

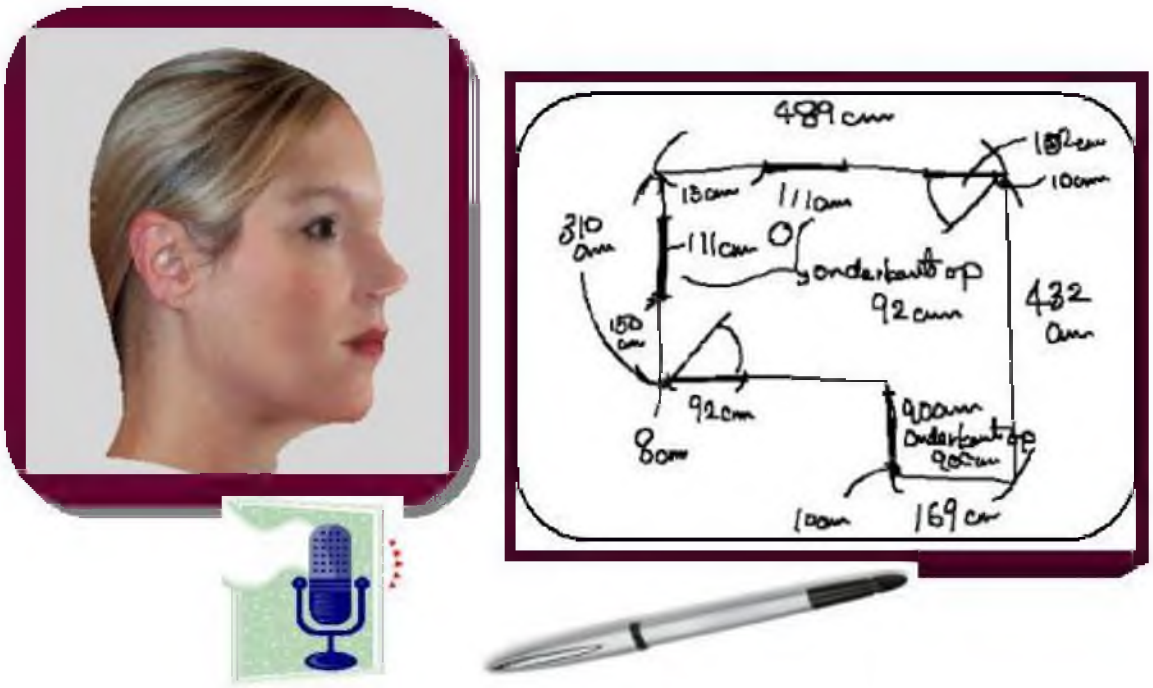

Figure 1. Overview of the bathroom design application. Shown is the part of the dialog where the user enters the size and dimension of the room.

After the ground plan of the room is entered, it can be decorated with tiles and sanitary equipment. Subsequently, the user can move through a 3D image of the design, and discuss possible changes. Figure 2 gives an impression of the image of a decorated room.

The choice of the bathroom design application in COMIC is, of course, inspired by ViSoft. However, it is easy to envisage many other applications of multimodal interaction with speech, pen, text and graphics in completely unrelated domains and areas. One domain that has received much attention from scientists working on multimodal interaction is a tourist support service (cf. SmartKom [4], and MUST [2]). These applications provide access to rich and partly unstructured sources of information about what a visitor (or perhaps also a group of visitors) can do in a city or region. The applications are designed to be really helpful, like a professional agent in a tourist information office, who can do much more than just answering factual questions asked by tourists, if only because one cannot easily ask questions about things one does not know that they exist. Such a service is likely to generate substantial extra business. Applications in which city maps play a pivotal role have also been at the centre of the research of Oviatt and her colleagues [5]. Simpler forms of multimodal tourist services, which use off-the-shelf PDAs connected through GPRS networks for driving directions and finding restaurants near the location of the user, are already on the verge of commercial deployment [6]. Somewhat related services guide users through a museum, providing in-depth information about specific objects, depending on the preferences and interests of the user.

Figure 2. Display of a proposal for a bathroom decoration. All items shown can be discussed and changed interactively.

In the remainder of this paper we will discuss the cognitive research and technology development that are needed to build applications such as the multimodal supportive bathroom design tool and other services that provide help to users who are not in the position to ask very specific questions. We will also elaborate on some of the logistics issues that must be taken into account when developing and deploying *e*Work and *e*Commerce devices. Last but not least, we will touch upon the issue of what might be suitable business models for *e*Work and *e*Commerce services.

