Children's Gesture and Speech in Conversation with 3D Characters

Stéphanie BUISINE — LIMSI-CNRS, Orsay, France; LCPI-ENSAM, Paris, France; stephanie.buisine@limsi.fr

Jean-Claude MARTIN — LIMSI-CNRS, Orsay, France; LINC-IUT de Montreuil, France; martin@limsi.fr

Niels Ole BERNSEN — NISLab, University of Southern Denmark; nob@nis.sdu.dk

Abstract

This paper deals with the multimodal interaction between young users (children and teenagers) and a 3D Embodied Conversational Agent representing the author HC Andersen. We present the results of user tests we conducted on the first prototype of this conversational system and discuss their implications for the design of the second prototype and for similar systems.

1 Introduction

The EU project NICE (Natural Interactive Communication for Edutainment) system enables interaction through speech and gesture with the famous author Hans Christian Andersen (Bernsen, Charfuelàn, Corradini, Dybkjær, Hansen et al., 2004) and some characters from his fairytales. The present paper focuses on the part of the interaction that takes place in Andersen's study, where users engage in conversation with a 3D embodied agent representing the author (Figure 1). The Andersen character is capable of conversing with users about himself, his life and his fairytales, about the world in his time, and about some of the interests shared by today's children and teenagers. Users can also explore Andersen's study, select objects to be found there and converse with Andersen about these objects. Conversation with Andersen is in English.

Figure 1: HC Andersen in his study

The Andersen system gives rise to several research issues on multimodal user input, such as: to what extent do users spontaneously combine speech and gesture into multimodal constructions? What kind of temporal and semantic integration of modalities does the system have to manage?

Previous research has shown that the use of multimodal constructions may depend on the interaction style which may be, e.g., spatial, verbal, numerical (Oviatt, 1996) and on the cognitive load related to task difficulty (Oviatt, Coulston & Lunsford, 2004). Some reliable patterns of temporal integration have also been found (Oviatt, Coulston, Tomko, Xiao, Lunsford et al., 2003). However, some features of our system make it difficult to anticipate users' behavior in our particular context: we are developing a conversational application while most of multimodal systems described in the literature are command-based and task-oriented. Moreover, our users address an embodied agent while multimodal systems are usually non-personified graphical user interfaces. Finally, our system is intended for use by a special group of users, i.e., the 9 to 18 years old, whereas users were adults in most of previous research.

Some data are available in the literature on children's spoken or multimodal interaction with an embodied agent (Oviatt, 2000 ; Xiao, Girand & Oviatt, 2002). We also conducted some preliminary experiments before the development of the present system (Buisine & Martin, 2003, in press) but all this data was limited to interaction with 2D characters and Wizard-of-Oz simulations. Here we report on user tests on a quasi-functional prototype, and analyze users' multimodal input behavior in a conversational 3D application.

2 Method

2.1 System

In the NICE system, recognition and interpretation of users' multimodal input is ensured by several modules (Figure 2):

- The Speech Recognition (SR) and Natural Language Understanding (NLU) modules process spoken input.
- The Gesture Recognition (GR) module recognises the shapes of movements. Three shapes were recognized: pointing gestures, circles and lines.
- The Gesture Interpretation (GI) module, whose role is to relate the recognized shape to the 3D graphical environment in order to detect which object(s) the user targeted. However, the GI module detects only some of the objects in Andersen's environment, i.e., the ones defined as referable. The first prototype included 21 such objects, such as pictures on the wall, objects on the writing desk, etc.
- The input fusion (IF) module integrates spoken and gestural inputs. In the first prototype, the IF was based solely on a temporal coincidence criterion. As it has been observed in similar settings that gesture often precedes speech (Oviatt, De Angeli & Kuhn, 1997), the IF, upon receiving a gesture frame, waits for speech during 3 seconds while it does not wait for gesture upon receiving input speech.

The prototype was functional except for the speech recognition module which was wizard-simulated: users' speech was transcribed online and transmitted to the natural language understanding module which processed the transcription and sent the resulting semantic representation to Input Fusion. The IF took care of any temporal fusion with gesture input and sent the result to the Character Module which manages the conversation.

