Incorporating Measures of Intermodal Coordination in Automated Analysis of Infant-Mother Interaction

Lauren Klein*
University of Southern California
kleinl@usc.edu

Victor Ardulov
University of Southern California
ardulov@usc.edu

Yuhua Hu
University of Southern California
yuhuahu@usc.edu

Mohammad Soleymani
Institute for Creative Technologies
University of Southern California
soleymani@ict.usc.edu

Alma Gharib
University of Southern California
agharib@usc.edu

Barbara Thompson
Michigan State University
thom1756@msu.edu

Pat Levitt
Children's Hospital Los Angeles
plevitt@chla.usc.edu

Maja J. Matarić
University of Southern California
mataric@usc.edu

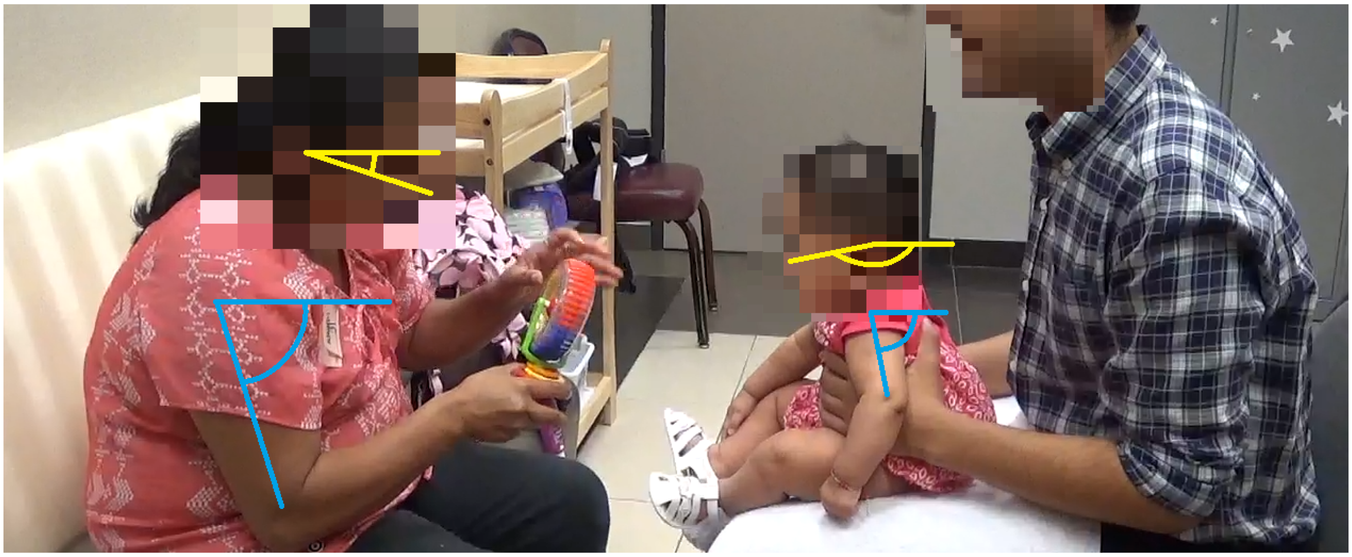

Figure 1: Infant-Mother Interaction Setup. Yellow and blue angles demonstrate the markers used to calculate head and arm pitch angles, respectively.

ABSTRACT

Interactions between infants and their mothers can provide meaningful insight into the dyad's health and well-being. Previous work has shown that infant-mother coordination, within a single modality, varies significantly with age and interaction quality. However, as infants are still developing their motor, language, and social skills, they may differ from their mothers in the modes they use to communicate. This work examines how infant-mother coordination across modalities can expand researchers' abilities to observe meaningful trends in infant-mother interactions. Using automated feature extraction tools, we analyzed the head position, arm position, and vocal fundamental frequency of mothers and their infants during the Face-to-Face Still-Face (FFSF) procedure. A de-identified dataset including these features was made available online as a contribution of this work. Analysis of infant behavior over the course of the FFSF indicated that the amount and modality of infant behavior change evolves with age. Evaluating the interaction dynamics, we found that infant and mother behavioral signals are coordinated both within and across modalities, and that levels of both intramodal and intermodal coordination vary significantly with age and across stages of the FFSF. These results support the

ICMI '20, October 25–29, 2020, Virtual event, Netherlands
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418870

significance of intermodal coordination when assessing changes in infant-mother interaction across conditions.

CCS CONCEPTS

• **Applied computing → Health care information systems**; **Health informatics**.

KEYWORDS

infant-mother interaction; intermodal coordination; behavioral signal processing

ACM Reference Format:
Lauren Klein, Victor Ardulov, Yuhua Hu, Mohammad Soleymani, Alma Gharib, Barbara Thompson, Pat Levitt, and Maja J. Matarić. 2020. Incorporating Measures of Intermodal Coordination in Automated Analysis of Infant-Mother Interaction. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3382507.3418870

1 INTRODUCTION

Infants rely on supportive communication with their caregivers for healthy cognitive and emotional development [15]. Researchers and therapists observe infant-caregiver interactions to assess child development and identify appropriate interventions [17]. Typically developing infants attempt to engage their parents through motor or vocal babbling, and the parents reciprocate with their own attention, gestures, or vocalizations. This is a key process known as "serve-and-return" interaction, and is characteristic of healthy infant-caregiver relationships [15].

