Comparing the Effect of Different Styles of Voice on Children's Engagement with a Virtual Robot: A Preliminary Study

Romain Vallée
Enchanted Tools & LIG - CNRS
Paris, France
email: romain@enchanted.tools

Lucas Prégaldiny
Enchanted Tools
Paris, France
email: lucas@enchanted.tools

Véronique Aubergé
LIG - CNRS
Grenoble, France
email: veronique.auberge@univ-grenoble-alpes.fr

Émilie Cénac
Sème!
Rennes, France
email: emilie.cenac0@gmail.com

Serge Tisseron
IERHR
Paris, France
email: serge.tisseron@gmail.com

Olivier Aycard
LIG - CNRS
Grenoble, France
email: aycardol@univ-grenoble-alpes.fr

***Abstract*—This paper aims at understanding the influence of prosody during a child-virtual robot interaction. We provide and analyze an experiment including 30 children aged 6 to 10, who interact with several virtual robots in a video game. This preliminary study highlights the impact of voice on children, as they tend to prefer an expressive voice using non-lexical vocal elements rather than an acted voice simulating stereotypical synthetic voices.**

***Index Terms*—human-virtual robot interaction, human-agent interaction, socio-affective prosody, trust with children, non-verbal speech.**

I. Introduction

Prosody (the sum of phonic elements such as intonation, pitch, rhythm, vocal timbre etc.) is a key element of human interaction, and thus a major issue in human interaction with a physical or virtual communicating machine [1]. Prior studies have shown that the "breathy voice" prosodic factor, characterized by a lax vocal tract and a very relaxed control of the glottis, is an intrinsic marker of relational proximity [2]–[4]. It has been shown that the prosody of speech — generated by the gestures of the vocal tract — and holistically the prosody of gestures of the communicative body elements orients and feeds the nature of the relationship between humans — and thus between prosody-emitting machines and humans [5]. The prosodic artifacts of speech synthesis systems, without any analysis of the markers and relational effects of these artifacts [4], have gradually entered into everyday life until the massive diffusion of these synthetic voices in GPS systems and in voice assistants embodied by "smart speakers", such as Amazon Echo or Google Home. These voices, whose characteristics are more and more often in the "breathy" stereotype [6], seem to be becoming cultural references for the voices of virtual or robotic agents, especially among adults. The prosody of these Alexa-like voices (i.e., breathy, intimate, etc.) is distinct from prosodies of vernacular situations such as play, especially in adult-child situations. The type of relationship these Alexa-like voices build with humans has not been extensively studied and is poorly understood, even though these voices have demonstrated their enduring appeal. Researchers such as Tisseron [7] and Sparrow [8] warn about the ethically toxic effects of voices that trigger an illusion of intimacy and trust invariable to any situation, without any other expressive mark.

Yet, as Vinciarelli et al. [9] or Hofstetter and Keevallik [10] have shown, non-lexical speech primitives (i.e., not containing words) convey relational roles, attitudes, intentions, mental states, emotions, moods and other socio-affects, and build a relationship by guiding its value (e.g., the altruistic relationship without dominance [11], [12]). Some speech synthesis systems offer voices which include non-lexical elements, and some robots implement them (such as Paro or Spoony), but without relying on a fine understanding of these vocal elements and their effects on the engagement and relational nature evoked, in consistency with the lexicalized prosody.

This paper proposes to the Human-Computer Interaction community a reflection on the still under-researched importance of speech posody in virtual agents and robots. In Section II, we introduce the research question and motivation behind the study. In Section III, we present the experimental design and methodology. In Section IV, we report the results of the study and discuss our indicators and potential biases. Finally, in Section V, we draw provisional conclusions and outline directions for future research.

II. Objectives

The long-term goal of our research is to understand how strong and how is established (with what nature - in particular, trust) the engagement in the interaction between a human and a robot, through the strong and weak prosodic signals conveyed within an overall relationship by all elements of the body, including the vocal tract. The adjective *prosodic* is assumed

 ISBN: 978-1-68558-078-0

here in its extension to all body signals, not restricted to the speech prosody. In this very first step presented here, we focus only on the voice (no variation of other body gestures), and only on the invariable *breathy voice* stereotype (Alexa type), in comparison to a non-breathy but overly expressive voice (to oppose it very clearly to the always "breathy" and confident voice, friendly but unresponsive to interactional changes) with or without non-lexical vocal elements. In order for this caricature to be ecologically relevant, and also because the "breathy" stereotype is essentially intended for adults, we proposed this prosodic contrast on a virtual robot interacting with children during a playful task - a pretextual cooking game.

