Children's and Adults' Multimodal Interaction with 2D Conversational Agents

Stéphanie Buisine [1,2] **and Jean-Claude Martin** [1,3]
{buisine,martin}@limsi.fr

(1) LIMSI-CNRS
BP 133,
91403 Orsay Cedex, France

(2) LCPI-ENSAM
151 bd de l'Hôpital,
75013 Paris, France

(3) LINC-Univ. Paris 8, IUT de
Montreuil, 140 rue Nouvelle France,
93100 Montreuil, France

ABSTRACT
Few systems combine both Embodied Conversational Agents (ECAs) and multimodal input. This research aims at modeling the behavior of adults and children during their multimodal interaction with ECAs. A Wizard-of-Oz setup was used and users were video-recorded while interacting with 2D ECAs in a game scenario with speech and pen as input modes. We found that frequent social cues and natural Human-Human syntax condition the verbal interaction of both groups with ECAs. Multimodality accounted for 21% of inputs: it was used for integrating conversational and social aspects (by speech) into task-oriented actions (by pen). We closely examined temporal and semantic integration of modalities: most of the time, speech and gesture overlapped and produced complementary or redundant messages; children also tended to produce concurrent multimodal inputs, as a way of doing several things at the same time. Design implications of our results for multimodal bidirectional ECAs and game systems are discussed.

Author Keywords
Multimodal Interface, Embodied Conversational Agents, Behavioral Corpus, Adults vs. Children.

ACM Classification Keywords
H5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Input devices and strategies, Interaction styles, User-centered design*.

INTRODUCTION
This study was conducted for a project of a multimodal game system with Embodied Conversational Agents (ECAs). An ECA is an interface represented on the screen by a human or cartoon-like body and aimed at being conversational in its human-like behaviors: generation of verbal and nonverbal output, management of turn-taking, feedback and repair functions, and also recognition and response to verbal and nonverbal input [1]. Our system is being designed for a wide range of users, i.e. children from about 9 years old to young adults. These potential users will interact with ECAs by speech and 2D gesture in both conversational and task-oriented scenarios.

Previous research on multimodal interfaces has shown that a preliminary study of users' spontaneous behavior helps in designing efficient and robust systems [5]. Some reliable patterns of multimodal Human-Computer Interaction have been found, but in general users' behavior will depend on the task that has to be achieved and on the interaction style (e.g. spatial, verbal, numerical [4]).

Regarding multimodal input, our project raises two main issues:

- What characterizes multimodal input with ECAs? We will study the naturalness and social aspects of verbal interaction with ECAs. We will also examine multimodal patterns and compare them to the literature on multimodal interfaces without ECAs.
- Are there differences between children's and adults' multimodal behavior with ECAs?

Neither multimodal interaction with ECAs nor children's behavior have been studied to any great extent. However, Xiao et al. [9] conducted an experiment related to both these issues, in which they analyzed children's multimodal behavior with ECAs in a pedagogical application. Our study provides more data on children's interaction with ECAs, as well as a fine-grained semantic analysis of their multimodal constructions. We think that a semantic analysis of each modality is indeed necessary for the system to perform relevant and accurate fusions (e.g. to remove duplicates from redundant multimodal constructions, or detect conflicts between modalities). Moreover, our study compares for the first time in the literature to our knowledge child and adult behavior in the same application.

We setup a Wizard-of-Oz experiment in which children and adults were invited to interact with 2D ECAs by means of speech and 2D pen gestures. The context of interaction was a game including both conversational and task-oriented activities.

CHI 2005, April 2–7, 2005, Portland, Oregon, USA.
ACM 1-59593-002-7/05/0004.

Figure 1. One user playing the game.

METHOD

Participants

Two groups of French-speaking users participated in the experiment:

- 7 adults (3 male and 4 female users, age range 22 – 38)
- 10 children (7 male and 3 female users, age range 9 – 15).

The two groups were equivalent regarding their frequency of use of video games. An additional adult user was excluded from the analysis because he had guessed the system was simulated.