In the first year of life, serve-and-return interaction is often characterized by asymmetric modalities, meaning that the infant and mother differ in their communication methods. For example, a younger infant may respond to their mother's vocalizations by moving their body, but as their own vocalization skills develop, they may mirror their mother's behavior [11]. Building on this observation, this work outlines a method for evaluating changes in intermodal coordination across conditions.

Prior work has established that the coordination of behaviors is a fundamental building block for the higher-level dyadic processes present in serve-and-return interactions. In their recent review of infant-mother dyadic processes, Provenzi et al. [16] described coordination as a form of low-level contingent engagement that enables dynamics such as synchrony and attunement. This holds true for adults as well; the coordination and coregulation of low-level behaviors such as speech rate can support improved collaboration outcomes [19]. In infant-mother interactions, fluctuations between coordinated and uncoordinated interaction states are key to enabling synchronous communication [11]. Higher levels of infant-mother synchrony are associated with more positive outcomes for the child [11]. Therefore, this work introduces metrics for identifying and evaluating intramodal and intermodal infant-mother coordination as a foundation toward more complex phenomena such as synchrony.

Recent advances in automated feature extraction and signal processing have enabled computational analyses of coordination between infant and mother behavioral signals [7, 10–13, 21]. These automated approaches have been proposed to evaluate features that influence infant-mother communication across multiple time scales, and to support quantitative measurements of dyadic processes for clinical observation. Pose, body movement, facial expressions, and vocal prosodic features have each been used to inform models of dyadic processes and characterize infant-mother interactions. Since infant and mother turns may incorporate vocalizations, gestures, or both, the ability to include multiple types of behavioral signals is essential to analyzing interaction dynamics [11, 15]. Additionally, as infants grow and expand their motor, language, and social skills, their communication strategies may change [1, 15], rendering analyses that depend on a single behavioral signal insufficient. Yet, existing research in this area has focused largely on computational models evaluated on symmetric interaction modalities.

The Face-to-Face Still-Face (FFSF) procedure [22] is one of the most widely used experimental procedures for observing infant-caregiver interaction in a research setting [11, 16]. Past work has used the FFSF paradigm to evaluate the effects of infant age, developmental disorders, and maternal depression on early communication [1]. These factors vary widely in their impact on interaction dynamics; therefore, it is necessary to have a detailed characterization of an infant-mother interaction in order to assess risk for multiple health and developmental outcomes.

Leveraging video and audio data from infant-mother dyads participating in the FFSF procedure across multiple ages, this work outlines the advantages of intermodal coordination for tracking changes in interaction dynamics. Data were collected from 57 infant-mother dyads who visited the research site between one and five times, at 2, 6, 9, 12, and 18 months after the infant's birth. Building on the methods introduced by Boker et al. [3] and utilized by Hammal et al. [7], we evaluated coordination across three modes, namely, head pose, arm angle, and vocal prosody, across age and stages of the FFSF procedure. As a contribution of this research, de-identified features were made publicly available at https://github.com/LaurenKlein/mother-infant-interactions so that others may extend this research. The major findings of this work are as follows:

1. Infant behavior changes during the FFSF procedure measured with head pose, arm pose, and vocalization signals each had unique trends across age.
2. Significant levels of coordination between infant and mother behavioral signals were found not only between the same behavioral signals, but across behavioral signals.
3. Metrics evaluated across asymmetric behavioral signals identified trends in infant-mother coordination beyond those identified by metrics evaluated using symmetric behavioral signals.

These results are consistent with prior observations that infants rely on communication across modes [11], supporting the value of intermodal infant-mother coordination metrics, upon which models of higher-level communication methods can be built.

2 BACKGROUND

Over the past four decades, research interest in infant-caregiver dyadic processes has grown, with the goal of understanding both causal and predictive relationships between these processes and developmental outcomes for the infant [16]. Coordination and synchrony have emerged as main concepts of interest, due to their effects on healthy social and emotional development [11]. Reviews focusing on infant-mother dyadic processes [5, 11, 16] have proposed working definitions of synchrony which include two key factors: 1) *coordination* of behaviors of each partner across time, and 2) *intermodality*, coordination of behaviors using different or multiple modes of communication.

Prior work has introduced a variety of computational approaches to address the temporal dynamics of infant-mother behavioral coordination. Messinger et al. [13] quantified the transition dynamics between infant and mother smiling states, showing that as age increases, infants smile more reliably in response to their mothers' smiles. Hammal et al. [7] tracked infant and mother head pose and movement during the FFSF procedure using windowed cross-correlation with peak-picking as a model of dyadic coordination, finding significantly more head movement correlation peaks in the Play stage than in the Reunion stage. Using global movement as the primary measurement, Leclère et al. [10] tracked pauses and overlaps between the infants' and mothers' actions during turn-taking in a play activity, finding these features to be significant predictors of risk for neglect. Mahdhaoui et al. [12] describe a method for using infant-directed speech to better understand how mothers interact with infants who develop Autism Spectrum Disorder (ASD).