In this interaction protocol, we used three different vocal profiles, all acted by the same female comedian [13]:

- Voice A - Colloquial enunciation, aiming for a playful, dynamic style, exaggeratedly child-like cartoon voice: modal or tense voice (tensed), fast-paced, high pitched (mean fundamental frequency F0 = 320 Hz)
- Voice B - Same instructions and prosodic values as voice A, but including non-lexical socio-affective vocal primitives consistent with the global prosody (vocal bursts, grunts, onomatopoeia, etc.) (mean F0 = 320 Hz)
- Voice C - An acted voice simulating "stereotypical" synthetic voices (e.g., Alexa), i.e., globally breathy without attitude variations and without non-lexical vocal elements: systematically breathy voice, slow rhythm, lower pitch (mean F0=250 Hz)

Our research hypothesis is as follows: In a playful cooking task involving children interacting with virtual robots, a robot using an expressive voice and non-lexical speech elements will elicit more engagement, trust or interest than a robot using only lexical speech elements and a constant prosodic modality. An ethical issue of this work will be to measure the nature and strength of the installed relationship in order to make explicit in future commercial product, how the robot engages and bonds with the people it interacts with (note that as far as we know, none of the currently available conversational agents warn about the nature and the strength of the relationship created with their users). Thus, according to the specific uses (e.g., health or service) and the targeted audience (e.g., fragile, young or elderly), these warnings will be taken into account so that the ethical validity or invalidity of the implementation can be determined contextually.

III. MATERIALS AND METHOD

In this section, we describe the division of participants in three groups and the procedure used to gather data, involving a vocal interactive game.

A. Participants

The participants of our research were voluntary children visiting the Cité des Sciences et de l'Industrie in Paris. Thanks to the Laboratoire des Usages en Technologies d'Information Numériques (LUTIN), we recruited 30 children between the ages of 6 and 10. The gender distribution was 35% girls / 65% boys, with an average age of 7.8 years (± 1.7 SD). We set up 3 groups of 10 children. The A/B group interacted only with voice profiles A & B, the B/C group with the profiles B & C , and the A/C group with the profiles A & C. Presenting only 2 voice profiles per child seemed necessary to limit the child's cognitive load. All interactions occurred in French, and were translated to English for this article.

B. Method

a) Description of the procedure for a participant The research protocol described below was submitted to the multi-disciplinary ethics committee of the Grenoble Alpes University (CERGA), which analyzed and evaluated it positively. Two experimenters intervene in the experimentation room: one in charge of guiding the child through the task, and another in charge of the "Wizard of Oz" procedure, giving the child the impression that the virtual robots were understanding the child's requests. The child sits on a chair in front of a screen where the game is projected. Shown in Fig. 1, the game consists of making a virtual cooking recipe with the help of two virtual robots who bring to the player the 6 ingredients needed to make a chocolate cake.

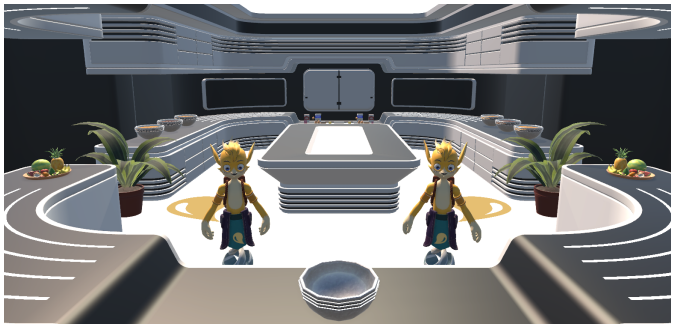

Fig. 1. Screenshot of the game, with the two virtual characters.

