2D Gestural and Multimodal Behavior of Users Interacting with Embodied Agents

Martin Jean-Claude [1 & 2] Buisine Stéphanie [1] Abrilian Sarkis [1]
(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 Nouvelle France, 93100 Montreuil, France.
{martin, buisine, sarkis}@limsi.fr http://www.limsi.fr/Individu/martin/

Abstract

When we communicate between humans, we usually bring into play several communication modalities such as speech, gesture, posture, gaze. These modalities are involved in simultaneous and bi-directional threads of communication. Current Human-Computer Interfaces and Embodied Conversational systems are still limited regarding this simultaneous bi-directional communication. In this paper, we describe an edutainment application we are working on, the actual aspects of perception (interpretation of user's 2D interface gesture detected via a pen device and input fusion module), and an experimental study we have already done on perception and ECAs in which part of the system was simulated by an experimenter. We conclude on some requirements we propose for balanced perception and action in ECAs.

1 Introduction

ECAs use multimodal output communication i.e. speech and nonverbal behaviors, such as arm gesture, facial expression or gaze direction. In some of these systems, the input from the user is limited to the classical keyboard and mouse combination to interact with agents (Koda and Maes 1996; André and Rist 2001). Other ones have been developed with speech input (Mc Breen and Jack 2001), which might be indeed an intuitive way to dialog with ECAs.

The goal of the project NICE[1] is to enable users to interact with conversational characters using 2D gestures and speech (Bernsen 2003). The application area is edutainment: the users are supposed to both learn and play when interacting with the system. The combination of speech and 2D gestures seems indeed to be an interesting combination of modalities when interacting with an ECA in an environment allowing communication about graphical objects.

[1] www.niceproject.com

Section 2 summarizes one of the two experimental study on perception and ECAs we have already presented in (Buisine et al. 2003; Buisine and Martin 2003). Section 3 describes the modules we are currently developing. We conclude on a discussion on the requirement for balanced perception and action in ECAs.

2 Evaluation study

We might expect from experimental studies of multimodal input interfaces (Oviatt 1996) that subjects prefer and are more effective when using more than one input modality. Yet, this hypothesis has to be experimentally grounded in the case of communication with ECAs. A few systems combining ECA and multimodal input were developed (Cassell and Thorisson 1999; Wahlster et al. 2001), but experimental evaluation of such systems is still an issue. So far, a few studies have been conducted to test the usefulness of ECAs or the impact of different output features (Dehn and van Mulken 2000; Mc Breen and Jack 2001; Moreno et al. 2001; Craig et al. 2002). However, as far as we know, the effect of input devices and modalities has not been much investigated in the context of the interaction with ECAs. On this point, we think that since ECAs are supposed to include a conversational dimension, the input mode should be considered as an integral part of the ECA. Therefore, intuitive ECAs should be multimodal not only in output and but also in input. In this paper, we will study whether bi-directionality of multimodality actually enhances the effectiveness and pleasantness of interaction in an ECA system.

2.1 Method

A bi-directional multimodal interface was tested with the Wizard-of-Oz method, which consists in simulating part of the system by a human experimenter hidden from the user. This type of simulation enabled us to disregard technical difficulties raised by speech and gesture understanding during the experiment (currently impossible unless numerous behavioral data are previously collected). Such a protocol for collecting behavioral data has already been used in the field of multimodal input interfaces without ECAs (Oviatt et al.

1997; Cheyer et al. 2001). Our experiment uses the 2D cartoon-like Limsi Embodied Agents that we have developed. Their multimodal behavior (e.g. hand gestures, gaze, facial expression) can be specified with the TYCOON XML language.

The game starts in a house corridor including 6 doors of different colors. Only three doors open onto a room and the three remaining ones are locked. The rooms are: a library, a kitchen and a greenhouse, each of them being inhabited by an agent. In the corridor, a jinn asks the subject to go to different rooms, meet people and fulfill their wishes. Agents' wishes oblige the subjects to bring them objects missing in the room where they are. Therefore, subjects have to go to other rooms, find the right object and bring it back to the agent. In order to elicit dialogues and gestures, many objects of the same kind are available, and the subject has to choose the right one according to its shape, size or color (Figure 1).

Figure 1: Screendump of the 2D game application.