Researchers have leveraged automated analysis of communication dynamics to build mobile or robotic systems capable of providing feedback or intervention during interaction with young children. TalkLime, developed by Kim and Yim [18], alerts parent to potentially problematic behaviors, such as interrupting their child or speaking too fast. Caregivers who received this feedback saw an increase in child-initiated utterances during conversations. While these approaches focus solely on speech signals, which may be infeasible for young infants, they demonstrate the potential for automated coordination analysis in understanding and improving communication dynamics. Gilani et al. [14] developed an interactive system to demonstrate sign language to deaf infants, using eye gaze and thermal infrared imaging to infer when the infant was attentive and ready to learn and communicate. Results showed promising levels of engagement from infants, though the activity involved infant-system interaction rather than infant-caregiver interaction.

Past research demonstrates the potential for developing quantitative measures of coordination, using computational models evaluated on autonomously extracted features, which can support clinical evaluation of infant health and risk for adverse outcomes. Existing approaches have addressed the first criteria of synchrony by identifying patterns in the coordination of infant and mother behavioral signals over time. This work explores how addressing the second criteria, intermodality, can expand the knowledge gained from automated assessments of infant-caregiver coordination within a given interaction paradigm.

3 DATA COLLECTION

3.1 Participants

57 infant-mother dyads were recruited from the local community. All data were collected at the Children's Hospital Los Angeles, under IRB protocol CHLA-15-00267. Each dyad was invited to participate at 2 months, 6 months, 9 months, 12 months, and 18 months after the birth of the infant, to complete the FFSF procedure. A total of 229 interactions were completed.

3.2 Experimental Procedure

At each visit, dyads engaged in the the FFSF procedure. Infants and mothers sat 2-3 feet apart, as shown in Figure 1, with the infant sitting on a researcher's lap or worn on the front of a researcher in a baby carrier. The procedure started with the Play stage, in which the mother and infant were given a basket of toys to play with. This stage lasted 2 minutes, and was followed by a 2-minute Still-Face stage, in which the mother was asked to maintain eye contact with the infant but not interact with the infant or express facial affect. If the infant was fussy and crying for a continuous period of thirty seconds, the Still-Face stage was terminated early. When the Still-Face stage ended, another 2-minute Play stage (called Reunion) started, in which the mother resumed play with her infant. All FFSF procedures were recorded at 30 frames per second from a profile view.

Interactions were excluded from analysis if the infant or mother were unable to engage in one or more stages of the FFSF procedure. Seven interactions were excluded because the infants fell asleep during the procedure. Eight interactions were excluded because the infants were too fussy to reach the Still-Face stage or the Reunion stage without a significant break from the procedure. Additionally, one interaction was excluded because the mother did not disengage from her infant during the Still-Face stage, and one interaction was excluded because the mother was using her phone for an extended period of time during the procedure. During one visit, the Play stage lasted three minutes rather than two minutes, and was excluded. During five of the interactions the camera was restarted, resulting in missing data. After these exclusions, 206 interaction videos remained for analysis.

3.3 Feature Extraction

Serve-and-return interaction can involve visual attention, gestures, and vocalizations, driving our choice of feature extraction methods.

3.3.1 *Video Features.*

Consistent with prior work in this area [1, 7], head and arm positions of the dyad were measured throughout the interactions. Head orientation was one of the original social signals measured during the FFSF procedure [22], and Stiefelhagen et al. [20] demonstrated head orientation to be a good indicator of visual attention.

Hand and arm movements are involved in manipulation of toys shared by the mom and infant; Tronick et al. [22] identified arm positions and movements as behaviors of interest during the FFSF procedure. Our work focuses on the position of the upper-arm closest to the camera, as these were consistently identifiable within the video frame. As the forearms and hands of the infant and mother were less frequently visible in the video frame, their positions were not considered in this analysis.

Pose features were extracted from the videos using the open-source software OpenPose [4]. In order to maintain invariance to the position of the dyad within the frame, the size of the infant, occasional readjusting of the camera, or slight repositoning of the infant by the researcher, we measured the angles between joints, rather than individual joint positions. As the videos were filmed in profile, the pitch of the infant's and mother's head and upper-arm were measured as proxies for head and arm positions. Head angle, or pitch, was approximated as the angle between the horizontal and the line connecting the detected nose position and ear position. Upper-arm pitch was measured as the angle between the horizontal and the line between the detected elbow and shoulder joints. These angles are shown in Figure 1.

In 10 videos, either the mother or infant could not be reliably tracked due to significant occlusions, or due to family members sitting too close to the dyad and preventing confident discrimination between the detected features of each person. The remaining dataset of 196 videos (D_{video}) includes interaction videos from 54 of the 57 infant-mother dyads who participated in the study.

3.3.2 Audio Features. Due to its connection to arousal and infant developmental status [8, 9], vocal fundamental frequency (F0) was extracted for both infant and mother, using Praat [2] open-source software for speech-processing. Praat default settings are based on adult speech; the range for detecting infant F0 was adjusted to between 250 and 800 Hz based on settings suggested by Gabrieli et al. [6] for analyzing infant vocal fundamental frequency. The F0 is sampled at 100 Hz, and subsampled to 30 Hz when relating F0 to pose.