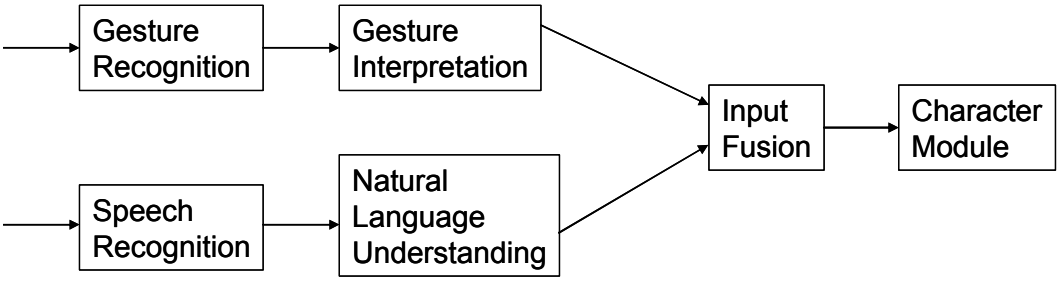

Figure 2: Part of the NICE software architecture

2.2 Participants

The user test involved 9 boys and 9 girls aged 10 to 18 years. They were all Danish except for a 17 years old Scotsman. Most of the test was conducted in English.

2.3 Apparatus

Users were wearing a microphone headset so that they could interact with HC Andersen by speech. They could also directly designate objects through 2D gestures: half of the participants used the system with a tactile screen and half with a mouse. The gesture trace was always displayed on the screen, whatever the input device.

Users could also navigate in the environment by controlling HC Andersen's locomotion with the keyboard arrow keys. Although navigation is not considered as a communicative modality by the system, we have studied this other kind of input and its use during conversation because it may have interesting design implications. In the first prototype version used in the user test, Andersen was not capable of autonomous locomotion.

2.4 Procedure

The application and the interaction devices were introduced in Danish. Users were then invited to engage in free conversation with Andersen. No instruction was given concerning the domains and topics of conversation known to Andersen or on how speak with Andersen.

After 15 minutes of free-style interaction, users were given a list of scenarios of interaction, such as talk to Andersen about his family, or find out which games Andersen likes. They had 20 more minutes for achieving the scenarios of their choice.

Users were rewarded with cinema tickets for their participation.

2.5 Data collection and analysis

For each user, we collected the NLU, GR, GI, and IF log files, which represent more than 9 hours of interaction. We also video-recorded 25% of the tests in order to evaluate the modules by comparing the logged data with users' actual behavior.

We went through the log files by means of a home-made software and submitted the data to statistical analysis using SPSS. The sample of tests for which we had both the log files and the video recordings was manually annotated with Anvil (Kipp, 2001): log files were imported into this tool as annotations so that we could accurately collect system failures (GR, GI, IF modules). The coding scheme we wrote for this analysis enabled us to indicate the occurrence of failures, their cause as well as further comments.

3 Results

3.1 Log files analysis

According to the log files, the test corpus includes 81% speech-only input, 18% gesture-only commands and 1% multimodal input. However, log file analysis also shows that numerous gestures were not processed because they targeted non-referable objects. Taking these non-processed gestures into account, the use of gesture amounts to 39% of the corpus.

The input device proved to strongly influence the use of modalities ($F(1/10) = 3.83$; $p = .08$): in the mouse corpus, speech and gesture were used with equivalent rates, whereas in the tactile screen corpus, speech represented 70% of inputs (Figure 3).

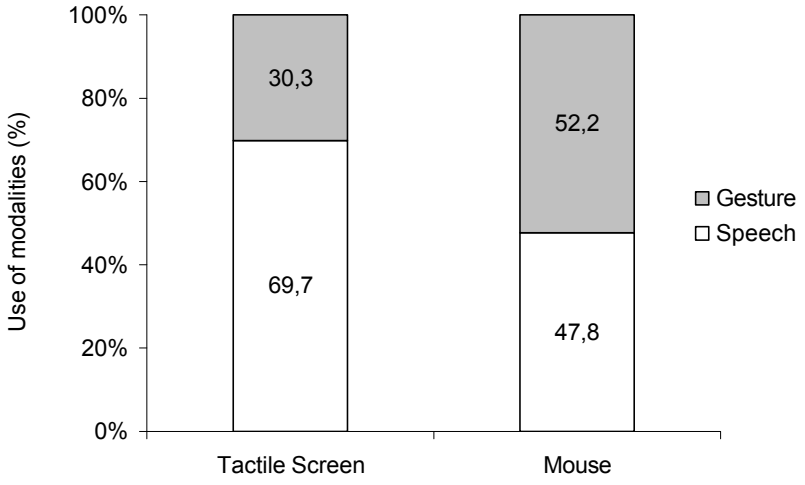

Figure 3: Use of modalities as a function of the gestural input device

The input device did not influence the shape of gestures: most of them were recognised as pointing (44%) or circles (33%). Each logged gesture shape was manually labelled to evaluate gesture recognition and identify unanticipated gesture shapes: this analysis showed that 84% of all shapes were correctly recognized by the GR module. Some noisy surrounding gestures were observed, such as parts of a circle or the contour of a 3D graphical object.