Two groups of subjects participated in the experiment: 7 adults (3 male and 4 female subjects, age range 22 – 38) and 10 children (7 male and 3 female subjects, age range 9 – 15). The two groups were equivalent regarding their frequency of use of video games. An additional adult subject was excluded from the analysis because he had guessed the system was partly simulated. The Wizard-of-Oz device was described in (Buisine et al. 2002). The 2D graphical display included four rooms, four 2D animated agents and 18 moveable objects (e.g. book, plant). Loudspeakers were used for speech synthesis with IBM ViaVoice. However, the wizard simulated speech and gesture recognition and understanding. The wizard could modify either the game environment (switch to another room, move objects), or the agents' spoken and nonverbal behaviors. For this purpose, the wizard interface contained 83 possible utterances (e.g. "Can you fetch the red book for me?"), each of them associated with a series of nonverbal behaviors including head position, eyes expression, gaze direction, mouth shape and arm gestures. Nonverbal combinations were defined with data from the literature (Calbris and Montredon 1986; Calbris and Porcher 1989; Cassell 2001). Arm gestures included the main classes of semantic gestures: emblematic, iconic, metaphoric, deictic, and beat (Cassel 2000). In addition to these pre-encoded items, the wizard could type a specific utterance and could associate it with a series of nonverbal cues extracted from the existing basis.

Subjects had to carry out successively two game scenarios: one scenario in a multimodal condition (in this case they could use speech input, pen, and combine these two modalities to play the game) and another scenario in a speech-only condition. The order of these conditions was counterbalanced across the subjects. The two scenarios were equivalent in that they involved the same agents, took place in the same rooms and implied the same goal to achieve. Only wishes differed from one scenario to the other (objects that had to be found and returned to the agents were different). After each scenario, subjects had to fill out a questionnaire giving their subjective evaluation of the interaction. This questionnaire included four scales: perceived easiness, effectiveness, pleasantness and easiness to learn. At the end of the experiment, subjects were explained that the system was partly simulated.

The 34 recorded videos (two scenarios for each of the 17 subjects) were then annotated. Speech annotations (segmentation of the sound-wave into words) were done with PRAAT[2] and then imported into ANVIL (Kipp 2001) in which all complementary annotations were made. Three tracks are defined in our ANVIL coding scheme (Figure 2a):

- Speech: every word is labeled according to its morpho-syntactic category;
- Pen gestures (including the three phases: preparation, stroke and retraction) are labeled according to the shape of the movement: pointing, circling, drawing of a line, drawing of an arrow, and exploration (movement of the pen in the graphical environment without touching the screen);
- Commands corresponding to the subjects' actions (made by speech and/or pen). Five commands were observed in the videos: get into a room, get out of a room, ask a wish, take an object, give an object. Annotation of a command covers the duration of the corresponding annotations implied in the two modalities and is bound to these annotations.

Annotations were then parsed by Java software we developed in order to extract metrics that were submitted to statistical analyses with SPSS[3] (see Figure 2b).

