Chapter 1

MITILI REPORT 2020

1 Introduction

Learning language at a young age is key for children's early literacy development, which in turn is crucial for later academic success (M. M. Páez & López 2007; Hart & Risley 1995; Dickinson & McCabe 2001). A major problem faced by many children, particularly from low socioeconomic status (SES) families, is limited exposure rich language during adult-child conversations at home. For instance, studies have reported that low-SES parent/child conversations tend to be less frequent and of shorter duration when compared with those of their higher-SES counterparts (Hart & Risley 1995; ROWE 2008). Low-SES children also tend to receive more directives from their parents, delivered in shorter utterances with less diverse lexical items, and are asked fewer open-ended questions (Hart & Risley 1995; ROWE 2008; Hoff 2003). In contrast, high-SES parents are more likely to negotiate with their children and teach them about the world (e.g., objects and events) through conversation (Lawrence & Shipley 1996). High-quality, social, responsive, and facilitative interactions between parents and their children is especially crucial for young children's language development (**Tamis-LeMonda96**; Roseberry et al. 2013; Tamis-LeMonda et al. 2001). Deficits in these early experiences can have a detrimental effect on young children's development of early language and literacy skills. Studies report that low-SES preschool-age children have significantly smaller vocabularies than their high-SES counterparts. Substantial disparities also exist in their vocalizations (Gilkerson et al. 2017). These differences often get magnified over time once children enter school (Hart & Risley 1995).

This "participation gap" is now widely recognized between families with different SES backgrounds, particularly as it pertains to the active participation of parents (or adults) in children's literacy learning process (Neuman & Celano 2012a).

A 10-year-long observational study by Neuman & Celano (2012a) found out that children from a low-SES and a high-SES neighborhoods used digital educational resources (computers and literacy software) in very different ways, even though the two groups had adequate access in terms of the amount of time and number of resources, and both invested a similar level of effort in using them for learning (Neuman & Celano 2012b,a). Specifically, parents from higher-SES neighborhoods used computer games as a tool to actively facilitate lessons on literacy learning with their children. In contrast, parents with from lower-SES neighborhoods tended to let their children use the computers on their own, without parental participation or facilitation, even when their children were struggling.

There is urgency in figuring out how to enrich social and conversational interactions between parents and children from lower-SES households. Recognizing this, a variety of early childhood interventions have been designed and implemented with the ultimate goal of promoting children's cognitive skills, language development and school readiness via fruitful parent-child interactions (Deutscher et al. 2006; Lugo-Gil & Tamis-LeMonda 2008; Rodriguez & Tamis-LeMonda 2011). Although these programs can be successful and effective, they are also costly and time-consuming so they do not scale well. Many do not take place in a home setting where most conversations happen naturally.

We argue that AI-augmented learning technologies have great potential in fostering and enriching parent-child interactions. Today, there are many e-books and educational apps designed for children, and some are designed to support parental participation (McNab & Fielding-Barnsley 2013; Takeuchi & Stevens 2011). However, the area of parent-child learning technologies that improve the quality of parent-child conversation is still largely under-explored. Very few technologies have been designed to support rich adult-child interactions, say to proactively facilitate dialogic storytelling between the two stakeholders in the here and now (Chang & Breazeal 2011). In addition to promoting children's learning, we believe that guided parental involvement in children's learning process can also boost parents' motivation and self-efficacy in their children's education (Hoover-Dempsey & Sandler 1995).

Given the great promise of facilitative technologies for parent-child dialogic storytelling (where parent and child not only read but actively converse about the story, asking and answering questions, commenting on the narrative, etc.), our long-term research goal is to develop an early childhood language intervention aimed at promoting and guiding parent-child interaction through a social agent. Our ultimate goal is to implicitly coach and empower low-SES parents to reduce this participation gap – and close the learning disparity between low-SES

and high-SES families. We envision that the social agent will be designed to participate in triadic activities – actively engaging with the parent and child in a joint story reading activity – to promote the interactivity between all of them during story time.

However, before designing a social agent facilitator, we need to have a comprehensive understanding of 1) how the parents and children engage in story reading activities in a natural setting, and 2) how the parent-child relationship impacts their story reading styles (i.e., both verbal and nonverbal communication). As a first step toward designing a robot facilitator, this study aims to uncover the interaction dynamics between a parents and their children in a co-reading activity. In this report, we describe our study design and the novel multi-modal dataset we collected of parent-child story reading interactions. We present the measures we used to assess diverse aspects of the parent-child relationship, as well as other factors related to children's language development (e.g., home literacy environment). Lastly, we use automated video analysis methods to extract body pose of both parents and children during these story reading interactions. We present our analysis that identifies interesting correlations between participants' body pose and the social/emotional relationship between parent and child.

2 Study Design

2.1 Participants

Thirty-four families with children between the ages of 3-7 years old were recruited for our study in the greater Boston area. In each family, one parent and one child participated in the study activities. Three families withdrew from the study without completing the full procedure for reasons not related to the study. Thus, a total of 30 families completed the full study and we included their data in our quantitative analysis (Table. 1; Table. 2).

Table 1: Gender identity of the participant families in our dataset for quantitative analysis.

	parent's gender	child's gender
Female	22	10
Male	8	20

Table 2: English proficiency of the participant families in our dataset for quantitative analysis.