We then studied users' multimodal input behavior. All in all, the log files included 7 multimodal constructions, in which modalities were always semantically unrelated, i.e. addressed concurrent topics. For instance, a user would be talking about the fairytale The Ugly Duckling while

designating the picture of Jenny Lind by gesture; another one would be saying "oh I remember that now" about the previous picture described by Andersen while at the same time gesturing on another picture to get a new description. Subsequent analysis of the video recordings revealed that our IF module actually did not work properly. Indeed, since speech recognition was simulated by a wizard who sometimes had to spend quite some time typing what the user said, all semantic frames from natural language understanding were slightly delayed, which proved incompatible with a fusion system only based on temporal coincidence. The result was that the inputs processed as multimodal by the system were all irrelevant and actual multimodal inputs were missed.

3.2 Video corpus analysis

The joint analysis of log files and video recordings enabled us to:

- Detect and characterize system failures;
- Further study multimodal behavior, e.g., the semantic integration of modalities, which was not implemented in this first prototype;
- Study the use of user-controlled Andersen navigation and its relation to conversation.

These issues are addressed in the following subsections.

3.2.1 System failures

The annotation of gesture recognition (GR) log files evidenced a failure rate of 12.8%. Failures such as recognition of wrong shape had no impact on the processing of user input, because it did not challenge gesture interpretation. On the other hand, the wrong processing of multi-stroke gestures did disturb gesture interpretation: multi-stroke gestures, such as a circle formed by several non-continuous segments or a cross formed by two separate gesture strokes, produced several independent GR frames although they were part of a single input. This sometimes made the system respond several times to what was semantically one and the same input.

The annotation of gesture interpretation (GI) log files showed a failure rate of 25.6%. These failures were mainly due to non-recognition of some referable objects. The video corpus enabled us to identify these problematic objects. We could also list the non-referable objects that were mostly selected by the users, and which may thus have priority when selecting new referable objects for the second Andersen system prototype. We may notice that we did not observe any simultaneous gesturing on several objects in the corpus.

3.2.2 Multimodal behavior

In the annotated videos, we found 61% of speech-only input and 36% of gesture-only input. The multimodal behavior observed in the videos corresponded to 3% of all user input. This included different semantic patterns, such as complementarity, e.g., "what is this?" + gesture on a picture, redundancy, e.g., "I want to know something about your hat" + gesture on HCA's hat, and concurrency, e.g., "How old are you?" + gesture on a vase.

We also observed inconsistencies regarding the plural/singular properties of gestural and spoken behavior, e.g., "tell me about these two" + gesture on a single object. Such examples illustrate the need to consider the perceptual properties of graphical objects in the fusion process and, more generally, the complexity of multimodal input in a conversational system. For instance, a picture

showing two people can be legitimately referred to both in the singular, as in, e.g., “What is this?”, and in the plural, as in, e.g., “Who are these?”.

Regarding the temporal integration of modalities, there were as many simultaneous constructions, i.e., constructions showing temporal overlap between modalities, as sequential constructions with no overlap between modalities. Gesture always preceded speech in sequential constructions.

Table 1 presents details on the 8 multimodal constructions collected in the video corpus.

