Can a Signing Virtual Human Engage a Baby's Attention?

Setareh Nasihati Gilani
David Traum
University of Southern California
sngilani@ict.usc.edu
traum@ict.usc.edu

Rachel Sortino
Grady Gallagher
Gallaudet University
rachel.sortino@gallaudet.edu
grady.gallagher@gallaudet.edu

Kailyn Aaron-lozano
Cryss Padilla
Gallaudet University
kailyn.aaron-lozano@gallaudet.edu
cryss.padilla@gallaudet.edu

Ari Shapiro
University of Southern California
shapiro@ict.usc.edu

Jason Lamberton
Gallaudet University
jason.lamberton@gallaudet.edu

Laura-Ann Petitto*
Gallaudet University
laura-ann.petitto@gallaudet.edu

ABSTRACT

The child developmental period of ages 6-12 months marks a widely understood "critical period" for healthy language learning, during which, failure to receive exposure to language can place babies at risk for language and reading problems spanning life. Deaf babies constitute one vulnerable population as they can experience dramatically reduced or no access to usable linguistic input during this period. Technology has been used to augment linguistic input (e.g., auditory devices; language videotapes) but research finds limitations in learning. We evaluated an AI system that uses an Avatar (provides language and socially contingent interactions) and a robot (aids attention to the Avatar) to facilitate infants' ability to learn aspects of American Sign Language (ASL), and asked three questions: (1) Can babies with little/no exposure to ASL distinguish among the Avatar's different conversational modes (Linguistic Nursery Rhymes; Social Gestures; Idle/nonlinguistic postures; 3rd person observer)? (2) Can an Avatar stimulate babies' production of socially contingent responses, and crucially, nascent language responses? (3) What is the impact of parents' presence/absence of conversational participation? Surprisingly, babies (i) spontaneously distinguished among Avatar conversational modes, (ii) produced varied socially contingent responses to Avatar's modes, and (iii) parents influenced an increase in babies' response tokens to some Avatar modes, but the overall categories and pattern of babies' behavioral responses remained proportionately similar irrespective of parental participation. Of note, babies produced the greatest percentage of linguistic responses to the Avatar's Linguistic Nursery Rhymes versus other Avatar conversational modes. This work demonstrates the potential for Avatars to facilitate language learning in young babies.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; **Empirical studies in HCI**.

*Principal Investigator, Corresponding

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
IVA '19, July 2–5, 2019, Paris, France
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6672-4/19/07.
https://doi.org/10.1145/3308532.3329463

KEYWORDS

Empirical Studies, Social Impact, Multi-agent Interaction

ACM Reference Format:
Setareh Nasihati Gilani, David Traum, Rachel Sortino, Grady Gallagher, Kailyn Aaron-lozano, Cryss Padilla, Ari Shapiro, Jason Lamberton, and Laura-Ann Petitto. 2019. Can a Signing Virtual Human Engage a Baby's Attention? . In *ACM Int'l Conference on Intelligent Virtual Agents (IVA '19), July 2–5, 2019, Paris, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308532.3329463

1 INTRODUCTION

Many AI systems have been designed for facilitating language learning by adults, and to a lesser extent, children [6, 11, 37]. However, there is a significant paucity of work on AI systems for young infants despite the widely understood critical importance that this developmental period has for healthy language and cognitive growth, and related reading and academic success [24, 29]. Children have proven to be a challenging population to design language learning technology, and some technologies designed for children have been shown not to facilitate language learning. Many studies have shown that children who receive linguistic stimuli from television do not learn as much as those who receive the same linguistic stimuli from live adults [13, 14, 32]. Our particular interest is young babies who lack the necessary language exposure in early life. Based on discoveries of brain-based "critical periods" of human development for language (e.g., ages 6-12 months; [2, 25, 29, 38]), a growing body of neuroscience research has identified the potentially devastating impact that minimal language exposure during this particular period of child development can have on all children's linguistic, cognitive, and social skills, be they hearing or deaf infants [24, 25, 29, 38]. As such, technology that can help fill a language-exposure gap can have a tremendous impact for social good in these populations, especially in young deaf babies who can experience dramatically reduced or no access to usable linguistic input during this period.

One recently introduced AI system, called RAVE (Robot, AVatar, thermal Enhanced language learning tool), was designed specifically for babies within the age range of 6-12 months [20, 21, 34]. RAVE consists of two agents: a virtual human (provides language and socially contingent interactions) and an embodied robot (provides socially engaging physical cues to babies and directs babies' attention to the virtual human). The virtual human is an expressive agent (both in facial expressions and posture) that can produce a natural signed language as linguistic input. Although there is

much literature on comparison of having a virtual human versus a robot in different applications [16–18], there are very few works that benefit from having a virtual human and a robot at the same time [1]. RAVE brings together science from multiple disciplines to explore the potential for technologies such as functional Near Infrared Spectroscopy (fNIRS) brain imaging (measures the baby's higher cognition), thermal IR imaging (measures the baby's emotional engagement), robotics, and virtual humans in an attempt to positively influence the learning process. Building on fNIRS studies of infant brains and language processing across infancy, one unique feature of RAVE is the specific linguistic nature of the avatar's language, which contains phonetic-syllabic rhythmic temporal patterns. These patterns precisely match the infant brain's peaked sensitivity to language patterns during the critical period of language learning [25–27, 29, 30].

