Multimodal Dialogue Management for Multiparty Interaction with Infants

Setareh Nasihati Gilani
Institute for Creative Technologies
University of Southern California
sngilani@ict.usc.edu

David Traum
Institute for Creative Technologies
University of Southern California
traum@ict.usc.edu

Arcangelo Merla
Dept. of Neuroscience & Imaging Sciences
University G. d Annunzio
arcangelo.merla@unich.it

Eugenia Hee
Dept. of Computer Science
University of Southern California
hee@usc.edu

Zoey Walker
Brain & Language Lab for Neuroimaging, BL2
Gallaudet University
zoey.walker@gallaudet.edu

Barbara Manini
Brain & Language Lab for Neuroimaging, BL2
Gallaudet University
b.manini@uea.ac.uk

Grady Gallagher
Brain & Language Lab for Neuroimaging, BL2
Gallaudet University
grady.gallagher@gallaudet.edu

Laura-Ann Petitto
Brain & Language Lab for Neuroimaging, BL2
PhD in Educational Neuroscience (PEN) Program
Gallaudet University
laura-ann.petitto@gallaudet.edu

ABSTRACT

We present dialogue management routines for a system to engage in multiparty agent-infant interaction. The ultimate purpose of this research is to help infants learn a visual sign language by engaging them in naturalistic and socially contingent conversations during an early-life critical period for language development (ages 6 to 12 months) as initiated by an artificial agent. As a first step, we focus on creating and maintaining agent-infant engagement that elicits appropriate and socially contingent responses from the baby. Our system includes two agents, a physical robot and an animated virtual human. The system's multimodal perception includes an eye-tracker (measures attention) and a thermal infrared imaging camera (measures patterns of emotional arousal). A dialogue policy is presented that selects individual actions and planned multiparty sequences based on perceptual inputs about the baby's internal changing states of emotional engagement. The present version of the system was evaluated in interaction with 8 babies. All babies demonstrated spontaneous and sustained engagement with the agents for several minutes, with patterns of conversationally relevant and socially contingent behaviors. We further performed a detailed case-study analysis with annotation of all agent and baby behaviors. Results show that the baby's behaviors were generally relevant to agent conversations and contained direct evidence for socially contingent responses by the baby to specific linguistic samples produced by the avatar. This work demonstrates the potential for language learning from agents in very young babies and has especially broad implications regarding the use of artificial agents with babies who have minimal language exposure in early life.

ICMI '18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3243029

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; **Empirical studies in HCI**; • **Computing methodologies** → *Planning and scheduling*; *Multi-agent planning*; • **Applied computing** → *Education*;

KEYWORDS

Human-Computer Interaction, Multi-Agent Interaction, Multimodal Interaction Design, American Sign Language, Eye-tracking, Thermal Infrared (IR) Imaging, Augmentative learning aids.

ACM Reference Format:
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 *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3242969.3243029

1 INTRODUCTION

Most dialogue system technology is aimed at enabling natural language dialogues with competent language performers. Even language teaching systems generally strive to introduce a new language to competent users of another language. In this paper we describe the dialogue management approach for a very different

population: infant L1 language learners who have yet to achieve language competence. The system, called RAVE (Robot, AVatar, thermal Enhanced language learning tool), focuses on multiparty social interaction and learning elements of a visual sign language, American Sign Language (ASL) [13]. Two differently embodied agents (a robot and a virtual human avatar) engage in the interaction with the infant (and sometimes a parent). The design of the RAVE interface was described in [37], here we focus on dialogue management to support contingent interaction.

The main long-term goal of RAVE is to develop an augmentative learning aid that can provide linguistic input to facilitate language learning during one widely recognized critical developmental period for language (ages 6-12 months; e.g. [26]). Exposure to the patterns of language during this period facilitates infants' acquisition of the phonetic-syllabic segments unique to their native language, vocabulary, syntactic regularities, and ultimately, letter-to-segment mapping vital to early reading and academic success [29]. This is particularly important for infants who might not otherwise receive sufficient language exposure. Of particular concern are deaf babies, many of whom are born to parents who do not know a signed language. Rather than receiving minimal language exposure, this circumstance leaves these deaf babies with no access to an accessible language for well into the second year of life when intensive auditory language training typically begins. To support this long-term goal, our system was designed to engage all infants with early-life minimal language exposure [31] (hence, our testing of both hearing and deaf babies described in section 5).

As a first step, we ask whether the system can engage infants' attention, and whether the system can elicit contingent conversational behaviors from infants. If so, then we will have identified a novel dialogue system with the potential to facilitate language learning in young infants. In order to accomplish this, the dialogue management routines are designed to attract the baby's attention, infer when it is appropriate to provide linguistic stimuli, and to provide contingent reactions to baby initiatives to maintain the dialogue as a socially contingent interaction.

Multiple inputs to the system's perceptional component were used, including an eye-tracker (a behavioral measure of attention) and a thermal camera and thermal infrared (IR) imaging (a computational psychophysiological measure of change in ANA/autonomic nervous system response [9, 43]). The use of thermal IR imaging constitutes one unique design feature. It enabled us to track a baby's changes in emotional engagement when interacting with the agents, as indicated by their ANA responses (e.g., a baby's parasympathetic response indicated prosocial engagement as compared with a sympathetic response associated with disengagement or distress [22]). Based on our fNIRS brain imaging of infant neural responses to specific language patterns as concomitant with ANA responses, thermal IR algorithms were created as input to the system to permit the identification of when infants were "ready to learn" even before they had the capacity to produce language. This was crucial to our dialogue management system, as it provided the central triggering mechanism as to when the agent should start or cease a socially contingent conversational turn.