[2] http://www.fon.hum.uva.nl/praat/

[3] http://www.spss.com/

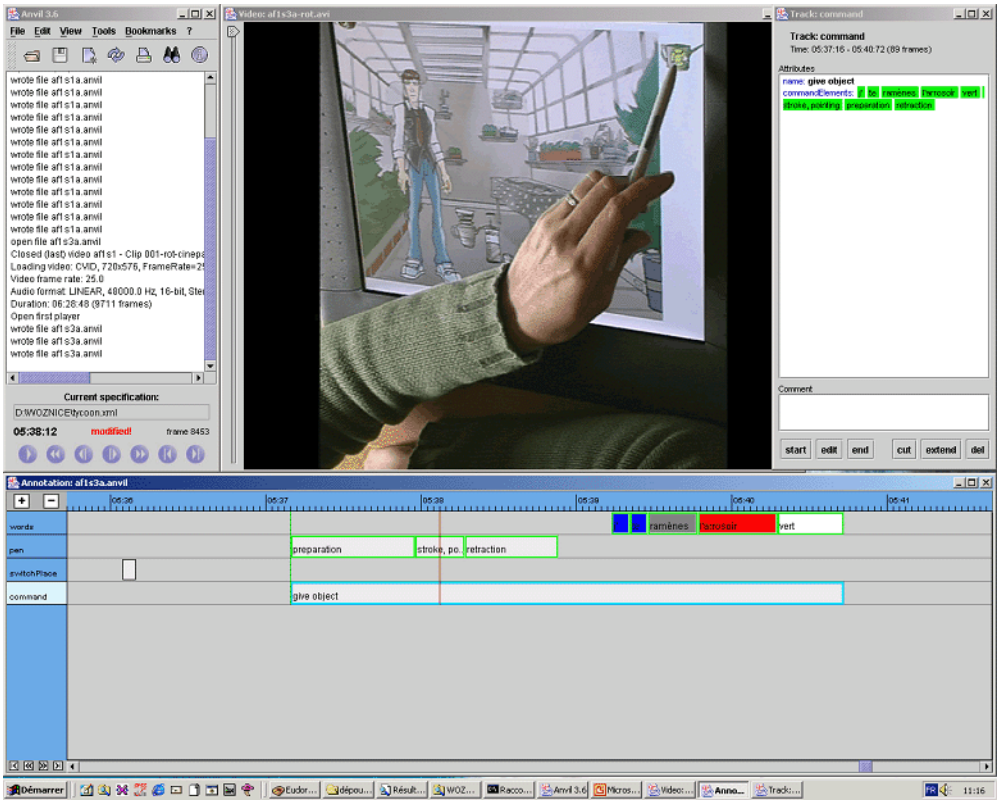

Figure 2a: Screenshot of the annotation of a multimodal behavior occurring during the interaction with a 2D conversational agent. The annotation software used is Anvil. The subject said "Iam bringing you the green watertank"**.**

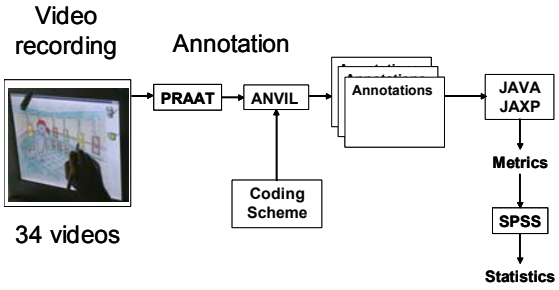

Figure 2b: Annotation and analysis process.

2.1.1 Data quantification and analyses

Metrics extracted from annotations (total duration of scenario, use duration of each modality, morpho-syntactic categories, shapes of pen movements) as well as subjective data from the questionnaires were submitted to analyses of variance using age, gender and condition-order as between-subject factors, and condition and commands as within-subject factors.

Factorial analysis and multiple regressions were performed with the following variables: total duration of scenario, use duration of speech, use duration of pen, age, perceived easiness, effectiveness, pleasantness and easiness to learn.

2.2 Results

2.2.1 Unidimensional analyses

The main effect of input condition (speech-only vs. multimodal) proved to be significant ($F(1/9) = 70.05$, $p<0.001$) and showed that multimodal scenarios were shorter (307.80s +/-88.71) than speech-only scenarios (437.19s +/-129.42). No main effect of between-subject factors (age, gender or order) was observed.

A main effect of input condition ($F(1/9) = 57.81$, $p<0.001$) showed that speech was used longer in the speech-only condition than in the multimodal condition. For multimodal scenarios, we studied the use of speech, pen, and their overlap (simultaneous use). Pen proved to be the interaction mode the most used ($F(2/18) = 14.44$, $p<0.001$). However, an interaction between age of subjects and modality ($F(2/18) = 5.91$, $p = 0.031$, see Figure 3) suggests that this main effect is due to children's behavior. Indeed, there was no significant difference between use duration of speech and pen for adults ($F(1/3) = 0.31$, NS) whereas this difference appeared to be significant for children ($F(1/6) = 7.51$, $p = 0.034$). Moreover, use duration of speech was not different between children and adults in the multimodal condition ($F(1/9) = 0.69$, NS), just like in the speech-only one ($F(1/9) = 0.26$, NS). Overlaps between speech and pen use were particularly short.

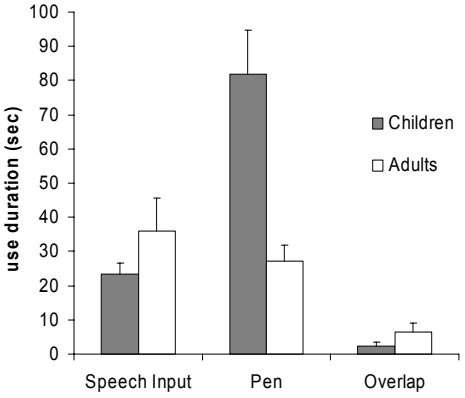

Figure 3: Mean use duration of each modality in multimodal condition as a function of age.