While RAVE appears to be using the best available research to inform its design, there is still the question of whether it could be used to facilitate babies' language learning. Is there evidence that the babies' behaviors are influenced and/or facilitated by the avatar's behaviors? Is there a principled and predictable relationship? In this paper, we perform a quantitative evaluation of the RAVE system with babies. In particular, we focus on the baby's interaction with the avatar, that is providing multiple kinds of social and linguistic behavior. We asked the following research questions:

(1) Do babies attend to the avatar and respond to its communicative behaviors?
(2) Can babies with little or no exposure to ASL distinguish among the avatar's different conversational modes (Linguistic Nursery Rhymes; Social Gestures; Idle/nonlinguistic postures; 3rd person observer)?
(3) If they can distinguish between different Avatar's roles, how do they react to different roles? Can an avatar stimulate babies' production of socially contingent responses, and crucially, nascent language responses?
(4) What, if anything, is the impact of the presence or absence of parents' participation in the conversational interaction?

Below we report the results from an experimental study using the RAVE system in order to evaluate the system's performance regarding the above questions, with the ultimate goal of evaluating an avatar's potential to engage the baby's attention.

2 BACKGROUND AND MOTIVATION

Language is the principal system of expression and communication for humans and arguably the most prominent cognitive and cultural tool that distinguishes human beings from other species. Acquiring language commences from birth aided by multiple factors, including brain-based sensitivities to aspects of the specific rhythmic patterning of human languages, observation, and engagement in social interactions with the outside world [5]. Language exposure plays an important role in infants' early development of linguistic abilities. Ages 6-12 months is widely recognized as a critical developmental period for language [25, 29, 36] It is during this period that babies acquire essential phonetic-syllabic segments unique to their native language, which make possible their ability to acquire their native language's vocabulary, discern their language's distributional and syntactic regularities, and crucially, to engage in letter-to-phonetic segment mapping in early successful reading [15, 29]. In early life, select brain sites participate in early human language learning (such as the Planum Temporale in the Superior Temporal Gyrus [31]), which are sensitive to specific rhythmic temporal patterns at the nucleus of phonological structure found in all languages (spoken and signed) [2, 25, 31, 38]. Exposure to these patterns is crucial for the development of this brain sites to support later healthy language, phonological, reading, and cognitive growth [25, 29]. Children deprived of this early exposure specifically during the ages of 6 to 12 months may face dire consequences such as delays in cognitive, linguistic, reading, and social skills which may last for years [33, 38] with accompanying devastating lifelong impact of reading and academic success [29, 36].

Intriguingly, these developmental brain sensitivities also exist in deaf babies learning a natural signed language, and it develops on the identical maturational time table as hearing babies [26, 27, 30]. This universal brain sensitivity enables young sign-exposed babies the early life language input that permits them to build a sign phonological system vital to letter-to-sign-phonetic segment mapping in successful reading acquisition [29]. All babies who miss exposure to the patterns of their natural language in early life (be it a signed or a spoken language) are rendered at risk for language and reading delays spanning life [25, 29].

Given that 91.7% of young deaf babies are born to non-signing families (hearing) [8], in these families, quickly learning a new signed language can become a challenge for the parents. There are some speech-based interventions such as cochlear implants designed to make available spoken language to the young deaf baby [9, 41]. However, most of these tools cannot be deployed until the ages of 18-24 months. While efforts have begun to implant children at younger ages (from ~8 months), precise adjustments, tuning of the device, as well as intensive speech training, still typically begins after ages 18-24 months and proceeds for months into years thereafter [22]. Thus, this is well past the early critical period for learning phonological units, phonological segmentation, categorization and mapping, and sequencing distributions - all vital to optimal, healthy language learning and reading. As such, there is a pressing opportunity for AI technology that can provide signed language input in the critical period of 6-12 months [20, 21, 34].

3 THE RAVE SYSTEM

The RAVE system includes two behavioral agents (a physical robot and a virtual human avatar on a screen) that can provide visual behaviors, as well as several sensor devices: an eye-tracker, thermal camera, and an interface for indicating communicative baby behaviors. Detailed description of the system's constituent components and dialogue algorithms are presented in [34], and [20], respectively. A preliminary evaluation of the system has been presented in [21]. Here we summarize and briefly identify the deployment of this system to motivate our experimental design. To that end, we review the components, and briefly describe the behavior selection procedure.

3.1 Agents

The avatar provides the linguistic stimuli to the baby. It was built using a real-time character animation system [35], and facial scans

(a) (b) (c) (d)

Figure 1: Frames of Avatar doing the BOAT Nursery Rhyme.
The four frames were selected (in order from left to right) from a fluid video clip of avatar signing where each frame represents a silent sign-phonetic-syllabic contrastive unit as produced with the hands in the ASL visual nursery rhyme "BOAT-ON-WAVE". In formal linguistic analyses, these contrastive phonetic-syllabic units are notated as follows: 1a /B/+low; 1b /B/+modulation+high; 1c /5/+modulation+high; 1d /5/+modulation+low. These phonetic-syllabic linguistic units are not produced in isolation like a list. Instead, they are bound into fluid movements that form rule-governed, grammatical clausal, phrasal, and syntactic constructions in all natural languages, here ASL.

from a Light Stage [4]. Avatar behaviors were built by motion capturing a real human deaf native signer of ASL.
The robot is based on the open-source Maki platform from Hello Robot [23]. The main purpose of the robot is to gain the baby's attention and to shift the baby's gaze to the avatar. Prior research focusing on the robot component of RAVE demonstrated the success of this outcome [34]. Greater detail about the robot design and impact are presented in [34].

