Interactive Narration with a Child: Avatar *versus* Human in Video-Conference

Alexandre Pauchet[1]✉, Ovidiu Şerban[1]✉, Mélodie Ruinet[2], Adeline Richard[2], Émilie Chanoni[2]✉, and Mukesh Barange[1]✉

[1] Normandie Univ, INSA Rouen Normandie, LITIS, 76000 Rouen, France
{surname.lastname}@insa-rouen.fr
[2] Normandie Univ, UNIROUEN, PSY-NCA, 76000 Rouen, France
{surname.lastname}@univ-rouen.fr

Abstract. This article reviews a part of the data collected in a "Wizard-of-Oz" environment, where children interact with a virtual character in a narrative setup. The experiment compares children's engagement depending on the narrator type: either a piloted virtual character or a human in video-conference. The results show that engagement exists, but the modality of the interaction feedback varies in the two contexts.

1 Introduction

Designing a virtual environment, where the participants can interact without any difficulty, is very challenging. Particularly, introducing an autonomous dialogue-based virtual character (or Embodied Conversational Agent -ECA- [1]) increases the expectations of the human participants, up to the point where they can be disappointed by the agent's capabilities [4]. Among the various applications of ECAs, interactive storytelling is a growing scientific field since 2010 [2]. It includes situations from a reproduction of the familiar parent-child narration situation to a new form of user's experience with story generation. Nevertheless, interactive storytelling aims at improving the user's immersion, pleasure, feeling of control, believability of the virtual characters and interaction engagement [7].

A few experiments exist regarding interactive storytelling with children (e.g. [5–7]), but unfortunately they do not characterize standard data in child-agent interaction, such as the average response time (latency) of the child. This article proposes an interactive environment with a virtual character, centered around a familiar story telling activity so that the children feel comfortable. As the dialogue component of ECAs remains a technical difficulty [3], our environment is based on the Wizard of Oz (WoZ) paradigm, so that the collected data expresses what can be expected of "natural" interaction. We aim at answering the following research question: Is child-agent interaction different from child-adult interaction? This is studied from the interaction engagement perspective. We propose an experimental study to compare child ↔ avatar[3] interaction *versus* child ↔ human in video-conference.

[3] In the following, '*avatar*' refers to a virtual character driven during a WoZ experiment. A '*virtual character*' can be either an avatar or an ECA.

2 Experimental Study: Avatar *vs* Video-Conference

2.1 WoZ scenario for interactive narration with children

The chosen story is "The lost ball", illustrated by 15 images to support the narration. The narration is constructed as a sequential scenario. Several parallel branches are integrated to give the illusion of an open story, although all the children's answers generates the same comments or explanations. For example[4], "*Oh god, where will the ball fall? Do you know it?*" is used to induce an interaction that always lead at the end to the following statement: "*Booyah! Look at the ball! It's stuck on the roof!*. Moreover, as a dialogue is never completely predictable, a set of free-context utterances has been added. It consists in a series of statements not directly linked to the context of the story, such as: "*OK*", "*You are right*", "*Shall we continue?*". They can be used to manage the dialogue and force the interaction to focus back on the story.

Six interactive errors were included to assess the children's attention ($A1$ and $A2$), emotional understanding ($E1$ and $E2$) and comprehension ($C1$ and $C2$). For example: 1) at $C1$, the narrator makes a semantic mistake by saying "the boy throws his *carrot* on the roof", instead of *boot*; 2) at $E2$, the narrator makes an emotional error: he says "oh, look at the teacher. Is he singing?", while he is currently shouting at the children; and 3) $A1$ presents a joint attention problem, with a black screen during 3s, while the narrator is describing the scene.

The story is split into two parts, narrated by two different actors, which enables the cross comparison of the interactions during the two conditions.

2.2 Participants and Procedure

20 children (6-8 years, average: 7.7) participated in this study and each session lasted approximately 20 minutes. 90% were familiar with animated virtual characters and 60% with occasional web-cam usage.

Various dependent variables were collected: the numbers of words, phrases and words per phrase; the mean disfluency rate for hundred words; the number of long pauses ($> 2s$); the response delay (latency) between avatar/adult's utterances and child's utterances; the number of out-of-context interactive phrases used by the narrator; the number of Emotional Mimics (EM - laughs, smiles, pouts,...) of the child, the number of Spontaneous Verbal Responses (SVR), as any verbal interaction initiated by the child and the number of Expected Response after a direct Verbal Question (ERVQ) from the narrator.

2.3 Results: child engagement and modality of interaction

Table 1 provides some quantitative measures of child engagement in interaction. Children use more words and longer sentences when interacting with the avatar than with the human in video-conference ($t_{words} = 0.883, p = 0.681 > 0.05$, *ns*

[4] All the presented utterances are translated from French.

Table 1. Children's engagement in interaction.