Apparatus

We used 2D cartoon-like ECAs whose multimodal behavior (speech, lip movements, facial expressions, gaze direction, arm and hand gestures) was specified with a low-level XML language. The 2D graphical display included four rooms, four ECAs and 18 moveable objects (e.g. books, plants…, see Figure 1).

The users could speak and make direct on-screen gestures on an interactive pen display[1]. The Wizard controlled ECAs' behavior (mainly by launching pre-encoded utterances[2] with relevant nonverbal cues specified in accordance with the literature) and the game environment.

All the users were video recorded.

Scenario and Procedure

Each user had to fulfill a wish for three Agents. The first step in the scenario was to meet the Agents and ask them their wish: it basically consisted in bringing them an object missing in their room (e.g. bring a lamp or a book to the Agent in the library). Then the user had to visit the rooms to find the right objects, take them and bring them back to the Agents. We did not give users any indications or instructions about how to interact with the Agents.

At the end of the experiment, users were told that the system was simulated.

Video annotation

The videos were annotated using PRAAT[3] for speech transcription and ANVIL [3] for all remaining annotations (verbal syntax, gesture, commands, social behaviors). Metrics were extracted by an in-house Java software application and submitted to statistical analyses with SPSS[4].

RESULTS

Verbal behavior

Users' utterances averaged 6-words (SD=4.53): there was no difference between adults and children, and no difference in the length of utterances in the speech-only situation.

Each utterance was labeled according to the naturalness of syntax: natural utterances concerned wordings similar to those used in Human-Human conversation. For example, "May I take the red book please" or "Take it" were both labeled as natural, whereas "Take book" or "Out" were judged as non natural. Using this analysis, we found that 91% of users' utterances with Agents would be appropriate to Human-Human interaction. There was no effect of users' age on this percentage.

We also annotated social cues in users' speech: this included politeness ("hello", "please", "thank you"…) as well as feedbacks on Agents' speech or actions: overall there were 2 social cues per minute in users' behavior. Once again this value was influenced neither by users' age nor by the input condition (speech-only or multimodal).

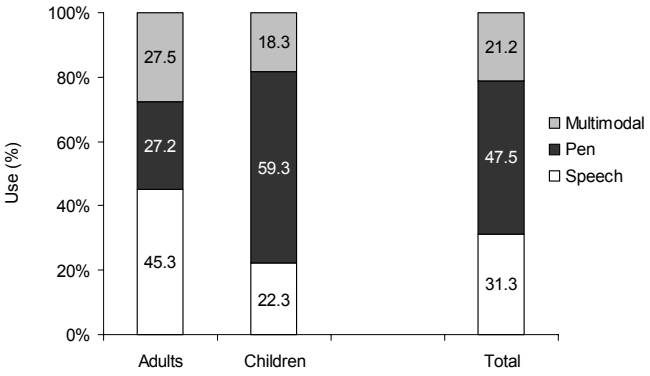

Figure 2. Use of modalities by adults and children, and for the whole sample.

[1] Wacom Cintiq 15X

[2] Synthesized with IBM ViaVoice in French.

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

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

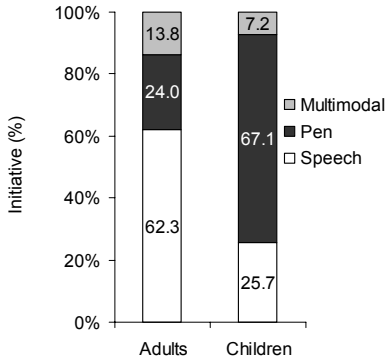

Figure 3. Modalities used for taking initiatives in the interaction.

Use of modalities

Gesture-only was the most widely used modality (47.5% of inputs). Speech-only accounted for 31.3% of inputs and multimodality for 21.2%. As shown in Figure 2, adults used more speech than children ($F(1/15) = 8.24$, $p = 0.012$) and reciprocally, children used more pen gestures than adults ($F(1/15) = 11.73$, $p = 0.004$). The difference was not significant for the use of multimodality.