3.2 Perceptual Modules

Multiple real-time sensory inputs were used to assess a baby's state of engagement (i.e., attentional, emotional/arousal) to facilitate a socially contingent interaction:

(1) Eye gaze is used as a measure of behavioral response of attention. It is categorized as either looking at the Robot, looking at the avatar, looking somewhere in between them, or directed to something else. A Tobii Pro X3-120 [40] was used to capture the baby's eye gaze at the rate of 120 Hz.
(2) The use of thermal Infrared (IR) imaging, facilitates monitoring the baby's changes in emotional/arousal and attentional engagement as indicated by their Autonomic Nervous System (ANA) responses; i.e., parasympathetic and sympathetic [10, 39] and it is used as a trigger as to when the agents should provide linguistic stimuli to the baby.
(3) A human observer interface was used to capture the baby's communicative and social behaviors. This feedback from the baby was another input to the system's dialog manager.

3.3 Avatar Behaviors

For the purposes of evaluating the ability of the system to engage in socially contingent interaction with the baby, we focus the analysis on the avatar's different conversational modes, including categories for noncommunicative behavior, social dyadic and triadic (including the robot) behaviors, and those that contain developmentally appropriate linguistic features. The categories used are as follows:

(1) **Idle behaviors ("Idle")** are nonlinguistic/nonsigning, and non socially communicating neutral bodily postures, e.g., arms at side with typical slight body shifting). This behavior typically occurred when the robot has the floor and is engaging with the baby, and avatar is looking at the robot or the baby as a 3rd-party conversationalist.
(2) **Nursery Rhymes ("NR")** are linguistic stimuli (with specific rhythmic temporal patterns), such as the "BOAT-ON-WAVE"[1] nursery rhyme in ASL. [2]
(3) **Social Gestures ("Social")** include universal social routines (e.g., BYE-BYE, HI), conversational fillers (e.g., Affirmative Head Nod), and/or short lexical phrases such as YES!
(4) **3-Way behaviors ("3-Way")** are avatar's communicative interactions that were directed to both the baby and the robot, such as "LOOK-AT-ME" (grammatically inflected in the grammar of ASL to include both the baby in second person role and the robot in third person role.

3.4 Agent Behavior Selection

The design of the social contingency uses perceptual modules of the system rather than having a fixed protocol. The dialogue management module is constantly updating its internal state based on the sensory input signals as well as the feedback/callback signals from agents. A rule-based decision system is used to output signals that are sent to the avatar and the robot. In other words, the avatar would adjust its behavior according to the babies' behavioral responses and attentional/emotional engagement in order to maintain a socially contingent interaction and would provide linguistic input to the baby upon seeing engagement from the baby and proof of its

[1]The formal linguistic notation of natural signed languages, such as ASL, uses glosses showing approximate English translations in capital letters and appear here in these original cross-linguistic sign-phonetic analyses originated and designed by senior author/P.I., L.A.Petitto.
[2]While the ASL NR is unique to the ASL language and Deaf culture, a rough semantic neighbor in the English language would be "Row-Row-Row-Your Boat" a simple repetitive rhythmic rhyme with approximate versions in many languages worldwide.

Figure 2: Experimental Setup (Side View)

attention (via the triggering from the thermal IR imaging). Detailed explanation about this system is presented in [20].

4 DESIGN OF NURSERY RHYME BEHAVIORS

Linguistic Patterns provide the vital linguistic stimuli for the baby. Nursery Rhymes were constructed with the identical rhythmic temporal patterning that matched the infant brain's specific neural sensitivity to that rhythmic temporal patterning [2, 26, 27, 29, 30]. All Nursery Rhymes were built with the maximally-contrasting rhythmic temporal patterning in 1.5 Hz alternations [26, 27]. Specific phonetic-syllabic contrasts that infants first begin to perceive and produce in language development (ages 6-12 months) were used. These include 3 maximally-contrasting phonetic hand primes in ASL: /5/, /B/, /G/ with contrastive transitions /B/⇒/5/, /5/⇒/F/, /G/⇒/F/, plus allophonic variants. Below we provide some examples of the Nursery Rhymes as per formal analyses in the formal discipline of Linguistics analyses for ASL, which had baby-appropriate lexical meaning with their respected action patterned sequences:

- **BOAT**[3]**(Phonetic-Syllabic units /B/, /5/)**
  (1) BOAT (/B/, double bounce=noun; palms in/+ low center)
  (2) BOAT-on-WATER (/B/+modulation, palms in/+ high center)
  (3) WAVE (ROLLING) (/5/+SAME modulation, palms out/+ high center)
  (4) WAVE (ROLLING) (/5/+SAME modulation, palms down/+ low center)
- **PIG (Phonetic-Syllabic unit: /5/)**
  (1) PIG (/5/, Chin)
  (2) PET (/5/, called "center space" in Linguistic sign notation)
  (3) HAPPY (/5/ + double-handed, Chest)
- **FISH (Phonetic-Syllabic unit: /B/ (allophonic))**
  (1) FISH (/B/, "center space")
  (2) FINS (/B/+double-handed, Head)
  (3) SWIMS (away) (/B/, Cross-Body)
- **CAT(Phonetic-Syllabic units: /5/; /G/allophonic; /BENT5/; /F/)**
  (1) Grandma has red cat [/5/⇒/G/] and [/G/⇒/F/]
  (2) Grandma has white cat [/5/⇒/BENT5/] and [/BENT5/⇒/F/]

5 EXPERIMENT PROTOCOL

To address the questions about the impact of the avatar behaviors on babies, we designed an experiment whereupon babies interacted with the system in a controlled setting. While previous investigations were conducted with over 40 babies focusing on RAVE system's functionality [20, 21, 34], the present study provides a first-time evaluation focusing specifically on the babies and the relationship between their behaviors and the avatar's behaviors.
4 babies participated in an intensive case study. One of the babies had been exposed to ASL (sign-exposed) and 3 had no previous exposure to a signed language (non-sign-exposed). Given our hypothesis regarding the universal nature of the rhythmic temporal patterning underlying human language phonological organization, which was specifically built into the avatar's linguistic stimuli, a key design feature of the present study was that non-sign-exposed babies would constitute a powerful test of this hypothesis. While it would have been ideal to have a larger sample size, it has been well established among scientists in this field (those studying signing/non-signing deaf/hearing children) that the vulnerability and rarity of this population renders traditional sample sizes unrealistic. Because of their theoretical power, smaller sample sizes have routinely appeared in prominent publications involving the rare sign-exposed infants (e.g., [27]; T=3 sign-exposed infants, [30], T=2 sign exposed babies).