As the infants, mothers, and researchers all spoke or vocalized during Play and Reunion stages, speaker diarization was performed by two trained annotators. 10% of the videos were processed by both annotators, and a Cohen's Kappa value of 0.83 was achieved, indicating high agreement. Vocalizations were labeled as 'mother and infant', 'infant', or 'other', with remaining vocalizations attributed to the mother. Timer sounds indicating the beginning or end of a stage, and occasional instances of a researcher or passerby speaking, were annotated as 'other' and were not considered in later analysis. At some points, infants and their mothers vocalized at the same time. Mothers tended to vocalize more often and more loudly (except when the infants were crying); therefore, the vocal fundamental frequencies measured during these times were attributed to the mothers. The amount of time with both an F0 value and an annotation of 'infant' was recorded as the total duration of vocalization for each infant. Only vocalizations with F0 were considered for this measurement, to prevent the detection of unvoiced sounds such as the infant's breathing.

Some of the toys in the experiment played loud music throughout the interaction, interfering with the F0 measurement; therefore, interactions where music was played were excluded from audio-based analysis. After excluding videos with music, 68 interactions remained for audio feature analysis. This subset of the data (D_{audio}) includes data from 39 of the original 57 infant-mother dyads. Dyads who did not play music typically included 2- or 6-month-old infants. As the musical toys were commonly used, most dyads are represented in D_{audio} once or twice; only five infants are represented at more than two ages. A breakdown of each dataset by age is shown in Figure 2. De-identified features from both datasets can be found via the link in Section 1.

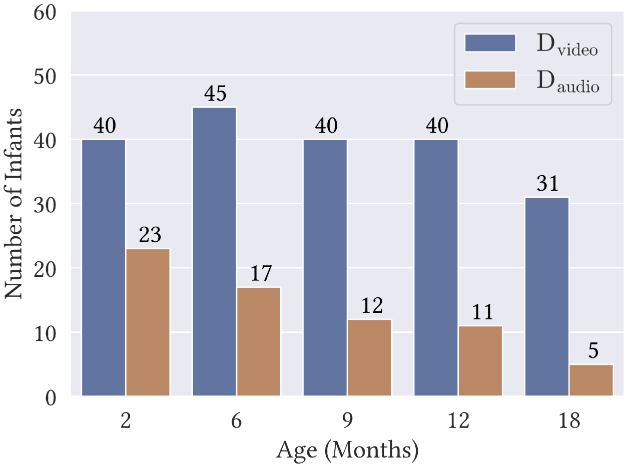

Figure 2: Number of interactions in each dataset. D_{video} is the set of interactions with reliable video data. D_{audio} is the set of interactions with reliable audio data. D_{video} consists of data from 54 dyads, and D_{audio} consists of data from 39 dyads.

4 DATA ANALYSIS

To evaluate the benefit of measuring intermodal behavioral coordination, we measured both the infants' individual behavior changes, as well as changes in infant-mother coordination, across age and stages of the FFSF procedure. As analysis of the infants' behavior changes is typical when conducting the FFSF procedure [1], this allowed us to confirm whether our results were consistent with previous research findings and developmentally-relevant phenomena. Additionally, differences in the infants' tendencies to change their behavior across the three types of behavioral signals support the need for measures of coordination that account for these trends.

4.1 Infant Responses to Still-Face Paradigm

To investigate infant behavior change, infant behavioral signals were aggregated by age and stage of the FFSF procedure. Using these values, we compared how infant responses to the Still-Face paradigm differed across age, and compared our results with expected infant behavior. The variances of the head and arm angles were calculated to approximate the total amount of infant head and arm movement during each stage of the FFSF procedure. This process was conducted separately for each infant at each age. To compare vocal behavior across stages, we calculated the percent of time during which the infant was vocalizing in each stage.

4.2 Infant-Mother Coordination

4.2.1 Windowed Cross-Correlation with Peak-Picking. In this work, we modeled behavioral coordination using windowed cross-correlation with peak-picking, as described by Boker et al. [3]. This approach has been used by Hammal et al. [7] in the context of the FFSF to evaluate changes in head-pose coordination between 4-month-old infants and their mothers from Play to Reunion. Windowed cross-correlation with peak-picking is characterized by its ability to track changes in the temporal dynamics of an interaction [3], making it suitable for modeling the fluctuating patterns of serve-and-return interaction.

Windowed cross-correlation estimates the peak strength and time lag of correlations between two signals at successive windows. A sliding window of length w_{max} is applied to the two signals at steady increments w_i. To account for varying temporal relationships between signals, Pearson correlation coefficients (R) are calculated across multiple lags, with a maximum lag value t_{max}. To evaluate the lagged correlation values between a mother's and infant's behavioral signals, the window of the infant's behavioral signal was shifted between $-t_{max}$ and $+t_{max}$. For an infant's signal I and mother's signal M, the pair of windows W_I and W_M considered after k window increments and at lag t were selected as shown in Equations 1 and 2, below:

$$W_I(k,t) = [I_{kw_i+t}, I_{kw_i+1+t}, ...I_{kw_i+w_{max}-1+t}] \tag{1}$$

$$W_M(k) = [M_{kw_i}, M_{kw_i+1}, ...M_{kw_i+w_{max}-1}] \tag{2}$$

Pearson correlation coefficients were calculated at each lag, producing a series R of correlation values shown in Equation 3:

$$\begin{split} R = [r(W_I(k,-t_{max}), W_M(k)), r(W_I(k,-t_{max}+1), W_M(k)), \\ ...r(W_I(k,t_{max}), W_M(k))] \end{split} \tag{3}$$

where $r(W_I, W_M)$ calculates the Pearson correlation coefficient between the two windowed signals, W_I and W_M.

