Interactive narration with a child: impact of prosody and facial expressions

Ovidiu Șerban, Mukesh Barange, Sahba Zojaji, Alexandre Pauchet, Adeline Richard, Emilie Chanoni

Interactive Narration with a Child: Impact of Prosody and Facial Expressions

Ovidiu Șerban, Mukesh Barange
Sahba Zojaji, Alexandre Pauchet
Normandie Univ, INSA Rouen Normandie, LITIS
76800 Saint-Étienne-du-Rouvray, France

Adeline Richard
Émilie Chanoni
Normandie Univ, UNIROUEN, PSY-NCA
76000 Rouen, France

ABSTRACT

Intelligent Virtual Agents are suitable means for interactive storytelling for children. The engagement level of child interaction with virtual agents is a challenging issue in this area. However, the characteristics of child-agent interaction received moderate to little attention in scientific studies whereas such knowledge may be crucial to design specific applications.

This article proposes a Wizard of Oz platform for interactive narration. An experimental study in the context of interactive storytelling exploiting this platform is presented to evaluate the impact of agent prosody and facial expressions on child participation during storytelling. The results show that the use of the virtual agent with prosody and facial expression modalities improves the engagement of children in interaction during the narrative sessions.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; *Human computer interaction (HCI)*;

KEYWORDS

Child-Agent Interaction, Interactive Narration, Prosody, Facial Expression, WoZ

ACM Reference Format:
Ovidiu Șerban, Mukesh Barange, Sahba Zojaji, Alexandre Pauchet, Adeline Richard, and Émilie Chanoni. 2017. Interactive Narration with a Child: Impact of Prosody and Facial Expressions. In *Proceedings of ICMI '17*. ACM, Glasgow, United Kingdom, 8 pages. https://doi.org/10.1145/3136755.3136797

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) [8]) increases the expectations of the human participants, up to the point where they can be disappointed by the agent's capabilities [24]. Building such an environment is even more difficult when young children are involved, since they are still developing their linguistic and interaction competencies. Providing children with a non-disturbing environment which offers natural reactions from the included ECAs, becomes critical.

Among the various applications of ECAs, interactive storytelling is a growing scientific field [12]. It includes situations that corresponds from a reproduction of the familiar parent-child narration up to a new user experience with story generation according to the interaction. Nevertheless, interactive narration aims at improving the user's immersion, pleasure, feeling of control, believability of the virtual characters and interaction engagement [46].

A few experiments exist regarding interactive storytelling with children (e.g. [31, 39, 46]), but unfortunately they do not enable to determine and characterize standard data in child-agent interaction, such as the average response time (latency) of a child interacting with an ECA. In this article, we propose 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 remain a technical difficulty [18, 44], we propose a Wizard of Oz (WoZ) platform in order to constitute desired virtual environment and interaction scenario to evaluate preliminary components. Thus, our environment is based on the WoZ paradigm, so that the collected data expresses what can be expected of 'natural' interaction. We aim at answering the following research question: Do different modalities of virtual agent have any impact on child-agent interaction? This question is studied regarding interaction engagement. We therefore propose an experimental study to compare the effects of different agent modalities (e.g. prosody and facial expression) on the child ↔ avatar[1] interaction during interactive storytelling sessions.

This article is organized as follows. The next section focuses on related work regarding the field of interactive narration with children and some existing WoZ experiments; then, we describe the WoZ platform used for the experiments as well as the narrative scenario; the section labeled "Experimental Studies" describes in detail the results obtained during the conducted experiment; finally the last section provides some concluding remarks.

2 RELATED WORK

Face-to-face interaction between a user and an interactive virtual character introduces the concept of ECA [8]. Unfortunately, the conversational skills of existing ECAs remain limited [18, 44]. As the expectation toward a virtual character's capabilities depends on the level of details in the animation, the user is usually deceived by the real ECA's abilities and the interaction becomes very unnatural. This phenomenon, known as the "uncanny valley" [24], affects the user's empathy towards the virtual agent and therefore the user's

[1] 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.

ICMI '17, November 13–17, 2017, Glasgow, United Kingdom

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3136797

experience [4]. To overcome this issue, the interactive agent must be designed so that it reacts according to the user's frustration [16], shows empathy [29, 35] and responds to the situation at the appropriate moment [37].

The more general influence of a virtual character on human perception is formalized as the controversial "persona effect" [21]. Pedagogical studies [25] and Serious Games [36] have shown a link between the presence of an animated character and the children's performances, whereas Miksatko et al. [23] conclude that no such impact exists. Finally, a WoZ experiment has shown a strong influence of an ECA on the performance of autistic patients [14].

In the following, we focus on child-agent interaction with an interactive narration perspective and the WoZ paradigm as a solution to overcome the current limitations of ECAs. Measures of user engagement in interaction are explored as an evaluation of ECAs.

2.1 Child-agent narrative interaction

Digital interactive storytelling proposes a new form of narration in which the user plays a direct role in the story and can interact with the narrator or with the story characters [12]. Interactive storytelling frameworks have been designed with virtual characters (e.g. [9]) and robots (e.g. [20]) following the same objectives: to establish a natural interaction with a believable artificial storyteller or characters, increase the user's interaction engagement and therefore propose a pleasant user experience. The design of an expressive storyteller becomes both a possibility and a necessity [46].