Babies were seated on their parent's lap facing the system (Figure 2). Multiple cameras were used to record the baby (and the parent) from different angles. Each baby's experimental session lasted until the baby became distracted or entered a fussy state in which case we immediately ceased the session. The experiment consisted of several steps: upon arrival, the baby and the parent were greeted and introduced to the robot and the avatar; which has been proved to be useful [19]. Next, a calibration process (for thermal IR Imaging and Tobii eye tracking systems), followed by the interaction session.

To make the baby feel comfortable and involved in this multiparty interactions and also to introduce the agents as conversational partners, we begin the experiment with a familiarization episode with the help of an experimental assistant who interacts with the agents. At the beginning, the assistant talks as well as signing to the robot to wake him up. The robot wakes up, lifts his head, blinks, sees the baby and nods as an acknowledgment of baby's presence. Then it turns toward the avatar. Avatar sees the robot, turns to him, nods, then turns back to baby and waves to the baby. Avatar takes the floor and signs HELLO and GOOD MORNING to the baby to begin the interaction. At this point the assistant signs GOODBYE to the baby and the agents, and departs from the experiment room, leaving the baby to interact with the system.

The avatar's socially contingent interaction session with the baby began after the assistant left the room (Condition 1). At approximately 2.5 minutes into the experimental session, parents were permitted to interact as per their natural inclination (Condition 2). Throughout the experiment, parents wore sunglasses which were meant to block the technology from recording eye-tracking artifacts from their eyes. Note that none of the perceptual components were monitoring the parent, so none of the agent's behaviors were contingent directly on the parent [20].

[3]The sequence of frames of this Nursery Rhyme is depicted in Figure 1.

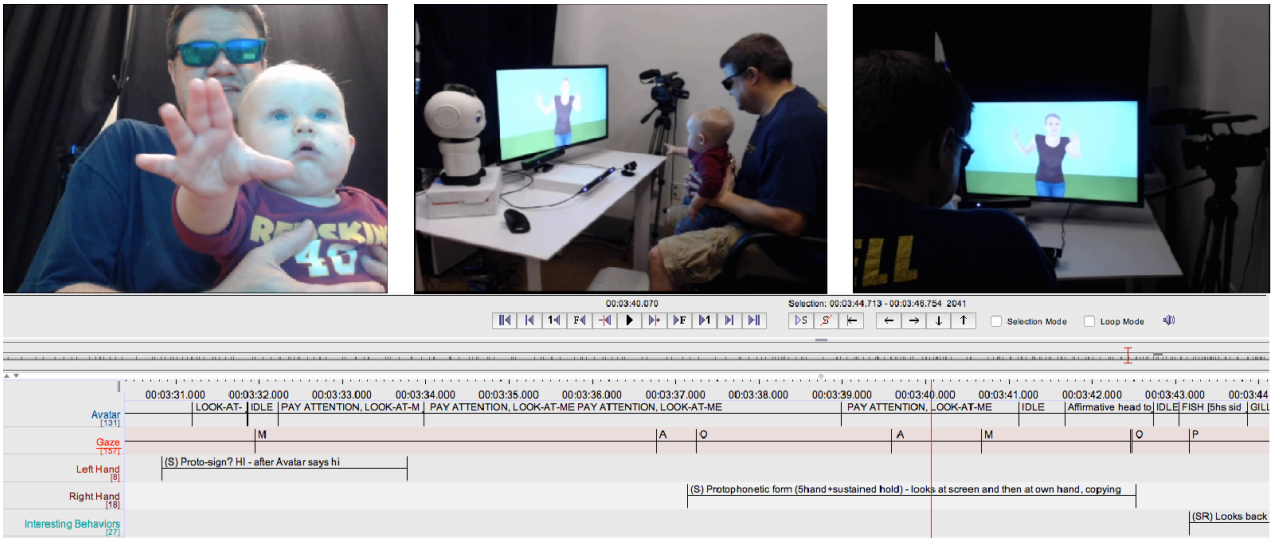

Figure 3: Annotation example using ELAN

6 RESEARCH QUESTIONS AND EVALUATION METRICS

Nasihati Gilani et al. [20, 21], Scassellati et al. [34] report observations of several kinds of baby's spontaneous behavioral responses to the Avatar conversational modes. All baby spontaneous responses were produced to the RAVE system with one exception, when baby turned from RAVE to the parent and back to RAVE thereby exhibiting social referencing or shared/joint visual attention. (Here, babies check to ensure that parents are jointly seeing/looking at a referent in question.) The babies' spontaneous responses cohered around the following three types of behaviors:

(1) **Linguistic Responses ("ling")** include manual babbling, the production of manual proto sign-phonetic units, proto-signs, and imitations of signs (i.e., the baby imitates or copies what it sees the Avatar is producing);
(2) **Social/Gestural Responses ("S/G")** include pointing, waving, clapping hands or attempts to copy the agents' behaviors, or social referencing;
(3) **Sustained Visual Attention ("SVA")** indicates the baby being visually transfixed on the agents for atypically extended periods for infants, defined as greater than one second for this study.

