Design Principles for Cooperation between Modalities in Bi-directional Multimodal Interfaces

Stéphanie Buisine [1] **Jean-Claude Martin** [1&2]
(1) LIMSI-CNRS BP 133, 91403 Orsay Cedex, France.
Tel: +33.1.69.85.81.04. Fax: +33.1.69.85.80.88.
(2) LINC-Univ. Paris 8, IUT de Montreuil, 140 Rue de la Nouvelle France, 93100 Montreuil, France
{buisine,martin}@limsi.fr http://www.limsi.fr/Individu/martin

AUTHORS' BACKGROUNDS

Stephanie Buisine graduated in Experimental Psychology and Ergonomics. She is currently preparing a PhD in Cognitive Psychology at LIMSI-CNRS on ways of carrying out experimental evaluations of multimodal HCI.

Jean-Claude Martin is Associate Professor in Computer Science at Paris 8 University. He is a researcher at LIMSI-CNRS where he has been working on the multimodality of Human-Computer Systems since 1991. His topics of research include typologies of cooperation between modalities, analysis of users' multimodal behavior, software tools for the specification and development of multimodal interfaces and embodied conversational agents.

ABSTRACT

In this article, after briefly reviewing some previous work on the design of multimodal HCI, we present two projects, the first results obtained from experimental studies, and draw conclusions on design principles for multimodal interfaces.

ON THE DESIGN OF MULTIMODAL HCI

Classical design principles for HCI (e.g. Mayhew, 1999) recommend conducting iterative evaluations at different stages in the design process. However, evaluations and available guidelines mainly focus on output characteristics, because input devices are often *a priori* fixed as mouse and keyboard. In multimodal interfaces design, evaluations must deal with both input and output devices, and test reciprocal influences they have on each other. Concerning input from the user, potential usefulness of multimodality has been shown in several studies (see Martin et al., 1998 for a review). Behavioral analysis methods have also been used to categorize types of cooperation between modalities (Martin et al., 1998 & 2001). Some aspects of multimodal output have been studied in the context of HCI with embodied conversational agents (ECA). For example, the relevance of the presence of agents was tested in several applications (Craig et al., 2002; Moreno et al., 2001) and influence of agents' properties on different subjective variables was assessed (Granström et al., 2002; Koda & Maes, 1996; McBreen et al., 2001; McBreen & Jack, 2000 & 2001; Wonish & Cooper, 2002). Finally, general principles underlying the development of multimodal HCI were also described (Benoit et al., 2000; Oviatt, 2002). In the next section, we briefly present two projects for which we carried out experimental studies. The first results we obtained constitute a basis for formulating a few additional principles.

ILLUSTRATIVE PROJECTS

The IST-NICE[1] (Natural Interactive Communication for Edutainment) project is aimed at conceiving a conversational game for children and adolescents based on multimodal input (speech and gesture) and multimodal output (embodied conversational agents). Concerning gesture from the user, a 2D pen input was initially chosen because it seemed likely to meet conversational goals and less constraining than 3D gestural input devices. A Wizard-of-Oz experiment was carried out within an experimental methodology framework to collect multimodal behavioral data from the users and test the effectiveness of interaction. Adults and children were videotaped while interacting with 2D animated agents in a game application (Buisine et al., 2002). Each subject performed a multimodal scenario (allowing use of speech and/or pen gesture on the screen) and a speech-only scenario. The results confirm the usefulness for multimodal input whatever the subjects' age and gender: multimodal scenarios proved to be shorter and rated as being easier than speech-only scenarios. Moreover, multimodality homogenized ratings of easiness across all the participants better than speech-only condition (Buisine & Martin, submitted). Additional results showed that gesture interaction was more important for children than for adults, both in terms of quantity and variety. A factorial analysis also showed that the use of pen by children was associated with high ratings of pleasantness. Finally, analyses combining speech and pen gestures showed that certain commands were mainly performed by pen gestures and others by speech. The collected data are currently being exploited to build the multimodal language model for the design of the real system.