We studied the modalities used for taking the initiative in the interaction (i.e. talking or acting before the Agent, or changing the topic of a conversation) because it helps modeling the course of the scenario. The modality preference as a function of users' age appeared even stronger in this case (see Figure 3: difference in use of speech $F(1/15) = 9.84$, $p = 0.007$; difference in use of pen $F(1/15) = 12.12$, $p = 0.003$).

Multimodal behavior

We collected 117 multimodal constructions. Most of them (64%) were produced for taking or giving objects: usually users made these actions by gesture and conjointly used speech for notifying or asking the Agents about it. Users thus spontaneously enhanced these task-oriented actions with conversational and social aspects.

Semantic integration of modalities

For the analysis of semantic integration, we distinguished five patterns:

- Redundancy between speech and gesture (e.g. "I would like to go to the pink room, please." + pointing on the pink door).
- "Classical" complementarity (e.g. "Take this." + pointing on a book).
- "Dialogical" complementarity: we created this category for cases in which the user started an action by speech, waited for the Agent's answer or feedback, and ended the action by gesture (e.g. "Can I take a cake?" Agent answered "Help yourself" and the user pointed to the piece of cake). Such constructions are multimodal in the users' viewpoint, but they can be processed as separate inputs by the system and related afterwards.

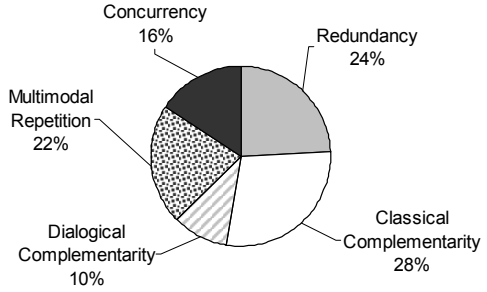

Figure 4. Semantic integration of modalities (in the whole sample).

- Concurrency (e.g. "Hello" + pointing on a cake; or "I want the coffee machine" + exploring the desk, although the coffee machine was on the floor).
- Multimodal repetition, when the user had to repeat the same command but chose to switch modality.

Figure 4 presents the repartition of multimodal constructions according to these categories. Adults and children did not differ regarding the semantic patterns they used, except for concurrent constructions. Indeed, 18 concurrent constructions out of 19 in the corpus were produced by the children.

Temporal integration of modalities

Regarding temporal integration, we adopted Oviatt's categories [6,7]: multimodal constructions are coded as simultaneous when there is an overlap between speech and gesture; otherwise they are coded as sequential.

The study of temporal integration is relevant only for redundant and complementary constructions (concurrent constructions are simultaneous by nature; dialogical complementarity and multimodal repetitions are always sequential): among them, 73% were simultaneous and 27% sequential. These rates are not different for redundant and complementary constructions.

We examined whether users individually adopted a dominant temporal pattern [5]. Ten of our users (5 adults, 5 children) produced at least 2 redundant and/or complementary constructions: 4 adults and 2 children followed a simultaneous dominant pattern (with mean consistency of 88%). Three children had a sequentially dominant pattern (89% of consistency). The remaining adult had no dominant pattern. In 89% of cases, the dominant pattern was predictable, i.e. it was used in the very first multimodal construction.

For sequential constructions, the intermodal lag ranged from 0.1 to 1.8 seconds (mean = 0.6; no difference between adults and children). Gesture preceded speech in 50% of sequential multimodal constructions.

DISCUSSION

One of our goals was to characterize users' multimodal behavior with ECAs. The spontaneous Human-Human syntax we observed seems to be specific to the interaction

with ECAs: previous studies on multimodal interfaces without ECAs describe a rather simplified syntax. Besides, utterances were not shorter in multimodal than in speech-only condition, which is also contrary to what Oviatt [4] observed with a classical multimodal interface. This may be due to the presence of ECAs or to the conversational aspects of our application. Explicit social behaviors (as opposed to unconscious ones sometimes observed [8]) may also be specific to interaction with ECAs.