Note that these categories are not mutually exclusive. A baby can exhibit SVA, that is be visually transfixed on the avatar and simultaneously be producing social/gestural responses or linguistic responses. Producing visually transfixed attention, social gestures and especially linguistic behaviors are an indication that system is successful at soliciting babies' interaction. Frequency analyses of the baby's behaviors throughout the experiments provided us with a good insight of the babies' behavioral pattern.

We can now operationalize the main research questions raised in section 1. Regarding the first question (do babies attend to the avatar and respond to its communicative behaviors?); one possibility is that babies do not see the avatar, or the agents collectively, as interesting social interlocutors or respond to them at all. Another possible outcome is that the infants may enter an agitated mode upon confronting an unknown (or "strange") situation such as the RAVE system [7]. We use the percentage of baby's responses to the avatar as a metric to evaluate the overall system's impact and performance in terms of engaging the babies.

The second question asked whether babies can differentiate among the different avatar conversational modes even though it's unlikely that these young babies understand the semantic content of the ASL language productions (i.e., vocabulary meanings, syntax, etc.). If so, this would corroborate the now-classic studies in infant language processing that demonstrates their ability to discriminate categorically among classes of linguistic units (such as phonetic-syllabic units) in different languages based on their contrastive patterning (peaked between ages 6-12 months; [2, 25, 27]). Here, we examined the baby's response rate to the avatar's different conversational modes.

The third question we asked is of particular scientific interest concerning the mechanisms that drive early language learning: does the avatar's specifically linguistic productions garner the baby's attention, and in particular, does the avatar's linguistic productions garner linguistic responses from the babies? We hypothesize that it is the linguistic patterning that is important in the avatar's productions, not its modality of language production and reception (here, signed; [25, 26]). Specifically, we claim that since we are correctly hitting on just the right temporal patterning in the avatar's productions, then all babies would be engaged by the avatar's language productions over other social and communicative conversational modes - indeed, even in babies who were never exposed to a signed language. We hypothesized that they would react with more linguistic content when the avatar was in this category, as compared to when the avatar was in its other conversational roles.

Finally, the fourth question concerns whether having the parent intervene in the conversational interaction is beneficial in terms of facilitating the system's overall language learning goals, or would it

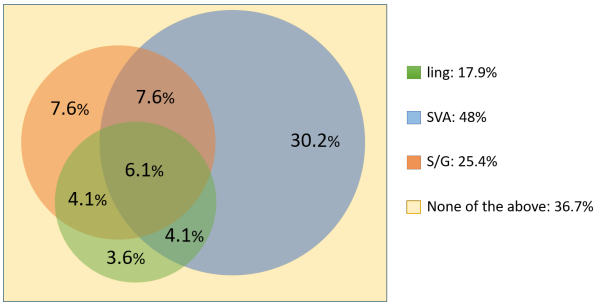

Figure 4: Frequency of Baby's Behaviors

have an adverse effect? Perhaps babies would feel more comfortable when they find themselves in a familiar and natural situation in which their parent is part of the interaction and acknowledging their social referencing other than standing still and not reacting to any of their behaviors (which is definitely not a routine for parents). On the other hand, the intervention from the parent might be distracting for the baby and steal the attention from the avatar; as a result, babies may turn to parents for interaction instead of engaging with the system. The first metric to assess this is the overall response rate across conditions. Furthermore, studying the distribution of baby responses across conditions would give us detailed insight on the parent's impact on this social interaction.

7 RESULTS

The video recordings of the babies' full range of behavioral responses to the Avatar were transcribed and coded by trained experts in the field of developmental cognitive neuroscience, child development, linguistics and sign-linguistics with reliability checks (initial r=0.83, post discussion r=1.00). ELAN [3] was used for annotating the baby's behavior and marking the times of avatar and robot's behaviors. A screenshot of the tool along with different tiers is shown in Figure 3. Table 1 gives an overview of the analysis of the four subjects. Here, we present the results of our analyses in two parts. First, we show the interactions between the baby and the avatar, the babies' specific categories of spontaneous behavioral responses, and their relative frequencies. Second, we show the corresponding analyses regarding the parents, and the impact of parent's intervention on baby's behaviors toward the system.

7.1 Baby and Avatar

In answer to questions 1 & 2 above (do babies attend to the avatar?; do they differentiate among its conversational modes?), a frequency analysis of responses to avatar behaviors was conducted using the categories of behavioral responses of babies stated in section 6 (linguistic; sustained visual attention; social/gestural responses). Analysis was done based on the occurrences of specific behaviors as its the convention in child developmental sciences. Overall, babies responded to more than 60% of avatar's behaviors ($M = 61.8, SD = 6.9$). Figure 4 shows a Venn diagram of the four discrete categories of baby behavioral responses to the avatar as well as their relative frequencies. The overlapping portions show cases where the baby responded in more than one way to the same avatar behavior. As shown, the babies' transfixed sustained visual attention (SVA)

Subject ID	Previous Exposure	Age	Experimental Session
1	No	9m14d	266s
2	Yes	7m5d	296s
3	No	9m4d	222s
4	No	8m26d	168s

Table 1: Subjects' previous exposure to sign language, age at time of testing, and the length of the experimental session

constitutes the biggest portion of babies' behavioral responses to the avatar (48% overall). Next, we studied the relationship between the Avatar's behaviors and the baby's response rate. Results show that babies' responses were not equally distributed across different types of Avatar's behaviors. Babies response rate was 37% to Avatar Idle mode, 85% to Nursery rhyme , 75% to social gestures and 84% to avatar's 3-way behaviors. Note that the distribution of avatar behaviors was also not uniform: 13% of the avatar's behaviors were NR, 13% 3-way, 36% were social, and the remaining 38% were idle.