There has been significant work in the child-agent narrative interaction. Fearnot! system uses a narrative interactive system to educate on bullying issues [2], and the MIXER system extends their work to the understanding of multiculturalism with virtual agents [3]. Furthermore, Porteous & al. have proposed an interactive narrative system to support cognitive psychology experiments in story understanding [34]. However, in the child-agent interaction context, the influence of the agent has not been well studied yet. An experiment realized by S. Oviatt has shown that children, aged between 6 and 10, use a more fluent discourse with less irregularities in their speech (called "*disfluencies*") when speaking to an agent compared to an adult [31]. Moreover, the children are very intrigued by their new partner and accept more easily the dialogue interaction. Ryokai & al. also proposed a study using an interactive ECA named Sam, in a tutoring scenario [39]. This study illustrates the efficiency of the social implication of Sam toward children, enabling them to acquire new words and complex linguistic structures.

Similar studies were conducted with robots. The same level of engagement was observed with autistic children [19], in tutoring context [15] or precocious cognitive development [47]. The potential is similar for both types of interlocutors, wether in educational or narrative studies. Studying the conversational, educational or narrative experiment context with a robot is rather complex, therefore using a virtual character offers a favorable alternative.

2.2 Wizard of Oz methodology

ECA are composed of a collection of components (speech-to-text, user's affect recognition, dialogue management, animation player, ...) interlocked in such a way that it is generally impossible to calibrate and evaluate them independently. Moreover, the existing dialogue components remain insufficiently robust [44] and current interactions still appear too rigid and jerky, as ECAs can only interpret a user's utterance once it is completed to be pronounced [18]. To tackle these issues, an iterative methodology is often adopted: numerous research groups begin their baseline design using the WoZ paradigm, in order to constitute an initial interaction corpus and/or to evaluate any preliminary dialogue component (e.g. the SEMAINE project [22]). This methodology can also be reused afterwards to improve an initial model. The WoZ paradigm is mainly exploited to evaluate multi-modal interactions [1], as it enables the collection of data in a controlled environment. During a WoZ experiment, the user believes that he is interacting with a complex and autonomous system, while he is currently interacting with another human piloting the agent. This gives the illusion of a natural interaction with an AI system.

The Children Interactive Multimedia Project (CHIMP) [28] is one of the first design guidelines using a WoZ methodology to design multimodal interfaces for children. Later, DiaWOZ-II proposes a text interface for a tutoring system. [48] creates a web interface to simulate dialogue scenarios in a restaurant. [27] includes a dialogue model based on a finite-state machine, driven by a pilot, with the possibility to add more states at run time. The SEMAINE Project also started with a Wizard of Oz set-up [22], which allowed them to design and test the interaction model exploited in the final version. More recently, Yildirim et al. [49] describes a WoZ methodology to detect emotional features on children, while interacting with an agent. Similarly, Chaspari et al. [10] measured the enjoyment based on acoustic features on autistic children. In the same context, Rachel ECA [26] is a system used to analyze the nature of interaction between an autistic child, his parent and the agent.

Unfortunately, most of the WoZ experiments cannot be easily reused in a different context [13, 30].

2.3 User engagement in interaction

User engagement is an indicator of interaction success [6]. In dialogical interaction, a virtual agent must create user engagement to maintain an enjoyable experience [32]. User engagement in dialogue can be defined according to two main characteristics: 1) "*the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction*" [32] and 2) "*the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake*" [41]. During human-agent interaction, user engagement is maintained with communicational signals [38]. They include for instance smiles, mutual gazes, shared glances [38] and disfluencies [31] as synchronization signals. The use of such signals is necessary but not sufficient to support user engagement. With no affective and interaction feedback, the speaker can be uncomfortable and disengages, and the conversation loses its natural interactivity. Anyway, a measure of the user engagement in dialogue can be the number of words, sentences, mimics and gestures performed during interaction [5].

Response latencies have been used in numerous studies to measure cognitive effort (e.g. [42]) or user engagement [5], but no precise data is given. The delay between utterances for a natural conversation between adults is [100ms–600ms] [11]. Stivers & al. analyze more precisely adult response latency in a yes-no set questions that confirmed this hypothesis [43]. Only few works

focus on child latency. Casillas & al. studied child turn-timing and found effects of linguistic processing on response delay. The answer complexity (type of information and grammatical complexity) impact the child reponse latency. With complex answers, children had longer latencies: simple yes-no answers generally had shorter latencies (442ms) than wh- answers (765 ms). When children had more material in their answers, they had longer latencies: complex answers increase the response latency up to 587ms for a yes-no complex answer and 948ms for a wh-complex answers [7].

2.4 Sum-up and discussion

Frameworks dedicated to interactive narration, in which the user interacts naturally with a virtual storyteller, are becoming a hot topic. The design of a narrative ECA able to interact efficiently with a child during a storytelling session is therefore of great interest, as children are obviously the prime target for such applications. To assess the interaction performance of such an ECA, baseline data must be known for comparison. Unfortunately, most of the existing child-agent experiments do not provide sufficient information concerning the interaction characteristics between a child and a virtual character (e.g. [5, 7]).

In the following, we focus on child-agent interaction when the agent is a virtual narrator during interactive storytelling sessions. Our objective is to compare child-agent interaction in term of user engagement. We therefore present an experiment designed to collect and analyze a corpus of child interactive narrations with a virtual character. To alleviate the "uncanny valley" difficulty, we design our own WoZ platform and interactive narration scenario. We focus on response latency, disfluency rate, number of words and phrases, and number of smiles as measures of child engagement during interaction. In this experiment, we aim at evaluating the impact of facial expression and prosody in child engagement while interacting with a narrative ECA.

3 EXPERIMENTAL SET-UP: A NARRATIVE WOZ

In this section, the storytelling scenario and the narrative (WoZ) platform are introduced. The platform is designed to collect child-avatar interaction data.