The dependent variable "mean number of each modality" was selected to investigate the use of modalities as a function of commands. Thus, a separate analysis of variance was carried out for each command and this showed large differences in modality use from one command to another. For example, the "ask wish" command proved to be mainly performed by speech ($F(2/18) = 21.99$, $p = 0.001$), whereas "take an object" and "give an object" were preferentially made with the pen (respectively $F(2/18) = 14.61$, $p = 0.002$ and $F(2/18) = 4.94$, $p = 0.046$). The "get into a room" command was also mainly performed with the pen ($F(2/18) = 24.27$, $p = 0.001$), but an age*modality interaction indicated that this effect was attributable to the children ($F(2/18) = 7.40$, $p = 0.023$, see Figure 4). Indeed, the main effect of modality is not significant for adults ($F(2/6) = 4.57$, NS) whereas it is for children ($F(2/12) = 26.21$, $p = 0.001$).

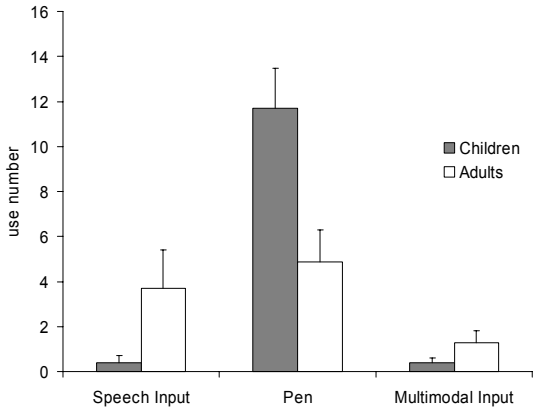

Figure 4: Mean use number of each modality for "get into a room" command as a function of age.

Concerning the "get out of a room" command, an age*modality interaction ($F(2/18) = 5.90$, $p = 0.020$, see Figure 5) reveals that adults preferred to use speech rather than pen ($F(1/3) = 12.31$, $p = 0.039$) whereas children equally used these two modalities ($F(1/6) = 1.40$, NS).

Percentages of morpho-syntactic categories observed during the experiment (whatever the subject group and the input condition) are listed in table 1. The "locution" category gathers expressions such as "Hello", "Bye", "Please", "Thank you", "OK", etc. This category was the most frequently used. We investigated the effect of the subjects' age on each morpho-syntactic category annotated. Although the total number of words used by adults and children was not different ($F(1/9) = 2.66$, NS), adults proved to use significantly more verbs at the indicative mood ($F(1/9) = 11.44$, $p = 0.008$), more articles ($F(1/9) = 6.83$, $p = 0.028$), more adverbs ($F(1/9) = 5.12$, $p = 0.05$) and more pronouns ($F(1/9) = 4.79$, $p = 0.056$) than children.

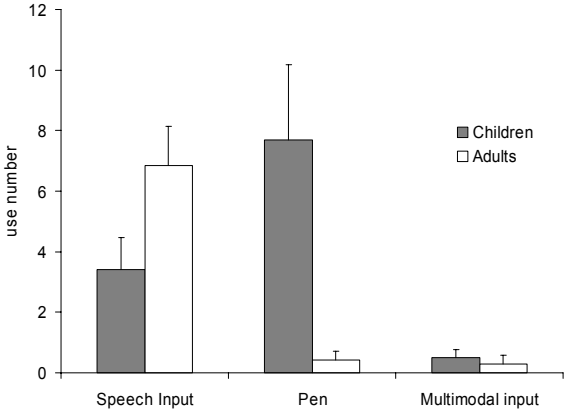

Figure 5: Mean use number of each modality for the "get out of a room" command as a function of subjects' age.

Morpho-syntactic category	Total number of occurrences	Percentage
Locutions	990	21.9 %
Verbs	871	19.3 %
Substantives	731	16.2 %
Pronouns	695	15.4 %
Adjectives	516	11.4 %
Articles	466	10.3 %
Conjunctions	141	3.1 %
Adverbs	104	2.3 %

Table 1: Morpho-syntactic categories used during the experiment.

Table 2 contains the total number of occurrences and percentages of each of the five observed shapes of pen movement. Pointing appears to be the main way subjects used the pen. The subjects were not trained for the pen prior to the experiment.