Following from question 3 above (does the Avatar's linguistic behavior impact linguistic productions in the baby?), as a first step in our analyses, we observed that the babies produced their greatest percentage of spontaneous responses to the Avatar when the avatar was producing linguistic Nursery Rhymes. Babies produced spontaneous behavioral responses to 85% of the Avatar's Linguistic Nursery Rhymes, 84% of the 3-Way conversational turns, 75% of the Avatar productions when in Social Gesturing conversational turn, but only 37% of times when the Avatar was idle. The babies' responses to the Avatar's *actions* (NR, Social, 3-Way) were significantly more compared to when the avatar was in its idle mode ($t = 3.35, p = 0.01$). Thus, the babies do appear to attend to and to respond to the Avatar's different conversational modes, with the babies' greatest percentage of responses being when the Avatar was producing Linguistic Nursery Rhymes.

Further to question 3 above, we conducted a frequency analysis of the different baby behaviors in response to the avatar's behaviors. Figure 5 shows the rate of each baby behavior in response to each category of avatar's production. Note that the bars in each category do not necessarily need to add up to 1, because sometimes the baby responds with multiple response types, as shown in Figure 4.
As shown in Figure 5, the babies responded differently when the avatar was in the Linguistic Nursery Rhyme conversational mode versus other modes (Social, Idle, 3-Way). The babies produced the largest percentage of linguistic responses to the avatar's Linguistic Nursery Rhymes (31% to Nursery Rhymes vs 10% to Idle, 19% to Social Gestures, and 20% to 3-Way). Further, the babies' responses to the avatar's Linguistic Nursery Rhymes (over the avatar's other conversational turn types) involved them to be largely riveted into a state of fixed and Sustained Visual Attention (77%). Of theoretical significance, there appears to have been a principled relationship between the avatar's socially contingent communicative turn types and the babies' specific responses. This relationship implies that the avatar was indeed having a linguistic impact on the baby.

7.2 Parent's Intervention

To address question 4 (impact of parental intervention), we analyzed the different baby behaviors across the two conditions. Interestingly, babies responded to 80% of avatar's behaviors in Condition

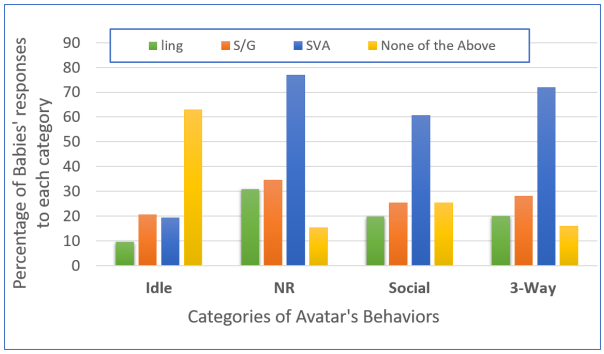

Figure 5: Babies' categorical responses to different avatar behaviors

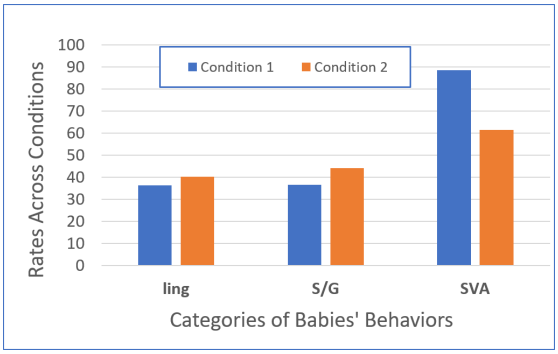

Figure 6: Frequency of baby response types in absence (condition 1) or presence (condition 2) of parental involvement

1 versus 60% in Condition 2 ($t = 2.22, p < 0.05$). This decrease is mainly due to a significant decrease in the babies sustained visual attention, SVA ($t = 4.3, p < 0.005$). This finding makes sense since in condition 2, parents were acknowledging and interacting with the baby, whereupon babies would naturally look more at the parent thereby exhibiting fewer instances of sustained attention toward the avatar.Figure 6 shows the distribution of baby responses across the two conditions. As shown, there is a significant increase in the percentage of babies' linguistic behaviors across Condition 1 vs Condition 2 ($t = 2.4, p < 0.05$). This is a very interesting finding, as it indicates that parents' interactions may have the potential to augment the language learning impact of RAVE. Apart from parental impact, the present pattern of change from Condition 1 to Condition 2 may imply that the infant is evidencing aspects of learning (to be further explored).

8 DISCUSSION

The driving theoretical question of the present paper was to understand whether an artificial agent (the ASL signing Avatar) had the potential to facilitate language learning in young babies. In particular, we asked whether the avatar's linguistic productions in signed language would spontaneously trigger linguistic responses from all babies irrespective of being exposed, or not, to a signed language (due to the shared linguistic structures universal to all world languages). To address this, we studied the impact that a signing avatar had on young babies' spontaneous behavioral responses. We were especially interested if a young baby would even detect the avatar's different communicative modes, as the avatar was projected onto a flat screen. To be sure, the results indicate that the babies were indeed able to detect the differences among the avatar's communicative modes even though they viewed all on a flat screen.