2 BACKGROUND, MOTIVATION AND GOALS

Acquiring language starts from birth through a concert of factors including the maturation of the neural systems that support language processing, observation, and engagement in social interactions [5]. By around ages 6-10 months, hearing babies learn the finite set of sound phonetic units and phonemic categories of their native languages [19]. Children who do not have sufficient language exposure during this critical period are at risk for delays in cognitive, linguistic, and social skills that can span life [29, 36]. Interestingly, deaf babies with exposure to visual sign language follow a similar pattern of phonological development in sign language, even though the units are silent and produced on the hands [27, 42], including a manual homologue to vocal babbling [28, 30].

However, Higgins [7] reports that 91.7% of the deaf individuals come from families where both parents are identified as hearing [38], where learning a new signed language quickly can be challenging. Interventions such as cochlear implants [8] exist that can allow some access to spoken language [48], but most of them cannot be deployed until the ages of 18-24 months, which is past an early critical period for learning basic phonological units. Thus, many deaf infants might be among this at-risk population due to insufficient language input in early life. Our work aims to provide visual language input to infants in the critical age period for phonological morphological development.

In particular, Petitto et al. [27] found that specific rhythmic temporal frequencies of language are important. To capture the attention of infants, we need to provide a linguistic stimulus that matches the rhythmic temporal patterning and facial expressiveness of a natural sign language, such as ASL. In ASL, crucial grammatical information is communicated via systematic (rule-governed) patterned changes in handshape and specific grammatical modulations of space and movement. The rhythmic temporal patterning binds sign phonetic-syllabic segments into signs, signs into sign phrases and clauses, and signed sentences [29], as well as grammatical facial expressions and body shifting [34]. The use of a virtual character on a screen gives the benefit of having an expressive agent that has the manual dexterity and obligatory facial expressiveness to produce sign language samples as linguistic input [11, 12, 32, 39, 47].

3 SYSTEM AND ARCHITECTURE

We adopted a complex multi-party design, involving multiple heterogeneous agents, linguistic stimuli tailored to the target population, and several types of sensory inputs. The system includes two agents (a physical robot and a virtual human) that can provide visual behaviors, as well as several sensor devices for perceptual input: an eye-tracker, thermal camera, and a Kinect.

Our hypothesis was that to capture the attention of infants, we should provide a linguistic stimulus that matched the rhythmic temporal patterning found in all natural languages, including natural signed languages, such as ASL. Petitto et al. [27] found that babies are sensitive to specific rhythmic temporal frequencies of language in early life. Specific rhythmic temporal patterning binds phonetic-syllabic segments into words, phrases, and clauses in spoken languages, with identical processes occurring in signed languages, whereupon specific rhythmic temporal patterning binds sign-phonetic units into signs, signs into sign phrases and clauses,

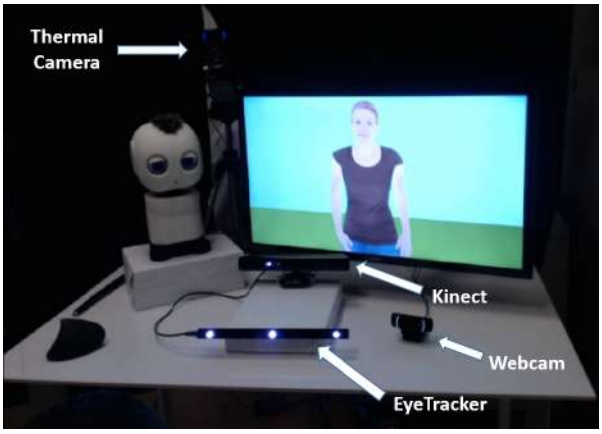


Figure 1: Physical deployment of system components from the front view

and signed sentences [29]. Grammatical information is also communicated in ASL via systematic (rule-governed) patterned changes in handshape, eye gaze, grammatical modulations of space and movement, and grammatical body shifting and crucially, facial expressions [11, 12, 32, 34, 39, 47]. Thus, our use of a virtual character on a screen was a key design feature that provides the benefit of having an expressive agent that produces the optimal rhythmic temporal patterning vital to acquire the phonological building blocks of language, manual dexterity, and obligatory facial expressiveness to produce signed language samples as linguistic input.

We use a robot since it is a physically-embodied agent, which provided a mechanism to engage the baby, a locus for facilitating attention to the virtual human, and a means to introduce a more natural social conversational setting whereupon agents and baby can occupy varying conversational roles. It has been found that a physical robot can evoke interest and social responses from young babies [3, 23], but even robots that have been designed specifically to act as signed language tutors [2, 14, 16, 46], often cannot support the full range of manual dexterity, fluidity, rhythmic temporal language patterning, and facial expressiveness required [15, 21, 41]. Hence our reasoning to use a physical robot to act as an initial target for infant attention. We also predicted that contingent interactions between the robot and the virtual human helps establish each agent as socially-interacting conversational partners rather than objects. Our design to use a combined robot+avatar was further supported by previous studies that found the exclusive use of video-recording and playing back visual language is unlikely to work [17, 18, 35]. Kuhl et al. [20] found that exposing American infants to human speakers of Mandarin Chinese reversed a decline in perception of Mandarin phonetic segments, but exposure only to audio or to audio-visual recorded stimuli did not.