	Words	**Phrases**	**Lengths**	**Ctx**	**Disf.**	**Pauses**	**Latency**	**EM**	**SVR**	**ERVQ**
Video-conf	62.7	20.9	2.9	5.9	7.7	0.5	1955	11	8	19
Avatar	74.8	22.0	3.1	6.5	6.0	2.5	2182	4	17	14

Words: mean number of words; *Phrases*: mean number of phrases; *Lengths*: mean number of words by phrases; *Ctx*: the number of out-of-context interactive phrases used by the narrator; *Disf.*: mean disfluency rate for 100 words; *Pauses*: mean number of long pauses ($> 2s$); *Latency*: response delay (in ms) between avatar/adult's utterances and child's utterances. After an interactive error, *EM*: number of emotional mimics; *SRV*: number of spontaneous verbal responses; *ERVQ*: number of expected responses after a direct verbal question.

and $t_{phrases} = 0.846, p = 0.359 > 0.05,\ ns$). When interacting with the avatar, the average size of the child's sentences increases slightly ($t_{length} = 0.445, p = 0.881 > 0.05,\ ns$). Moreover, the number interactive phrases "triggered" by the narrator after an out-of-context utterance from the child enables to evaluate the children's spontaneous interaction. The results show that the avatar pronounces more additional sentences than the human in video-conference ($t_{out-of-context} = 0.653, p = 0.545 > 0.05,\ ns$).

Concerning the quality of the oral interactions, the disfluency rates shows that the children use more disfluencies and shorter pauses with human in video-conference than with the avatar ($t_{disfluencies} = 1.153, p = 0.277 < 0.05,\ ns$ and $t_{pauses} = 1.775, p = 0.076 > 0.05,\ ns$). Concerning the high number of pauses with the avatar, we verified that this was not due to a lack of attention from the child in the post-experiment survey. All the participants have correctly answered when asked to describe some specific elements of the story. Finally, children have a longer latency when responding to the avatar compared to the human in video-conference ($t_{latency} = -0.741, p = 0.468 > 0.05,\ ns$). In each case, the latency is far higher than the known standard for children, even for complex questions. Therefore, the fact that the interaction is mediated seems to impact more the latency than using a virtual agent.

When focusing on the stimulated interactive situations, Table 1 also includes the children's reactions after an interactive error (EM, SVR and ERVQ). Our analysis shows that, after an interactive error, children communicate more with the human in video-conference than with the avatar ($EM + SVR + EVRQ$). However, they react more spontaneously with the avatar ($EM + SVR$), ($h^2 = 0.02,\ p = 0.362 > 0.05,\ ns$). Children also address the situation differently depending on the narrator: they prefer to communicate spontaneously with the avatar by verbal responses (SVR) rather than non verbal (EM) (significant, $p = 0.014 < 0.05$). Moreover, they also prefer to interact with the adult in video-conference using non verbal res¡ponses (significant, $p = 0.009 < 0.05$). These results are consistent with our previous results on verbal answers: the children use less words, less and shorter phrases with the video-conference than with the avatar.

2.4 Sum-up of the Results and Discussion

It seems that children are able to adapt to a virtual character and engage in the interaction: the number of children's interventions with the avatar does not statistically differ from that with the video-conference. However, communication with the avatar appears more spontaneous, more verbal and does not seem as natural as with the human, as they try to adapt their discourse to their interlocutor. We have also noted a less hesitant, clearer and more assured speech from the children toward the avatar, confirmed by the disfluency rate measures.

The fact that the verbal modality is preferred with the avatar suggests that this modality is of particular importance for multi-modal interactive systems. Unfortunately, the transcription remains one of the biggest problems in ECAs, due to transcription time and errors.

3 Conclusion

In this article, we have presented a narrative WoZ experiment that compares child $\leftrightarrow$ avatar interaction with child $\leftrightarrow$ human in video-conference. The main result is that children were engaged in narrative interaction with a virtual character in a different way but not less valuable than with a human in video-conference. In other words, any ECA dedicated to child-agent narrative interaction have to be designed so that the verbal understanding should be implemented with great care as the children seem to favor this modality.

As the children seem to have enjoyed their experience, using an avatar driven in a WoZ or in an ECA can be of great interest to psychologists. Dialogue or interaction models could be designed and tested to evaluate, for instance, children's language acquisition. Moreover, the particularities of a virtual narrator could also be exploited from a therapeutic point of view, such as children with communication difficulties, offering rehabilitation tools adapted to their difficulties of interaction.

References

1. Cassell, J.: Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. Embodied conversational agents pp. 1–27 (2000)
2. Crawford, C.: On Interactive Storytelling. New Riders Games (2013)
3. Kopp, S., van Welbergen, H., Yaghoubzadeh, R., Buschmeier, H.: An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing. Journal on Multimodal User Interfaces 8(1), 97–108 (2014)
4. Mori, M.: The uncanny valley. Energy 7(4), 33–35 (1970)
5. Oviatt, S.: Talking to thimble jellies: Children's conversational speech with animated characters. In: Proceedings of ICSLP'00. pp. 67–70 (2000)
6. Ryokai, K., Vaucelle, C., Cassell, J.: Virtual peers as partners in storytelling and literacy learning. Journal of computer assisted learning 19(2), 195–208 (2003)
7. Theune, M., Linssen, J., Alofs, T.: Acting, playing, or talking about the story: An annotation scheme for communication during interactive digital storytelling. In: Proceedings of ICIDS'13. pp. 132–143 (2013)