	Native	Bilingual	English Language Learner
Parent	14	10	6

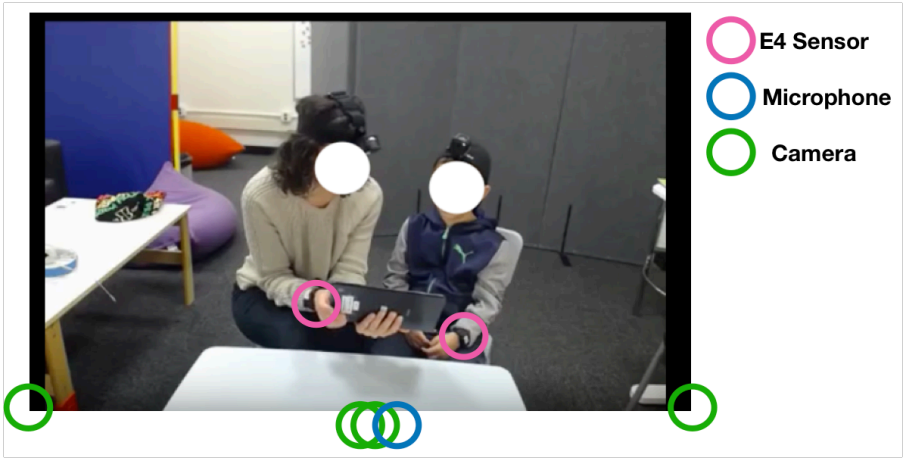


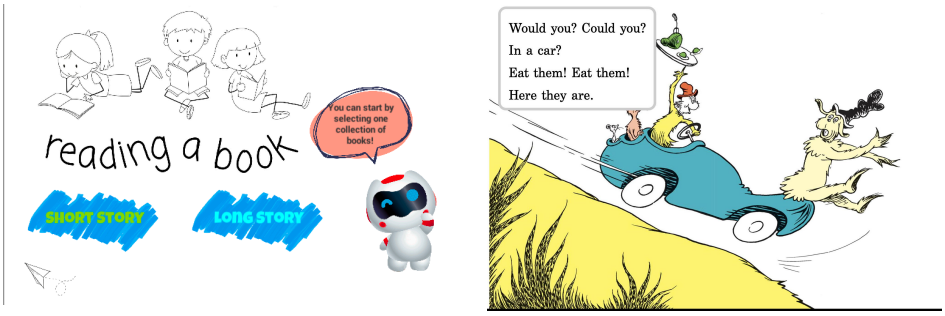
Figure 1: Parent-child interaction scenario in the lab setting.

2.2 Protocol and Procedure

The study protocol consisted of three parts in the following sequence: 1) a 45-minute in-lab session at MIT where parent-child pairs read stories together for 20 minutes, then the parent filled out surveys for another 20 minutes, 2) a session in the participants' home where they engaged in two 20-min story reading activities on an assigned Android tablet, and 3) one 45-minute in-lab session for 45 minutes that was similar to the first in-lab session. During the in-lab sessions, parent-child pairs sat next to each other as shown in Fig.1. Families that completed all three sessions were given $75 as their compensation.

2.3 Materials

A digitized version of our storybook corpus on a touchscreen tablet was used for both the in-lab and home sessions (Fig. 2). The storybook corpus consists of around 30 storybooks recommended by early childhood education experts and teachers. Each story lasts from 3 minutes to 15 minutes. Stories shorter than 5



(a) App landing page.

(b) An example of a story page on the screen.

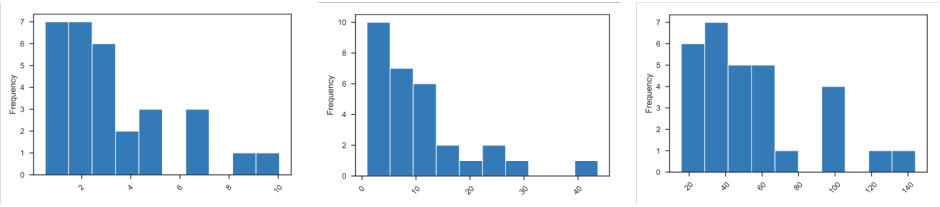
Figure 2: A digitized version of the storybook corpus on a touchscreen tablet. The books are divided into two categories based on story length.

minutes were categorized as ”short stories” and the rest as ”long story.” During the story reading activity, the parent and child could select any books they wanted to read from the corpus.

2.4 Measurements

During the in the in-lab sessions, audio and video was captured during parent-child co-reading using the microphone and cameras installed in the story reading station (Fig. 1). Four cameras were used to capture different angles of the dyadic interaction (i.e., frontal view, birds-eye view, parent-centered view, and child-centered view). The audio recordings were sent to a professional transcription service (Rev.com) to obtain textual annotation of recorded speech. For the in-lab sessions, parents and children were asked to wear unobtrusive wearable sensors (E4 sensors from Empatica.com) on their wrists. The E4 sensor measures physiological arousal (reflected in electrodermal activity), body temperature, and heart rate variability. These bio-signals are good predictors of a variety of affective states. In the home deployment, only the audio was recorded from the tablet's built-in microphone.

We also collected demographic information. Each parent filled out surveys on a touchscreen tablet reporting their social-economic status, home literacy environment, parenting styles (Kamphaus & Reynolds 2006), a parental theory of mind assessment (Warnell & Redcay 2014), and a child's temperament and behavior questionnaire (Putnam & Rothbart 2006).



(a) Number of questions asked by the child

(b) Number of questions asked by the parent

(c) Number of conversational turns between parent and child

Figure 3: The conversational interaction dynamics between a parent and a child was assessed in terms of the number of questions asked by a child, the number of questions asked by a parent, and the number of conversational turns between the two. The time window was 10 minutes of story reading randomly selected from the in-lab story reading sessions for each family.