3.1 WoZ scenario for interactive narration with children

To collect a proper corpus of 'natural' interaction of a child with a virtual character, three options can be considered: 1) a completely open dialogue set-up; 2) a non-linear scenario, with a story adapted to each participant; and 3) a fixed scenario with timings and gestures synchronized according to the child's reactions. The first option is challenging due to the current transcription errors and dialogue management [31]. The second set-up requires multiple pilots to perfectly synchronize the story with the emotional feedback, gestures and speech management according to the child's reactions. Our final choice is therefore a fixed branched scenario, augmented with free-context outputs adapted to unpredictable situations. Thus, the cognitive load of the pilot should be lowered, enabling him to concentrate on the child's reactions rather than on the scenario and on the pilot HMI.

The story chosen for this experiment is "The lost ball", that describes a school boy who plays with his ball before the class. The ball is kicked on the class roof and the boy and his friends try to recover the ball by throwing a boot, a school bag and a scarf. When they enter into class, the ball is not recovered and the various objects are still on the roof. Finally, during the class, a huge storm blows all the things off the roof, enabling their recovery. The story is illustrated by 15 images, offering a good level of details to support the narration.

The narration is constructed as a sequential scenario. Several parallel branches opened with questions are integrated in order to give the illusion of an open story, although all the child's answers generates the same comments or explanations. For example[2], "*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.

3.2 WoZ platform

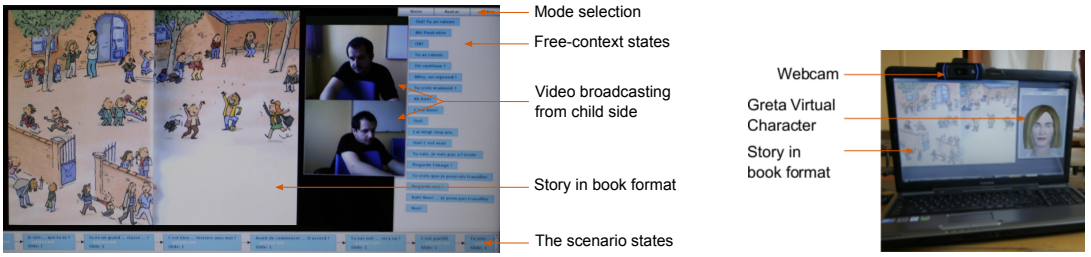
Our system, called OAK (Online Annotation Kit), extends the SEMAINE project [22, 40], embedding new components to fully control the architecture. Thanks to a video-conference system, OAK enables the wizard to see the child's activity and interact with him by piloting an avatar. OAK is therefore based on three major elements: 1) the SEMAINE Platform [40], which proposes a component-based communication system; 2) the Greta virtual characters [33] which are part of the SEMAINE project; and 3) OAK, which consists in a pilot graphical interface (see figure 1(a)), and the system view at the user level (see figure 1(b)).

The interface presented figure 1(a) is used by the pilot. It has a scenario area, which presents the whole collection of possible states. A state can be executed at any time, as often as necessary. On the right, the free-context library is displayed. On top, a menu that allows the selection of the experiment mode is present: *video*[3], *avatar* or *none*. In the later case, only the story slides are displayed.

The child interface consists of a representation of the narrative virtual agent using a Greta virtual character as avatar and the story image (figure 1(b)). We used as virtual agent a humanoid "talking head" that can display emotional facial expressions. It comes from the Greta system [33] of the SEMAINE project [40]. SEMAINE provides four sensitive artificial listeners with different emotional traits: Poppy (happy and positive), Obadiah (gloomy and sad), Spike (argumentative and angry) and Prudence (pragmatic and sensitive). Poppy was selected as the more suitable to interact with children in a storytelling context. According to the describing criteria proposed in [4] for an affective animated agent, in our experiment Poppy is: a 3D semi-autonomous animated face (representation of the agent) that display basic emotions (types of emotions expressed) in accordance with its prosody and the utterance (representation of emotional expression). The transition between two emotions

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

[3] OAK also offers a video-conference mode which is not exploited in the experiment presented in this article.



(a) The pilot view of OAK (b) The user's view of OAK: a narrative avatar

Figure 1: The human-machine interfaces of OAK

exposed is smooth, as the agent dynamically compute intermediary states (method for expression of emotion). All in all, our virtual character provides positive back-channels during the interaction using smiles, blinks and head nodes.

The story in book format is the common element of all the views (avatar view in the user's interface, as well as in the pilot's HMI). The images are synchronized among the pilot and user sides. To enable deictic gestures, the pilot can use the mouse to point at the story images.

The video stream from the child's side is sent to the pilot view and recorded simultaneously. The video conference set-up exploits multiple communication channels, built with the GStreamer [45] toolkit.

All the components of OAK are fully customizable, with independent XML based configuration files. The actions are translated into BML [17] code by an action interpreter and forwarded to the avatar. One can note that the Greta component can therefore be replaced by any virtual character or robot designed to interpret BML.

3.3 Examples of interaction

Figures 2 and 3 present some interactions extracted from the collected corpus to illustrate how the scenario evolves for standard answers or for unexpected responses from the child. In figure 2, the pilot drives the interaction following the main scenario. When reaching the question, he continues according to the reaction of the child (here, the first statement). As a scenario cannot predict all the possible interactions, out-of-context answers can be triggered to focus on the story. In figure 3, the child spontaneously speaks and therefore the pilot activates the most appropriate answer.