Given that analysis of variance was not relevant for these data (because of numerous missing values), we performed a Wilcoxon-Mann-Whitney test (non-parametric method) on each shape of movement with age as between-subject factor. Children globally made more gesture than adults ($Z = -3.18$, $p = 0.001$). In

particular, they proved to use more circling movements ($Z = -2.17$, $p = 0.03$), to point more ($Z = -2.10$, $p = 0.036$) and tended to explore more than adults ($Z = -1.84$, $p = 0.066$). Multimodal scenarios were evaluated easier than speech-only scenarios ($F(1/9) = 9.64$, $p = 0.013$). Moreover, the age*condition interaction ($F(1/9) = 8.31$, $p = 0.018$, see Figure 6) indicated that adults' and children's ratings of easiness were the same for multimodal scenarios ($F(1/9) = 0.17$, NS), whereas children found speech-only scenarios more difficult than adults ($F(1/9) = 9.78$, $p = 0.012$).

Shape of movement	Total number of occurrences	Percentage
Pointing	413	66 %
Circling	113	18.1 %
Exploration	53	8.5 %
Line	34	5.4 %
Arrow	13	2.1 %

Table 2: Shapes of movements used during the experiment.

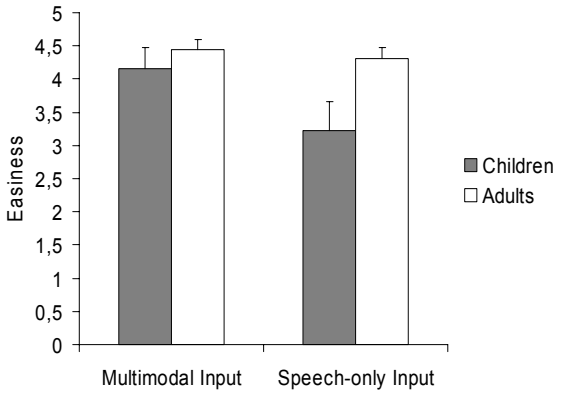

Figure 6: Mean ratings of easiness as a function of the input condition and the subjects' age.

The same kind of result appeared in a gender*condition interaction ($F(1/9) = 6.73$, $p = 0.029$, see Figure 7) which showed gender differences on ratings of easiness for speech-only scenarios ($F(1/9) = 8.04$, $p = 0.02$, female subjects' ratings being lower) but not for multimodal scenarios ($F(1/9) = 0.16$, NS).

The analysis of the three other subjective variables (perceived effectiveness, pleasantness and easiness to learn) yielded no significant results.

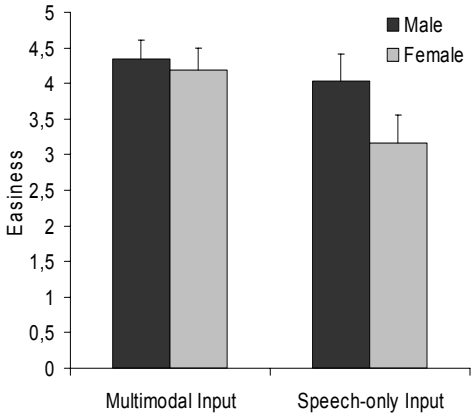

Figure 7: Mean ratings of easiness as a function of the input condition and the subjects' gender.

2.2.2 Multidimensional analyses

A factorial analysis with principal component extraction was carried out to seek a link between all variables collected during multimodal scenarios (total duration, use duration of speech, use duration of pen, age, perceived easiness, effectiveness, pleasantness and easiness to learn). The so-called extracted components actually represent axes that best summarize a set of data. Here, three components appeared to account for 75.6% of the total variance. Table 3 presents correlations between variables and these components. Grey cells highlight strongest correlations.

	Components		
	1	2	3
Total duration	-0.912	-6.5E-02	-0.104
Speech duration	1.7E-02	-0.816	-0.264
Pen duration	-0.424	0.682	0.135
Age	0.520	-0.582	0.398
Easiness	0.828	0.116	-0.270
Effectiveness	0.172	0.171	0.906
Pleasantness	0.434	0.503	-0.411
Learning	0.848	0.239	6.2E-03

Table 3: Correlations between variables and components.