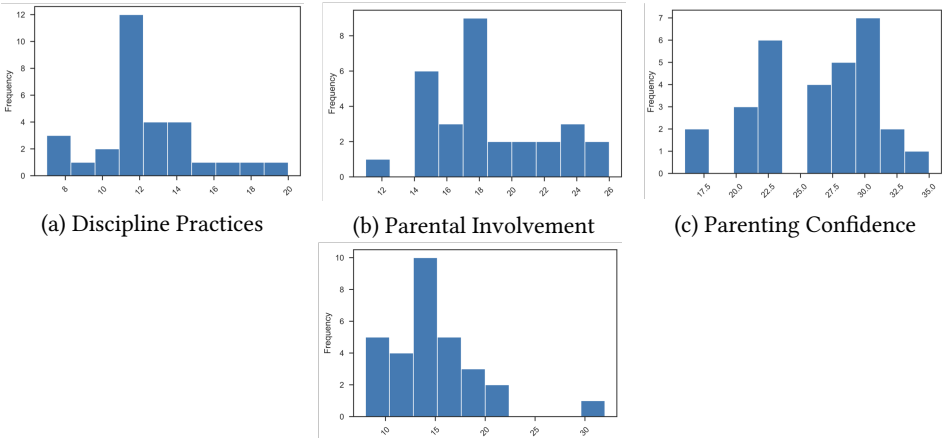
3 Data Analysis and Results

3.1 Question Asking Behavior

To characterize the conversational interaction dynamics between each parent-child pair during co-reading, we used three measures: 1) the number of questions the child asked, 2) the number of questions the parent asked, and 3) the number of conversational turns between the two. We define the number of conversational turns as the number of listener-speaker turns that a parent and child exchange when reading stories together. To annotate the video, we randomly selected a 10-minute interaction from the in-lab story reading sessions for each family, and calculated the three aforementioned measures within that time window. The distribution of each measure's results across 30 families is displayed in Fig. 3. All three distributions are skewed toward the left, indicating that the story reading for a very small subset of families were much more interactive than the rest of the families.

3.2 Parenting Styles

Adult participant's parenting style was assessed using self-report survey questions from the Parenting Relationship Questionnaire (PRQ) (Kamphaus & Reynolds 2006). The PRQ questionnaire is comprised of multiple dimensions of the parent-child relationship such as parental discipline practices, parental involvement, parenting confidence, and relational frustration. The distribution of each PRQ dimension across the 30 families is shown in Fig. 5. These results show that the



(a) Discipline Practices (b) Parental Involvement (c) Parenting Confidence

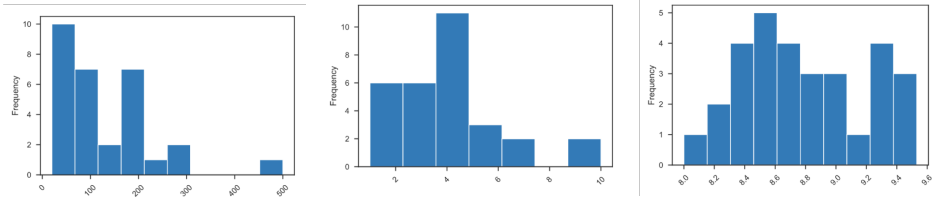
(d) Relational Frustration

Figure 4: Distribution of each parenting style scale from Parenting Relationship Questionnaire across 30 families.

four parenting style dimensions share distinct distribution patterns with parenting confidence having the most widely-spread distribution. This suggests that the families in our study have very diverse parenting styles, and such diversity gives greater motivation for understanding how parenting styles impact story co-reading interaction between parents and children.

3.3 Child's Home Literacy Environment and Temperament

The home literacy environment was measured in terms of the number of children's books in the home (Fig. 5a), and the amount of time (in hours) at home that someone reads to the child each week (Fig. 5b). A child's temperament was measured using Child's Behavior Questionnaire (CBQ) completed by the child's parent (Putnam & Rothbart 2006). The CBQ is comprised of measures for surgency, negative affect, and effortful control. To obtain an overall summary of the child's temperament, we summed up the individual score of each of the three sub-scales with a high score indicating negative temperament. Our CBQ results indicate that participant distribution was not heavily skewed toward either extreme, and the child participants had a wide range of temperaments.



(a) Number of children's books at home.

(b) Amount of time (in hours) at home that someone reads to the child each week.

(c) Child's temperament measured by Child Behavior Questionnaire (CBQ).

Figure 5: Distributions of child's home literacy environment and temperament across 30 families

3.4 Body Pose Extraction

3.4.1 Background

Nonverbal behavior is expressed through body movement as surveyed in Kleinsmith & Bianchi-Berthouze (2013); Karg et al. (2013); Zacharatos et al. (2014); Stephens-Fripp et al. (2017); Noroozi et al. (2018). Expressive movement is categorised into four types: communicative (e.g. gestures), functional (e.g. walking), artistic (e.g. choreography), and abstract (e.g. arm lifting), where a single or a combination of these types represent an affect (Karg et al. 2013). For example, anxiety is linked to expanded limbs and torso, fear is linked to elbows bent, and shame is linked to bowed trunk and head (Kleinsmith & Bianchi-Berthouze 2013). In computational movement detection, actions are detected through body models, image models and spatial statistics, where grammar, template and temporal statistics are extracted (Weinland et al. 2011). In Zacharatos et al. (2014), a focus on low-level and high-level features, as well as features from coding systems for body movement emotion recognition was given. Efforts have been made to create a consensus reliable coding system, where the two main body coding systems are Body Action and Posture (BAP) and Labal Movement Analysis (LMA).