3. Cognitive Research

Human-human communication is so efficient by virtue of the fact that both participants bring an enormous amount of overt and tacit knowledge to the task. In a typical conversation few, if any, disturbing speech recognition and understanding errors occur. The participants in the dialog seem to know exactly when they can speak and when they must listen. People can effortlessly and accurately tell the difference between supporting back-channel noises, such as 'yes, I see', or 'uh uh', which are meant to encourage their interlocutor to continue talking, from highly similar noises that are intended to interrupt the speaker. Also, humans can effortlessly combine speech and pen to convey information, and at the same time they interpret each other's body language.

Although that may perhaps seem somewhat surprising at first sight, we know very little about the perceptual and cognitive processes that humans use to make conversations so efficient. Fortunately, the technical and computational infrastructure needed to conduct in-depth analyses of face-to-face interaction is now emerging. Communication scientists have started designing taxonomies of relevant communicative act and gestures that must be detected and annotated to support such analyses. From the little research that has been done so far it appears that many gestures, body movements, gazes and glances can only be understood correctly in conjunction with each other and especially in conjunction with the contents of the speech utterances (or the words and numbers written with the pen). To

investigate the crucial interdependencies between the parallel communication channels, COMIC has designed communication scenarios in which channels can be switched off selectively and asymmetrically: to simulate a situation in which the computer is not able to understand body language and eye contact, we can prevent one of the dialog partners from seeing the body of the other (while the second subject can see the first). This allows us to investigate whether and how it becomes more difficult for the 'impaired' subject to understand the intentions of the other subject. By doing so, we can build models of the minimum amount of information that is needed for efficient and pleasant interaction.

In face-to-face interaction people can transmit information in a seemingly redundant manner. In the bathroom design application, one can provide data about sizes and dimensions in speech or writing. However, it is not known under what conditions people provide such information in several parallel channels, and how this redundancy makes the communication more robust and efficient. It may well be that people have personal preferences for specific combinations of information channels. It is widely assumed that some persons have a preference for verbal and audio information, while others are more inclined to use visual information. However, if non-verbal behaviour can only be interpreted if we can combine several channels, such as gaze, head movement and body posture, the information in the parallel channels is only seemingly redundant.

3.1 Developing design guidelines

Understanding the usage of parallel information channels in human-human communication is necessary for developing models that predict the efficiency of human-machine communication as a function of the degree to which the machine is able to understand the behaviour of the user. These models can then be converted into concrete guidelines for the design of multimodal applications, which take the cognitive capabilities of the human users into account, as well as the limitations in the capability of the machine to recognise and understand the intentions of a user. Due to a lack of understanding of the 'natural' use of parallel modalities when users perform their daily professional and entertainment tasks, virtually all multimodal systems have been developed with a trial-and-error approach. Trial-and-error does not guarantee that a close-to-optimal result will be reached. Moreover, it is extremely expensive, especially in the case of multimodal interaction, because of the very large number of possible combinations of interaction strategies that can be followed, even in cases where the number of input/output devices is relatively small. Therefore, design guidelines based on sound cognitive models will make the development of future multimodal applications much more cost-effective. The design process will converge after a lower number of iterations, and the result will be systems that are easier and more pleasant to use.

An issue that deserves special attention in this context is the extent to which a computer application should behave like a human being. It is well known that many people attribute personalities to computer systems. However, it is not known under which conditions a highly natural humanlike avatar will enhance the experience users have in interaction with supportive computer applications. Nor is it known exactly how natural such an avatar should look and behave to obtain the effect. In COMIC we investigate which parts of a synthetic face must move in order for mild mood changes of the avatar to be perceivable and believable for the users. In this way we can draw up specifications for the design of future multimodal applications that use avatars to make a service more appealing.

4. Issues for the future of multimodal services

4.1 Standardisation Issues

At present, the number of operational multimodal services is still very small. In the previous section we have argued that this type of applications is very difficult and expensive to design. However, that is not the only explanation for the slow take up of an interaction strategy that seems extremely promising. In a previous R&D project in the domain of telecommunication services [2] we have seen that the absence of reliable wide coverage mobile IP networks adds significantly to the problem. Other problems that interfere with the development and deployment of multimodal services include the lack of commonly agreed multimodal communication protocols, as well as the lack of standard terminals that support those protocols [3].

Another issue that complicates the commercial deployment of multimodal services is the lack of standards for the specification and development of those services. However, several attempts are being made to develop such standards, mostly in the context of the WWW consortium, such as the Multimodal Interaction Working Group [7] and SALT Forum [8].