Figure 1 shows the physical deployment of the hardware components from the front view. Multiple webcams were used to record the experiment from different angles. A snapshot of the experiment is shown in Figure 2. As seen in Figure 2, the infant was sitting on a parent's lap, facing the avatar's monitor and the robot. In this section we give a brief description about each component and then advance to the architecture of the system.

Figure 2: Multiparty interaction between Avatar, Robot, and infant from multiple viewpoints

3.1 Components

A detailed description of each of the components is beyond the scope of this paper. We focus on a brief description to highlight their role in the multiparty dialogue and provide more details on the dialogue manager in Section 4.

3.1.1 ***Avatar.*** The avatar was constructed by capturing a native ASL signer inside of a photogrammetry cage. Facial scans were also captured using a Light Stage [4] and the 3-D avatar was built using a real-time character animation system described in [40].

3.1.2 ***Robot.*** The robot is based on the open-source Maki platform from Hello Robot [25]. The robot has an articulated head (pan left/right, tilt up/down), articulated eyes (pan left/right, tilt up/down) and eyelids (open/close). Scassellati et al. [37] present more detail about the robot design.

3.1.3 ***Eye-tracker.*** Infant's gaze direction as a behavioral response was used as an input to the system. A Tobii Pro X3-120 [44] was used to capture the baby's eye-gaze at the rate of 120 Hz using a customized python script. 4 different area of interests (AOI) were defined: Robot, Avatar, In-Between and Outside. AOI coordinates were defined in relation to the infant's point of view, as shown in Figure 3. We took into account the AOIs as well as the fixation on the target as an indicator of baby's focus of interest. We performed a majority vote paradigm every half second (60 data points) to determine the area of interest. A calibration process was done in the beginning of the experiment to accommodate the program to the physical setup and the relative coordination of the baby's eyes, targets and the tracker.

3.1.4 ***Thermal Camera and Thermal Infrared Imaging.*** We used facial thermal patterns and dynamics for capturing the infant's internal state which is used to determine when the infant is engaged with the interaction [9]. Thermal infrared imaging, by harnessing the body's naturally emitted thermal irradiation, enables cutaneous temperature recordings to be measured noninvasively, ecologically, and contact free [24]. To calculate information about the infant's affective state, we used the nose tip's average temperature as it was extracted from each frame thus obtaining a temperature signal in real time. The classification of the infant's internal state was built on foundational studies linking the modulation of nose tip temperature and human psychophysiological states, with whereupon

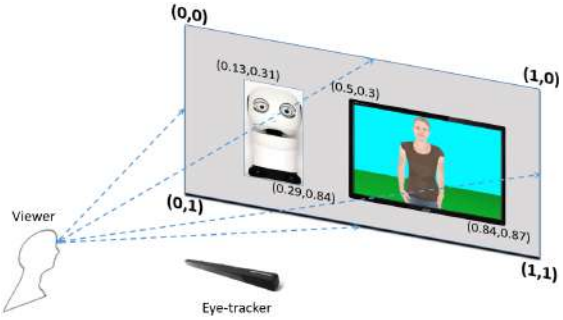


Figure 3: AOI regions from infant's perspective.

positive emotional responses such as interest and engagement are associated with nasal temperature increases (or parasympathetic response) while a decrease in temperature corresponds to sympathetic responses such as distress and disengagement [9, 22]. Thermal IR imaging was performed by means of a digital IR thermal camera FLIR A655sc [10] (640 x 480 microbolometer FPA, NETD: < 30 mK @ 30 °C, sampling rate: 50 Hz).

3.1.5 ***Baby Behavior***. It is crucial to have a visual perception component to capture the communicative and social behaviors of the infant such as hand clapping, pointing, reaching, etc. Current tracking systems such as Kinect [49], use models that are trained on adult anatomy and do not work properly on infants due to their fundamental differences in posture and relative proportions of body parts. Furthermore, the baby is sitting on a parent's lap, so this will bring additional complications for tracking systems which are mostly trained on full body postures. To address these issues, we collect Kinect data for future analysis, with the hope of eventually collecting enough data for future customizing models to conform to our specific needs with respect to the experiment setup. But as a first step, we adopted an interface, shown Figure 4, that is used by an expert observer to indicate relevant infant behaviors to the rest of the system in real time .

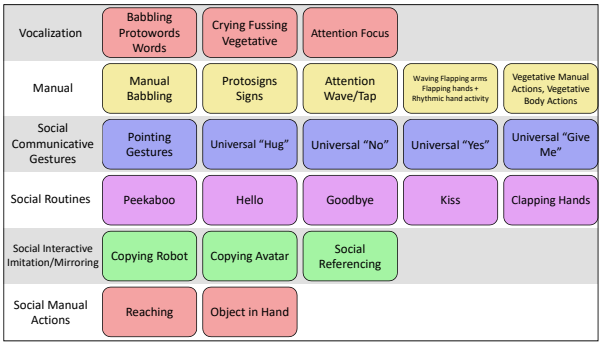


Figure 4: Observer interface for baby's behaviors

3.2 Software Architecture

Controllers for the hardware components were running on three separate machines, using multiple programming languages. We used a publisher-subscriber model [33] to facilitate communication between the components where ActiveMQ [1] was used as a message passing server. As shown in Figure 5, the perceptual components (eye-tracker, thermal imaging and behavior recognizer) send their messages to the server (publishers). These messages are subscribed to by the dialogue controller to update the information state and send messages to the Robot and Avatar (subscribers), directing them to perform communicative behaviors.