BAP is a micro-description of body movement proposed by Dael et al. (2012), where body movements are described on an anatomical level, a form level, and a functional level. Body behaviours and actions are described through body parts and joints movement, location and orientation. Automatic coding and annotation of BAP were proposed in Velloso et al. (2013), where a full body motion tracking suit is used. However, such sensor suits are not suitable for remote body behaviour analysis. To the best of our knowledge, there is no automatic remote recognition of BAP coding.

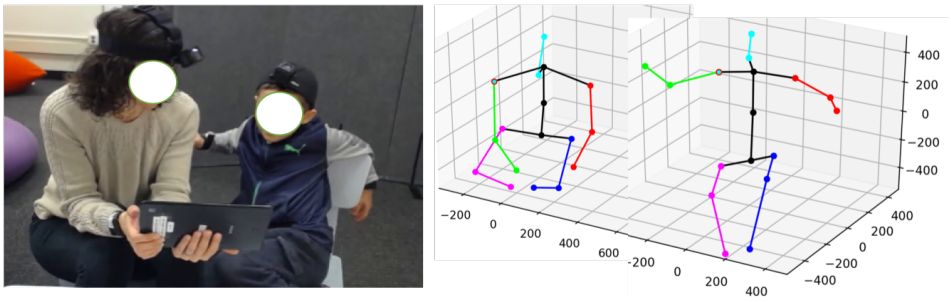


Figure 6: Body pose extraction

Dancing motion was coded by Laban and used for recognising performers' emotion from their body motion (Aristidou et al. 2015). A motion capture system of 8-cameras and a special suit were used to collect performed dance movements. Laban divides human motion into four components: body, effort, shape and space. In Aristidou et al. (2015) study, some of the Laban features are automatically calculable and were selected for body emotion classification. The results reached up to 98% in mutli-class categorical emotions such as anger (i.e. desecrate emotions).

Continues emotion recognition in theatre performance using LMA was proposed in Senecal et al. (2016), where a Microsoft Kinect device was used for motion capture. The initial results were promising, and showed a potential for advancing emotion recognition using LMA.

Beside the body coding systems described above, static and dynamic features and their combination were used for affect recognition from body behaviour (Noroozi et al. 2018). These features could also be represented as geometrical and appearance features.

Using dynamic features, body posture extracted features proved to be valuable in recognising acted emotions as unimodal and when combined with other modalities (Kessous et al. 2010). Statistical measures were calculated from the dynamic computer vision features such as quantity of motion (QoM), contraction ondex (CI), fluidity, velocity, and acceleration, where the classification results were these body-based features produced better recognition results than speech and facial expression classification results. A 3D joint Euler rotation was recorded in Kleinsmith et al. (2011) for affective posture recognition, where the automatic recognition achieved a comparable results to the human observers.

3.4.2 Our Approach to Body Pose Estimation

Inspired by BAP and LMA body movement features, we aimed at extracting body movement features from both the parent and the child that would facilitate analyzing their individual body gestures as well as their interaction with each other. Both BAP and LMA represent the body gestures in three dimensional state. Therefore, with the advancement of computer vision and deep learning techniques, we utilized 3D body joint localization described in Tome et al. (2017). The 3D location of the body joints are estimated from the detected 2D joints from images. The body pose estimation is applied to each frame in the video individually to extract the 3D location of 17 joints (see Fig. 6).

The raw data of the located 17 body joints are used to extract features that are then use for body-motion behaviour analysis. However, before the features could be extracted, normalisation techniques should be considered to account for both within and between participants variation. That is, the distance between the participant's body and the camera while moving is variable (during approaching and moving away from the camera), as well as the body size differences (height, shoulder size, etc.) between the patent and the child. Normalisation assures reliable measures of the extracted features with comparability for analysis.

In this work, we use the distance between the sternum point location and the collarbone (clavicle) points to normalise the distance between the other points (their distance is divided by the distance between two given points). We selected these two points for the normalisation since they are rigid, which make them robust from continuous, sudden and skewed movements. Normalisation is performed in each frame, then outlier detection using Grubbs' test for outliers (Grubbs 1969) was used to remove the frames with skewed measures (e.g. erroneous joint location).

Once the low-level data are processed, we can do high-level features extraction. We extract individual body pose and movement from the parent and the child, as well as interaction and synchrony features from both bodies in relation to each other. Assuming that the camera's focal length is in the middle of the frame and there is no lens distortion, we estimate individual body orientation (pitch, roll and yaw) by solving the Direct Linear Transform (DLT) followed by Levenberg-Marquardt optimization. We also calculate the bodies' rotation in relation to each other by calculating the Euler angles. Several features based on distances are also extracted such as the distance between the child and the parent bodies. Moreover, we calculate the distance between the child hand to their face, other hand, body, parent's face, parent's body, and parent's hands, and vise versa. This process results in 32 features for each frame (i.e. low-level). In addition, from these 32 low-

level features we calculate the derivatives (velocity and acceleration) for each consecutive frame.

Using overlapping windows of 5 seconds and 50% overlap, we extract functional features (i.e. high-level) form these low-level body gestures for each session. The functional features are: minimum (min), maximum (max), range, average, standard deviations (std), variance (var), skewness, kurtosis, peaks and valleys. For each window slice, this process produces a total of 960 functional features that are extracted from the head pose (32 low-level features ×3 (the feature and its two derivatives) ×10 statistical measures).