[1] http://www.niceproject.com

The RNRT-iTV [2] (Interactive Television) project is intended to develop an interactive television interface including multimodal input. In this respect, both speech and pen input seem likely to provide direct designations of items. To test this hypothesis and the usefulness of multimodality in input, a Wizard-of-Oz experiment was held in which participants had to perform TV-program search scenarios within three modality conditions: interaction by speech input, by pen input, and multimodal interaction (speech and pen input). The output device consisted of a web site including text and graphics, to which verbal error messages had been added. Preliminary analyses of the obtained behavioral corpus showed that the syntax of verbal commands was very simple: subjects used the labels displayed on the interface and did not build complex sentences. In the multimodal condition, most of the subjects used only one of the two modalities and appreciated choosing it in accordance with their preferences.

SUGGESTED DESIGN PRINCIPLES

From the results obtained in these two experimental studies, we propose the following principles:

- **Enable the use of the same modalities in input and output:** The symmetry principle for speech (speech must be bi-directional) could be transferred to multimodality. For example, we observed that interaction with conversational agents could benefit from gestural input, particularly for children (Buisine & Martin, submitted). When users interact with an ECA both in input and output by speech, gesture, and sometimes facial expressions or body posture, the symmetry concerns the modality of interaction. But it can also be extended to the characteristics of communication: if the interface makes use of 3D gestures via an ECA, users should also be able to use 3D gestures. Moreover, observed cooperations in the users' modalities (e.g. redundancy between speech and gesture…) should elicit similar modalities combinations in the ECA.
- **Use multimodal cues both in input and output to improve speech turns:** The use of multimodal cues is likely to improve speech turns. This principle has proved to be relevant in output when the user interacts with an agent (e.g. Cassell & Vilhjalmsson, 1999; Gustafson, 2002), but it could also be adapted to input. Recognition of facial expressions, gaze direction and/or non verbal speech could indicate when the user wants to take or give turn in the conversation (Thórisson, 1999). For example, as we observed with the pen for children, gestural exploration could mean that the user is quite lost and that the system should take the initiative in the dialog.
- **Use appropriate outputs to induce multimodal input behavior which is easier to process:** Appropriate outputs may induce multimodal input behavior which is easier to process. For example, labels displayed in the interface can be spontaneously used in speech input, which will limit vocabulary and thus facilitate vocal recognition.
- **Use modalities appropriate for the user:** Multimodality requires paying even more attention to users' profile than is normally the case in the design process. For example, age must be taken into account in speech and motor preferences and capabilities.
- **Adapt the recognition system according to the observed cooperations between modalities:** Task analysis and preliminary tests could shed light on "intuitive" cooperation between modalities in a given application. For example in one of our study, modalities appeared to cooperate by specialization (some commands were mainly performed by one modality). In this case, the adaptation of the multimodal recognition system could enhance its effectiveness and robustness. Conversely, for commands in which no specialization arise, the recognition system would allow freedom of choice between modalities and adapt to users preferences.

CONCLUSION AND SUGGESTIONS FOR THE WORKSHOP

An experimental approach in the context of conception projects is likely to provide both applied and general results. The latter can be exploited in general HCI design specification, as our recommendations to integrate input and output in the design of multimodal HCI. In case of intuitive multimodal interfaces with ECA, multimodality in input should be considered as part of the ECA and not dissociated during either development or evaluation phases. Bidirectionality and simultaneity of communication must be better integrated, for example with speech turns.

However, such design principles need to be integrated in a larger framework and confronted with other results. In other respects, principles arising from different protocols and contexts should be classified according, for example, to dimensions of multimodality or goals of designers (in terms of system performance, user satisfaction, etc.).

ACKNOWLEDGMENTS

The work described in this paper was financed by the IST-NICE project (http://www.niceproject.com) and the RNRT Interactive Television project (http://cpn.paris.ensam.fr/tvi/). The authors wish to thank their partners in these projects.

