2022 31st IEEE International Conference on Robot and
Human Interactive Communication (RO-MAN)
August 29 - September 2, 2022. Naples, Italy.

2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) | 978-1-7281-8859-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/RO-MAN53752.2022.9900811

How a Social Robot's Vocalization Affects Children's Speech, Learning, and Interaction

Lauren L. Wright[1], Aditi Kothiyal[1], Kai O. Arras[2] and Barbara Bruno[1]

***Abstract*—A wider incorporation of robots into classrooms is hampered by current technological limitations on full autonomy in social robots. Automated speech recognition, for example, a key enabler for vocal communication, is still unable to perform with sufficient accuracy. Past studies have shown that humans adjust their speech patterns to accommodate less skilled interlocutors. If such a response holds in human-robot interactions as well, we may be able to exploit it to lessen the burden on social robots and enable rich, autonomous vocal communication. In this paper we explore whether a robot's speaking ability could have an impact on children's speech patterns, learning, and engagement by designing an interaction where a child and a robot collaborate on a Tower of Hanoi puzzle. Sixteen children aged 7-14 completed this collaborative task partnered with a social robot that communicated with either high verbal (full sentences), low verbal (short phrases or single words), or non-verbal (sound-based utterances) vocalization. While we found no significant impact on children's speech patterns or learning due to the robot's method of vocalization, children in the non-verbal condition had a significantly lower perception of the robot's intelligence along with higher rates of providing feedback and more instances of undoing its moves. This suggests that a link may exist between a robot's perceived speaking ability and children's confidence in that robot's overall intelligence and capability in a collaborative task, as well as their empathy towards a peer they perceive as less skilled in the task.**

***Index Terms*—social robotics, human-robot interaction, education, sonic interaction**

I. Introduction

Social robots have the potential to transform education by providing individual instruction for students whose needs are unmet by the traditional models of learning and instruction [1]. Verbal communication is a cornerstone of interaction between humans, and speech especially plays a critical role in learning and development [2]. Social robots in education thus need to be capable of *sonic*, or sound-based, communication in some form. Sonic communication requires two features from a robot: the ability to speak or convey information through sound, and the ability to understand vocal communication from the human interlocutor. While in humans these two abilities are tightly coupled, the same cannot be said for social robots [3]. The former ability is quite simple to provide with existing technology. Text-to-speech enables a robot to speak with the same complexity of diction and grammar as a human, if not the same expressive capability. However, understanding speech with an equal degree of complexity is much more difficult, especially when it comes to the speech of children. Children are more likely to have speech impediments and irregularities, tend to enunciate less, and use improper grammar – all of which contribute to poor automated speech recognition [4]. This asymmetry in social robots' sonic communication abilities has been identified in the literature as a possible factor negatively affecting human-robot interaction [3].

In order for social robots to be adopted as a widespread educational tool, we must find a way to facilitate autonomous verbal interaction: our study aims to contribute to this goal, following the rather unconventional approach of trying to bring the children's expressiveness *down-to-par* with the robot's ability to understand. Indeed, while efforts are being made to increase the language processing capabilities of AI [4], in this paper we address the problem from the side of the human interlocutor and propose to exploit a phenomenon known to occur in human-human interactions where a person lowers their level of speech to accommodate a less capable partner [5]. Intuitively, if we can engineer the interaction with a robot in a way that affects the speech patterns of the human such that they become more compatible with the capabilities of automated speech recognition systems, we can enable richer autonomous interactions without depending on unrealized technological ability. At the same time, however, children's explanations or elaborations while learning, to themselves or others, is known to play an integral role in understanding [2]. Consequently, a manipulation which may lower the child's speech complexity during a collaborative task may also negatively impact their learning.

To explore the feasibility of such a solution, in this study we investigate whether manipulating the speaking ability of the robot has any effect on a child's (i) level of speech complexity, (ii) learning, and (iii) perception of the robot and engagement in the interaction with it, through a version of the Tower of Hanoi logical thinking puzzle wherein children collaboratively solve the puzzle with a fully autonomous social robot.

II. Related Work

A. Social Robots in Childhood Education

The potential efficacy of using social robots in education has been studied for decades. While most research envision the robot in the role of a tutor, or teacher [1], a number of studies in this domain take advantage of the *protégé effect*

This project has received funding from the Swiss National Science Foundation through the National Centre of Competence in Research (NCCR) Robotics.

[1]L.L. Wright, A. Kothiyal and B. Bruno are with the Computer-Human Interaction in Learning and Instruction (CHILI) Lab, Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland. Corresponding author's email: mailto:lauren.wright@alumni.epfl.chlauren.wright@alumni.epfl.ch.

[2]K.O. Arras is with Bosch Corporate Research, Stuttgart.

wherein it is theorized that children can learn better through the act of teaching a peer [6]. In fact, positioning the robot as a peer in the interaction has produced many successful results in educational robotics experiments. Chen et al. [7] showed that in a reciprocal peer learning task with a social robot, engaging in both tutor and tutee roles can be beneficial for vocabulary acquisition, while Gordon et al. [8] explored how a robot's displayed curiosity towards a child's teaching could heighten the child's own curiosity.