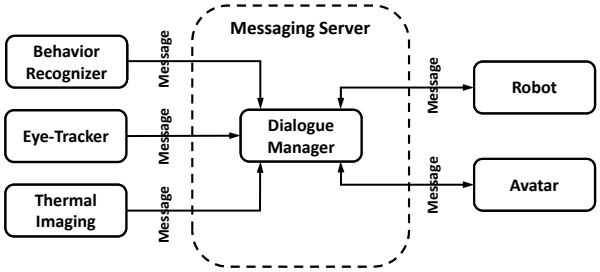


Figure 5: Logical overview of system components

4 DIALOGUE MANAGEMENT

The dialogue manger has three main goals:

(1) To engage the baby by participating in interactive dialogues.
(2) To maintain the engagement (sustained engagement).
(3) To promote engagement-rendered responses from the baby.

It uses input signals from the perception modules to update its information state [45] and choose new actions. Here, we give more details on multimodal perception signals, different output behaviors for agents, and finally the protocol for deciding on actions.

4.1 Input signals

These are input signals received from perceptual components and agents as well as the internal dialogue manager signals. Note that in our design, we have no direct perceptual monitoring of the parent, and no Avatar/Robot actions are contingent directly on the parent.

- **Area of Interest (AOI)** is the signal received from the eye-tracker component with discrete values for 4 different areas of the baby's eye gaze: Robot, Avatar, Between and Outside.
- **Readiness-To-Learn** is the signal received from the Thermal Imaging system with 5 discrete values: very negative (sustained decrease in attention, sympathetic), negative (non-sustained decrease in attention, sympathetic), very positive (sustained increase in attention, parasympathetic), positive (non-sustained increase in attention, parasympathetic), and a None signal which shows the signal's absence because of not detecting a reliable signal from the baby.
- **Baby-Behavior (BB)** is the signal received from the human observer interface about the baby's social and communicative behaviors. As seen in Figure 4 the input signals are classified into several categories such as vocalization, social communicative gestures, social routines and social manual actions. There are a total of 23 distinct states for this variable.
- **Component State Signals** come from the Avatar and Robot indicating their state of action: when a requested behavior

Agent	Category	Behaviors
Avatar	Conversational Fillers	Nod, Gaze forward/right/left, Head Shake, Contemplate, Think, Toss
	Social Behaviors	Wave, Hello, Peekaboo, Go Away / Come Back
	Question Solicitation	What?, What's Wrong?, What's That?, Ready? (To Robot/Baby)
	Linguistic Patterns	Good Morning, Look at Me! (To Robot/Baby), Boat, Pig, Fish, Cat
Robot	Fillers and Social Behaviors	Nod, Hide/Unhide, Gaze Forward/Right/Left, Startle, Blink, Sleep, Wake Up

Table 1: Robot and Avatar Primitive Behaviors

has started, ended or if there were any errors or exceptions during the execution of a specific behavior.

- **Timing Signals** are initiated from the dialogue manager itself. The DM tracks when events of different types have happened and sets up automatic signals that can change behaviors, e.g. if nothing interesting has happened recently.

4.2 Output commands

There are two different control levels of actions for the agents, described below: Primitive Behaviors and Action Sequences.

4.2.1 ***Primitive Behaviors.*** These are defined as atomic actions of the agents which are single behaviors that cannot be interrupted; such as nodding or a single nursery rhyme. Table 1 shows a list of primitive behaviors for the Avatar and Robot.

The virtual human's different language samples comprised different conversational/communicative social functions that are commonly used in Infant-Adult conversations. These functions include nursery rhymes, social routines, questions, conversational fillers, soothing responses, social affirmations and negation, solicitations and conversationally neutral idling. These were grouped as follows:

(1) **Conversational Fillers and Social behaviors** Conversational Fillers are short lexical items or phrases that assure the addressee that the conversational partner is attending and still "in" the conversation. They are like social punctuations e.g., YES! or THAT!, which are full lexical items in ASL. Social behaviors (or, social routines) are standard gestures that are widely used with infants, such as PEEKABOO.
(2) **Question Solicitation** such as ASL signs WHAT? or WHAT'S THAT? are used when the infant is in a sympathetic state.
(3) **Linguistic Patterns** provide the vital linguistic stimuli for the baby. All Nursery Rhymes were constructed with the identical rhythmic temporal patterning that matched the infant brain's specific neural sensitivity to that rhythmic temporal patterning [27, 29]. The identical overall rhythmic temporal patterning that all Nursery Rhymes were built with was this: maximally-contrasting rhythmic temporal patterning in 1.5 Hz alternations [27, 28]. Inside this temporal envelope were specific phonetic-syllabic contrasts, including 3 maximally-contrasting phonetic hand primes in ASL that human infants first begin to perceive and produce in language development: /5/, /B/, /G/ with contrastive transitions /B/=>/5/, /5/=>/F/, /G/=>/F/, plus allophonic variants. The Nursery Rhyme patterns were produced such that each had baby-appropriate lexical meanings. Below we provide an example of one of the four Nursery Rhymes, though each were designed with the same canonical structure.
BOAT (Phonetic-Syllabic units /B/, /5/)
(a) BOAT (/B/, palm in)
(b) BOAT-on-WATER (/B/+ movement modulation, palm up)
(c) WAVE (/5/+same movement modulation, palm down)

4.2.2 ***Action Sequences.*** We define an action sequence as a plan for a timed sequence of primitive actions by the agents that will be executed in order as planned. An example of an action sequence is the triad social greetings between the agents: (1) Avatar turns toward the Robot. (2) Robot turns toward the Avatar. (3) Avatar and Robot both nod to each other. (4) Avatar signs LOOK-AT-ME to both baby and Robot. (5) Avatar signs READY? to both baby and Robot. (6) Avatar turns back and looks at baby.