The first component contrasts age, perceived easiness and easiness to learn with total duration of the scenario: this means that older subjects (within our sample) rated the interaction easier to play and to learn and performed scenarios quicker. In the same way, the second component shows that subjects who mostly used the pen also gave high ratings of pleasantness, made little use of speech and were a young age. Finally,

perceived effectiveness strongly correlates with the third component, but no other variable is linked to it.

Multiple regression analyses confirmed that some of the subjective ratings could be predicted from values of behavioral metrics. Indeed, these metrics (total duration of scenario, use of speech, use of pen, and age) provide a good regression model to predict perceived easiness ($F(5/11) = 3.74$, $p = 0.032$), in which total duration of scenario is the most important variable ($t(16) = -3.01$, $p = 0.012$). Moreover, using behavioral metrics, easiness to learn could also be predicted ($F(5/11) = 6.40$, $p = 0.005$), particularly by the total duration of scenario ($t(16) = -5.18$, $p<0.001$) and the duration of pen-use ($t(16) = 2.51$, $p = 0.029$). However, behavioral metrics fail to provide a good model of perceived effectiveness and pleasantness.

2.3 Discussion on evaluation

Concerning total duration of scenarios, time spent on the task usually constitutes a measure of efficiency. Yet, the user may spend extra time with the ECA because he likes it or finds it interesting (Ruttkay et al. 2002). However, in our results, time spent was longer in the speech-only scenario and subjects rated this condition as being more difficult. Thus, our results suggest that multimodality in input facilitates interaction, as it was previously observed in interfaces without ECA (Oviatt 1996). Moreover, multimodality seems to homogenize ratings of easiness better than speech-only condition. This globally highlights the usefulness of multimodal input when a subject, whatever his age and gender, interacts with an ECA.

One of the strongest age effects yielded by our results concerned the use of pen, significantly more important for children. Furthermore, the factorial analysis showed that the use of pen by children was associated with high ratings of pleasantness. These results underline that children enjoy direct gesture interaction and exploration. Thus, speech-only ECA game applications might not be so relevant for children, even if pleasantness is not reducible to the interaction mode, as shown by the multiple regression. Previous work comparing the use of each modality on a multimodal interface (Guyomard et al. 1995; Mignot and Carbonell 1996; Siroux et al. 1997) tended to show that speech was used more than gesture. However, given that in these cases, gesture modality was a tactile screen (maybe less engaging than pen) and that users were exclusively adults, these results might not be comparable to ours.

Table 2 indicates that users in the multimodal condition made frequent use of the pointing gesture. This may be evidence for transfer from traditional point and click interfaces. In other words, maybe users did well because they simply used their everyday WIMP interface experience.

Finally, factorial analysis and multiple regression also showed that perceived effectiveness was not linked to any of the metrics we collected. This subjective variable does not seem to be influenced by the interaction mode. Conversely, easiness to play and to learn the interaction are strongly linked to the duration of scenarios and the use of pen.

Our experiment showed that as far as morpho-syntactic categories were concerned, there were not large differences between children's and adults' spoken behavior for this task, and that locutions (i.e. invariable familiar expressions) constituted the most frequently used category. The analyses we performed on spoken behavior were quite limited compared to other studies where variables such as disfluencies were analyzed (e.g. (Oviatt 1996; Oviatt 2000)).

Our approach is based, on the one hand, on a methodological process stemming from Experimental Psychology, which has seldom been followed before in this domain (Dehn and van Mulken 2000): setting-up of a factorial design and experimental groups, controlled and standardized procedure, and toolkit of statistical methods. On the other hand, this study was equipped with a series of computer-aided analyses including PRAAT, ANVIL, SPSS, and Java software. Besides being useful for our specific application, these results are likely to be exploited in the general framework of ECA evaluation and specification. Indeed, results obtained with inferential statistical methods can be generalized to the whole populations from which the subjects' samples were extracted.

3 Modules processing 2D Gesture and Input fusion

Regarding multimodal input fusion, it is common to find separate early processing of each modality such as speech and gesture recognition in separate modality specific modules. The next processing steps differ from one system to another. The results of these early processing may afterwards be used for further monomodal processing at a higher level (spoken only natural language understanding) followed later by a "late fusion" of the input modalities, such as in the Quickset system (Cohen et al. 1999). In other systems, modality integration may appear earlier in the architecture in a "multimodal interpreter" or "media fusion" module achieving integrated multimodal syntactical and semantic processing out of the different modalities: the Smartkom system (Wahlster et al. 2001), and the IBM RIA and Embassi systems. Some system use more loosely connected architectures in which the components exchange messages when needed without being constrained by a feed forward architecture such as the Chameleon platform which uses a blackboard in (Broendsted et al. 2001), or the Open Agent Architecture using a central facilitator brokering the messages between the agents (Cheyer and Martin 2001).