[2] http://cpn.paris.ensam.fr/tvi/

REFERENCES

Benoit, C., Martin, J.C., Pelachaud, C., Schomaker, L. & Suhm, B. (2000). Audio-Visual and Multimodal Speech Systems. In: D. Gibbon (Ed.). Handbook of Standards and Resources for Spoken Language Systems - Supplement Volume.

Buisine, S., Abrilian, S., Rendu, C., Martin, J.C. (2002). Towards experimental specification and evaluation of lifelike multimodal behavior. Proceedings of the workshop “Embodied conversational agents - let's specify and evaluate them!”, July 16, 2002, Bologna, Italy, in conjunction with the 1st international joint conference on Autonomous Agents & Multi-Agent Systems.

Buisine, S., Martin, J.C. (submitted). Experimental evaluation of bi-directional multimodal interaction with conversational agents.

Cassell, J., Vilhjalmsson, H. (1999). Fully embodied conversational avatars: making communicative behaviors autonomous. Autonomous Agents and Multi-Agent Systems, 2, 45-64.

Craig, S. D., Gholson, B., & Driscoll, D. (2002) Animated Pedagogical Agents in Multimedia Educational Environments: Effects of Agent Properties, Picture Features, and Redundancy. Journal of Educational Psychology. 94, (428-434).

Granström, B., House, D., Swerts, M. (2002). Multimodal feedback cues in human-machine interactions. http://fdlwww.kub.nl/~krahmer/workshp_files/pros2002_09.doc

Gustafson, J. (2002). Developing multimodal spoken dialogue systems. Empirical studies of spoken Human-Computer Interaction. Doctoral dissertation.

Koda, T., Maes, P. (1996). Agents with faces: the effects of personification of agents. In: Proceedings of HCI'96, London, UK, pp. 98-103.

Martin, J.C., Grimard, S., Alexandri, K. (2001) On the annotation of the multimodal behavior and computation of cooperation between modalities. Proceedings of the workshop on " Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents ", May 29, 2001, Montreal, in conjunction with The Fifth International Conference on Autonomous Agents. pp 1-7

Martin, J.C., Julia, L. & Cheyer, A. (1998) A Theoretical Framework for Multimodal User Studies. *Proceedings of the Second International Conference on Cooperative Multimodal Communication, Theory and Applications (CMC'98)*, 28-30 January 1998, Tilburg, The Netherlands.

Mayhew, D.J. (1999). The usability engineering lifecycle: a practitioner’s handbook for user interface design. Morgan Kaufmann.

McBreen, H., Anderson, J., Jack, M. (2001). Evaluating 3D embodied conversational agents in contrasting VRML retail applications. In: Proceedings of the workshop Multimodal communication and context in embodied agents. 5th International conference on Autonomous Agents. Montreal, Canada.

McBreen, H., Jack, M. (2000). Empirical evaluation of animated agents in a multi-modal e-retail application. Proc. AAAI Fall Symposium on Socially Intelligent Agents, November 2000.

McBreen, H., Jack, M. (2001). Evaluating humanoid synthetic agents in e-retail applications. IEEE Transactions on Systems, Man and Cybernetics, vol. 31 (5), pp. 394-405, 2001.

Moreno, R., Mayer, R.E., Spires, H.A., Lester, J.C. (2001). The case for social agency in computer-based teaching: do students learn more deeply when they interact with animated pedagogical agents? Cognition and Instruction, 19, 177-213.

Oviatt, S.L. (2002). Breaking the robustness barrier: recent progress on the design of robust multimodal systems. In: M. Zelkowitz (Ed.). Advances in Computers, vol. 56, Academic Press.

Thórisson, K. R. (1999). A Mind Model for Multimodal Communicative Creatures and Humanoids. International Journal of Applied Artificial Intelligence , 13(4-5), 449-486.

Wonish, D., Cooper, G. (2002). Interface Agents: preferred appearance characteristics based upon context. In: Proceedings of the workshop Virtual Conversational Characters: Applications, Methods, and Research Challenges. In conjunction with HF2002 and OZCHI2002. 29th November, 2002 Melbourne, Australia.