The educational domains for which social robots are typically used include (second) language and vocabulary learning [9], [10], [11] and, to a lesser extent, training computational thinking skills [12]. In many such cases, the robot's embodiment and ability to physically interact with the child and the task environment are neglected, possibly due to the difficulty of ensuring a safe physical interaction and a smooth learning experience. Robots used in these experiments are often limbless and/or do not interact much with their environment, relying on tablets or other digital devices to support and mediate their interaction with the child [1]. Such limitations make these types of robots less suitable for physical tasks, which prevents them from being as useful in tasks for kinesthetic learning. Hood et al. [13], whose setup is also one of the first incorporating the protégé effect in a human-robot interaction setting, had one of the few studies where the embodiment of the robot was essential to the learning task, as the robot simulated "learning" handwriting from the child.

In addition to the stated limitations in their physical agency, many social robots used for educational purposes rely on the *Wizard of Oz* approach, where the robot does not act autonomously but is instead controlled remotely by a human [1]. While this approach is useful for advancing the understanding of children's perceptions and responses to robots, it does very little to advance the technology needed to fully take advantage of this new understanding in a non-experiment setting.

We argue that if social robots are to become reliable enough to be adopted for use in education they need to be fully embodied and autonomous, otherwise they provide little benefit for long term use in the classroom. Embodiment in this context refers to a robot's ability to interact physically with its environment, as opposed to a mostly static robot that lacks functioning arms with end effectors. These robots should ideally leverage their embodiment to be adaptable to different activities which cater to different learning outcomes. Full autonomy is an essential objective, as without it a robot will never be an asset in education, either as an assistant to a teacher lessening that teacher's burden or as a peer/protégé to a learner. This goal is the reason why we sought to design a learning task in which a fully embodied and autonomous social robot would interact with a child.

B. Sonic Human-Robot Interaction

Using vocal communication with a robot appears to be an obvious basis for HRI to successfully integrate robots into dynamic human environments. This approach is hampered, however, by the shortcomings in automatic speech recognition and language processing. Despite significant progress in the last decade driven by deep learning methods and an increasing availability of data, state-of-the-art speech recognition algorithms still have word error rates of approximately 5% [14], which underperforms compared to a human transcriber. Further, the ability of such systems to incorporate context and derive meaning from speech is still far from human-level performance.

At the same time, studies indicate that human-like appearance and communication raise higher expectations in the robot's capabilities – Wuth et al. [15] found that users expected more capability from robots that used text-to-speech vs. beeping noises – and that not fulfilling such expectations can lead to disappointment and even to an overly pessimistic perception of the robot's actual competencies, as found by Paepcke and Takayama [16].

Managing expectations, therefore, may be the key to aiding autonomous interaction with social robots which, in the context of this work, motivates us to also investigate non verbal utterances instead of language as the only vehicle for communication. Prior work on sonic human-robot interaction include Scheeff et al. [17] who use semi-abstracted vocalizations that sound like muffled speech combined with a French horn, Kismet's child-like utterances generated with a speech synthesizer [18] or the sound synthesis system by Schwenk and Arras [19] that allows internal or external stimuli to alter the sound or speech synthesis process in real-time. Further work using semantic-free utterances include Read and Belpaeme [20], Ritschel et al. [21], and Robinson et al. [22].

C. Effects of Vocalization in Interactions

While the effects of a robot's vocalization on human-robot interactions are not yet well understood, a number of hypotheses can be made based on how vocalization affects human-human interactions. *Entrainment* in verbal interactions is a well documented phenomenon where people mimic each other in speech [23], [24]. This mimicry can extend beyond prosody and pitch to affect the sophistication of speech: in interactions between adults and children, adults typically lower their level of speech to adjust to the child's capability [5]. It thus stands to reason that a humanoid robot may be able to elicit similar entrainment from a human interlocutor, as long as vocalization is the vehicle for communication.

D. Synthesis

Building on the key concepts highlighted above – the need for fully embodied and autonomous social robotics in education, sonic HRI and the entrainment phenomenon occurring in verbal interactions – the contribution of this paper is the investigation of an under-explored hypothesis that sheds light on the effects of a robot's vocalization on human-robot interactions. In order to investigate the hypothesis, we develop and present a system, embodied in the robot Reachy, in which all necessary components have been integrated to conduct the experiments autonomously without human intervention. Further, the creation and use of abstracted sound-based utterances to this end is, to the best of our knowledge,

Fig. 1. Experimental setup from the child's viewpoint, showing Reachy behind the Tower of Hanoi puzzle, mid-solution. The child sits across from Reachy where they can both reach the game to play collaboratively. Sessions were recorded using the webcam mounted to the right of Reachy's head in the image. The researcher was seated behind the monitor with their face and upper body blocked from the child's direct view.

new. The results extend our understanding of what it takes – or what can be removed – to make social robots more autonomous and effective for education in particular and HRI in general.