For the NICE project, we have developed a first version of a Gesture Recognition and Interpretation module. In a first step, the Gesture Recognition module receives 2D input events from the graphical application and the recognition of the gesture shape is achieved

with back-propagation Neural Networks. In a second step, the interpretation consists in producing a simple semantic representation including possible associated referred objects. This semantic representation will be sent by the Gesture Interpretation module to the Input Fusion module for future integration with spoken events (Corradini et al. 2003). These modules have been tested firstly in a 2D simple game environment simulated scenario and are currently integrated and under test with the 3D characters and other modules developed by the other NICE partners.

4 Future directions : requirements for an ECA in the loop

The study we presented here was quite limited to 2D interface gesture captured via a pen tablet which do not enable the capture of communicative gestures such as iconic 3D gestures.

Human-human communication is not limited to mono-directional transmission of information (Ossimitz; Abric 1996): it is at any time bi-directional. Studies and systems considering bi-directionality of communication in ECA have mostly studied turn taking and feedback (Cassell and Thorisson 1999; Thórisson 2002).

Achieving bi-directional ECAs induce the following requirements: detecting subtle signs of turn giving (e.g. detecting user's hesitations and taking initiative), being continuously interruptible by the user during the real-time production of the agent's multimodal behavior, progressive processing of user's input, progressive and real-time computation of the ECA's behavior, build shared meaning representations (in the ECA systems this « shared » feature of communication is seldom represented). Current limitations of monomodal processing of each input and output modality make such requirements hard to meet but should benefit of the consideration of theoretical issues in communication theories and experimental studies.

Acknowledgments

The work described is supported by the EC's Human Language Technologies Programme, Grant IST-2001-35293 in the project NICE (Natural Interactive Communication for Edutainment) www.niceproject.com. The support is gratefully acknowledged.

References

Abric, J.-C. (1996). *Psychologie de la communication*, Armand Colin.

André, E. & Rist, T. (2001). "Controlling the Behaviour of Animated Presentation Agents in the Interface: Scripting vs. Instructing." *AI Magazine* 22(4): 53-66.

Bernsen, N. O. (2003). When H.C. Andersen Is Not Talking Back. *Intelligent Virtual Agents*, Springer-Verlag Heidelberg. 2792 / 2003**:** 27 - 30.

Broendsted, T., Dalsgaard, P., Larsen, L. B., Manthey, M., McKevitt, P., Moeslund, T. B. & Olesen, K. G. (2001). The IntelliMedia WorkBench - an environment for building multimodal systems. *Advances in Cooperative Multimodal Communication: Second International Conference, CMC'98, Tilburg, The Netherlands, January 1998, Selected Papers'*. H. B. a. R.-J. Beun. Berlin, Germany, Springer Verlag**:** 217-233.

Buisine, S., Abrilian, S. & Martin, J.-C. (2003). Evaluation of individual multimodal behavior of 2D embodied agents in presentation tasks. *Workshop "Embodied conversational characters as individuals", Marriot, A., Pelachaud, C., Ruttkay, Z. (Eds), 2nd International Joint Conference on Autonomous Agents & Multiagent Systems (AAMAS'03)*, Melbourne, Australia, 15th july

Buisine, S., Abrilian, S., Rendu, C. & Martin, J.-C. (2002). Towards Experimental Specification and Evaluation of Lifelike Multimodal Behavior. *Workshop on "Embodied conversational agents - let's specify and evaluate them!", in conjunction with The First International Joint Conference on "Autonomous Agents & Multi-Agent Systems"*, Bologna, Italy, 16 July

Buisine, S. & Martin, J.-C. (2003). Experimental Evaluation of Bi-directional Multimodal Interaction with Conversational Agents. *Proceedings of the the Ninth IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'2003)*, Zürich, Switzerland, September 1-5, IOS Press, 168-175. 1-58603-363-8

Calbris, G. & Montredon, J. (1986). *Des gestes et des mots pour le dire*, Paris: Clé International.

Calbris, G. & Porcher, L. (1989). *Geste et communication*, Paris: Hatier.