Herein lies one of the important findings of the present study concerning the nature of the brain-based mechanisms that govern human language acquisition. Most (if not all) of these babies did not understand the meanings of the avatar's productions (be they its general communications or linguistic signs) and all avatar productions involved repetitive movements. Nonetheless, the data provide support for the hypothesis that the babies were perceptually discriminating among the avatar's four distinct categories of productions. How? Here, the babies' differential responses to the avatar's categories of productions suggest that the babies' perceived differences among the avatar categories, crucially, based on factors outside of any understanding of the meanings or attraction to repetitive movement. That the babies specifically exhibited peaked behavioral responses to only one category of avatar productions, the Linguistic Nursery Rhymes, over all others is especially revealing. Of all four avatar production categories, the linguistic category was the only one in which we built in the specific rhythmic temporal patterning unique to phonetic-syllabic units in natural language phonology, and to which human infant brains have been discovered to possess maturationally time-locked peaked sensitivity [25–27, 29, 30, 38]. Thus, rather than being attracted to the meanings or general movements of the avatar productions before them, we hypothesized that all babies (deaf and hearing) were differentiating among the avatar's communicative modes based on differences in their +/- relation to the rhythmicity of language phonetic-syllabic (phonological) structure [2, 25, 26]. The present findings provide support for this, confirming that our avatar had hit squarely on those patterns. This finding is powerfully corroborated by studies showing that babies demonstrate riveted attention to the phonological patterns in their native language as well as in a foreign (non-native) language over other patterns of acoustic stimuli, even acoustic stimuli built to closely resemble the rhythmic timing properties of speech [12, 15, 29].

The present findings also suggest that the dialogue management had achieved a level of verisimilitude to social contingency found in natural parent-baby discourse [28]. Beyond the importance of social interactions, the role of social contingency in early language acquisition will be further pursued in our future work along with work analyzing the robot's role in the system. Nonetheless, all of the babies' appeared to be captivated by the avatar, and exhibited spontaneous engagement with the avatar. In conclusion, the present work provides a novel demonstration of the potential for avatars to facilitate language learning in young babies.

ACKNOWLEDGMENTS

This work was supported by the W.M. Keck Foundation (PI: Petitto), and National Science Foundation (IIS-1547178, PI: Petitto).

REFERENCES

[1] Ron Artstein, David Traum, Jill Boberg, Alesia Gainer, Jonathan Gratch, Emmanuel Johnson, Anton Leuski, and Mikio Nakano. 2016. Niki and Julie: a robot and virtual human for studying multimodal social interaction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 402–403.

[2] Stephanie A Baker, Roberta Michnick Golinkoff, and Laura-Ann Petitto. 2006. New insights into old puzzles from infants' categorical discrimination of soundless phonetic units. *Language Learning and Development* 2, 3 (2006), 147–162.

[3] Hennie Brugman, Albert Russel, and Xd Nijmegen. 2004. Annotating Multimedia/Multi-modal Resources with ELAN.. In *LREC*.

[4] Paul Debevec. 2012. The light stages and their applications to photoreal digital actors. *SIGGRAPH Asia* 2, 4 (2012).

[5] Amy Sue Finn. 2010. *The sensitive period for language acquisition: The role of age related differences in cognitive and neural function.* University of California, Berkeley.

[6] Ewa M Golonka, Anita R Bowles, Victor M Frank, Dorna L Richardson, and Suzanne Freynik. 2014. Technologies for foreign language learning: a review of technology types and their effectiveness. *Computer assisted language learning* 27, 1 (2014), 70–105.

[7] David J Greenberg, Donald Hillman, and Dean Grice. 1973. Infant and stranger variables related to stranger anxiety in the first year of life. *Developmental Psychology* 9, 2 (1973), 207.

[8] P Higgins. 1980. *Outsiders in a hearing world.* SAGE Publishing.

[9] William F House. 1976. Cochlear implants. *Annals of Otology, Rhinology & Laryngology* 85, 3 (1976), 3–3.

[10] Stephanos Ioannou, Vittorio Gallese, and Arcangelo Merla. 2014. Thermal infrared imaging in psychophysiology: potentialities and limits. *Psychophysiology* 51, 10 (2014), 951–963.

[11] Jiyou Jia. 2009. An AI framework to teach English as a foreign language: CSIEC. *AI Magazine* 30, 2 (2009), 59.

[12] Peter W Jusczyk, Derek M Houston, and Mary Newsome. 1999. The beginnings of word segmentation in English-learning infants. *Cognitive psychology* 39, 3-4 (1999), 159–207.

[13] Marina Krcmar. 2011. Word learning in very young children from infant-directed DVDs. *Journal of Communication* 61, 4 (2011), 780–794.

[14] Marina Krcmar, Bernard Grela, and Kirsten Lin. 2007. Can toddlers learn vocabulary from television? An experimental approach. *Media Psychology* 10, 1 (2007), 41–63.

[15] Patricia K Kuhl. 2004. Early language acquisition: cracking the speech code. *Nature reviews neuroscience* 5, 11 (2004), 831.

[16] Kwan Min Lee, Younbo Jung, Jaywoo Kim, and Sang Ryong Kim. 2006. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International journal of human-computer studies* 64, 10 (2006), 962–973.

[17] Maja J Mataric. 1997. Studying the role of embodiment in cognition. *Cybernetics & Systems* 28, 6 (1997), 457–470.

[18] Monica Meijsing. 2006. Real people and virtual bodies: How disembodied can embodiment be? *Minds and Machines* 16, 4 (2006), 443–461.

[19] Andrew N Meltzoff, Rechele Brooks, Aaron P Shon, and Rajesh PN Rao. 2010. "Social" robots are psychological agents for infants: A test of gaze following. *Neural networks* 23, 8-9 (2010), 966–972.