III. Experiment Design

A. Research Questions

In this study we want to explore the effects of a robot's speaking ability on a child's speech, learning, and engagement during a collaborative task. To this end we have three primary research questions for our investigation:

RQ1: Speech Do children alter their habits in speech and verbalization in response to the robot's speaking ability?

RQ2: Learning Does the robot's speaking ability influence children's learning of how to solve the Tower of Hanoi task?

RQ3: Interaction Is the children's engagement with and perception of the robot during the collaborative task affected by the robot's speaking ability?

B. Task Design

We chose the Tower of Hanoi puzzle for this study as it is a logical thinking task that challenges children's working memory and can be easily adapted to a collaborative turn-taking task [25]. In the Tower of Hanoi puzzle, there are three posts and a set of rings of different sizes, as shown in Figure 1. At the beginning of the puzzle these rings are stacked in order of size on one of the posts in a pyramid, with the largest ring on the bottom and the smallest on top. The goal is to rebuild the pyramid onto another post, maintaining the size order, with two rules: only one ring may be moved at a time and a larger ring can never be placed on a smaller one.

C. Session Outline

At the beginning of each filmed experiment session, the child was introduced to the humanoid robot Reachy[1], shown in Figure 1, and had the Tower of Hanoi rules explained to them by a researcher. Then children participated in a pre-test, framed as teaching Reachy how to play the game. In this pre-test, children were asked to play through the three ring version of the Tower of Hanoi puzzle on their own while explaining to Reachy. Though Reachy did not speak or vocalize during the pre-test, it indicated interest and attention by randomly raising or lowering its head to gaze at the child or the game respectively. This pre-test allowed us to benchmark the child's understanding of the game, their speech complexity and use of gestures and eye contact before the child gained any insight into the robot's speaking capabilities.

During the subsequent collaborative task, children were asked to alternate taking turns with Reachy in solving a four ring version of the puzzle, with Reachy taking the first move. Children were told that Reachy has difficulty grasping the rings and would be indicating its move vocally and by pointing first to the ring it wanted to move and then pointing to the destination post, leaving the child to complete the move for Reachy. They were instructed to verbally indicate that they had finished moving the ring for Reachy, and to wait until Reachy lifted its head to look at them before making their move. At the end of the child's turn, they should again tell Reachy that they've completed their turn. Children were told that they may have to repeat themselves to Reachy, as it doesn't always hear them the first time. Children were not told in advance whether Reachy would be speaking, however throughout the experiment Reachy vocalized according to a randomly assigned vocalization condition. In addition to vocalization while indicating its move, Reachy would also interject depending on situations and events during the game.

After the child and Reachy successfully solved the four ring Tower of Hanoi puzzle, the child was asked to work through the three ring version on their own in a post-test. The post-test was again framed as teaching Reachy, with the justification that robots, like us, learn better through repetition. As in the pre-test, Reachy did not vocalize but moved its head between gazing at the child and at the game. Afterwards, the child was guided through the Godspeed questionnaire by the researcher and also asked general questions about their experience and perceptions of Reachy. In total a session would last approximately 20 minutes.

This experiment operates as a reciprocal learning task. The pre-test and post-test are framed as teaching activities wherein the child is asked to teach Reachy the proper way to solve the Tower of Hanoi puzzle using only three rings. However, during the collaborative task, Reachy is programmed to always make the optimal next move. As such, though Reachy does not explicitly instruct the child, it might end up correcting non-optimal moves made by the child or provide hints at key junctures of the game. In fact, having Reachy make the first move serves as a hint to the child to

[1]https://www.pollen-robotics.com/reachy/

complete the optimal move sequence, as moving the smallest ring to the wrong post on the first turn already means the number of moves required to solve the puzzle is more than the optimal solution.

D. Experimental Conditions

Three experimental conditions were tested concerning Reachy's level of vocalization. In every turn, Reachy would communicate during its move as it indicated which ring to move and which post to move it to. Other instances of communication from Reachy were dependent upon the child's actions during the game and also whether they were speaking to Reachy in turn, as shown in Table II.

- In the *high verbal* condition, Reachy's responses are created to be full sentences which conveyed an appropriate response to the event trigger.
- In the *low verbal* condition, the robot's responses are reduced to the minimum possible way to convey the same idea. For example, Reachy's offer to help, triggered whenever a child takes more than 20 seconds to make their move, is "Do you want me to help you?" in the high verbal condition, translated into "Need help?" in the low verbal condition.
- In the *non-verbal* condition Reachy communicates via non-linguistic utterances. For each trigger in the reduced set of low verbal responses, we designed two to four sound candidates that convey its meaning to the best of our knowledge: A muffled low-overtone sound with decreasing pitch for the expression of an error (e.g. event 1 in Table II, "Negative2NotRight"), a periodic slightly randomized sound sequence to indicate that the robot is "thinking" (e.g. event 6, "Thinking1Hmmm") or an overtone-rich sound with a raising pitch envelope for offering help (e.g. event 3, "Question2HelpYou").

