Experimental Evaluation of Bi-directional Multimodal Interaction with Conversational Agents

Buisine Stéphanie [1] & Martin Jean-Claude [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/research/projects/lea/

Abstract: In the field of intuitive HCI, Embodied Conversational Agents (ECAs) are being developed mostly with speech input. In this paper, we study whether another input modality leads to a more effective and pleasant "bi-directional" multimodal communication. In a Wizard-of-Oz experiment, adults and children were videotaped while interacting with 2D animated agents within a game application. Each subject carried out a multimodal scenario (speech and/or pen input) and a speech-only scenario. The results confirm the usefulness of multimodal input, which yielded shorter scenarios, higher and more homogeneous ratings of easiness. Additional results underlined the importance of gesture interaction for children, and showed a modality specialization for certain actions. Finally, multidimensional analyses revealed links between behavioral and subjective data, such as an association of pen use and pleasantness for children. These results can be used for both developing the functional prototype and in the general framework of ECA-systems evaluation and specification.

Keywords: Evaluation, multimodal behavior, conversational agent, experimental psychology.

1 Introduction

Amongst current research in the field of intuitive Human-Computer Interaction, Embodied Conversational Agents (ECAs) are interesting from a usability and intuitive point of view. ECAs use multimodal output communication i.e. speech and nonverbal behaviors, such as arm gesture, facial expression or gaze direction (Cassell et al. 2000).

In some of these systems, the input from the user is limited to the classical keyboard and mouse combination to interact with agents (e.g. Pelachaud et al. 2002). Other ones have been developed with speech input (e.g. Mc Breen and Jack 2001), which might be indeed an intuitive way to dialog with ECAs. However, one may wonder whether other input modalities would lead to an even more intuitive "bi-directional" multimodal communication. We might expect from experimental studies of multimodal 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 (e.g. Cassell and Thorisson 1999), 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 (see Dehn and van Mulken 2000 for a review; McBreen 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 never been 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.

This study was conducted in the context of a game conception currently in progress in the NICE[1] (Natural Interactive Communication for Edutainment) project. 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. Demonstration samples of the XML control of these agents are available on the web[2].

Section 2 describes the experimental method. Section 3 presents the results, which are discussed in section 4.

2 Method

2.1 Participants

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.

2.2 Apparatus

The Wizard-of-Oz device was composed of two computers (see Figure 1). PC#1, which ensured the presentation of the game to the subject, was connected with a Wacom Cintiq 15X interactive pen display allowing direct on-screen input with a pen. 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.

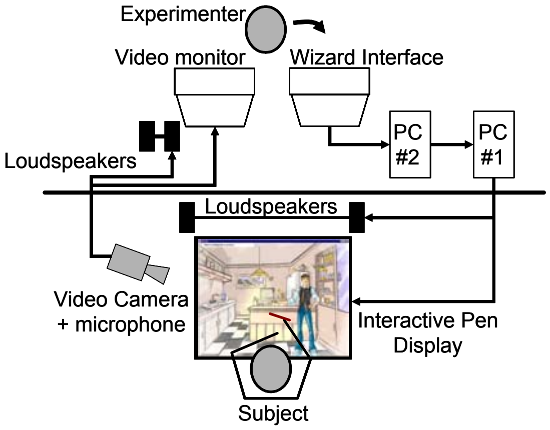

Figure 1: Experimental device.

A digital video camera ensured video and audio recording of the subject's behavior and was connected to a monitor and a loudspeaker in another room. This device let the wizard know what the subject was doing and saying and enabled her to manage the interaction. 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 on PC#2 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 (e.g. Calbris and Porcher 1989). 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.

2.3 Scenario

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

[1] http://www.niceproject.com NICE IST-2001-35293.
[2] http://www.limsi.fr/Individu/martin/research/projects/lea/

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 (ex: three different books). This task requires dialogues with characters.

2.4 Procedure

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.

2.5 Video annotation

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[3] and then imported into ANVIL (Kipp 2001) in which all complementary annotations were made. Three tracks are defined in our ANVIL coding scheme:

- 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[4] (see Figure 2).

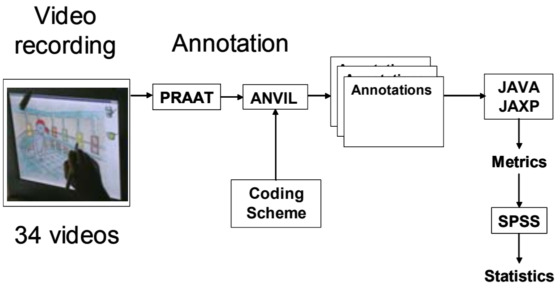

Figure 2: Annotation and analysis process.

2.6 Data quantification and analyses

2.6.1 Unidimensional 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.

2.6.2 Multidimensional analyses

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.