4.2 Technology issues

Until now, few examples of multimodal interaction systems have left academic (and industrial) research labs. Typically, the processing in these advanced demonstrators is so heavy that long delays are incurred between a user action and a system response. However, COMIC aims at latency-free processing, to allow for uninhibited multimodal interaction. To reach that goal we have carefully designed the system architecture, as well as the specifications of all individual operating modules, avoiding unnecessary complexity. The first implementation of the complete architecture (shown in Figure 3), obtained one year after the start of the project, did indeed show virtually latency-free operation.

Several technology issues have emerged as potential bottlenecks in previous experiments in multimodal interaction. The first is the functionality and the performance of the speech recogniser. It has appeared that little remains from the hypothetical advantages of speech-based multimodal interaction if the recogniser makes too many errors. It is not exactly known what 'too many' means, but the proportion of errors must be very small. This is especially difficult since natural interaction implies that users can speak whenever they think that this is appropriate from a communication point of view. In practice this means that the system must listen continuously, even when it is speaking itself. This constitutes a major challenge for the speech recogniser in such systems. In COMIC we intend to provide a recognition scheme that is able to deal with a continuously open microphone. At the same time, that recognition scheme must be able to forward its output to the subsequent modules as soon as a meaningful utterance has been processed, even if the user does not stop speaking. This is necessary to obtain latency-free operation.

The second major issue is related to pen input processing. In an application such as the bathroom design system the pen can be used for at least three different purposes: for pointing (as a substitute for a mouse), for handwriting (for example dimensions of a wall), and for drawing. Natural interaction requires that the system can automatically detect the mode in which the pen is being used. This is very different from typical drawing and painting software, where one must always select a mode from a palette before entering some object.

COMIC intends to advance the state-of-the-art in processing speech and pen input on the level of the linguistic content of the messages as well as on the paralinguistic level. Both speech and pen input contain –often involuntary- details that can be used to estimate

the moods of the user. Specifically, we intend to use this type of information to estimate how certain the user is about specific statements. Another aspect where COMIC will develop advanced technology is Dialogue Management (DM). The intelligence of the DM must be able to support problem solving, even in situations where the user is not certain about the full functionality of the service, nor of the details of the problem under discussion. Finally, COMIC is investigating advanced technology for output presentation in multimodal systems. Here the emphasis is on the use of a naturalistic presentation agent, in combination with graphical and textual rendering of the information.