OAK: Oh! no! He had thrown his carrot on the roof!
OAK: Ummm ... I'm not sure ... this is called a carrot, right?
CHILD: No. It's a boot
OAK: [*In case of a valid given by the child*] Ah! Yes! You're right, this is a boot!
OAK: [*In case of an no answer or invalid answer*] Something's wrong. I think this is a boot.

Figure 2: Standard scenario with two narrative choices

OAK: And the bell rings the end of the playtime.
CHILD: But the boot is on the roof!
OAK: [*Out of context answer*] True! You're right!

Figure 3: An out-of-context answer from the narrator

4 EXPERIMENTAL STUDY: IMPACT OF PROSODY AND FACIAL EXPRESSIONS

This section presents the experimental study carried out to assess the impact of facial expressions and prosody on child-agent interaction, using the previously described WoZ platform.

4.1 Method and material

The objective of this experimental study is to compare the effects of different agent modalities of behavior on child-agent interaction in the context of an interactive storytelling, using the previously described WoZ platform. We have considered three different experimental conditions. These conditions are based on the presence or absence of the facial expressions and prosody of the avatar in the aforementioned context. Therefore three different versions of the avatar were tested (NoM: no mimics, NoP: no prosody or MP: mimics and prosody) for an interactive storytelling session. . For the purpose of this experiment, we define the *No Prosody* context as the absence of any rhythm, stress or intonation on the agent's utterances. In this context, the agent has a monotonous voice, by disabling all the voice inflections in the speech synthesizer. Similarly, the *No Mimics* context suppresses any facial and head gestures from the agent's animation library. In this context, the agent has only lip movement to animate the speech, but no head movement, smiles or any other facial gestures. These conditions were selected to make the interaction as pleasant as possible

In order to evaluate children engagement with virtual avatar, we have defined following two hypotheses:

- **Hypothesis-1:** Virtual avatar with *Prosody* and *Facial expression* modalities improves the engagement of children in interaction during the narrative sessions.
- **Hypothesis-2:** *Prosody* is more important than *facial expression* as a modality of the virtual avatar for the interaction with children during narrative storytelling.

The second hypothesis is particular to child-agent interaction. Prosody is a pragmatic language acquisition indispensable to adapt its linguistic behaviors to context. The use and understanding prosody development starts in utero and continues from 5 to 13 year old. Concerning facial expressions, although the recognition of emotional facial expressions is acquired between 2 and 3 years, the ability to infer simple emotional states from the facial expression does not emerge before 3 years. It is only at about 6-8 years of age that children succeed in differentiating and naming expressions of surprises. These elements are in favor of an importance

of prosody at an earlier stage of social language acquisition than facial expression understanding.

4.1.1 Participants. Since the aim of a narrative virtual agent is to tell the story to children in an interactive manner, we recruited 50 children (6 to 11 year-old) from an elementary school. These children were all native French speakers and were placed in front of the avatar for a narrative session with the basic narrative scenario (i.e. without any interactive error). An overlapping population between condition could not be setup due the age of the participants and their lack of attention to perform under strict directions.

4.1.2 Data Collection. In order to evaluate the effects of different modalities of agent on children, various dependent variables were collected: the numbers of words, phrases and words per phrase, collected from manual transcription; the mean disfluency rate for hundred words; the response delay (latency) between avatar's utterances and child's utterances; 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. All these features were manually annotated on the videos collected during the experiment using Oviatt's [31] protocol.

The child may decide to use gestures, mimics or verbal responses to provide an answer to a question. The feedback may be quite complex and, in practice, all these modalities are combined. For example, the children were often smiling or laughing when told that the boot landed on the roof of the building. All these points in the story line are "diversion", designed with the sole purpose to generate emotional responses from the user. We considered that the smile begins during the hint of the smile until the total relaxation of the large zygomatic generating a total disappearance of the smile. Smile is a social act which, we believe, reflects a desire to show cooperation in interaction, but also a satisfaction reaction to the comic characters of history.

4.1.3 Procedure. The evaluation process involves three steps. In the first step, before the experiments, we obtained individual parental consent under the constraint of retaining children children's anonymity. Furthermore, these experiments were supervised by the local school personnel. The experiment was conducted in two rooms. In the first room the child was placed in front of a laptop screen and in the other room the pilot was installed with another laptop that could control the narrative story thanks to our WoZ platform (see fig 1). In the second step, children were told that a lady will tell the story to them and then they listen the interactive story "The lost ball" which has been described in the section 3.1. Each participant performed the experiment with the avatar having one of the three controlled conditions:

1) avatar with full modalities (MP)
2) avatar with no prosody (NoP)
3) avatar with no mimics (NoM).

After the experiment, children follow one post-hoc interview where they can give their opinion and comments.

4.1.4 Design Analysis. In this study, since we have 50 children, we randomly associated them with one of the evaluation conditions, therefore 16 children with the avatar having full interaction modalities (MP), 17 children with the avatar having no mimics (NoM) and the other 17 children with the avatar having no prosody (NoP) experimental conditions.

After the experiments, the data was manually transcripted from the video recordings of experiments (video recordings of user view and the recoding of participant during interaction) provided by our WoZ platform. Furthermore, these videos were manually processed to analyze the facial expressions (number of smiles) of each child during the narrative interaction. Given the multiple dependent measures, we use one-way analysis of variance (ANOVA) to determine whether the mean of a dependent variable is the same in two or more unrelated, independent groups. Furthermore we apply post hoc tests (Bonferroni correction) to identify which specific groups were significantly different from each other. The significant level for all of the analysis was set to 0.05.

4.2 Results: child engagement in interaction

The goal of the experiment is to assess the engagement of the children in three different modalities of the agent behavior.