3.5 Head Pose Extraction Approach

Simple behaviours such as head movement could reflect cues about mood, emotions, personality, or cognitive processing (Heylen 2006). More specifically, Leclère et al. (2016) studied head movement during child-parent interaction sessions, where they focused on mutual engagement (percentage of time spent face to face or oriented to the task). Monadic Phases coding system is a measure for child-parent synchrony, where head movement and orientation are indicators of the level of interaction synchrony (Moore & Calkins 2004).

Automatic head pose estimation, once the face is detected, using computer vision techniques has been surveyed in Murphy-Chutorian & Trivedi (2008). Such methods include template matching, where the head image is matched to the nearest group of images that approximate the head pose, which requires a huge variety of head pose images. Deformable models are another method for head pose estimation, including the AAM, where specific facial points are labelled and trained to create a 2D model, then the head pose is estimated using the direction of the first principal component of the principal components analysis (Murphy-Chutorian & Trivedi 2008). Another method of head pose estimation is determining the geometry between local features, such as the eyes, mouth, and nose tip. Most of these methods could estimate one or more of the three dimensions of the head pose (pitch, yaw and roll). Recently, Bulat & Tzimiropoulos (2017) proposed an estimation of the 3D location of each 2D face landmarks using deep learning techniques. In this work, we utilized the 3D estimation of facial landmarks proposed in Bulat & Tzimiropoulos (2017) to extract head pose features from both the parent and the child (see Fig. 7).

Similar to the process of body pose feature extraction, head pose and movement are extracted. We extract functional features from individual heads orientation, orientation towards each others, and the head distances. The process produce a total of 300 functional features are extracted from the head pose (10 low-

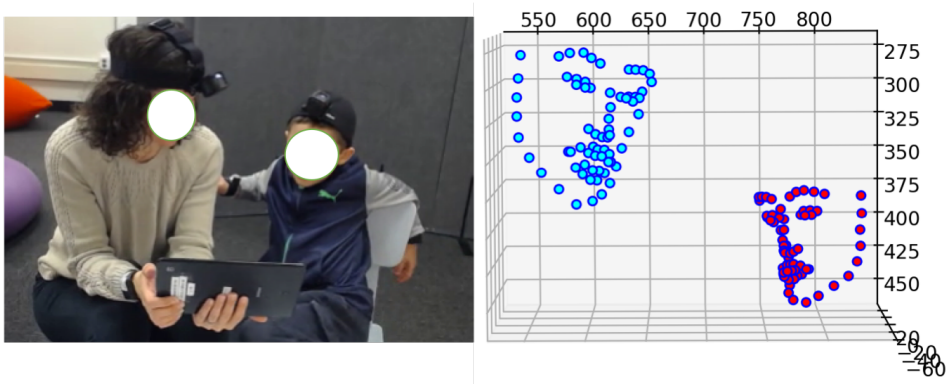


Figure 7: Head pose extraction

level features ×3 derivatives ×10 statistical measures) for each window slice (5 seconds with 50% overlap).

3.6 Correlation Analysis

For this exploratory phase, we selected a slice of 10 minutes from the middle of each sessions to extract the nonverbal behavior for the analysis. In this work, the total number of extracted features from the nonverbal behavior is 1260, which is very high to perform correlation analysis with corrected values for multiple testing. To overcome this substantial array of features, we were inspired by pattern recognition field, where feature selection techniques are used when the number of extracted features is higher than the sample size to reduce the feature list. We chose two techniques of feature selection, which are F-test and Mutual Information, which assess the relevancy and redundancy of features in relation to the class. These techniques are suitable for regression analysis, where the label (dependent variable) values are continues. However, as with most feature selection techniques, these techniques are effected by the sample values. That is, using different subset of the same data selects different features. Therefore, we performed these two techniques in cross-validation manner using 10-fold, where a random split of 80% of the samples is used in each turn for feature selection to ensure a robust final set of selected features. We set the number of features to be selected in each round as 20% of the total features (i.e. 252 features). The features that been consistently selected in all 10 rounds are then selected. This is performed for all dependent variables (see the first column of Table 3). The average number of features passing this condition for each dependent variables is 21 features.

Table 3: Significant Correlation Analysis of Nonverbal behaviour after Feature Selection Techniques

Factor	Nonverbal Behaviour	Movement	Statistical Function	r	p-val
Parentes Education	Parent body	Roll	Average acceleration	-37.2	0.047
CBQ	Parent body	Pitch	Max	-50.1	0.006
		Pitch	Range	-46	0.012
PRQ	Parent body	Pitch	Max acceleration	38.2	0.041
Discipline Practices	Child body	Roll	Min	43.9	0.017
Involvement	Both body	Pitch	Skewness	-39.9	0.032
	Child body	Yaw	Skewness of velocity	-45.2	0.014
Parenting Confidence	Parent body	Pitch	Kurtosis	38.5	0.039
		Pitch	Skewness	-40.7	0.028
Relational Frustration	Parent body	Yaw	Average acceleration	-45.7	0.013
Child Questions	Parent body	Yaw	Max of velocity	-36.8	0.050
		Pitch	Min	45.3	0.014
		Pitch	Peaks	-37.8	0.043
		Pitch	valleys	-38.8	0.037
		Pitch	Peaks of velocity	-37.9	0.043
		Pitch	Peaks of acceleration	-36.8	0.049
		Pitch	valleys of acceleration	-37.1	0.047
		Yaw	valleys	-37	0.048
	Child body	Pitch	valleys of velocity	-36.7	0.050
		Pitch	valleys of acceleration	-36.8	0.050
		Roll	Peaks of velocity	-36.8	0.050
Parent Questions	Parent body	Pitch	Max acceleration	-44.5	0.016
		Pitch	Range acceleration	-42.5	0.022
	Child body	Pitch	Max acceleration	-38.9	0.037
Turns	Parent body	Roll	Skewness	-39.1	0.036
		Pitch	Min	37.5	0.045
		Roll	Variance	45.6	0.013
		Roll	Kurtosis	-51.6	0.004
		Yaw	Std	45.3	0.014
		Yaw	Variance	45.6	0.013
		Yaw	Kurtosis	-51.6	0.004
		Yaw	Std acceleration	39.2	0.035