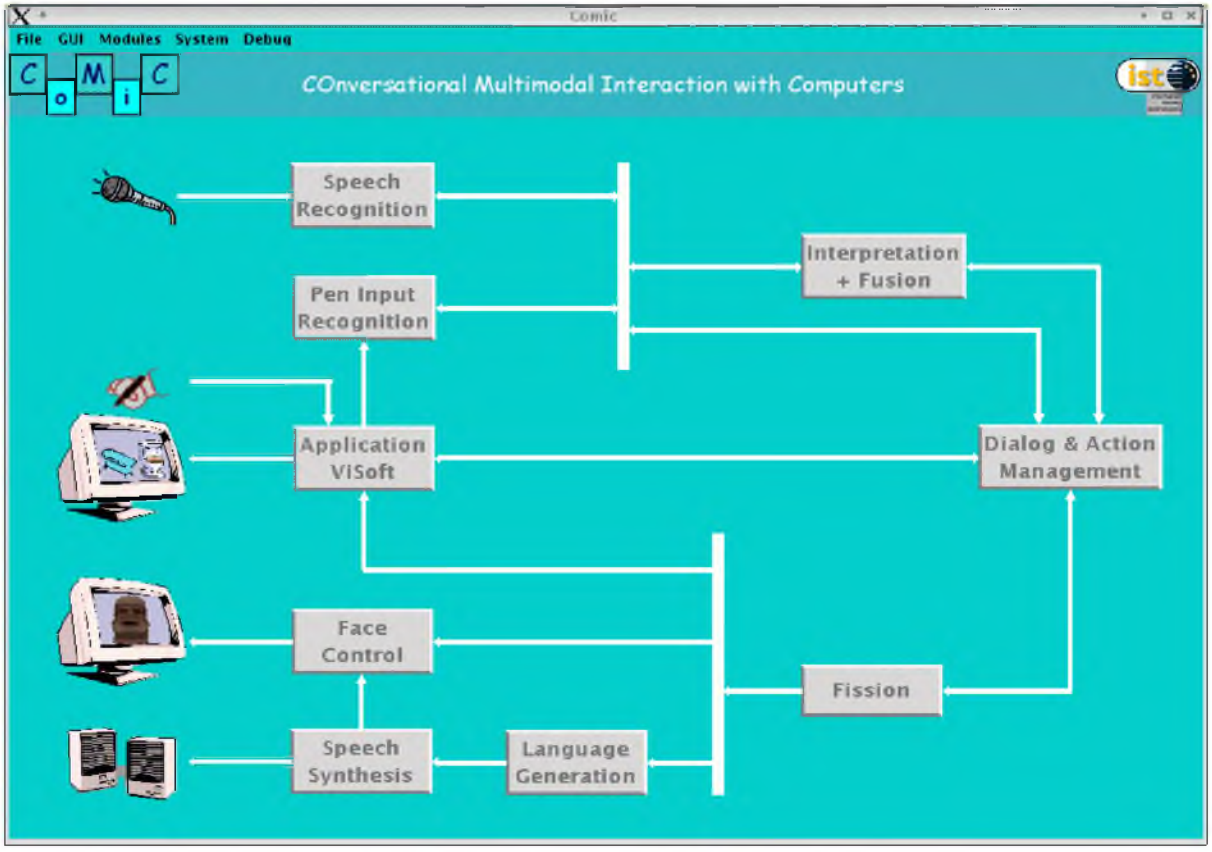

Figure 3. Architecture of the COMIC/ViSoft bathroom design application

4.3 Logistics issues

Designing multimodal interfaces is already difficult on its own, developing multimodal services is even more difficult. An in-depth analysis of the service creation process conducted in the MUST project has shown (and this finding has since been confirmed by several independent investigations of major market research companies) that successful service creation must encompass the complete service chain, and it must do so in an integrated manner. Fielding the bathroom design service, to give just one example, requires the integration of the product databases of all relevant manufacturers. Obtaining the right to access the databases from a large number of competing manufacturers is so difficult that it turns the technical problem of converting incompatible data formats to some common representation into an easy task. The same holds for tourist and travel information services: to make these attractive for paying customers many different dynamic databases must be integrated.

The logistic problems do not end when the databases are made accessible and an interface has been designed and implemented. The lack of standardisation of terminals, I/O

devices, and communication protocols promotes the development of proprietary solutions, which are very costly to develop, and quite unattractive for the customers, because of the risk that an expensive terminal can only be used for a small number of services from only a single service provider.

4.4 Business models

Multimodal services will require the development of new business models. For example, the bathroom design service might also be offered via Internet. However, in doing so the question arises how the companies that make the investments to develop the service can make profit of their investments. Locking the service up in a paid-for Internet site is probably not a good solution.

The potential of multimodal interaction to revolutionize services is not limited to fields like architectural design in general or bathroom design in particular. In the same way, the need for creative business models will be apparent in virtually all services.

5. Conclusion

In this paper we have used the FP5 project COMIC to explain the state-of-the-art in conversational multimodal interaction for *e*Work and *e*Commerce applications. We have addressed technological, logistics and business issues that remain to be solved in order to unleash the full potential of natural multimodal interaction. In tackling the technological problems, operational models of human-human interaction can help to show the right direction. Especially the recognition, understanding and reasoning capabilities of multimodal systems need to be improved. To enable successful commercial deployment of multimodal services, due attention must be paid to standardisation, logistics and business issues.

Acknowledgement

The work of the first author was supported by the EU project COMIC, IST-2001-32311.

References

[1] http://www.hcrc.ed.ac.uk/comic/
[2] http://www.eurescom.de/public/projects/P1100-series/p1104/default.asp
[3] L. Boves, D. Jouvet, J. Sienel, R. de Mori, F. Bechet, L. Fissore, and P.Laface, ASR for automatic directory assistance: the SMADA project," In: Proceedings: ESCA ITRW Automatic Speech Recognition. Challenges for the new millennium. Paris: LIMSI-CNRS, 2000, pp. 249-254.
[4] W. Wahlster, SmartKom: Fusion and Fission of Speech, Gestures, and Facial Expressions. In Proc. First International Workshop on Man-Machine Symbiotic Systems, Kyoto, Japan, 2002, pp. 213-225.
[5] Oviatt, S.L. Multimodal interfaces. In: J. Jacko and A. Sears (Eds.) The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, Lawrence Erlbaum Assoc., Mahwah, NJ, 2003, pp. 286-304.
[6] Leiboh, L. and Kaplun, Y. Multimodal Interfaces. In: August de los Reyes, Gregory P. Burch, et al. (Eds.) Flash Design for Mobile Devices: A style guide for the wireless revolution, J. Wiley & Sons, 2002.
[7] http://www.w3.org/2002/01/multimodal-charter.html.
[8] http://www.saltforum.org/