The experimenter first asks the child a few open-ended questions before starting the game (about their background with video games, cooking, robots) and then at the end about their experience of the game. It is made clear to the child that there are no right or wrong answers and that they must answer according to what they have seen or felt in the game. When the game starts, both virtual robots welcome the child at the same time. The experimenter then asks the child to pick (using a counting-out rhyme) the robot which will start the task and fetch the first ingredient. The robots then interact successively with the child to fetch ingredients. The game unfolds at the child's own pace, through their vocal statements, on which the "Wizard of Oz" experimenter bases the triggering of the different sequences of the game. At the end of the game, both robots say goodbye to the child at the same time. Note that when the two robots speak at the same time (only at the beginning and end of the game), a slight delay (about 100 ms) between the 2 voices is used, to make the voices more distinguishable. The average duration of the game is about 4 minutes. The progress of the task is recorded by a webcam

 ISBN: 978-1-68558-078-0

(audio and video) located under the projection screen. Finally, two boxes physically present in the experimentation room are placed under each virtual robot projected on the screen, each box containing stickers representing the character under which it is located (the characters differ only by the direction of the symbol which decorates their apron). At the end of the experiment, the experimenter asks the child to take a sticker of the virtual robot they prefer from one of the two boxes, to keep a souvenir of the game. This forces the child to make a choice between the two virtual robots presented (contrary to the preliminary questions asked to the child where the answers can be multiple without requiring a final choice). In this choice between two stickers, we also wish to mobilize the body and to see the choice which results from this request which is not only addressed to the mind (as it was the case in the previous questions where the body was not solicited). We believe that the reasons that lead a child to choose one of the virtual robots over the other may not be easily conscientizable and that the conscious representations that the child has of the robot do not necessarily induce a given behavior towards the robot.

b) Avoiding potential biases Apart from the differences in speech profiles between the two virtual robots, we made sure that the robots' animations and the elements of the game's virtual kitchen decor were as symmetrical as possible. Moreover, different potential biases (related to the child's age, gender or laterality, their number of interactions with each robot. . .) have been identified. We will be vigilant to neutralize at best these biases in future analyses.

IV. RESULTS AND DISCUSSIONS

All the children said they enjoyed the game, 95% said they appreciated the help they got from the robots and 75% appreciated the help provided by the first experimenter. These responses support a real involvement of the children in this research protocol.

A. Perceived differences in voices

85% of the children noticed that the two virtual robots did not speak in the same way. When the first experimenter asked what is different for the children who perceive a difference, the answers were mostly about pitch differences between the voices, as shown in Tab. I. As a reminder, the actual fundamental frequency (F0) of C is lower than the one of A and B. Some of the children who noticed a difference in pitch assumed that the robot with a higher pitched voice was a female, and that the other one was a male. Other less salient factors are reported by the children, for instance the laughter included in some utterances of voice B, and an "alien" characterization of voice C. Although some qualifiers (e.g., *enthusiastic* (A), *laughing* (B), *cheerful* (B), *softer* (B), *nicer* (B), *better* (B)) have a more positive valence than others (*shouting* (B), *unpleasant* (B), *alien* (C)), it is difficult to identify children's preferences from the data we collected.

B. Potential indicators of trust

We wished to pay particular attention to the relationship of trust in which the social signals of the robot (virtual in this game and physical in the future) could engage children. To begin the evaluation of this potential trust, we did not explicitly ask questions using the notion of trust, in order to avoid any bias, especially with children. We addressed it indirectly; when asked, "Would you lend a precious object or a toy to one of the robots?" (Q5 in Tab. II), children responded positively, for one particular virtual robot or both, 70% of the time. When a child's response was positive for only one virtual robot, that virtual robot was also chosen by the child 75% of the time when making the final sticker choice. When asked "Would you let one of the robots enter your bedroom?" (Q6), children responded positively for one or both virtual robots in 70% of cases. When a child's answer was positive for only one virtual robot, that virtual robot was also chosen by the child 75% of the time in the final choice of a sticker. If we compare the results of questions 5 and 6 in group B/C, more children are willing to let robot B into their room than to lend it their toys. This trend is reversed for robot C. It would therefore be interesting to distinguish more carefully in a future research two axes of trust; one axis of *centrifugal* trust in lending an object to the other, and a second axis of *centripetal* trust in letting the other into the home.

TABLE I
QUALIFIERS FOR EACH VOICE BY NUMBER OF OCCURRENCES (NB)

Voice A		Voice B		Voice C	
qualifier	**nb**	**qualifier**	**nb**	**qualifier**	**nb**
higher than C	3	higher than C	3	lower than B	3
lower than B	2	higher than A	2	lower than A	2
higher than B	2	softer than C	1	male	2
female	2	nicer than C	1	bigger than B	1
high	1	smaller than C	1	slow	1
fast	1	better than C	1	alien	1
enthusiastic	1	lower than A	1	not too fast nor too slow	1
		laughing	1		
		shouting	1		
		unpleasant	1		
		cheerful	1		
		female	1		

reading hint: voice A is reported as higher than voice C by 3 children

C. Preference between voices

We will now look specifically at the data collected in each group A/B, A/C and B/C. Among all the questions asked to each child, 6 questions allow determining whether or not the child expresses or not a preference towards a virtual robot. The final choice of the sticker also gives information about the child's preference. All these answers and choices are summarized in Tab. II to evaluate the engagement, trust or interest that a child has toward virtual robots.