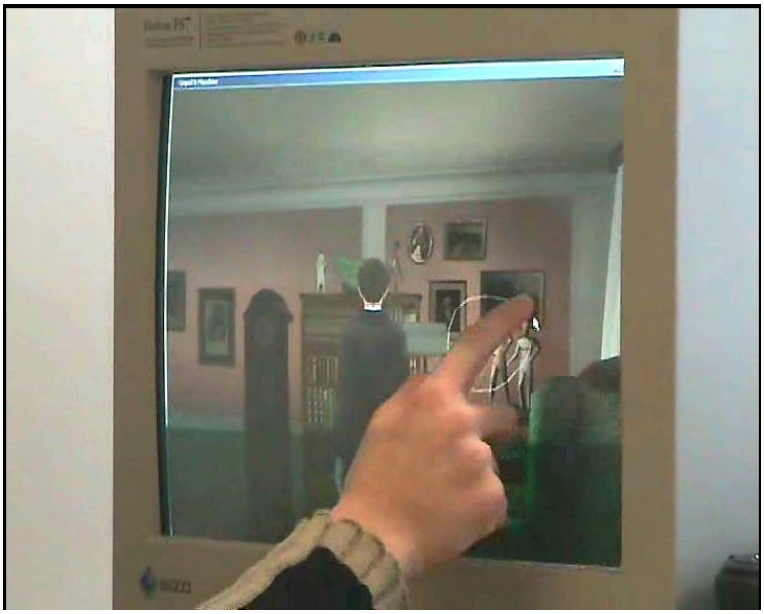

Figure 4: Encircling gesture preceding the question “Do you have anything to tell me about these two?”

3.2.3 Navigation combined with conversation

Since the camera follows HC Andersen, users’ navigation consisted in controlling the locomotion of the agent. The video corpus revealed that the activity of navigating in the environment represented no less than 32% of the total interaction time.

Most of the time, users navigated while Andersen was speaking, but they also often navigated while speaking themselves: 40% of the speech-only inputs in the video corpus were produced during navigation.

Table 1: Multimodal constructions collected in the video corpus. The (1) lag is the delay between end of 1st modality and start of 2nd modality. The (2) lag is the delay between end of 1st modality and end of 2nd modality.

Temporal pattern	**Inter-modal lag (1)**	**Inter-modal lag (2)**	**Object**	**Gesture shape**	**Verbal input**	**Semantic pattern**
Sequential: Gesture – Speech	1 sec.	2 sec.	Picture of Coliseum	Circle	What's this?	Complem-entarity
Simultaneous	X	X	Picture of HCA's mother	Circle	What's that picture?	Complem-entarity
Simultaneous	X	X	Hat	Circle	I want to know something about your hat.	Redundancy
Sequential: Gesture – Speech	2 sec.	4 sec.	Statue of 2 people	Circle	Do you have anything to tell me about these two?	Complem-entarity
Simultaneous	X	X	Statue of 2 people	Point	What are those statues?	Complem-entarity
Sequential: Gesture – Speech	1 sec.	4 sec.	Picture	Circle	Who is the family on that picture?	Complem-entarity
Sequential: Gesture – Speech	0.1 sec.	3 sec.	Picture	Circle	Who's in that picture?	Complem-entarity
Simultaneous	X	X	Vase	Circle	How old are you?	Concurrency

A side effect of the navigation mode was that users had to view Andersen from behind when they wanted to navigate. The extensive use of navigation may have challenged the concept of face-to-face conversation between Andersen and the user. The navigation mode may also have challenged agent believability because the coordination between Andersen's verbal (conversation) behavior and his nonverbal (turn his back to the user) behavior was partly removed. Andersen's non verbal expressivity in terms of facial expressions and gestures also disappeared from the user's visual field most of the time. For example, during the conversation HCA happens to say "Now it is your turn…" and simultaneously point in front of him. This gesture is obviously intended for the user but most of the time it missed its target.

It seems safe to conclude that there is an inconsistency between Andersen's autonomy as a speaking, gesturing, and facially expressive conversational agent, on the one hand, and the fact that the user can determine Andersen's locomotion and orientation vis-à-vis the user. Since Andersen's conversational autonomy is fundamental, the only solution would seem to be to endow him with autonomous locomotion.

4 Conclusions

Speech input was the most widely used modality, which was predictable given the conversational nature of the scenario. However, the gestural input interaction should not be neglected since it represented more than one third of users' inputs (39% in log files, 36% in the video corpus). This rate is impressive given that gestures were weakly reinforced in this test. Since no feedback was given to gestures towards non-referable objects, 65% of all gestures appeared not to be processed by the system. In the post-test interviews, 5 users (3 girls, 2 boys) mentioned that they would like to have more referable objects in Andersen's study. Most other users were happy with the level of gesture affordance present in the first prototype (Bernsen & Dybkjær, 2004). We thus consider that our users showed great interest in using the gesture modality in the conversational context. We had previously observed an attraction of young users towards the use of gesture (Buisine & Martin, 2003, in press) and we assume that it may derive in part from a transfer of behavioral patterns from contemporary computer games which are usually played by gesture input.