The plot of correlation value as a function of lag was smoothed to reduce noise using a quadratic Savitzky-Golay filter with a moving window of 5 samples. The lag at which a peak correlation occurs was identified and considered for later analysis. This process is further illustrated in Figure 3.

The appropriate window size, window step size, maximum lag, and lag step size depend on the dataset. Additionally, establishing criteria for identifying peaks can help to reduce the number of erroneous peaks caused by noise. The next subsection discusses the selection of these parameters and criteria.

4.2.2 Parameter Selection. An inherent challenge of windowed cross-correlation across two different behavioral signals stems from the fact that the appropriate window size may be different for the two behaviors. Window sizes of 3-4 seconds are typical when analyzing motor movements. Hammal et al. [7] found 3 seconds to be an appropriate window size for analyzing infant and mother head pose coordination, while Boker et al. [3] used window sizes of 4 seconds to analyze head and arm gestures during dyadic conversations to account for the typical amount of time needed to produce and perceive these gestures. Based on these prior studies and our initial analysis, we used a window size of 3 seconds (90 samples) when analyzing arm and head pose coordination.

In contrast to head and arm movements, vocalizations occurred in much smaller windows of time. After interpolating between vocalizations that occurred less than 0.25 seconds apart, the average length of vocalizations by the mother was found to be between 0.5 and 1 seconds ($\mu = 0.61s, \sigma = 0.39s$). As each instance of windowed cross-correlation requires a single w_{max}, the difference in time scales was resolved by selecting the smaller of the two windows when modeling the coordination between vocal and pose signals. Since continuous 3-second-windows of vocalization cannot be measured reliably, we selected a 1-second-window for calculations involving vocalization. The correlations between signals was calculated provided the infant or mother vocalized for over 50% of the window period; otherwise, no correlation value or peak was reported. As this resulted in smaller window sizes for a portion of the data and therefore increased noise, we selectively considered only peaks corresponding to R values that were statistically significant at $p < 0.05$. Based on our window size and initial data analysis, a maximum lag value of 10 was selected.

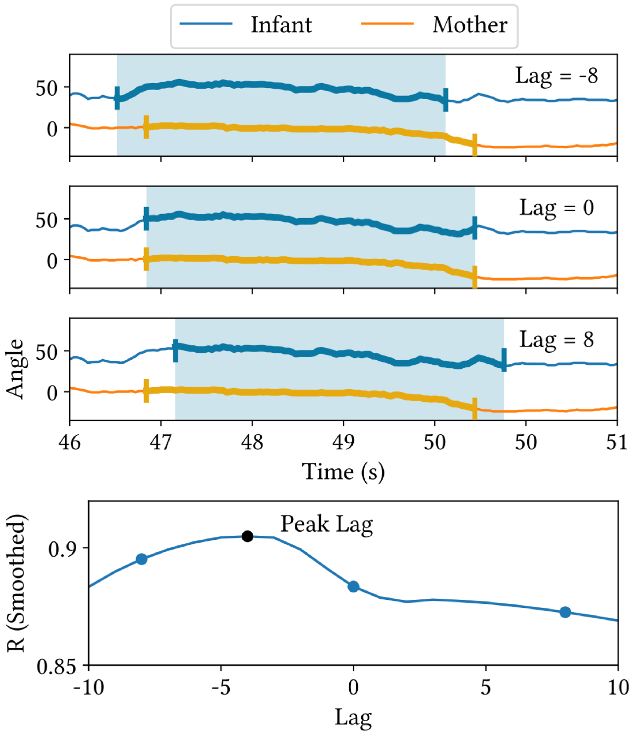

Figure 3: Illustration of windowed cross-correlation with peak-picking, evaluated for a single window on example signals. Top: a given time window is shifted across multiple lag values, changing the range of infant angle values considered. Bottom: The correlation values between windowed signals are plotted, and a peak lag value is identified.

4.2.3 Model Evaluation. Results of the cross-correlation analysis were evaluated in two ways. First, we evaluated whether two signals were significantly coordinated by comparing the number of peaks found between the true signals with the number of peaks found when each signal was randomly reordered. Behavioral signals that had significantly more peaks than their randomized counterparts were considered for further analysis.

To determine how dyadic coordination changed with FFSF stage and across ages, we calculated the variance of peak lag values for each infant, stage, and age. Boker et al. [3] noted that higher lag variance corresponded with interactions that are further from synchrony, and found that lag variance increased significantly following an interruption or communication barrier (in their case, the presence of ongoing loud noise). As a higher lag variance indicates the presence of lag values that are further from 0 and less stability in the interaction dynamics, a higher lag variance may also indicate less coordinated behavior in the dyadic interaction. We therefore used lag variance as an approximation of the level of coordination between two behavioral signals.

5 RESULTS

This work evaluates differences in behavior across both age and stage of the FFSF; we therefore considered statistical significance at $\alpha < 0.025$, using the Bonferroni correction for multiple comparisons.