First, we want to evaluate analyze the facial expression (smile) of the children during the interaction. A univariate ANOVA has been run on the children's smile in function of the condition type of avatar (i.e. NoM, NoP and MP). There is a significant effect of the type of avatar on the number of children smiles F(2,49)=8.241; p= .001). A post hoc comparison (Bonferroni) shows that children smile more often to an avatar with full interaction modalities (MP) than to one without prosody (NoP) (variable: smile; $p = .001$) suggesting that the children were more amused by the avatar with full modalities than the avatar with only one modality. Although, we found that there is no significant difference for the smile between MP and NoM ($p = .076$) as well as between NoM and NoP ($p = .220$), we can observed that the presence of *Prosody* in avatar utterance gives the positive impact (mean smile $NoP < NoM < MP$) during the interactive narration (figure 4).

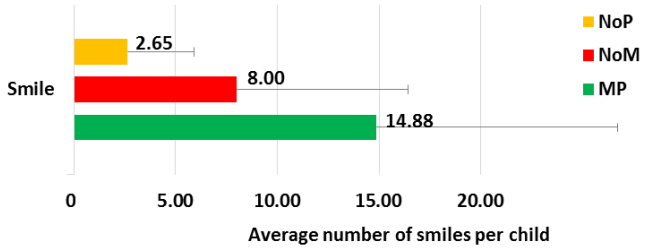


Figure 4: Average number of smiles per child.

The smile in interaction is an indicator of the interlocutor's engagement; the lack of smile in NoM and NoP condition in comparison with the MP condition reveals a less natural interaction. The children have to pay more attention to the content and thus are less available for the social part of the interaction.

A univariate ANOVA has been performed on the children's response latency in function of the condition (different type of avatar). There is an significant effect of the condition (type of avatar) on the latency (F(2,49)=3.381; p=.044). A post hoc comparison (Bonferroni) shows that the children answer more quickly to an avatar with all interactive modalities (MP) than to an avatar without prosody (NoP) ($p = .040$). Significant difference in the response latency may be due to the children adapt their speaker's

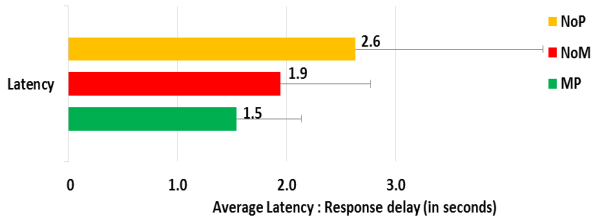


Figure 5: Latency: response delay (in seconds) between avatar's utterances and child's utterances.

interaction style (figure 5). However, the results are not statistically significant for MP and NoM ($p = 1.0$), as well as for NoP and NoM ($p = 0.315$). The results also indicate that the presence of prosody also reduces the response latency (mean response latency $MP = 1.54 < NoM = 1.94 < NoP = 2.63$).

The absence of prosody clearly impact the children's understanding of the narrator utterances; for instance, children have difficulties to discriminate questions. Results show in a same way, the impact of absence of mimics on children's understanding. Synthetics emotions (mimics) seems facilitate children's understanding. All these results support the *hypothesis-1* stating that the virtual avatar with *Prosody* and *Facial expression* modalities improves the engagement of children in interaction during the narrative sessions. Furthermore, regardless of having positive impact of prosody on children's smile and response latency, there is no statistically significant difference between the NoP and NoM. These results do not support the *hypothesis-2* that the *Prosody* is more important than the *facial expression* as a modality of the virtual avatar for the interaction with children during narrative storytelling.

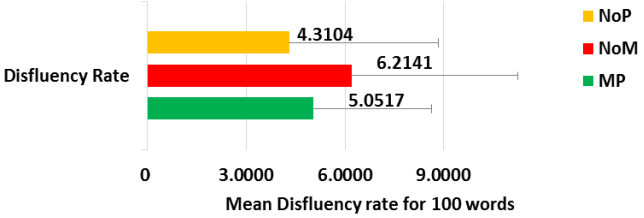


Figure 6: Disfleucy Rate for 100 words.

Concerning the quality of the oral interactions, we also run a univariate ANOVA on the mean disfluency rate in function of conditions (type of avatar) (figure 6). There is no statistically significant effect of the condition (type of avatar) on the disfluency rate F(2,49)= 15.471; p= 0.49). That is, the amount of fluent discourse with less irregularities used by the children during interaction with avatar having different modalities have no significant difference (mean disfluency rate MP=5.0517, NoM=6.2141, NoP=4.3104). However we can observe that the disfluency rate is higher when the children interact with the avatar having no mimics modality.

We now compare the frequency of response from children (figure 7). Due to the adaptation of the interaction, the free context scenario utterances have been used and that bring some differences on the numbers of questions. Therefore, we calculated a ERVQ ratio (in function of the number of questions in the interaction) A univariate ANOVA has been run for the ERVQ ratio on the condition variable (type of narrator) and show a statistically significant

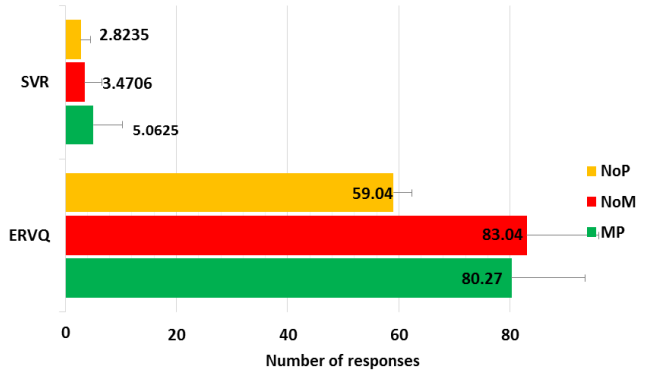


Figure 7: SVR: number of spontaneous verbal responses; ERVQ: number of expected responses after a direct verbal question.