Another example is the familiarization sequence which is executed at the beginning of the experiment and will be described in detail in section 4.3.

4.3 Interaction Protocol

At each point in the 3 way robot-avatar-baby interaction, the system has a sequence of actions as the current plan for the agents to execute. These are designed with the assumption that the baby will behave accordingly, but if the baby acts differently, the planned actions may get removed or get updated by a completely different plan, in order to maintain a socially contingent interaction. The only signal that causes an interruption in the execution of currently planned actions is the input signal coming from the baby behavior interface. In this case, the policy overrides the current plan with a new plan according to the new state of the baby.

Each set of input combinations leads to a sequence of actions from the Avatar and Robot. Theoretically speaking, considering only the 3 input variables coming from the perceptual components, we are looking at an information state space of $4 * 5 * 23 = 460$ possible combinations. However, not every combination is possible or likely to happen; but to build a completely reliable system all combinations should be considered. We used a rule based policy which will trigger specific sequences of behaviors based on predefined combination of variables. Figure 6 shows a highly abstract decision tree used as part of the policy in which many branches are aggregated with each other. Each branch consists of more fine-grained

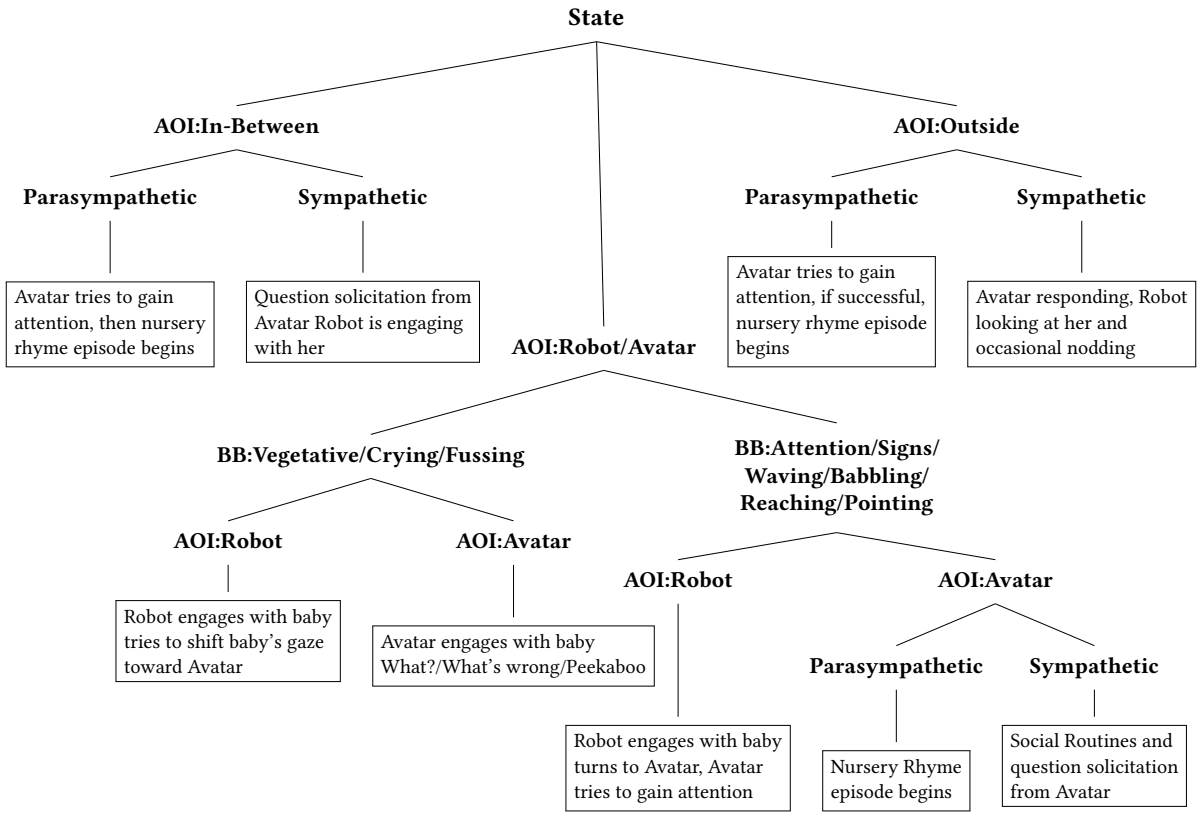


Figure 6: Summarized decision tree based on system variables

branches based on different input values for the baby-behavior, fixation of gaze, former executed plans and other state variables.

In order to make the baby familiar and comfortable with its surrounding, we begin the experiment with a familiarization episode. The goal of this period is to introduce the agents as conversational partners and make the baby feel involved in this multiparty interaction. This is a trace of what happens in the familiarization episode: At the start of the experiment, both Avatar and the Robot are in their idle and neutral form. The Robot's head is down with its eyes closed and Avatar is standing still looking forward. Robot wakes up from his sleeping position, sees the baby and nods as an indication that it has acknowledged the baby's presence and then turns to Avatar. Avatar looks at the robot and acknowledge it by nodding and then turns to the baby. Then the Avatar tries to gain the baby's attention by waving to it. The Avatar will sign GOOD MORNING toward both the Baby and the Robot to begin the interaction.