5.1 Infant Responses to Still-Face Paradigm

5.1.1 Infant Head Angle. Repeated measures ANOVAs revealed significant differences in the amount of head pitch variance across stages of the FFSF at ages 6 and 12 months, $p < 0.005$. Student's t tests were then used to identify differences between individual stages, with results reported in Figure 4. There was an increase in variance during Still-Face at 9 months as well, but this was not found to be statistically significant. Figure 4 shows the distribution of head pose variance measurements across ages and stages of the FFSF procedure.

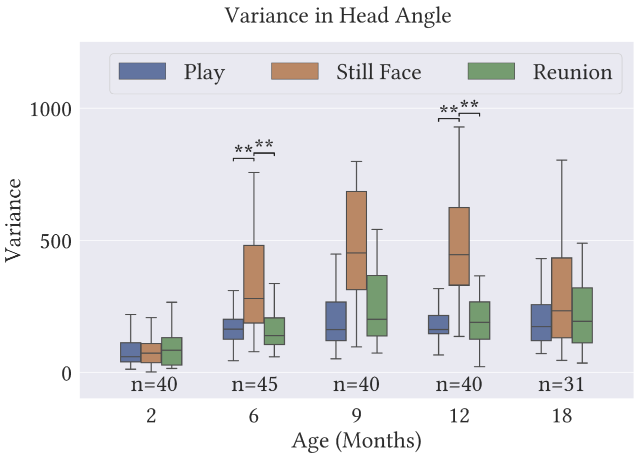

Figure 4: Variance in infant head pitch across FFSF stage and age. Bars and asterisks represent significant results for Student's t tests between individual FFSF stages with $p < 0.001$.

5.1.2 Infant Arm Angle. Figure 5 shows the distribution of arm angle variance across age and FFSF stage. Differences in variances between stages were not found to be significant; however, a linear mixed effects model with infant ID as a fixed effect found that the average amount of arm angle variance across all three stages increased with age, $p < 0.025$. This indicates an increase in the amount of infant arm movement with age.

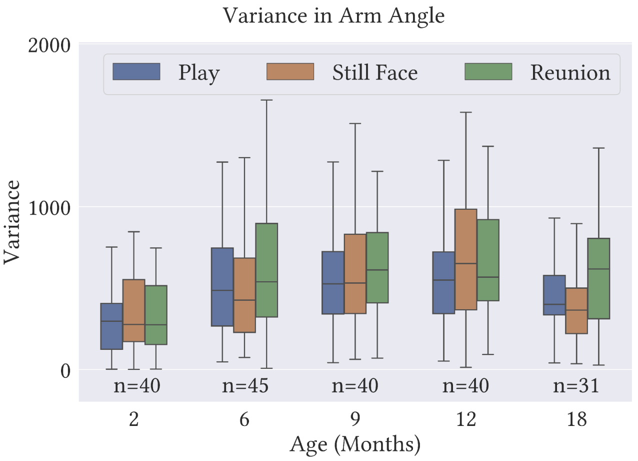

Figure 5: Variance in infant arm angle across stage and age

5.1.3 Infant Vocalizations. Repeated measures ANOVAs found significant differences in the amount of infant vocalization across the FFSF stages at 2 and 6 months of age, with $p < 0.005$. We also note increases in the amount of vocalization at 9 and 12 months, though these were not significant with the Bonferroni correction ($p = 0.048$, 0.035, respectively). Student's t tests were used to evaluate differences between individual stages at these ages. Infants at 6, 9, and 12 months of age vocalized more often during the Still-Face stage, while 2-month-old infants vocalized most during the Recovery stage. These trends are illustrated in Figure 6. No significant relationship was found between age and amount of vocalization. The amount of infant vocalization was the only metric with significant differences across stages of the FFSF procedure for 2-month-olds.

5.2 Infant-Mother Coordination

5.2.1 Testing for significant coordination. Each measure of cross-correlation incorporating two pose signals produced significantly more peaks than randomly reordered signals ($p < 0.001$), indicating significance with Bonferroni correction. Measures of cross-correlation incorporating the mother's vocal signals produced more peaks than random ($p < 0.005$) at every age except for 18 months, which had a sample size of only 5 dyads. The number of peaks was not significantly greater than random for measures incorporating the infant's vocalization signals. For pairs of signals showing significantly more peaks than random, this indicates that the ordered behaviors of the mother and infant were coordinated across time.

5.3 Lag Variance

For pairs of signals that demonstrated significant coordination, we compared the peak lag variance across FFSF stage and across age to identify trends in infant behavior across time. Student's t tests identified whether differences in lag variance between the Play and Reunion stages were significant, indicating a shift in the level of coordination across the two stages. These results are illustrated in Figure 7. Each pair of signals had a unique range of ages for which the lag variance could significantly distinguish between Play and Recovery. The *(mother head angle, infant arm angle)* pair was