The process of feature selection is critical to improve the classification or prediction of certain variable using the nonverbal behavior, which is a goal for future work, which will be discussed in Section 4 below. Nonetheless, we use the feature selection to highlight the most distinctive features, where we then execute the correlation analysis. We performed an exploratory data analysis through Pearson or Spearman pairwise correlation, depending on the normality of the feature, which was measures using Shapiro–Wilk test. Several nonverbal features were found to have statistically significant correlation with the dependent variables, which are listed in Table 3. For parenting styles (PRQ, Discipline Practices, Involvement, Parenting Confidence, and Relational Frustration), a correlation was found with several nonverbal behavior from the child and the parent, but mostly

the parent. Interestingly, CBQ was not found to be significantly correlated to any of the child behavior. This might be due to the strict process of feature selection, where child nonverbal behavior were redundant to the parent behavior. Nonetheless, parent moving their body in the y-axis (i.e. forward/backward) has significant negative correlation (i.e. moving forward) to CBQ, which is expected for the parent to attend to their child. Lastly, for the in lab sessions, the number of questions asked by the child and the parent, as well as the number of turns in the session are highly correlated with nonverbal behavior from both the child and their parent. This is expected as high interaction and engagement are associated with variety of nonverbal behaviors. Moreover, head and touching features were not selected by the feature selection techniques, which could indicate that these features are redundant to the general body movement. Even though these are positive and promising findings, they are exploratory and more in-depth analysis is needed to confirm them.

4 Next Steps and Future Work

With our collected dataset, we plan to do further analysis to understand the interaction dynamics between parent and child, and then design a robot that can perceive and facilitate parent-child interaction synchrony in future.

4.1 Annotating parent-child interaction quality

The quality of interaction between a parent and their child is crucial for the child's development and is considered predictive of children's behaviour (Lotzin et al. 2015). Assessing parent-child interaction is highlighted in the literature in order to develop early interventions for enhancing general child development and for improving specific outcomes such as social competence, cooperation, language and cognition (Peterson et al. 2007). Studies have investigated parent-child interaction in different contexts (e.g. free play) and with respect to age (e.g. infant to adolescent), as well as for different child physical and mental health conditions (e.g. disabilities). The quality of parent-child interaction, based on the contextual situation, could be evaluated as joint engagement (Adamson et al. 2018; Adamson et al. 2019), interpersonal synchrony (Leclère et al. 2014), emotional availability (Greenspan et al. 2001), and bidirectional mutuality (Funamoto & Rinaldi 2015), among others. For each one of these categories, measures have been developed and validated, where the measurements includes questionnaires, rating scales, or observational coding schemes.

We are very interested in understanding how the parent-child relationship (e.g., parenting styles) and other social-economic factors (e.g., home literacy environments) impact their co-reading interaction quality. Towards this goal, we have recruited four trained psychology students to annotate the collected videos of the families' co-reading interaction. Regarding the interaction quality coding, we chose the Joint Engagement Rating Inventory (JERI) (Adamson et al. 2018), as it quantitates and qualitates the interaction between the child and the caregiver during a joint activity, where verbal and nonverbal behaviors related to engagement are observed and rated (e.g. gazing toward each other or to a shared object). JERI contains more than 10 scales capturing diverse aspects of parent-child interaction, and we chose to code fore *Child Unengaged* and *Child Coordinated Engagement. Child Coordinated Joint Engagement* involves the child's engagement with the parent instead of their engagement with the story/tablet, whereas *Child Unengaged* captures the child's overall engagement with both the parent and activity. When the child is not actively attending to both parent and the reading activity (book/tablet), the child is considered as *unengaged*. The child's coordinated joint engagement will be rated as low if the child is engaging in the story listening or reading without attending to the parent and acknowledging his/her presence.

In addition to coding the direct parent-child interaction quality, we also annotated the affect of both parent and child separately in terms of valence and arousal. Then, combining their individual affects, we obtained four joint affects: (1) arousal/valence synchrony, (2) arousal synchrony and valence asynchrony, (3) arousal asynchrony and valence synchrony, and (4) arousal/valence asynchrony.

When annotating these joint engagement and affect scales, our trained coders viewed all the in-lab story reading videos and gave ratings every five seconds. According to JERI, child's engaged/unengaged state can last as short as three seconds. Thus, a short time window (i.e., five seconds) was selected for the annotation to produce continuous quality scales, which will enable us to observe the change patterns of the parent-child interaction quality throughout an entire story reading session. The annotation is done, and we are currently analyzing the patterns of their joint affect and engagement over time.