Table 2 shows a sequence of snapshots drawn from one interaction with a baby, along with different state variable values and an explanation of the state of the system at each point. This triad interaction between Avatar-Robot-Baby consists of the agents greeting each other (as participants in the conversation) and then the Avatar taking the floor and signing to the baby. Robot will nod to the Avatar occasionally to establish his role as a 3rd party conversationalist. We call this sequence a "Nursery Rhyme episode".

5 EVALUATION

The final dialogue manager routines described in the previous section were tested in interaction with 8 babies: 2 females and 6 males with average age of 9 months and 20 days (range 7-13 months) from whom 5 were hearing participants not sign-exposed, two hearing sign-exposed and one deaf sign-exposed. On first entering the room, babies were permitted to briefly touch the robot in a quick greeting interaction with the agents, as had been shown to be important in older children (~18 months old) interacting with robots [23]. Following this brief greeting exchange lasting less than 1 minute, the babies were seated on their parent's lap, facing the Avatar and Robot, as shown in Figure 2. Parents were instructed to wear sunglasses throughout the experiment to block the eye-tracker from registering their eyes. The session was divided into two conditions: (1) The parent did not participant in the agent-baby interaction (2-way); (2) the parent joined the agent-baby interaction (3-way). This provides additional opportunity for social referencing and conversational scaffolding involving Avatar-Baby-Parent triads.

We first asked the important scientific question as to whether RAVE can engage the infants' attention. Perhaps these young babies would not see the agents as interesting social interlocutors as predicted, but possibly boring objects, or worse, a source of agitation. When confronted with unknown humans and/or situations of novelty, babies at this age are prone to *stranger anxiety* [6]. Given that our 8 babies (age range 7-13 months) were within the onset period

of "stranger anxiety" (onset range 6-12 months), crying and fussing could have occurred. Thus, upon the babies' first contact with the Robot-Avatar system, the babies could have been interested in the agents, but they also could have fussed, become distracted, etc., at which point the interactional session would have been immediately ceased. None of these distracted behaviors were observed and instead, all 8 of the tested babies exhibited positive engagement behaviors, including: (1) Immediate visual engagement (locked attention) with the agents, (2) Sustained attention (persisting over time) and (3) Visually tracked (gaze following) the Avatar and Robot as they interacted with each other and the baby.

All 8 babies exhibited sustained engagement with RAVE lasting 4-5 minutes (average 3m40s; range 1m33s-4m56s). This is an atypical attention engagement window for very young infants. There was only one baby with the low engagement time (1m33s; > 2 standard deviations from the mean), but this baby entered the room very fussy. Although fussy on entering, she changed to riveted attention upon sight of the agents, and then slipped into a fussy state again at which point, we terminated the session. If we were to remove this outlier, the average sustained engagement time for the remaining babies is nearly 4 minutes (3m58s). Interestingly, this baby was an outlier for another reason. Her age (13mths;16days) fell outside of

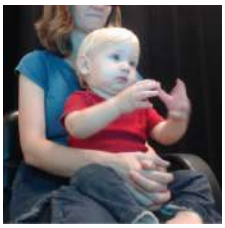
No.	AOI	Thermal	Baby	System	Description
1	Between	+			Baby is focusing, paying attention to the system. He is looking somewhere in between avatar/ robot. The goal of the system is to shift his gaze toward the Avatar
2	Avatar	++			Avatar tries to gain attention by signing LOOK-AT-ME. Thermal signal indicates that baby is in an engaged prosocial (parasympathetic) state (or "ready to learn"), thus the system transits to a nursery rhyme episode.
3	Robot	++			The agents nod after turning to each other. The goal is for the Avatar to acknowledge the robot as a 3rd party conversationalist in the interaction before she takes the floor and begin signing.
4	Avatar	++			Avatar begins a nursery Rhyme. Robot turns to her in the middle and nods towards her. Baby is copying the Avatar and is producing signs/proto-signs in response to the Avatar's linguistic input.
5	Neither	None			Baby turns to look at his mom to exhibit the classic social referencing behavior. Avatar signs ATTEND-TO-ME at the baby and tries to get back his attention.

Table 2: A sequence of snapshots drawn from a sample experiment showing different stages of the interaction. (Participant is a hearing male with no sign exposure and is 12 months and 1 day old)

our predicted window of peak infant engagement for RAVE (age range 6-12 months). Her performance thus provided preliminary support for our prediction that RAVE was most optimal for babies within the developmental period when they had peaked sensitivity to the rhythmic patterning of language, ages 6-12 months. The fact that the presence of the agents impacted all babies' preexisting emotional and/or attentional states for such durations is in itself remarkable, and invites us to understand why was it so.

We observed such sustained engagement in all babies, even hearing babies with no prior exposure to signed language, meaning that something about the avatar's productions was engaging to the babies even though they could not understand the meanings of the signs, with interesting group differences. For example, we found that our one baby with early bilingual language exposure (i.e., early ASL and early English exposure) had the greatest combined positive impact on its engagement span (longest experiment run time of 4m56s). This finding corroborates our earlier studies showing significant processing advantages afforded to babies and children with early bilingual language exposure [26].