effect of the condition (F(2,49)=11.84 ; p=.001). A post hoc comparison (Bonferroni) shows that the children answer more often to the direct questions of an avatar with all interactive modalities (MP) than with an avatar without prosody (NoP) (variable: ERVQ; $p = .001$). These differences suggest that the children were more engaged with the agent having full modality than with the agent having no prosody, and this result also support the hypothesis-1. Similarly, they also answer more often to the direct questions of an avatar with no facial expressions (NoM) than those without prosody (NoP) (variable: ERVQ; $p = .0001$). this result clearly supports the hypothesis-2. However there is no statistically significant difference for the number of answers from children in the case for the agent with full modalities (MP) and with no facial expression (NoM) ($p = 1.0$).

Furthermore, we applied a univariate ANOVA on the numbers of children's spontaneous verbal response (SVR) in function of conditions (type of avatar) (figure 7). There is no statistical significant effect of the condition (type of avatar) on the number of SVR sentences F(2,49)= 1.308; p= 0.281). However we can observe that the children preferably communicate spontaneously with the avatar having MP condition and utter less comments in the case of the absence of prosody (mean SVR MP=5.0625, NoM=3.4706, NoP=2.8235).

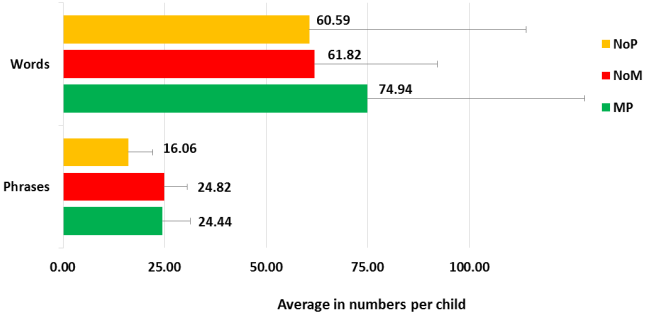


Figure 8: Phrases: mean number of phrases, Words: mean number of words.

In addition, we analyzed the trace of the dialogue between children and avatar (figure 8). A univariate ANOVA has been applied on the numbers of children's intervention in function of conditions (type of avatar). There is a statistically significant effect of the condition (type of avatar) on the number of children sentences F(2,49)=

10.87; p= ,001) but not on the number of words (F(2,49)=0.471; p=.627) nor the number of words by sentence (F(2,49)=1.173; p=.318). A post hoc comparison (Bonferroni) shows that children formulated more sentences during the interaction with agent with full modalities (MP) than that with no prosody (variable: Phrase; p = .001) and without mimics (variable: Phrase; p = .001). This result also supports the *hypothesis-1*. The children also constructs more sentences during the interaction with avatar with no facial expression (NoM) than with the interaction with the agent with no prosody (variable: Phrases; p = .0001) (mean number of phrases $NoM = 24.82 > MP = 24.44 > NoP = 16.06$). This result also supports the *hypothesis-2*. Furthermore, there is no statistically significant difference between the MP avatar and the NoM avatar (P = 1.0). Children seems to rely more on verbal modality when talking to the avatar. Moreover, we observed that the total number of words uttered by children is much higher in the interaction with agent having full modality than that with the agent having only one interaction modality (mean number of words $MP = 74.94 > NoM = 61.82 < NoP = 60.59$) (figure 9).

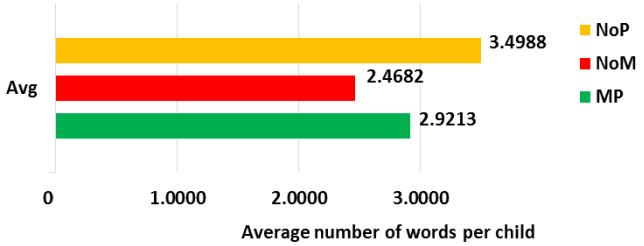


Figure 9: Mean number of words by phrases.

Moreover, an interesting result we observed is that although the children produce fewer sentences in NoP than in NoM or in MP, the sentences are long. That is, the children use more words per sentence in NoP than in NoM. The children may have think that the avatar with NoP is not able to interpret their expressions and utterances, therefore they used long sentences to convey their responses. However this result is not statistically significant, it would be interesting to analyze large corpus to validate the result.

4.3 User experience

During the post experiment interview, the children had very interesting comments which illustrates their perception of the virtual character. They found some differences between the agents with different modalities, such as the absence of microphone and headset on the virtual character. Others compared the avatar with a *toy* or *a lady made out of modeling clay*. In particular, 55% of children found the interaction with the avatar slower, but none of them believed that this was a problem. Moreover, all the children adapted very well to the avatar's rhythm and felt that this allowed them to speak. This effortless adaptation seems a good pledge to use an ECA without affecting the interaction.

4.4 Sum-up of the Results and Discussion

This study aims to analyze the effects of different agent modalities on the child-agent interaction in the context of an interactive storytelling. We can conclude that the children are more engaged with the virtual character having full modalities (prosody and mimics).

We tested two hypotheses. The first hypothesis states that the virtual avatar with *Prosody* and *Facial expression* modalities improves the engagement of children in interaction during the narrative sessions. This hypothesis is supported by the results. The results state that the children address more smile during the interaction with the avatar having full modalities. The results also showed that the children often respond to the agent with full modalities than to the agent with no prosody. However, the results indicate no significant difference for the expected responses and the latency in the cases of the agent with full modalities and with no mimics cases.