The clarity of the intended meaning of the high and low verbal responses was qualitatively verified by test users, who found no ambiguities. The sounds used in the non-verbal condition were designed iteratively. In each round, users were asked to associate each sound to a speech prompt category. The final selection included the candidates whose association was the most correct and least ambiguous. After the second round there was sufficient agreement to cease iteration.

E. Participants

This experiment was conducted over two weeks in individual sessions with 16 children aged 7-14[2], distributed between conditions as seen in Table I, by order of participation in the study. None of the subjects had heard about the Tower of Hanoi prior to participating. This age range was chosen to match the complexity of the Tower of Hanoi puzzle. Though it seems that the non-verbal condition skews younger than the other two conditions, a Kruskal-Wallis test did not show any significant difference between the groups with respect to age. Due to the voluntary method of recruitment, a gender imbalance was unavoidable.

[2]This work has been approved by the EPFL Human Research Ethics Committee (HREC No: 030-2021 / 06.04.2021 – Amendment to General Protocol HREC 051-2019 / 05.09.2019).

TABLE I
MEAN AGE AND GENDER DISTRIBUTION OF PARTICIPANTS

Condition	Mean Age	Median Age	Girls	Boys
High Verbal	10.6	11	1	4
Low Verbal	11.4	11	1	4
Non-verbal	9.7	10	0	6

F. Evaluation Metrics and Hypotheses

We expect that the robot's level of speech complexity would impact not only the children's speech patterns in response to the robot, but also their learning and their engagement with the robot as a peer in the interaction.

Concerning the children's speech patterns (RQ1), we hypothesize that if the robot demonstrates an inability to speak at a level equal to the child, the child would speak more simply in order to accommodate their partner. We expect this tendency towards accommodation to be most noticeable in the non-verbal condition, and least noticeable in the high verbal condition. The metric we use to measure such changes is the Mean Length of Utterance (MLU). MLU is a measure of language proficiency often used as a benchmark for early childhood language development [26], calculated by dividing the number of morphemes (smallest meaningful lexical unit in a language) by the number of utterances spoken. Utterances here are understood to be full sentences. The MLU does not measure the complexity of individual words. Though other metrics may better assess word complexity, many require higher fidelity audio recordings than were possible to obtain here. We also measure the word count to assess the amount that children are speaking. More formally, we hypothesize that:

H1: *Children will reduce their speaking complexity (i.e., their MLU) and quantity (i.e., their word count) between the pre-test and post-test the most in the non-verbal condition, slightly in the low verbal condition, and make no change in the high verbal condition.*

Concerning the learning outcomes (RQ2), we expect all children to improve their strategic planning between the pre-test and the post-test, by reducing the number of moves needed to solve the Tower of Hanoi with three rings. We expect to see a greater improvement in the children in the high verbal condition, due to the presence of more guided interaction. We thus anticipate that:

H2: *Children will show the most strategic improvement by reducing their number of moves in the high verbal condition, followed by low verbal and then non-verbal conditions.*

Lastly, regarding their interaction with the robot (RQ3), we anticipate children in the non-verbal condition to engage more with Reachy on a peer level. Since the non-verbal condition is the most "robotic" sounding, it might help manage expectations on Reachy's performance and engage children on an emotional level, where they feel more compelled to help Reachy in the collaborative task and in teaching Reachy during the post-test task. We expect the high verbal condition to have the least interactive engagement as it may begin to approach the uncanny valley [27]. Interaction and engagement are assessed by tracking eye contact and use of gestures, counting instances where feedback is offered to Reachy, while perception of the robot is assessed by tracking when Reachy's moves are undone and via the Godspeed questionnaire. We postulate that:

H3: *Children in the non-verbal condition will display the most engagement (i.e., increased eye contact, gesture usage, and feedback) and positive perception (i.e., fewer moves undone, higher Godspeed ratings) of the robot, while the high verbal condition will display the least.*

IV. System Architecture

The experimental setup is shown in Figure 1. Reachy is equipped with two cameras in its head, a microphone and speaker in its torso, and arms with 7 degrees of freedom. Reachy's torso and arms are roughly the size of a preteen, which makes it well suited for child-robot interaction studies. Reachy's overall form-factor is not overly humanoid; the antennas and asymmetrical eyes in particular give Reachy a distinctly "robotic" appearance.

The control software, written in python under ROS2, includes: (i) a ROS2 node managing Reachy's visual input, to retrieve the rings configuration during the collaborative task, (ii) components managing Reachy's sonic communication system, to perceive the child's commands and utterances and respond in accordance with the active condition, (iii) a solver for the Tower of Hanoi puzzle, (iv) a component responsible for all Reachy's movements and (v) an orchestrating node managing the interactions among all other nodes.