D. Calculation of the total score for each robot

In a first simplifying approach to evaluate the engagement, trust or interest that a child has towards a virtual robot, we assign an identical weight to the positive answers for the 6 questions above and to the final choice of the sticker. If a robot is chosen or preferred by X% for a question or for the choice

 ISBN: 978-1-68558-078-0

TABLE II
ANSWERS FOR ENGAGEMENT & TRUST-RELATED QUESTIONS

group	answer	Q1	Q2	Q3	Q4	Q5	Q6	sticker	score
A/B	none	90%	70%	90%	70%	10%	30%		
	both					40%	30%		
	A	10%	30%	10%	10%	40%	30%	**60%**	**26**
	B				20%	10%	10%	40%	15
A/C	none	80%	80%	80%	90%	40%	30%		
	both					40%	60%		
	A	10%	20%	20%		10%	10%	**70%**	**24**
	C	10%			10%	10%		30%	16
B/C	none	50%	30%	50%	40%	40%	30%		
	both								
	B	30%	20%	40%	20%	20%	50%	**80%**	**26**
	C	20%	50%	10%	40%	40%	20%	20%	20

Q1 Is there a robot who helped you more?
Q2 Is there a robot you liked more?
Q3 Is there a robot you understood better?
Q4 Is there a robot you preferred to talk to?
Q5 Would you lend a precious object or a toy to one of the robots?
Q6 Would you let one of the robots enter your bedroom?

of the sticker, we assign it X/10 points. Each group containing 10 children, this is equivalent to assigning one point to a robot each time it is chosen by a child. For example, the score for voice A in group A/C is 24 points, obtained as follows: 1 point (Q1 : 10% for only A) + 2 points (Q2 : 20% for only A) + 2 points (Q3 : 20% for only A) + 4 points (Q5 : 40% for both A and C) + 1 point (Q5 : 10% for only A) + 1 point (Q6 : 10% for only A) + 6 points (Q6 : 60% for both A and C) + 7 points (Sticker : 70% for A).

These data show an overall tendency for children to express a preference for vocal profile A (whether it is opposed to B or C), and to prefer profiles A and B when they are opposed to profile C. This is confirmed by grouping the results of groups A/C and B/C, which allows us to compare the so-called "expressive" voices (A and B) with the stereotypical voice C. If we aggregate the scores of A and B in groups A/C and B/C, and compare this score with the aggregate score for voice C in these two groups, we obtain a score of 50 for expressive voices versus 36 for the stereotypical voice, again showing a tendency for children to prefer expressive voices. These tendencies will naturally have to be confronted to a larger sample of participants. We also plan to study different kinds of embodiment and situations to get a better insight of the robustness and generalization of these tendencies.

E. Effect of non-lexical primitives

The comparison of the results obtained between groups A/C and B/C is revealing of the effect of non-lexical primitives since this is the only element of the game that changes between these two groups. The frequency of negative answers ("none") to questions 1, 2, 3 and 4 change substantially, with the average of these frequencies dropping by 40% when the non-lexical primitives are introduced (cf Tab. II). Even if the direct contrast in the A/B group did not induce a preference for the B voice, this drop in frequency seems to indirectly show a positive cleavage effect of the non-lexical primitives on the children that would be useful to study further with a larger number of children. This is consistent with the results of previous studies for adults [14], [15].

F. Relevance of holistic gestural behavior

In the protocol of the experiment, it is also important to notice the interest that the expression of the preference of a child towards a virtual robot is not only verbalized but more generally gesturalized: the final choice of the sticker of the preferred robot implies the physical displacement of the child and its gripping. This interest seems to be quite visible for the B/C group. Indeed, the answers to the 6 questions highlighted in Tab. II do not show a clear preference between the voices B and C, with a score of 20 for the virtual robot B and 18 for the virtual robot C (if we exclude the final choice of the sticker). Yet, a clear preference appears for the virtual robot B which is chosen at 80% by the sticker.

G. Potential biases

Further research will be needed to estimate the impact of potential biases in this study such as the robot design and its kinematics, the laterality of the child or the order of the questions.