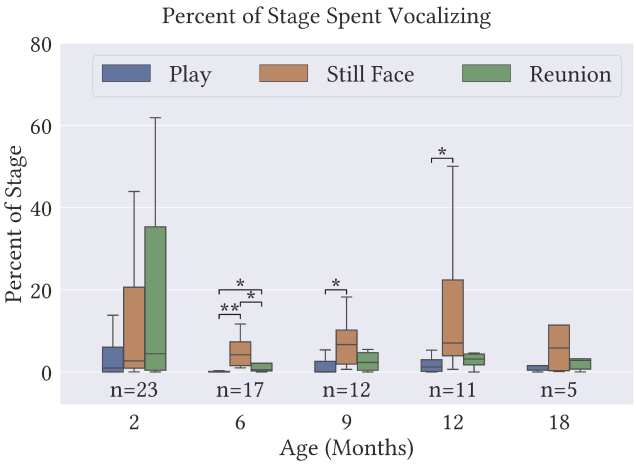

Figure 6: Amount of infant vocalization across stage and age. Bars and asterisks represent significant results for Student's t tests between individual FFSF stages, with * $p < 0.025$, and ** $p < 0.001$. While differences between individual pairs of stages were not significant at age 2 months, a repeated measures ANOVA reported a significant relationship between stage and amount of vocalization ($p < 0.005$).

the only signal pair to show significant differences in lag variance across stages for all age groups; however, each age group had at least two signal pairs for which lag variance differed significantly across stages. We note that intermodal signal pairs, specifically the *(mother head angle, infant arm angle)* and *(mother F0, infant head angle)*, were the only pairs of signals for which significant differences in lag variance were found between Play and Reunion for 2-month-old infants. Notably, all significant differences across stages showed an increase in lag variance from Play to Reunion, representing less consistent dynamics in the Reunion stage compared to Play.

A linear mixed models analysis with infant id as a random effect indicated a significant, negative relationship between age and lag variance for the *(mother head angle, infant head angle)*, *(mother arm angle, infant arm angle)* and *(mother head angle, infant arm angle)* pairs ($p < 0.025$). These trends are illustrated in Figure 8. Significant differences across age were not seen for the remaining pairs. While the max lag t_{max} was chosen empirically, the direction of change in lag variance remained negative across age and positive across experimental stage for t_{max} values between 8 and 14 inclusive, for the ages and signal pairs that yielded significant results.

6 DISCUSSION

Our results support the value of intermodal coordination metrics for achieving a better understanding of infant-mother interaction. Prior to this work, research in computational modeling of infant-mother interaction mainly focused on coordination across symmetric modes. In this work, the temporal dynamics of intermodal infant-mother coordination identified meaningful trends across conditions, showing significantly more variance after the stressful Still-Face stage, and at younger infant ages. These trends mirrored similar analyses evaluated on unimodal signals, but extended our ability to distinguish between stages of the FFSF at certain ages. For example, only intermodal signal pairs were significantly able to distinguish between coordination dynamics from Play to Reunion at 2 months of age. Moreover, the *(mother head angle, infant arm angle)* pair demonstrated a greater decrease over time than any other pair of signals. We anticipate that including intermodal measures of coordination will improve performance in automated evaluation of synchrony, and subsequently in predicting health outcomes.

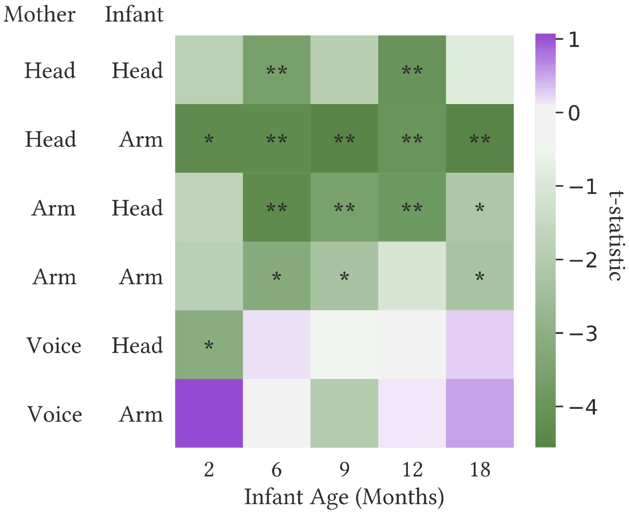

Figure 7: Student's t test statistic between lag variance distributions during Play and Reunion. Negative values indicate a higher mean lag variance during Reunion compared to Play. The key at the left indicates the behavioral signals which were input into the windowed cross-correlation model. Significance is reported with * $p < 0.025$, and ** $p < 0.001$.

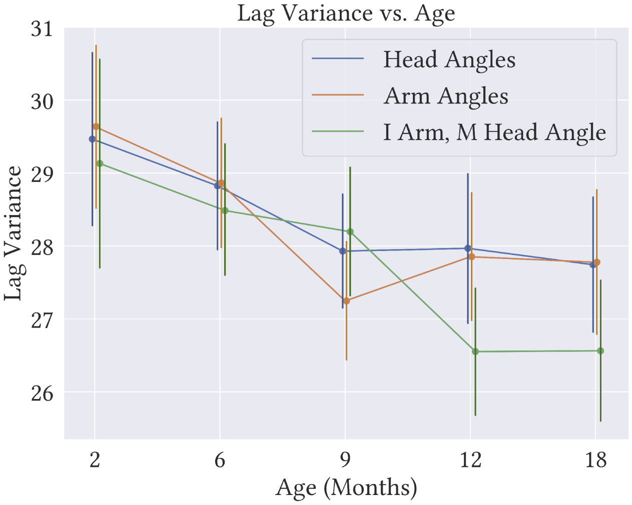

Figure 8: Median lag variance across age, with 95% confidence intervals.