The frequent use of gesture in our corpus (47.5% of inputs) is surprising regarding the literature on multimodal interfaces (13% in [2]; 17,5% in [7]; 10% in [9]). This may be due to our application domain (a game), to which users may have simply transferred their patterns of behavior: contemporary computer games are usually played by gesture input, and exploration is generally reinforced by visual feedbacks such as rollover effects.

Regarding semantic integration of modalities, the amount of concurrent multimodal constructions in our corpus is interesting because such integration is never reported in the literature. It seems to be specific to children, who may be prone to carry out in parallel the task-oriented and conversational aspects required by our scenario. This finding also confirms the usefulness of semantic analysis for multimodal fusion, because the system must be able to detect such constructions in which simultaneous modalities have to be processed separately and not to be merged.

Our results concerning the temporal integration fits well the patterns reported in the literature (predominance of simultaneity; presence, consistency and predictability of individual dominant patterns). But we did not observe a reliable precedence of any modality, while the literature usually reports that gesture precedes speech.

This body of results, as compared to previous literature, may raise some language and cultural issues, because our users were French, while other studies used English-speaking people. For example, users' utterances in our corpus were slightly longer than in previous studies (6-word length, against 4.8 [4] and 5 [2]): it is difficult to know to what extent it is attributable to the presence of ECAs or to the language of interaction. Likewise, could the amount of gestures and the type of multimodal constructions in our study be partly related to French culture? Cross-cultural experiments are needed to clarify this point.

Another goal of this study was to compare children's and adults' behavior. In this respect, the differences we observed concerned the playing attitude rather than the integration of modalities. Both groups spent the same time on the scenario, but children played it with more gestural exploration and direct actions (sometimes up to concurrent multimodal constructions), whereas adults were more prone to conversation. Children's preference for gestural interaction and initiatives might be linked to their playing habits, and also to a kind of shyness we noticed when they spoke to the ECAs. Thus, speech-only ECA systems might not be so comfortable for young children.

CONCLUSION

The main goal of this study was to collect behavioral data for guiding the implementation of a functional multimodal game with ECAs. This corpus actually helped us implement the semantic fusion algorithm and parameterize the temporal fusion.

Results reported in this paper on verbal and multimodal users' behavior also provided us with some clues for scenario design (e.g. regarding social and conversational behaviors, gestural exploration, or the management of multimodal concurrency). They may also be relevant to practitioners for the conception of systems similar to ours, i.e. intended to young users, with multimodal input, or ECAs in ouput.

ACKNOWLEDGEMENTS

This work was partly supported by the EU/HLT funded project NICE (IST-2001-35293). The authors thank C. Le Quiniou, M. Wolff (Paris-5 Univ.), and B. Turner (LIMSI-CNRS).

REFERENCES

1. Cassell, J., Sullivan, J., Prevost, S., Churchill, E. *Embodied Conversational Agents*, MIT Press, 2000.
2. Hauptmann, A.G. Speech and gestures for graphic image manipulation. In *Proc. CHI'89* (1989), pp. 241-245.
3. Kipp, M. Anvil - A generic annotation tool for multimodal dialogue. In *Proc. Eurospeech'01* (2001), pp. 1367-1370.
4. Oviatt, S.L. Multimodal interfaces for dynamic interactive maps. In *Proc. CHI '96* (1996), pp. 95-102.
5. Oviatt, S. L. Ten myths of multimodal interaction. *Communications of the ACM 42*, (1999), pp. 74-81.
6. Oviatt, S., Coulston, R., Tomko, S., Xiao, B., Lunsford, R., Wesson, M., Carmichael, L. Toward a theory of organized multimodal integration patterns during Human-Computer Interaction. In *Proc. ICMI'03*, ACM Press (2003), pp. 44-51.
7. Oviatt, S. L., DeAngeli, A. & Kuhn, K. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proc. CHI '97* (1997), pp. 415-422.
8. Reeves, B., Nass, C. *The Media Equation.* Cambridge University Press, New York, 1996.
9. Xiao, B., Girand, C., Oviatt, S.L. Multimodal integration patterns in children. In *Proc. ICSLP'02*, Casual Prod. Ltd (2002), pp. 629-632.