[20] Setareh Nasihati Gilani, David Traum, Arcangelo Merla, Eugenia Hee, Zoey Walker, Barbara Manini, Grady Gallagher, and Laura-Ann Petitto. 2018. Multimodal Dialogue Management for Multiparty Interaction with Infants. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 5–13.

[21] Setareh Nasihati Gilani, David Traum, Rachel Sortino, Grady Gallagher, Kailyn Aaron-Lozano, Cryss Padilla, Ari Shapiro, Jason Lamberton, and Laura-Ann Petitto. [n. d.]. Can a Virtual Human Facilitate Language Learning in a Young Baby?. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-19)*.

[22] Johanna Grant Nicholas and Ann E Geers. 2007. Will they catch up? The role of age at cochlear implantation in the spoken language development of children with severe to profound hearing loss. *Journal of Speech, Language, and Hearing Research* 50, 4 (2007), 1048–1062.

[23] Tim Payne. 2018. MAKI - A 3D Printable Humanoid Robot. https://www.kickstarter.com/projects/391398742/maki-a-3d-printable-humanoid-robot.

[24] Laura-Ann Petitto. in press. The Impact of Minimal Language Experience on Children During Sensitive Periods of Brain and Early Language Development: Myths Debunked and New Policy Implications. retrieved from http://petitto.net/wp-content/uploads/2014/04/Petitto_Minimal-Language-Experience_Final_Oct-6-2017.pdf.

[25] Laura-Ann Petitto, Melody S Berens, Ioulia Kovelman, Matt H Dubins, K Jasinska, and M Shalinsky. 2012. The "Perceptual Wedge Hypothesis" as the basis for bilingual babies' phonetic processing advantage: New insights from fNIRS brain imaging. *Brain and language* 121, 2 (2012), 130–143.

[26] Laura Ann Petitto, Siobhan Holowka, Lauren E Sergio, Bronna Levy, and David J Ostry. 2004. Baby hands that move to the rhythm of language: hearing babies acquiring sign languages babble silently on the hands. *Cognition* 93, 1 (2004), 43–73.

[27] Laura Ann Petitto, Siobhan Holowka, Lauren E Sergio, and David Ostry. 2001. Language rhythms in baby hand movements. *Nature* 413, 6851 (2001), 35.

[28] Laura Ann Petitto, Marina Katerelos, Bronna G Levy, Kristine Gauna, Karine Tétreault, and Vittoria Ferraro. 2001. Bilingual signed and spoken language acquisition from birth: Implications for the mechanisms underlying early bilingual language acquisition. *Journal of child language* 28, 2 (2001), 453–496.

[29] Laura-Ann Petitto, Clifton Langdon, Adam Stone, Diana Andriola, Geo Kartheiser, and Casey Cochran. 2016. Visual sign phonology: Insights into human reading and language from a natural soundless phonology. *Wiley Interdisciplinary Reviews: Cognitive Science* 7, 6 (2016), 366–381.

[30] Laura Ann Petitto and Paula F Marentette. 1991. Babbling in the manual mode: Evidence for the ontogeny of language. *Science* 251, 5000 (1991), 1493–1496.

[31] Laura Ann Petitto, Robert J Zatorre, Kristine Gauna, Erwin James Nikelski, Deanna Dostie, and Alan C Evans. 2000. Speech-like cerebral activity in profoundly deaf people processing signed languages: implications for the neural basis of human language. *Proceedings of the National Academy of Sciences* 97, 25 (2000), 13961–13966.

[32] Rebekah A Richert, Michael B Robb, and Erin I Smith. 2011. Media as social partners: The social nature of young children's learning from screen media. *Child Development* 82, 1 (2011), 82–95.

[33] Jenny R Saffran, Ann Senghas, and John C Trueswell. 2001. The acquisition of language by children. *Proceedings of the National Academy of Sciences* 98, 23 (2001), 12874–12875.

[34] Brian Scassellati, Jake Brawer, Katherine Tsui, Setareh Nasihati Gilani, Melissa Malzkuhn, Barbara Manini, Adam Stone, Geo Kartheiser, Arcangelo Merla, Ari Shapiro, David Traum, and Laura-Ann Petitto. 2018. Teaching Language to Deaf Infants with a Robot and a Virtual Human. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 553.

[35] Ari Shapiro. 2011. Building a character animation system. In *INTERNATIONAL Conference on Motion in Games*. Springer, 98–109.

[36] Arielle Spellun and Poorna Kushalnagar. 2018. Sign language for deaf infants: A key intervention for a developmental emergency. *Clinical pediatrics* 57, 14 (2018), 1613–1615.

[37] Glenn Stockwell. 2007. A review of technology choice for teaching language skills and areas in the CALL literature. *ReCALL* 19, 2 (2007), 105–120.

[38] Adam Stone, Laura-Ann Petitto, and Rain Bosworth. 2018. Visual sonority modulates infants' attraction to sign language. *Language Learning and Development* 14, 2 (2018), 130–148.

[39] M Teena and A Manickavasagan. 2014. Thermal infrared imaging. In *Imaging with Electromagnetic Spectrum*. Springer, 147–173.

[40] Tobii Eyetracker 2018. Tobii Pro X3-120. https://www.tobiipro.com/product-listing/tobii-pro-x3-120/.

[41] Blake S Wilson, Charles C Finley, Dewey T Lawson, Robert D Wolford, Donald K Eddington, and William M Rabinowitz. 1991. Better speech recognition with cochlear implants. *Nature* 352, 6332 (1991), 236–238.