3 Results

We describe the results in this section but we will discuss them globally in the next section.

3.1 Unidimensional analyses

3.1.1 Total duration of scenarios

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.

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

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

3.1.2. Use duration of each modality

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.

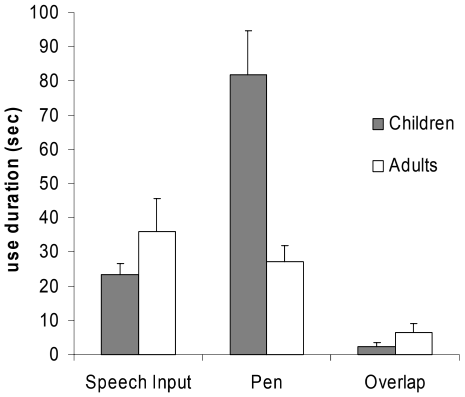

Figure 3: Mean use duration of each modality in the multimodal condition as a function of subjects' age.

3.1.3 Use number of each modality

This dependent variable 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).

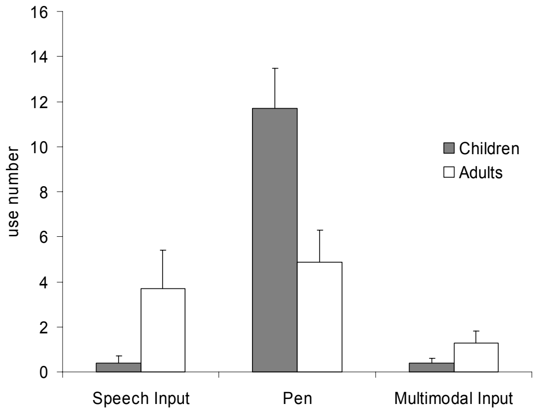

Figure 4: Mean use number of each modality for the "get into a room" command as a function of the subjects' age.

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).

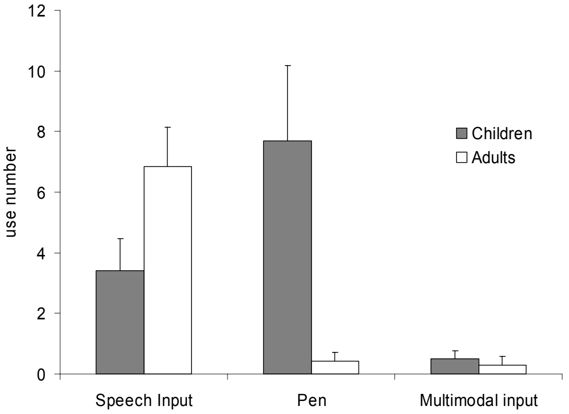

Figure 5: Mean use number of each modality for the "get out of a room" command as a function of subjects' age.

3.1.4 Morpho-syntactic categories

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.

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.

3.1.5 *Shapes of pen movements*

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.

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.

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).

3.1.6 *Subjective data*

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).

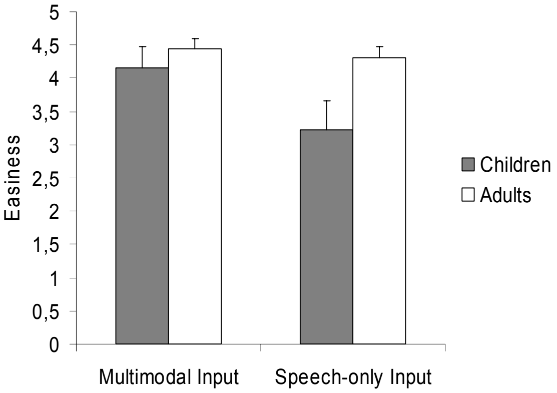

Figure 6: Mean ratings of easiness as a function of the input condition and the subjects' age.

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).

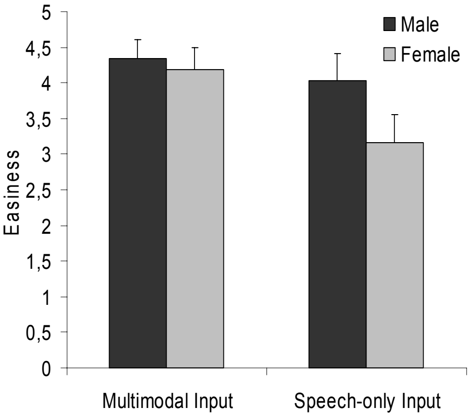

Figure 7: Mean ratings of easiness as a function of the input condition and the subjects' gender.

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

3.2 Multidimensional analyses

3.2.1 Factorial analysis

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.

3.2.2 Multiple regression

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.

4 Discussion

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; 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. For this reason, we intend to have in future similar experiments a third pen-only control condition.