4.2 Designing social robot's intervention on parent-child co-reading

We observed that it was common for a parent and their child to lose joint engagement and affective synchrony for a certain period of time when engaging in the co-reading activity. Some parent-child pairs can recover from this disengagement and asynchrony very fast by adapting how they communicate with

each other, while others seem to struggle to keep the dyadic experience uplifting throughout the interaction. For those families who are struggling, having a social robot as a facilitator in the interaction seems promising given its growing prevalence in people's daily life and technical capability to communicate with humans in a social/emotional manner.

Therefore, the key research question we are planning to investigate next is how a robot can produce social behaviors in real time to intervene at these "down moments" (e.g., disengagement, miscommunication, conflict) in parent-child co-reading, and to foster positivity (e.g., trust, proactivity, reciprocity) between the parent and child, which can last and transfer to other parent-child joint activities even after the robot-facilitated reading interaction.

In addition, our exploratory results show that the parent-child relationship and real-time reading interaction both vary largely across the 30 families. This finding gives us great motivation to design robot's intervention strategies that can be personalized to each family in real time. When compared with voice agents and computers that communicate with humans primarily through speech, social robotics supports nonverbal social communication with humans through a diverse set of multimodal social signals (e.g., motion, light, eye gaze, facial expression, body gesture). Social robots have greater capacity to personalize how they can engage in the parent-child interaction as a friendly peer that can help build rapport and reciprocity between the parent and child. Therefore, we plan to examine and compare how a robot's nonverbal and verbal behaviors impact both parent's and child's co-reading experience .

References

Adamson, L.B., R. Bakeman & K. Suma. 2018. The joint engagement rating inventory (technical report 25, 2nd ed.).

Adamson, Lauren B, Roger Bakeman, Katharine Suma & Diana L Robins. 2019. An expanded view of joint attention: skill, engagement, and language in typical development and autism. *Child development* 90(1). e1–e18.

Aristidou, Andreas, Panayiotis Charalambous & Yiorgos Chrysanthou. 2015. Emotion analysis and classification: understanding the performers' emotions using the lma entities. In *Computer graphics forum*, vol. 34, 262–276.

Bulat, Adrian & Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International conference on computer vision.*

Chang, Angela & Cynthia Breazeal. 2011. Tinkrbook: shared reading interfaces for storytelling. In *Proceedings of the 10th international conference on interaction design and children* (IDC '11), 145–148. Ann Arbor, Michigan: Association for Computing Machinery. https://doi.org/10.1145/1999030.1999047. DOI:10.1145/1999030.1999047

Dael, Nele, Marcello Mortillaro & Klaus R Scherer. 2012. The body action and posture coding system (bap): development and reliability. *Journal of Nonverbal Behavior* 36(2). 97–121.

Deutscher, Barbara, Rebecca R. Fewell & Michelle Gross. 2006. Enhancing the interactions of teenage mothers and their at-risk children: effectiveness of a maternal-focused intervention. *Topics in Early Childhood Special Education* 26(4). 194–205. DOI:10.1177/02711214060260040101

Dickinson, D. K. & A. McCabe. 2001. Bringing it all together: the multiple origins, skills, and environmental supports of early literacy. *Learn. Disabil. Res. Pract.* 16. 186–202.

Funamoto, Allyson & Christina M Rinaldi. 2015. Measuring parent–child mutuality: a review of current observational coding systems. *Infant Mental Health Journal* 36(1). 3–11.

Gilkerson, Jill, Jeffrey Richards, Steven Warren, Judith Montgomery, Charles Greenwood, D. Kimbrough Oller, John Hansen & Terrance Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology* 26. 1–18. DOI:10.1044/2016_AJSLP-15-0169

Greenspan, Stanley I, Georgia DeGangi & Serena Wieder. 2001. *The functional emotional assessment scale (feas): for infancy & early childhood.* Interdisciplinary Council on Development & Learning Disorders.

Grubbs, Frank E. 1969. Procedures for detecting outlying observations in samples. *Technometrics* 11(1). 1–21.

Hart, B. & T. R. Risley. 1995. *Meaningful differences in the everyday experience of young american children.* ERIC.

Heylen, Dirk. 2006. Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics* 3(03). 241–267.

Hoff, Erika. 2003. The specificity of environmental influence: socioeconomic status affects early vocabulary development via maternal speech. *Child development* 74 5. 1368–78.

Hoover-Dempsey, K. V. & H. M. Sandler. 1995. Parental involvement in children's education: why does it make a difference? 97(2). 310–331.

Kamphaus, R. W. & C. R. Reynolds. 2006. Parenting relationship questionnaire.

Karg, Michelle, Ali-Akbar Samadani, Rob Gorbet, Kolja Kühnlenz, Jesse Hoey & Dana Kulić. 2013. Body movements for affective expression: a survey of automatic recognition and generation. *IEEE Transactions on Affective Computing* 4(4). 341–359.

Kessous, Loic, Ginevra Castellano & George Caridakis. 2010. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces* 3(1-2). 33–48.

Kleinsmith, Andrea & Nadia Bianchi-Berthouze. 2013. Affective body expression perception and recognition: a survey. *IEEE Transactions on Affective Computing* 4(1). 15–33.

Kleinsmith, Andrea, Nadia Bianchi-Berthouze & Anthony Steed. 2011. Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41(4). 1027–1038.