1) Vision: The vision system uses the right hand camera in Reachy's head to identify the rings' locations. First the image is processed to produce individual binary masks based off of HSV color ranges tailored to each of the ring colors. Morphological operations are performed on each of the masks to remove noise and join disconnected 'blobs' corresponding to the same ring. Taking a contour around each refined mask, the contour centroid corresponding to each ring color can then be localized within the x-y pixel grid of the image. By segmenting the image based on the locations of the posts in same x-y coordinates, each ring's corresponding post location is determined. The ring locations are obtained for each frame of the video and averaged with a moving window over ten frames to account for losses due to lighting changes or occlusions. If a ring cannot be located, its location is denoted as 'unknown', to indicate that either an illegal move has occurred or the ring has been removed entirely.

2) Communication: The audio interface includes one node for listening and one for speaking. The listening node was implemented using the speech recognition library[3] and building on [28], and operates over four-second intervals. Upon successful speech recognition, the transcript is parsed for keywords indicating a change of turn, agreement, dissent, and requests for help, some of which are used to trigger the reactions listed in Table II. Speech that doesn't align to a keyword is classified as an "interruption" and handled by randomly picking one among a set of generic responses to maintain the illusion of back and forth communication.

[3]https://pypi.org/project/SpeechRecognition/

The speaking node either uses text-to-speech (TTS) from the pyttsx3 library (for the high and low verbal conditions) or directly plays .wav audio files (for the non-verbal condition). The TTS output is run through a diode ring synthesizer which adds a weak vocoder effect to Reachy's voice to make it more "robotic" [29]. The sounds used for the non-verbal condition were generated using analogue subtractive synthesis on a Arturia Microbrute synthesizer. This synthesis principle has comparably few parameters and a sound character commonly associated with the early synthesizers of the 1970s and 1980s which have also been used to give a voice to many popular movie robots and computers. In all three vocalization conditions, the specific audio spoken or played either illustrates the robot's move or is dependent upon event triggers in the state of the game as shown in Table II.

3) Game solver: The Tower of Hanoi is a puzzle well suited for a robot to solve as, given the current location of the rings, the optimal sequence of moves can be generated through recursion. The best next move is then used to control Reachy's movement during its turn.

4) Motion: Reachy's move sequence has Reachy raise its arm to a rest position above the table, point first to the source post, then to the destination post, then return to the rest position. Reachy's sonic description of its move is synchronized with the pointing movements, to enforce clarity. Lastly, Reachy's head is controlled to gaze towards the puzzle or towards the child depending on whose turn it is.

5) Overseer: The whole interaction is monitored and run by an overseer node which tracks the status of the game, dictating whose turn it is and which speech prompts need to be published for the speaking node based on what's occurring. Turn hand-offs are dictated by the child verbally acknowledging when they finish a move for themselves or Reachy. During Reachy's turn, the robot first gazes down at the board, to update its representation of the game status, and identifies the best next move to make. The overseer then has a 50% chance of triggering one of the three speech events indicating Reachy is uncertain (events 5 and 6) or needs help with its move (event 7), as long as the turn is not the first or last of the game, before proceeding to perform its move. While Reachy is moving and until the turn is handed back to the child, updates from the listener node are logged and handled. Once the child has completed Reachy's move, the overseer checks that the move made matches the one Reachy indicated. If this is not true, the error speech event (event 9) is prompted before switching to the child's turn. If there is no error, Reachy lifts its head and the overseer switches the turn sequence from Reachy to the child. If it was the last move and the game is finished, the overseer prompts Reachy to celebrate.

During the child's turn, the overseer continuously processes any updates from the listener node, acting accordingly. If the "help" keyword is received, indicating the child has asked for help (event 3), Reachy responds affirmatively and the overseer hands the turn back to Reachy. If the child takes longer than 20 seconds to complete their turn (event 4), Reachy is prompted to offer help. If the child agrees,

TABLE II
TABLE OF ALL SPEECH PROMPTS USED DURING THE COLLABORATIVE ACTIVITY.