Cassel, J. (2000). "More than Just Another Pretty Face: Embodied Conversational Interface Agents." *Communications of the ACM* 43(4): 70-78.

Cassell, J. (2001). "Embodied Conversational Agents: Representation and Intelligence in User Interface." *AI Magazine* 22(3): 67-83.

Cassell, J. & Thorisson, K. R. (1999). "The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents." *Applied Artificial Intelligence* 13(4-5): 519-538.

Cheyer, A., Julia, L. & Martin, J.-C. (2001). A Unified Framework for Constructing Multimodal Experiments and Applications. *Cooperative Multimodal Communication*. H. Bunt, Beun, R.J., Borghuis, T., Springer**:** 234-242.

Cheyer, A. & Martin, D. (2001). "The Open Agent Architecture." *Journal of Autonomous Agents and Multi-Agent Systems* 4(1): 143-148. http://www.ai.sri.com/oaa/

Cohen, P. R., McGee, D., Oviatt, S., Wu, L., Clow, J., King, R., Julier, S. & Rosenblum, L. (1999). "Multimodal interaction for 2D and 3D environments." *IEEE Computer Graphics and Applications* 19(4): 10-13.

Corradini, A., Mehta, M., Bernsen, N. O., Martin, J.-C. & Abrilian, S. (2003). Multimodal Input Fusion In Human-computer Interaction - On The Example Of The Nice Project. *NATO ASI 2003 "Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management"*, Narek center of Yerevan University, Tsakhkadzor , Armenia, 18-29 August, Kluwer

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.

Dehn, D. M. & van Mulken, S. (2000). "The impact of animated interface agents: a review of empirical research." *International Journal of Human-Computer Studies*(52): 1-22.

Guyomard, M., Le Meur, D., Poignonnec, S. & Siroux, J. (1995). Experimental work for the dual usage of voice and touch screen for a cartographic application. *ESCA Tutorial and Research Workshop on Spoken Dialog Systems*, Vigso, Denmark, 30 may - 2 june, 153-156.

Kipp, M. (2001). Anvil - A Generic Annotation Tool for Multimodal Dialogue. *Eurospeech'2001*,

Koda, T. & Maes, P. (1996). Agents with faces: the effects of personification of agents. HCI'96, London, UK, August 20-23, The British HCI Group, 98-103. ISBN 1-85924-119-0

Mc Breen, H. & Jack, M. (2001). "Evaluating humanoid synthetic agents in e-retail applications." *IEEE Transactions on Systems, Man and Cybernetics* 31(5): 394-405.

Mignot, C. & Carbonell, N. (1996). "Commandes orales et gestuelles: une étude empirique." *Techniques et Sciences Informatiques* 15(10): 1399-1428.

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.

Ossimitz, G. Information, Communication and Social Awareness.
http://www.uni-klu.ac.at/~gossimit/pap/go_icsa.pdf

Oviatt, S., De Angeli, A. & Kuhn, K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. *Human Factors in Computing Systems (CHI'97)*, New York, ACM Press, 415-422.

Oviatt, S. L. (1996). "User-centered modeling for spoken language and multimodal interfaces." *IEEE Multimedia* 3(4): 26-35.

Oviatt, S. L. (2000). Talking to Thimble Jellies: Children's conversational speech with animated characters. *International Conference on Spoken Language Processing (ICSLP'2000)*, Beijing, China, Chinese Friendship Publishers, 877-880.

Ruttkay, Z., Dormann, C. & Noot, H. (2002). Evaluating ECAs - What and how ? *Workshop on "Embodied conversational agents - let's specify and evaluate them!" in conjunction with The First International Joint Conference on "Autonomous Agents & Multiagent Systems" (AAMAS'02)*, Bologna, Italy, 16 July 2002

Siroux, J., Guyomard, M., Multon, F. & Remondeau, C. (1997). Multimodal references in GEORAL TACTILE. *Workshop "Referring phenomena in a multimedia context and their computational treatment" held in cunjunction with ACL/EACL'97*, Madrid, Spain, july 11th, 39-43.

Thórisson, K. R. (2002). Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action. *Multimodality in Language and Speech Systems*. D. H. B. Granström, I. Karlsson. Dordrecht, The Netherlands, Kluwer Academic Publishers**:** 173-207.

Wahlster, W., Reithinger, N. & Blocher, A. (2001). Smartkom: Multimodal Communication with a Life-Like Character. *EuroSpeech'2001*, Aalborg (Denmark),