V. CONCLUSIONS AND FUTURE WORK

This first experiment with a limited number of children allows us to draw several provisional conclusions, showing the interest of continuing this research:

- As expected, the voice seems to have a significant impact in a video game context intended for children, where many other parameters intervene (the visual aspect of the game, its playability, its novelty, the presence of an adult at the child's side...), which could have strongly limited this impact. The voice being one of the elements of the relational construction in the physical robot, we will have to work on the voice in coherence with the other modalities of expressivity of the robot to come.
- The factors of voices A and B do not place the child in the intimacy of voice C (which they sometimes describe as "polite", "friendly"), but in an expressiveness (e.g., "enthusiastic"), which attract their preference (remember that voice C imitates voice assistants like Alexa).
- Non-lexical speech primitives seem to have a determining role in the child's perception, installing a relational space different from the one installed by strictly lexical elements. This will guide further research devoted to analyzing more precisely the relational effects pointed by the inconsistency when comparing directly voice A and B.

These first results reinforce the importance of the choice of the voice prosody that can be given to a robot. The seductive power of stereotypical voice C, which creates the illusion of intimacy, gentleness, politeness, seems complex to analyze in its effects. In any case, even if children assign positive qualifiers to it, it does not attract their overall preference. More generally, the prosodies of all gestural modalities (gaze, head,

 ISBN: 978-1-68558-078-0

arms, navigation...) will be studied in order to establish an ethically justified choice.

These initial findings will lead to future experiments involving other virtual robots as well as physical robots currently under development at Enchanted Tools.

ACKNOWLEDGMENTS

We would like to thank all the people who participated in this experiment, the members of the Ethics Committee of the Grenoble-Alpes University (CERGA) and the Laboratoire des Usages en Technologies d'Information Numériques (LUTIN).

Thanks also to Enchanted Tools who funded this research.

REFERENCES

[1] N. Ward, "Prosody Research and Applications: The State of the Art, Keynote", Interspeech Conference, 2019

[2] C. Wightman and R. C. Rose, "Evaluation of an efficient prosody labeling system for spontaneous speech utterances", Proceedings of the Automatic Speech Recognition and Understanding Workshop, pp. 1837-1840, 1999

[3] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension", 15th ICPhS, pp. 2417-2420, August 2003

[4] V. Aubergé, "Gestual-facial-vocal prosody as the main tool of the socio-affective 'glue': interaction is a dynamic system", International workshop on audio-visual affective prosody in social interaction, Bordeaux, France, 2015

[5] V. Aubergé, "The Socio-Affective Robot: Aimed to Understand Human Links?", AVEC'2019: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, p.1, 2019

[6] M. Cohn and G. Zellou, "Prosodic differences in human-and Alexa-directed speech, but similar local intelligibility adjustments", Frontiers in Communication, vol. 6, no. 675704, 2021

[7] S. Tisseron, "Le Jour où mon robot m'aimera. Vers l'empathie artificielle", Albin Michel, Paris, 2015

[8] R. Sparrow, "Why machines cannot be moral", AI & SOCIETY, 36(3), pp. 685-693, 2021

[9] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain", Image and vision computing, vol. 27, no. 12, pp. 1743-1759, 2009

[10] E. Hofstetter and L. Keevallik, "'More than meets the eye': Accessing senses in social interaction", Video Based Studies of Human Sociality, vol. 4, no. 3, 2021

[11] Y. Sasa and V. Aubergé, "Socio-affective interactions between a companion robot and elderly in a Smart Home context: prosody as the main vector of the 'socio-affective glue"', SpeechProsody, pp.86-90, 2014

[12] G. De Biasi., V. Aubergé, L. Granjon, and A. Vanpé, "Perception of social affects from non lexical sounds", In Proceedings of VII GSCP International Conference: Speech and Corpora, Brazil, 2012

[13] The audio files used in the experiment are available at https://lpy-et.github.io/ACHI2023/

[14] N. Campbell, "Specifying affect and emotion for expressive speech synthesis", International Conference on Intelligent Text Processing and Computational Linguistics, pp. 395-406, Springer, Berlin, Heidelberg, 2004

[15] Y. Sasa and V. Aubergé, "Caractéristiques prosodiques de la 'glu socio-affective' de l'interaction face à face: un robot-compagnon médiateur d'un habitat intelligent pour personnes âgées", 3rd SWIP-Swiss Workshop on Prosody, pp. 185-196, 2014

 ISBN: 978-1-68558-078-0