The second hypothesis is that *Prosody* is more important than *facial expression* as a modality of the virtual avatar for the interaction with children during narrative storytelling. This hypothesis is not supported by the results. We found that although the children use more phrases during the interaction with the agents with no mimic and with full modalities, they use longer phrases for the interaction with the agent having no Prosody. One of the important observation is the absence of prosody in the agent modalities results in higher latency, longer phrases, and less smile for children during the interaction with virtual agent in the context of the interactive narration. However this observation needs to be validated with more experimental data. 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.

Finally, children interact with pleasure and spontaneously with virtual characters even when some of the interaction modalities are absent (lack of prosody or facial expressions). Nevertheless, facial expressions and prosody seem to play an important role in understanding the nature of an utterance, such as recognizing that a phrase is a question.

Currently we have performed experiments with 50 children using our WoZ platform. Although initial results are in favor with the use of agent with multi-modal interaction capability, it would be interesting to conduct more experiments in order to improve the precision of results. We further intend to conduct more experiments in order to compare the effects of different modalities of child ↔ agent interaction *versus* child ↔ human in video-conference interaction.

In this paper we have considered only smile, response latency, number of phrases etc. to evaluate the user engagement. We are also interested in evaluating the effects of other interaction modalities such as joint-attention (e.g. both agent and the child looks at the same object) and co-construction of joint-attention (e.g. child looks at an object then the agent also looks at the same object, agent points at an object and then the child also looks at that object).

5 CONCLUSION

In this article, we have presented a narrative WoZ environment dedicated to child-avatar interaction. We have also described an experiment that compares the impact of prosody and facial expression of avatar on children during an interactive storytelling context.

The main result is that children were more engaged in narrative interaction with a virtual character having full interaction modalities. In other words, any ECA dedicated to child-agent narrative

interaction have to be designed to take into account these particularities: 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, by offering them adapted interaction models.

ACKNOWLEDGEMENTS

This work was supported by the NARECA project (ANR-13-CORD-0015).