The second condition, where parents were permitted to join in as they would naturally, allows for baby's social referencing to be acknowledged and responded to. In fact, we also observed instances where (nonsigning) parents copied the Avatar's signs and encouraged the baby to react and interact with the Avatar (only that will be picked up by the Avatar to continue its cycles).

Our second question was whether the artificial agents can elicit socially interactive and socially contingent conversation with an infant, above and beyond babies' production of prosocial emotional engagement/sustained visual attention. In an intensive analysis of 4 of the 8 babies (as analyses involve hundreds of hours of frame-by-frame video transcription with trained experts, behavioral coding, and reliability checks), all four babies produced social interactions and/or solicitations to the agents (e.g., waving HI, pointing, reaching, etc.) and attempted to copy the avatar, either through attempts to copy the avatar's signs (and components of signs) or matching the avatar's rhythmic movements at the nucleus of its sign productions. This novel finding is noteworthy because most babies (3 of 4) were never exposed to signed language yet attempted to copy the Avatar's linguistic signed language productions, and as noted above, they did so without understanding the meaning of the avatar's signs. Crucially, the babies' powerful engagement with the avatar occurred even though the avatar is an artificial agent on a flat TV monitor. We also performed a detailed case-study with one of the babies (a 7 month-old hearing baby boy who was exposed to signed language/ASL). In particular, we examined: (1) Whether the baby performed age-appropriate proto-linguistic behavior? (2) Whether this was produced in a socially contingent sequence as solicited by the Avatar's linguistic behavior?

In pursuit of these questions, we first coded the videos of conversational interactions with respect to Avatar and baby behaviors, followed by reliability checks. The rigorous coding was done by trained coders in the field of child language. Every video was coded for the categories of social conversational turns and content (see Figure 4) along with the time marking in coding and total time length of coded segments. Regarding question (1), we see linguistic behavior from the baby in both conditions (with and without the parent joining in). The baby waved and produced proto-signs related two distinct Nursery Rhymes. Regarding question (2), the sign-productions in all cases appeared as socially contingent reactions to the Avatar. Baby proto-signs were produced within a few seconds of the Avatar producing the relevant Nursery Rhymes. Baby social behaviors, such as waving, were produced as a response to social routines such as the signs for HELLO or GOODBYE. Thus, we see that the agents performing dialogue routines, in reaction to continuous multimodal sensory updates, were successful in soliciting socially contingent conversation from the infant. This would suggest the potential viability for using this kind of system for language learning in young infants.

6 CONCLUSION AND FUTURE WORK

While our system shows much potential, there is still much to be done to achieve the ultimate goal of facilitating an infant to learn visual sign language. We have established that the system can engage infants and stimulate socially contingent rudimentary conversation using the rhythmic-temporal patterned language stimuli and the dyadic and multiparty social interactions. One strand of future work involves making the system more autonomous and robust. We would like to replace the observer GUI with automated perception of conversationally relevant baby behaviors, by training behavior recognition models using collected Kinect and video data.

We also want to streamline the hardware footprint and start to look at whether the system can be used repeatedly, outside the laboratory, such that persistent learning over time can be achieved and assessed. Also, the system can adapt itself to behave differently to each baby to accommodate to specific behavioral and learning patterns of each individual. Another focus is to extend the dialogue routines to focus more specifically on the critical Agent-Infant-Parent triad. In particular, we desire to look at whether even parents who don't previously know sign language can assist in the child's learning (and learn themselves). This work has the potential for vast societal benefits, given the potential of baby's interaction with this type of system to "wedge open" their language learning capacity during the critical period of phonetic-syllabic development until the baby can receive systematic language input [26].

ACKNOWLEDGMENTS

The primary funding for this research was from the W.M. Keck Foundation (Petitto, PI: "Seeing the Rhythmic Temporal Beats of Human Language"), the National Science Foundation "INSPIRE" (Petitto, PI: IIS-1547178, "The RAVE Revolution for Children with Minimal Language Experience During Sensitive Periods of Brain and Language Development"), and the National Science Foundation Science of Learning Center Grant at Gallaudet University (Petitto, Co-PI and Science Director: SBE 1041725, "Visual Language and Visual Learning, VL2"). We extend sincere thanks to our colleagues on the project, including Ari Shapiro, Andrew Feng, Brian Scassellati, Jake Brawer, Katherine Tsui, Melissa Malzkuhn, Jason Lamberton, Adam Stone, Geo Kartheiser, and Gallaudet University student research assistants Rachel Sortino, Kailyn Aaron-Lozano, and Crystal Padilla. We especially thank the families who so generously gave of their time and support to participate in this study.

REFERENCES

[1] ActiveMQ 2018. Apache ACTIVEMQ. http://activemq.apache.org. Accessed: 2018-04-25.

[2] Oya Aran, Ismail Ari, Lale Akarun, Bülent Sankur, Alexandre Benoit, Alice Caplier, Pavel Campr, Ana Huerta Carrillo, and François-Xavier Fanard. 2009. SignTutor: An Interactive System for Sign Language Tutoring. *IEEE MultiMedia* 16, 1 (2009), 81–93.

[3] Akiko Arita, Kazuo Hiraki, Takayuki Kanda, and Hiroshi Ishiguro. 2005. Can we talk to robots? Ten-month-old infants expected interactive humanoid robots to be talked to by persons. *Cognition* 95, 3 (2005), B49–B57.

[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] 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.