Lawrence, Valerie W. & Elizabeth F. Shipley. 1996. Parental speech to middle- and working-class children from two racial groups in three settings. *Applied Psycholinguistics* 17(2). 233–255. DOI:10.1017/S0142716400007657

Leclère, C, M Avril, S Viaux-Savelon, N Bodeau, C Achard, S Missonnier, M Keren, R Feldman, M Chetouani & David Cohen. 2016. Interaction and behaviour imaging: a novel method to measure mother–infant interaction using video 3d reconstruction. *Translational psychiatry* 6(5). e816.

Leclère, Chloë, Sylvie Viaux, Marie Avril, Catherine Achard, Mohamed Chetouani, Sylvain Missonnier & David Cohen. 2014. Why synchrony matters during mother-child interactions: a systematic review. *PloS one* 9(12). e113571.

Lotzin, Annett, Xiaoxing Lu, Levente Kriston, Julia Schiborr, Teresa Musal, Georg Romer & Brigitte Ramsauer. 2015. Observational tools for measuring parent–infant interaction: a systematic review. *Clinical child and family psychology review* 18(2). 99–132.

Lugo-Gil, Julieta & Catherine Tamis-LeMonda. 2008. Family resources and parenting quality: links to childrens cognitive development across the first 3 years. *Child development* 79. 1065–85. DOI:10.1111/j.1467-8624.2008.01176.x

M. M. Páez, P. O. Tabors & L. M. López. 2007. Dual language and literacy development of spanish-speaking preschool children. *J. Appl. Dev. Psychol* 28. 85–102.

McNab, K. & R. Fielding-Barnsley. 2013. Digital texts, ipads, and families: an examination of families' shared reading behaviours. 20. 53–62.

Moore, Ginger A & Susan D Calkins. 2004. Infants' vagal regulation in the still-face paradigm is related to dyadic coordination of mother-infant interaction. *Developmental Psychology* 40(6). 1068.

Murphy-Chutorian, Erik & Mohan Manubhai Trivedi. 2008. Head pose estimation in computer vision: a survey. *IEEE transactions on pattern analysis and machine intelligence* 31(4). 607–626.

Neuman, S. B. & D. C. Celano. 2012a. *Giving our children a fighting chance: poverty, literacy, and the development of information capital*. New York, New York, USA: Teachers College Press.

Neuman, S. B. & D. C. Celano. 2012b. Worlds apart: one city, two libraries, and ten years of watching inequality grow. *Am. Educ.* (Fall). 13–23.

Noroozi, Fatemeh, Dorota Kaminska, Ciprian Corneanu, Tomasz Sapinski, Sergio Escalera & Gholamreza Anbarjafari. 2018. Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*.

Peterson, Carla A, Gayle J Luze, Elaine M Eshbaugh, Hyun-Joo Jeon & Kelly Ross Kantz. 2007. Enhancing parent-child interactions through home visiting: promising practice or unfulfilled promise? *Journal of Early Intervention* 29(2). 119–140.

Putnam, Samuel P. & Mary K. Rothbart. 2006. Development of short and very short forms of the children's behavior questionnaire. *Journal of personality assessment* 87 1. 102–12.

Rodriguez, Eileen & Catherine Tamis-LeMonda. 2011. Trajectories of the home learning environment across the first 5 years: associations with childrenâ⊠⊠s vocabulary and literacy skills at prekindergarten. *Child development* 82. 1058–75. DOI:10.1111/j.1467-8624.2011.01614.x

Roseberry, Sarah, Kathy Hirsh-Pasek & Roberta Golinkoff. 2013. Skype me! socially contingent interactions help toddlers learn language. *Child development* 85. DOI:10.1111/cdev.12166

ROWE, MEREDITH L. 2008. Child-directed speech: relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language* 35(1). 185–205. DOI:10.1017/S0305000907008343

Senecal, Simon, Louis Cuel, Andreas Aristidou & Nadia Magnenat-Thalmann. 2016. Continuous body emotion recognition system during theater performances. *Computer Animation and Virtual Worlds* 27(3-4). 311–320.

Stephens-Fripp, Benjamin, Fazel Naghdy, David Stirling & Golshah Naghdy. 2017. Automatic affect perception based on body gait and posture: a survey. *International Journal of Social Robotics* 9(5). 617–641.

Takeuchi, B. L. & R. Stevens. 2011. The new coviewing: designing for learning through joint media engagement.

Tamis-LeMonda, Catherine S., Marc H. Bornstein & Lisa Baumwell. 2001. Maternal responsiveness and children's achievement of language milestones. *Child Development* 72(3). 748–767. http://www.jstor.org/stable/1132453.

Tome, Denis, Chris Russell & Lourdes Agapito. 2017. Lifting from the deep: convolutional 3d pose estimation from a single image. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 2500–2509.

Velloso, Eduardo, Andreas Bulling & Hans Gellersen. 2013. Autobap: automatic coding of body action and posture units from wearable sensors. In *2013 humaine association conference on affective computing and intelligent interaction*, 135–140.

Warnell, Katherine & Elizabeth Redcay. 2014. Spontaneous mentalizing captures variability in the cortical thickness of social brain regions. *Social cognitive and affective neuroscience* 10. DOI:10.1093/scan/nsu081

Weinland, Daniel, Remi Ronfard & Edmond Boyer. 2011. A survey of vision-based methods for action representation, segmentation and recognition. *Computer vision and image understanding* 115(2). 224–241.

Zacharatos, Haris, Christos Gatzoulis & Yiorgos L Chrysanthou. 2014. Automatic emotion recognition based on body movement analysis: a survey. *IEEE computer graphics and applications* 34(6). 35–45.