REFERENCES

[1] V. Aubergé, N. Audibert, and A. Rilliard. 2003. Why and how to control the authentic emotional speech corpora. In *Proceedings of Eurospeech'03*. 185–188.
[2] Ruth Aylett, Joao Dias, and Ana Paiva. An Affectively Driven Planner for Synthetic Characters.. In *16th International Conference on Automated Planning and Scheduling (ICAPS)*. 2–10.
[3] Ruth Aylett, Lynne Hall, Sarah Tazzyman, Birgit Endrass, Elisabeth André, Christopher Ritter, Asad Nazir, Ana Paiva, GertJan Höfstede, and Arvid Kappas. 2014. Werewolves, cheats, and cultural sensitivity. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 1085–1092.
[4] R. Beale and C. Creed. 2009. Affective interaction: How emotional agents affect users. *International Journal of Human-Computer Studies* 67, 9 (2009), 755–776.
[5] M. Black, J. Chang, J. Chang, and S. Narayanan. 2009. Comparison of Child-human and Child-computer Interactions Based on Manual Annotations. In *Proceedings of WOCCI'09*. 1–6.
[6] E. A. Boyle, T. M. Connolly, T. Hainey, and J. M. Boyle. 2012. Review: Engagement in Digital Entertainment Games: A Systematic Review. *Computer Human Behavior* 28, 3 (May 2012), 771–780.
[7] M. Casillas, S. C. Bobb, and E. V. Clark. 2015. Turn-taking, timing, and planning in early language acquisition. *Journal of child language* (2015), 1–28.
[8] J. Cassell. 2000. Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. *Embodied conversational agents* (2000), 1–27.
[9] M. Cavazza, F. Charles, and S. J. Mead. 2002. Interacting with Virtual Characters in Interactive Storytelling. In *AAMAS*. 318–325.
[10] Theodora Chaspari, Emily Mower Provost, Athanasios Katsamanis, and Shrikanth Narayanan. 2012. An acoustic analysis of shared enjoyment in ECA interactions of children with autism. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 4485–4488.
[11] S. Cheshire. 1996. Latency and the quest for interactivity. White paper commissioned by Volpe Welty Asset Management L.L.C for the Synchronous PersonâĂŞto-person Interactive computing Environments Meeting. (1996).
[12] Chris Crawford. 2013. *On Interactive Storytelling*. New Riders Games.
[13] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amier, and D. Heylen. 2008. The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In *Proceedings of LREC'08*. 1–4.
[14] O. Grynszpan, J.-C. Martin, and J. Nadel. 2008. Multimedia interfaces for users with high functioning autism: An empirical investigation. *International Journal of Human-Computer Studies* 66, 8 (2008), 628–639.
[15] J. Han, M. Jo, S. Park, and S. Kim. 2005. The educational use of home robots for children. In *Proceedings of RO-MAN'05*. 378–383.
[16] J. Klein, Y. Moon, and R. W. Picard. 2002. This computer responds to user frustration:: Theory, design, and results. *Interacting with computers* 14, 2 (2002), 119–140.
[17] S. Kopp, B. Krenn, S. Marsella, A. Marshall, C. Pelachaud, H. Pirker, K. Thórisson, and H. Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *Proceedings of IVA'06*. 205–217.
[18] S. Kopp, H. van Welbergen, R. Yaghoubzadeh, and H. Buschmeier. 2014. An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing. *Journal on Multimodal User Interfaces* 8, 1 (2014), 97–108.
[19] H. Kozima, C. Nakagawa, and Y. Yasuda. 2005. Interactive robots for communication-care: a case-study in autism therapy. In *Proceedings of RO-MAN'05*. 341–346.
[20] Q. A. Le, S. Hanoune, and C. Pelachaud. 2011. Design and implementation of an expressive gesture model for a humanoid robot. In *Proceedings of HUMANOIDS'11*. 134–140.
[21] J. C. Lester, S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone, and R. S. Bhogal. 1997. The Persona Effect: Affective Impact of Animated Pedagogical Agents. In *Proceedings of CHI'97*. 359–366.
[22] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. 2010. The SEMAINE corpus of emotionally coloured character interactions. In *Proceedings of ICME'10*. 1079–1084.
[23] J. Miksatko, K. H. Kipp, and M. Kipp. 2010. The persona zero-effect: Evaluating virtual character benefits on a learning task with repeated interactions. In *Proceedings of IVA'10*. 475–481.
[24] M. Mori. 1970. The uncanny valley. *Energy* 7, 4 (1970), 33–35.
[25] M. Moundridou and M. Virvou. 2002. Evaluating the persona effect of an interface agent in a tutoring system. *Journal of computer assisted learning* 18, 3 (2002), 253–261.
[26] Emily Mower, Chi-Chun Lee, James Gibson, Theodora Chaspari, Marian E Williams, and Shrikanth Narayanan. 2011. Analyzing the nature of ECA interactions in children with autism. In *Twelfth Annual Conference of the International Speech Communication Association*.
[27] C. Munteanu and M. Boldea. 2000. Mdwoz: A wizard of oz environment for dialog systems development. In *LREC*.
[28] Shrikanth Narayanan, Alexandros Potamianos, and Haohong Wang. 1999. Multimodal systems for children: building a prototype.. In *EUROSPEECH*.
[29] M. Ochs, C. Pelachaud, and D. Sadek. 2008. An empathic virtual dialog agent to improve human-machine interaction. In *Proceedings of AAMAS'08*. 89–96.
[30] M. Otto, R. Friesen, and D. Rösner. 2011. Message oriented middleware for flexible wizard of Oz experiments in HCI. In *HCI*. 121–130.
[31] S. Oviatt. 2000. Talking to thimble jellies: Children's conversational speech with animated characters. In *ICSLP*. 67–70.
[32] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi. 2005. A Model of Attention and Interest Using Gaze Behavior. In *Proceedings of IVA'05*. 229–240.
[33] I. Poggi, C. Pelachaud, F. Rosis, V. Carofiglio, and B. Carolis. 2005. Greta, a believable embodied conversational agent. *Multimodal intelligent information presentation* (2005), 3–25.
[34] J. Porteous, F. Charles, C. Smith, C. Cavazza, J. Mouw, and P. van den Broek. 2017. Using Virtual Narratives to Explore Children's Story Understanding. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 773–781.
[35] H. Prendinger and M. Ishizuka. 2005. The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence* 19, 3-4 (2005), 267–285.
[36] H. Prendinger, S. Mayer, J. Mori, and M. Ishizuka. 2003. Persona effect revisited. In *Proceedings of IVA'03*. 283–291.
[37] K. Prepin and C. Pelachaud. 2013. Basics of Intersubjectivity Dynamics: Model of Synchrony Emergence When Dialogue Partners Understand Each Other. In *Agents and Artificial Intelligence*. Springer, 302–318.
[38] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner. 2010. Recognizing Engagement in Human-Robot Interaction. In *Proceedings of HRI'10*. 375–382.
[39] K. Ryokai, C. Vaucelle, and J. Cassell. 2003. Virtual peers as partners in storytelling and literacy learning. *Journal of computer assisted learning* 19, 2 (2003), 195–208.
[40] M. Schröder. 2011. *The SEMAINE API: A component integration framework for a naturally interacting and emotionally competent Embodied Conversational Agent*. Ph.D. Dissertation. Saarland University.
[41] C. L. Sidner, C. Lee, N. Lesh, and C. D. Kidd. 2004. Where to Look : A Study of Human-Robot Interaction. In *Proceedings of IUI'04*. 78–84.
[42] A. W. Siegman. 1979. Cognition and hesitation in speech. *Of speech and time* (1979), 151–178.
[43] T. Stivers, N. J. Enfield, and S. C. Levinson. 2010. Question-Response Sequences in 10 languages. *Special Issue of the Journal of Pragmatics* 42 (2010), 1–5.
[44] W. R. Swartout, J. Gratch, R. W. Hill Jr., E. H. Hovy, S. Marsella, J. Rickel, and D. R. Traum. 2006. Toward Virtual Humans. *AI Magazine* 27, 2 (2006), 96–108.
[45] W. Taymans, S. Baker, A. Wingo, R. Bultje, and S. Kost. 2001. GStreamer application development manual. (2001).
[46] M. Theune, J. Linssen, and T. Alofs. 2013. Acting, Playing, or Talking about the Story: An Annotation Scheme for Communication during Interactive Digital Storytelling. In *Proceedings of ICIDS'13*. 132–143.
[47] C. Von Hofsten and K. Rosander. 2007. *From action to cognition*. Vol. 164.
[48] S. Whittaker, M. Walker, and J. Moore. 2002. Fish or Fowl: A Wizard of Oz evaluation of dialogue strategies in the restaurant domain. In *Proceedings of LREC'02*.
[49] Serdar Yildirim, Shrikanth Narayanan, and Alexandros Potamianos. 2011. Detecting emotional state of a child in a conversational computer game. *Computer Speech & Language* 25, 1 (2011), 29–44.