[7] P Higgins. 1980. Outsiders in a hearing world. (1980).

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

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

[10] IR thermal camera 2018. FLIR A655sc. https://www.flir.com/products/a655sc/. Accessed: 2018-04-25.

[11] Kabil Jaballah and Mohamed Jemni. 2013. A Review on 3D signing avatars: Benefits, uses and challenges. *International Journal of Multimedia Data Engineering and Management (IJMDEM)* 4, 1 (2013), 21–45.

[12] Michael Kipp, Alexis Heloir, and Quan Nguyen. 2011. Sign language avatars: Animation and comprehensibility. In *International Workshop on Intelligent Virtual Agents*. Springer, 113–126.

[13] Edward S. Klima and Ursula Bellugi. 1979. The signs of language.

[14] Hatice Kose, Neziha Akalin, and Pinar Uluer. 2014. Socially interactive robotic platforms as sign language tutors. *International Journal of Humanoid Robotics* 11, 01 (2014), 1450003.

[15] Hatice Kose, Rabia Yorganci, Esra H Algan, and Dag S Syrdal. 2012. Evaluation of the robot assisted sign language tutoring using video-based studies. *International Journal of Social Robotics* 4, 3 (2012), 273–283.

[16] Hatice Kose, Rabia Yorganci, and Itauma I Itauma. 2011. Humanoid robot assisted interactive sign language tutoring game. In *Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on*. IEEE, 2247–2248.

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

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

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

[20] Patricia K. Kuhl, Feng-Ming Tsao, and Huei-Mei Liu. 2003. Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences* 100, 15 (2003), 9096–9101. arXiv:http://www.pnas.org/content/100/15/9096.full.pdf

[21] Daniel Leyzberg, Samuel Spaulding, Mariya Toneva, and Brian Scassellati. 2012. The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 34.

[22] Barbara Manini, Daniela Cardone, Sjoerd Ebisch, Daniela Bafunno, Tiziana Aureli, and Arcangelo Merla. 2013. Mom feels what her child feels: thermal signatures of vicarious autonomic response while watching children in a stressful situation. *Frontiers in human neuroscience* 7 (2013), 299.

[23] 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.

[24] Arcangelo Merla. 2014. Thermal expression of intersubjectivity offers new possibilities to human–machine and technologically mediated interactions. *Frontiers in psychology* 5 (2014), 802.

[25] Tim Payne. 2018. MAKI - A 3D Printable Humanoid Robot. https://www.kickstarter.com/projects/391398742/maki-a-3d-printable-humanoid-robot. Accessed: 2018-04-25.

[26] 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.

[27] 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.

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

[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 and BL Neuroimaging. [n. d.]. The Impact of Minimal Language Experience on Children During Sensitive Periods of Brain and Early Language Development: Myths Debunked and New Policy Implications. ([n. d.]).

[32] Farzad Pezeshkpour, Ian Marshall, Ralph Elliott, and J Andrew Bangham. 1999. Development of a legible deaf-signing virtual human. In *Multimedia Computing and Systems, 1999. IEEE International Conference on*, Vol. 1. IEEE, 333–338.

[33] Ragunathan Rajkumar, Michael Gagliardi, and Lui Sha. 1995. The real-time publisher/subscriber inter-process communication model for distributed real-time systems: design and implementation. In *Real-Time Technology and Applications Symposium, 1995. Proceedings.* IEEE, 66–75.

[34] Judy Snitzer Reilly, Marina McIntire, and Ursula Bellugi. 1990. The acquisition of conditionals in Language: Grammaticized facial expressions. *Applied Psycholinguistics* 11, 4 (1990), 369–392.

[35] 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.

[36] 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.

[37] Brian Scassellati, Jake Brawer, Katherine Tsui, Setareh Nasihati Gilani, Melissa Malzkuhn, Barbara Manini, Adam Stone, Geo Kartheiser, Arcangelo Merla, Ari Shapiro, et al. 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.

[38] Jerome D Schein and Marcus T Delk Jr. 1974. The deaf population of the United States. (1974).

[39] Jerry Schnepp, Rosalee Wolfe, John McDonald, and Jorge Toro. 2013. Generating co-occurring facial nonmanual signals in synthesized American sign language. (2013).

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

[41] Michelle Starr. 2014. Toshiba's new robot can speak in sign language. https://www.cnet.com/news/toshibas-new-robot-can-speak-in-sign-language/. Accessed: 2018-04-25.

[42] 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.

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

[44] Tobii Eyetracker 2018. Tobii Pro X3-120. https://www.tobiipro.com/product-listing/tobii-pro-x3-120/. Accessed: 2018-04-25.

[45] David Traum and Staffan Larsson. 2003. The Information State Approach to Dialogue Management. In *Current and New Directions in Discourse and Dialogue*, Jan van Kuppevelt and Ronnie Smith (Eds.). Kluwer, 325–353.

[46] Pınar Uluer, Neziha Akalın, and Hatice Köse. 2015. A new robotic platform for sign language tutoring. *International Journal of Social Robotics* 7, 5 (2015), 571–585.

[47] Lynette van Zijl and Jaco Fourie. 2007. The development of a generic signing avatar. In *Proceedings of the IASTED International Conference on Graphics and Visualization in Engineering, GVE*, Vol. 7. 95–100.

[48] 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.

[49] Zhengyou Zhang. 2012. Microsoft kinect sensor and its effect. *IEEE multimedia* 19, 2 (2012), 4–10.