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 data concerning morpho-syntactic categories will be used for controlling the speech recognizer. In this respect, 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). However, our analyses combined speech and pen gestures, making possible for us to use the information collected on the shape of pen movements for developing a multimodal recognition system. The data collected concerning the use of speech and gestures will help improve our HCI design. For example, we now know that certain commands are mainly performed by pen gestures (e.g. "take an object" or "give an object" commands) and others by speech (e.g. for the "ask wish" command).

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.

5 Future work

We intend to carry out further analyses (e.g. spoken disfluencies, speech acts, multimodal language model) on the collected multimodal corpus and to complete this study with new experiments.

The data collected in this study are likely to provide a model of multimodal behavior of children and adults using this kind of ECA game application. Further annotations and analyses could be carried out for this purpose. Cooperation between speech and pen gestures should be further studied: for the moment, the only index collected was simultaneous use of both modalities, but other kinds of cooperation were observed. For example, when subjects were about to take an object by means of pen, they sometimes asked the agent for authorization beforehand. Speech hesitations and disfluencies will also be annotated, in order to control the recognition system. Finally, we intend to define an initiative index (e.g. number of times the subject spoke before the agent, number of questions he asked, etc.). Indeed, children seemed to rarely take the initiative during interaction, and this kind of data could be exploited for building new conversational scenarios. We could also study a way to evaluate believability of agents using quantitative and qualitative measures. The non-verbal behavior of the 2D agents is currently being extended in order to improve the current state of the agent (e.g. multimodal cues for giving turn or willing to take turn behavior, higher level XML specifications of dialog and emotion).

Other experiments will be held with the same evaluation platform. We intend to test the effect of the agents' graphical features (2D vs. 3D) and multimodal personality on behavioral and subjective variables. The novelty effect (Ruttkay et al. 2002) could be taken into account by evaluating the subject's behavior over time and over a series of sessions. The platform might also be useful for experimental studies of trust relations between subject and agents and of the contextual factors that lead the subject to delegate a task to the ECA or to do it by herself.

Although experimental evaluation of individual ECA is still at its early stages, systems involving teams of ECAs have appeared (Traum and Rickel 2002) where the user can press buttons via a touch screen to provide feedback on the team's collective presentation (Baldes et al. 2002) or speak to the agents (Cavazza et al. 2002). Experimental evaluation of the user's multimodal behavior when interacting with such teams of ECA is a future issue that we are willing to tackle within our methodology using different 2D agents and the forthcoming 3D agents.

Acknowledgments

This work was developed at LIMSI-CNRS and funded by the European Project NICE: http://www.niceproject.com/. Our agents were designed by C. Rendu and implemented by

S. Abrilian. The authors wish to thank their partners in the NICE project as well as W. Turner, N. Sabouret and the INTERACT reviewers for their useful comments.

References

Baldes, S., Gebhard, P., Kipp, M., Klesen, M., Rist, P., Rist, T. & Schmitt, M. (2002). The interactive CrossTalk installation: meta-theater with animated presentation agents. *Int. Workshop on "Lifelike Animated Agents: Tools, Functions, and Applications", in conj. with the 7th Pacific Rim Int. Conf. on Artificial Intelligence (PRICAI'02)*, 9-15, Tokyo, Japan.

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., Sullivan, J., Prevost, S. & Churchill, E. (2000). *Embodied Conversational Agents*, MIT Press.

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.

Cavazza, M., Charles, F. & Mead, S. J. (2002). Interacting with virtual characters in interactive storytelling. *First Int. Joint Conf. on Autonomous Agents and Multiagent Systems*, 318-325. Bologna, Italy, ACM Press.

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.

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 Res. Workshop on Spoken Dialog Systems*, pp.153-156, Vigso, Denmark.

Kipp, M. (2001). Anvil - A generic annotation tool for multimodal dialogue. *Eurospeech'2001*, 1367-1370.

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.

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., De Angeli, A. & Kuhn, K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. *Human Factors in Computing Systems (CHI'97)*, 451-422, New York, ACM Press.

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. *Int. Conf. on Spoken Language Processing (ICSLP'2000)*, 877-880, Beijing, China, Chinese Friendship Publishers.

Pelachaud, C., Carofiglio, V., De Carolis, B., De Rosis, F. & Poggi, I. (2002). Embodied contextual agent in information delivering application. *First Int. Joint Conf. on Autonomous Agents and Multiagent Systems (AAMAS'02)*, 758-765. Bologna, Italy, ACM Press.

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 Int. Joint Conf. on Autonomous Agents & Multiagent Systems (AAMAS'02)*, 168-175. Bologna, Italy.

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" in conjunction with ACL/EACL'97*, 39-43. Madrid, Spain.

Traum, D. & Rickel, J. (2002). Embodied agents for multi-party dialogue in immersive virtual worlds. *First Int. Joint Conf. on Autonomous Agents and Multiagent Systems (AAMAS'02)*, 766-773. Bologna, Italy, ACM Press.