Event	High	Low	Non
1- Child made illegal move	“I don’t think you can make that move”	“Hmm, not right?”	Negative2NotRight
2- Child made non-optimal move	“Is that the best move that you can make?”	“Best move?”	Negative2NotRight
3- Child has asked for help	“Okay, I can help, let me see”	“Okay!”	Question3HelpYou
4- Child is taking too long to move	“Do you want me to help you?”	“Need help?”	Question3HelpYou
5- Reachy is uncertain on its turn	“I’m not sure which move I should make”	“Hmm, not sure”	Thinking1Hmmm
6- Reachy is thinking before its turn	“Hold on, let me think for a minute”	“Hmm, thinking”	Thinking1Hmmm
7- Reachy wants help with its turn	“Can you help me choose? I can’t decide”	“Help me?”	Thinking1Hmmm
8- The puzzle is completed	“Hooray, we did it!”	“Hooray!”	Happy1Hooray
9- Something has gone wrong	“Oh no! Something is wrong”	“Oh no!”	Negative2NotRight
10- The child has answered a question	“Okay”	“Okay”	Question3HelpYou
11- Child speaks during their turn, randomly selected	“Tell me what you’re thinking?”	“Thinking?”	Question1WhatNext
	“Can you explain your next move to me?”	“Hmm, what next?”	Question1WhatNext
	“Have you thought about what move to make?”	“What next?”	Thinking2JustThinking
	“I’m happy we’re playing together”	“Fun!”	Happy1Hooray
	“I’m having fun working with you”	“Fun.”	Happy1Hooray
12- Child speaks during Reachy’s turn, randomly selected	“I’m just thinking about how I want to move”	“Hmm, thinking”	Negative1ItsHard
	“It’s my turn, give me time to think”	“Shh, my turn”	Negative1ItsHard
	“I’m not sure what I should do next”	“My turn now”	Thinking1Hmmm
	“Wow, choosing a move is a little hard”	“It’s hard”	Thinking1Hmmm
	“Are we close to solving it yet?”	“Almost finished?”	Thinking2JustThinking
13- First part of move	“Can you move the ring from here”	“Here”	Question2HelpMe
14- Second part of move	“to here”	“to here”	Question3HelpYou

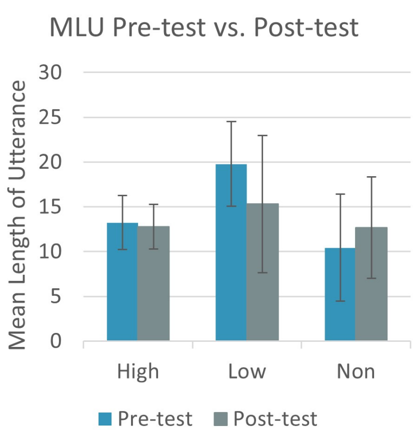

Fig. 2. Comparison of the average mean length of utterance in morphemes (MLU-m) between pre- and post-test, separated by vocalization condition (high-verbal, low-verbal, and non-verbal).

the turn gets handed back to Reachy. Otherwise Reachy continues waiting until the child hands off the turn in their own time. When the child communicates the end of their turn, the overseer switches the turn back to Reachy.

V. RESULTS

A. Speech

A Wilcoxon signed-rank test revealed no significant difference between the pre-test and post-test MLU values, in any condition. As shown in Figure 2, the high verbal condition had the least noticeable change as hypothesized. Low verbal showed the slight decrease we expected. However, the non-verbal condition showed an increase in MLU which contradicts our hypothesis entirely.

A comparison of the raw word counts between pre- and post-test reveals a perplexing trend: all conditions show an increase in word count between pre- and post-test. The high verbal condition had the largest change (mean difference = 31.2 words, std = 19.4) followed closely by low verbal (mean difference = 28.8 words, std = 56.1). Non-verbal condition showed a smaller increase (mean difference = 8.3 words, std = 41.5). A Kruskal-Wallis test comparing the average change in word count between conditions was non significant (p-value = 0.186). However, it is interesting that for the low verbal condition, while MLU trended down the average word count increased. This would suggest that though the utterances themselves were shorter, there were more of them being spoken. In contrast, the non-verbal condition saw an increase in both utterance length and word count.

B. Learning

The Tower of Hanoi has an optimal minimum number of moves needed to solve. For three rings, seven moves is optimal. While some children improved from the pre-test to the post-test, a Wilcoxon signed-rank test showed no significant change in the number of moves, in any condition. The mean change in number of moves was -0.6 in the high verbal condition, -0.2 in the low, and 0.2 in the non-verbal condition.

A possible contributing factor to this apparent lack of improvement is how the pre-test and post-test were framed to the children. They were told to work through the three ring version of the puzzle on their own, but to explain their moves to Reachy while they did so, with the goal of teaching Reachy. This may have led participants to focus more on their explanation than on the optimality of their moves.

Indeed, throughout the course of the experiment, children’s body language, facial expressions, and use of gestures served as external indications that reasoning and learning were taking place. Some children had visible “eureka” moments when they figured out the moves to solve the puzzle (see Figure 3, left). There were also instances of participants

Fig. 3. Example of a "eureka" moment (left). Participant covers a gasp of realization. Example of gestures used in planning a move sequence (right). The participant points out how the next three moves should occur.

miming out their next few moves, indicating the presence of strategic thinking. They often did this during the "dead space" throughout the interaction, either while Reachy was planning its move or right after they had completed their move but hadn't yet passed the turn back to Reachy, as shown in Figure 3 on the right.

C. Interaction

The evolution of the use of gestures and eye contact can be seen in Figure 4. Regardless of condition, the gesture and eye contact usage during the experiment is much lower than during the pre- and post-tests. However, Friedman tests found no significant difference between pre-test, experiment phase, and post-test among vocalization conditions.

General trends are still observable though. Mean normalized gesture count increases between the pre-test and post-test in all three conditions, maybe due in part to mimicry since during the experiment Reachy communicates its turn by pointing to the move it wants to make. Conversely, only the high and low verbal conditions see the same increase for eye contact. Mean normalized eye contact count decreases between the pre- and post-test in the non-verbal condition.