The physical device proved to influence the use of gesture: gestural interaction was more frequent in the mouse corpus than in the tactile screen one. This effect may be due to the strong experience users have of the mouse. Some users also seemed to use the mouse like a computer gaming device, i.e. frenetically clicking on graphical objects. Similar "mine-sweeping" mouse use from children sometimes appear during website navigation (Nielsen, 2002). The tactile screen may thus be more relevant for our purpose, i.e. providing a gestural device as part of natural and coherent conversation, in line with the use of 3D gesture in human-human conversation.

Regarding system failures, the user test enabled us to quickly improve our modules and implement some new features. For example, when several 2D gestures are made onto the same object and within the same temporal window, as when doing multi-stroke gestures, the gesture interpretation module merges them into a single semantic gesture. The user tests also highlighted new levels of complexity in the semantic and perceptual dimensions which must be handled in multimodal input fusion.

Some improvements still remain to be done, for example regarding feedback when the user gestures to non-referable objects. Given the extensive use of gesture and the lack of feedback, some users may think that the system is not working properly. We thus modified the communication between modules so that all gestures are signalled to the input fusion module, even if the gesture interpretation module fails to find a referable object. In response to such an event, the system could launch a question, such as "What did you touch?", or a feedback utterance, such as "I do not understand what you want" or "Try to point something else". We could also imagine that the system takes advantage of verbal input in case of a multimodal construction. For instance, "What's this?" + gesture on a non-referable object might elicit the answer "What are you talking about?".

Another area in need of improvement is the Andersen navigation mode. The user test data reveals a paradox in our system, i.e., that Andersen is controlled by the user as an avatar but has autonomous conversational behaviors at the same time. Some alternative navigation modes may be worth studying, e.g., one with two cameras at the same time on the screen, the first camera serving user navigation and the second being focused on Andersen.

Apart from being useful in the design process of multimodal input handling in the second prototype of our system, the results presented in this paper may also be relevant to practitioners working on similar applications, i.e., applications intended for young users, with multimodal speech/gesture input, and/or Embodied Conversational Agents output.

Acknowledgement

This work was supported by the EU Human Language Technologies programme under contract IST-2001-35293. We gratefully acknowledge the support.

5 References

Bernsen, N.O., & Dybkjær, L. (2004). Evaluation of spoken multimodal conversation. *Proceedings of the 6th International Conference on Multimodal Interaction (ICMI'04)*, New York: ACM Press, pp. 38-45.

Bernsen, N.O., Charfuelàn, M., Corradini, A., Dybkjær, L., Hansen, T., Kiilerich, S., Kolodnytsky, M., Kupkin, D., & Mehta, M. (2004). Conversational H.C. Andersen. First prototype description. *Proceedings of Tutorial and Research Workshop on Affective Dialogue Systems (ADS'04)*, Heidelberg: Springer Verlag: Lecture Notes in Artificial Intelligence vol. 3068, pp. 305-308.

Buisine, S., & Martin, J.C. (2003). Experimental evaluation of bi-directional multimodal interaction with conversational agents. *Proceedings of the 9th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'03)*, Amsterdam: IOS Press, pp. 168-175.

Buisine, S., & Martin, J.C. (in press). Children's and adults' multimodal interaction with 2D conversational agents. *Proceedings of the Conference on Human Factors in Computing Systems (CHI'05)*, New York: ACM Press.

Kipp, M. (2001). Anvil - A generic annotation tool for multimodal dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology (EuroSpeech'01)*, pp. 1367-1370.

Nielsen, J. (2002). Kids' corner: Website usability for children. Jakob Nielsen's Alertbox. http://www.useit.com/alertbox/20020414.html

Oviatt, S.L. (1996). Multimodal interfaces for dynamic interactive maps. *Proceedings of the Conference on Human Factors in Computing Systems (CHI'96)*, New York: ACM Press, pp. 95-102.

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

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

Oviatt, S.L., Coulston, R., & Lunsford, R. (2004). When do we interact multimodally? Cognitive load and multimodal communication patterns. *Proceedings of the 6th International Conference on Multimodal Interaction (ICMI'04)*, New York: ACM Press, pp. 129-136.

Oviatt, S.L., Coulston, R., Tomko, S., Xiao, B., Lunsford, R., Wesson, M., & Carmichael, L. (2003). Toward a theory of organized multimodal integration patterns during Human-Computer Interaction. *Proceedings of the 5th International Conference on Multimodal Interaction (ICMI'03)*, New York: ACM Press, pp. 44-51.

Xiao, B., Girand, C., & Oviatt, S.L. (2002). Multimodal integration patterns in children. *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP'02)*, Casual Prod. Ltd, pp. 629-632.