We also tracked how often the children provided feedback to Reachy during the experiment. Feedback was considered to be any commentary on how good or bad Reachy's move was, any encouragement or discouragement directed at Reachy, and any hinting or suggestions for moves. The evolution of mean normalized usage of feedback through the experiment can be compared among conditions as seen in Figure 4.

In all conditions feedback increased between the first half of the experiment and the second half. The non-verbal condition appears to have the largest increase, although high standard deviations make the validity of the trend inconclusive. A Kruskal-Wallis test comparing feedback throughout the experiment among the vocalization conditions did not show any significant difference.

Children were not told (and did not assume) that Reachy always knew the optimal move; we thus also measured the number of times they undid one of Reachy's moves to see how confident they were in their own and Reachy's strategic thinking. No child in the high verbal condition attempted to undo any of Reachy's moves. In the low verbal condition, one out of five participants chose to undo two of Reachy's moves. In the non-verbal condition, four out of six participants chose to undo one of Reachy's moves. Their willingness to disagree with the validity of a move shows that at least in the non-verbal condition the children are thinking critically about Reachy's moves. This, in combination with providing feedback, could indicate a relationship between the vocalization and the children's confidence in the robot's ability to think strategically.

TABLE III
GODSPEED QUESTIONNAIRE ITEMS AVERAGED BY CONDITION.

Godspeed Category	High	Low	Non
Anthropomorphism	2.7 ± 0.8	3.2 ± 1.0	2.9 ± 1.2
Animacy	3.1 ± 0.9	3.4 ± 1.1	2.9 ± 1.1
Likeability	4.4 ± 0.7	4.7 ± 0.5	4.4 ± 0.6
Perceived Intelligence	4.1 ± 1.0	$\mathbf{4.5 \pm 0.6}$	$\mathbf{3.9 \pm 0.9}$
Perceived Safety	3.7 ± 1.2	3.3 ± 1.2	3.9 ± 1.4

The Godspeed questionnaire provided a way to gain insight into the children's impressions of Reachy. The aggregated ratings for each experimental condition are reported in Table III. Reachy scores in the middle on anthropomorphism and animacy, slightly above middle on perceived safety, and high in likeability and perceived intelligence. The experimental conditions appear to yield no significant differences. A Kruskal-Wallis comparison produced a marginally significant p-value = 0.07 for perceived intelligence. Subsequent two way comparisons showed that the difference between the low verbal and non-verbal conditions on perceived intelligence was statistically significant (p-value = 0.025), with the non-verbal condition ascribing a lower perceived intelligence to Reachy.

VI. CONCLUSION

In this paper, we sought to explore the effect that a robot's ability to speak has on children's speech patterns, learning, and engagement in the interaction. Sixteen children aged 7-14 participated in a collaborative version of the Tower of Hanoi puzzle with a Pollen Robotics Reachy robot vocalizing at one of three vocalization conditions: high verbal, low verbal, or non-verbal. No significant conclusions could be made on whether the robot's speaking abilities affected the children's learning. However, contrary to the precedent of entrainment expected from human-human interactions, children in the non-verbal condition actually increased the quantity and complexity of their speech. This suggests that children decoupled the robot's ability to produce words from its ability to understand them. This finding, if confirmed in further studies, can have interesting repercussions on social HRI well beyond child-robot interactions. Also in the non-verbal condition, children displayed a significantly lower perception of Reachy's intelligence. This, combined with higher rates of providing feedback, more willingness to undo Reachy's moves and increased speech suggests that there may be a link

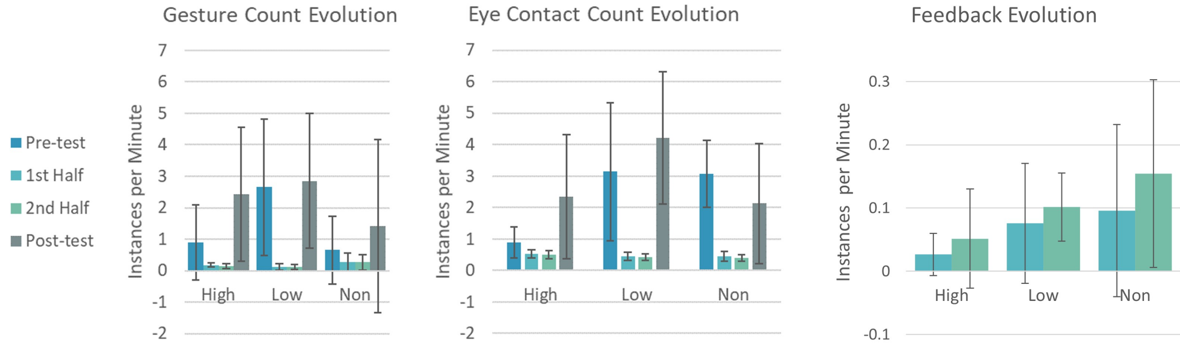

Fig. 4. Comparison of the mean normalized count of gestures and eye contact (middle) evolution through each phase of a full session, separated by vocalization condition. Comparison of the mean normalized count of feedback (right) through the collaborative session, separated by vocalization condition.

between a robot's speaking ability and children's confidence in that robot's overall intelligence and capability. Specifically, this suggests that children may have perceived Reachy as a less capable peer who could understand language but not yet speak, who nevertheless was someone worth teaching. This could be especially pertinent in learning contexts as a trigger for the protégé effect.

References

[1] W. Johal, "Research trends in social robots for learning," *Current Robotics Reports*, 2020.

[2] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*, 1978.

[3] A. Thomaz, G. Hoffman, and M. Cakmak, "Computational human-robot interaction," *Foundations and Trends in Robotics*, vol. 4, no. 2–3, 2016.

[4] P. Gurunath Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Computer Speech Language*, vol. 63, 2020.

[5] J. N. Bohannon III and A. L. Marquis, "Children's control of adult speech," *Child Development*, pp. 1002–1008, 1977.

[6] C. C. Chase, D. B. Chin, M. A. Oppezzo, and D. L. Schwartz, "Teachable agents and the protégé effect: Increasing the effort towards learning," *Journal of Science Education and Technology*, vol. 18, no. 4, 2009.

[7] H. Chen, H. W. Park, and C. Breazeal, "Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement," *Computers & Education*, vol. 150, 2020.

[8] G. Gordon, C. Breazeal, and S. Engel, "Can children catch curiosity from a social robot?" in *ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2015.

[9] J. Kory and C. Breazeal, "Storytelling with robots: Learning companions for preschool children's language development," in *IEEE Int. Symposium on Robot and Human Interactive Communication*, 2014.

[10] H. W. Park, M. Gelsomini, J. J. Lee, and C. Breazeal, "Telling stories to robots: The effect of backchanneling on a child's storytelling," in *ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2017.

[11] K. Ryokai, C. Vaucelle, and J. Cassell, "Virtual peers as partners in storytelling and literacy learning," *Journal of computer assisted learning*, vol. 19, no. 2, pp. 195–208, 2003.

[12] J. Nasir, P. Dillenbourg, U. Norman, and B. Bruno, "When positive perception of the robot has no effect on learning," in *IEEE Int. Conf. on robot and human interactive communication*, 2020.

[13] D. Hood, S. Lemaignan, and P. Dillenbourg, "When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting," in *ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2015.

[14] A. Stolcke and J. Droppo, "Comparing human and machine errors in conversational speech transcription," *arXiv preprint*, 2017.

[15] J. Wuth, P. Correa, T. Núñez, M. Saavedra, and N. B. Yoma, "The role of speech technology in user perception and context acquisition in hri," *Int. Journal of Social Robotics*, vol. 13, no. 5, 2021.

[16] S. Paepcke and L. Takayama, "Judging a bot by its cover: An experiment on expectation setting for personal robots," in *ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2010.

[17] M. Scheeff, J. Pinto, K. Rahardja, S. Snibbe, and R. Tow, "Experiences with sparky, a social robot," in *Socially intelligent agents*, 2002.

[18] C. Breazeal, *Designing sociable robots*. MIT press, 2004.

[19] M. Schwenk and K. O. Arras, "R2-D2 reloaded: A flexible sound synthesis system for sonic human-robot interaction design," in *IEEE Int. Symposium on Robot and Human Interactive Communication*, 2014.

[20] R. Read and T. Belpaeme, "People interpret robotic non-linguistic utterances categorically," *Int. Journal of Social Robotics*, vol. 8, no. 1, 2016.

[21] H. Ritschel, I. Aslan, S. Mertes, A. Seiderer, and E. André, "Personalized synthesis of intentional and emotional non-verbal sounds for social robots," in *Int. Conf. on Affective Computing and Intelligent Interaction*, 2019.

[22] F. A. Robinson, M. Velonaki, and O. Bown, "Smooth operator: Tuning robot perception through artificial movement sound," in *ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2021.

[23] Š. Beňuš, "Social aspects of entrainment in spoken interaction," *Cognitive Computation*, vol. 6, no. 4.

[24] R. Levitan, A. Gravano, L. Willson, Š. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*.

[25] M. C. Welsh, T. Satterlee-Cartmell, and M. Stine, "Towers of hanoi and london: Contribution of working memory and inhibition to performance," *Brain and Cognition*, vol. 41, no. 2, 1999.

[26] M. L. Rice, F. Smolik, D. Perpich, T. Thompson, N. Rytting, and M. Blossom, "Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 2, pp. 333–349, 2010.

[27] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robotics Automation Magazine*, vol. 19, no. 2, 2012.

[28] Y. Ma, "Help a humanoid robot understand my verbalised intentions!" 2021.

[29] R. Maure, "Help reachy recognise me, and refer to me and its environment with pointing gestures!" 2021.