The importance of observing behaviors across modes was also evident from the infants' individual behaviors, with each type of behavioral signal demonstrating a different trend. Increases in head movement followed an almost parabolic trend, with the largest changes at 6, 9, and 12 months. Our finding that infants reacted more strongly to the Still-Face at certain ages based on this metric is consistent with prior work using the FFSF procedure: as infants grow older, they gain motor skills and therefore their reaction to the Still-Face stage along certain metrics becomes more observable [1]. As the 2-month-old infants often needed to have their heads supported by the researcher, their ability to adjust their head movements may have been limited. Adamson et al. [1] also noted that after 9 months of age, infants are better able to distract themselves and therefore are less affected by their mother's ignoring behavior during the FFSF procedure. This was reflected in the smaller increase in head movement at 18 months as compared to 6, 9, and 12 months. The significance of the change in head pose variance at 12 months may be due to the lack of toys during the Still-Face stage. In many of the procedures, the mothers removed the toys from the infant's grasp during the Still-Face stage, often causing the infants to become fussy, resulting in more head movement.

Conversely, infants demonstrated increased arm movement with age, but not across FFSF stages. As infants used their arms during the Play and Reunion stages to touch and manipulate toys, the variance remained high during these stages. The amount of vocalization varied across stages at all ages except 18 months, and was the only feature to distinguish infant behavior between stages at 2 months. Given the evolution of communication behaviors with age, tracking behavior using multiple modalities is necessary to fully represent infant responses to stressful situations such as the FFSF procedure.

These results also highlight the impact of differences in communication abilities on relevant coordination metrics. While some intermodal signals were coordinated and showed meaningful changes in coordination across age, this did not necessarily imply that the reversed pair of signals demonstrated the same trend. For example, while the pairs including the mothers' vocalization signals were significantly coordinated, pairs including the infants' vocalization signals were not. This was likely because mothers vocalized and spoke more often, while infant vocalizations were relatively infrequent compared to the length of the interaction. While the *(mother head angle, infant arm angle)* pair showed significant trends across age, the *(mother arm angle, infant head angle)* pair did not. Further work should explore whether this pattern of asymmetric behavioral coordination may be generalized to domains where communication differences can pose a challenge, such as in human-robot interaction or during interactions with individuals with ASD.

The ability to detect intermodal coordination with computational methods allowed a more detailed view of infant-mother interaction at each age. Incorporating these metrics increases the number of features available when evaluating coordination between individuals. The expanded feature space may make models of higher-level phenomena such as synchrony more robust against missing data. This is supported by the presence of multiple signal pairs that produced similar differences in coordination patterns across FFSF stages at certain ages, especially at ages 6, 9, and 12 months. Given that pose and audio features can be occluded or obstructed by toys and siblings even in a lab setting, this robustness would be necessary if caregivers were to record these interactions in the home.

7 CONCLUSION

This work introduces intermodal coordination as a valuable feature for computational modeling of trends in infant-mother coordination. Given the fast-changing nature of infant communication abilities, evaluating these additional interactions and how they evolve with time may help to more accurately track higher-level dyadic processes such as synchrony. Widening the feature space by incorporating intermodal coordination could promote robustness in evaluating these processes. By identifying trends in infant behavior and dyadic processes across age and experimental condition, this work presents intermodal coordination as an essential feature in the computational analysis of infant-mother interaction.

8 FUTURE WORK

Evaluating coordination across conditions is a first step toward modeling interaction quality. Future work will leverage intermodal coordination metrics to model interaction quality over time for a given infant-mother dyad; this is an important feature for computational tools that can evaluate dyadic processes in a clinical setting. Differences between videos, including occlusions, varying camera angles, and inconsistencies in whether the toys were removed for the Still-Face stage made this analysis infeasible across individual pairs of signals. As these challenges will be difficult to mitigate in larger trials and in-the-wild studies, it may not be feasible for individual metrics of coordination to be used to predict trajectories across age for individual infants. Rather, future work will use these metrics as features to model changes in synchrony over time.

This work established the presence of meaningful intermodal coordination; while we evaluated a small set of interaction modes, we anticipate that this approach can be generalized to a larger set of behavioral signals. Future work will involve a similar study utilizing the FFSF procedure with the addition of a second camera, to capture both facial features and pose. We will incorporate directional microphones to capture higher resolution vocal data which can be more easily and automatically diarized. This will also facilitate distinguishing between infant crying and vocal babbling, as these forms of vocalization represent different emotional states. Finally, while we have shown that there exist significant coordination and trends across intermodal signals, we acknowledge that windowed cross-correlation may not be optimal for each interaction mode; this method should be compared with others that monitor state transition dynamics or delays and overlaps between actions in infant and mother turn-taking. We anticipate that exploring these methods will benefit the analysis of behaviors that occur with low frequency, including vocalizations.

ACKNOWLEDGMENTS

This research was funded by the National Science Foundation under grant NSF CBET-1706964, and by the JPB Foundation through a grant to The JPB Research Network on Toxic Stress: A Project of the Center on the Developing Child at Harvard University. The authors would like to thank Alma Martinez and Kathryn Arneaud for their contributions toward the study.

REFERENCES

[1] Lauren B Adamson and Janet E Frick. 2003. The still face: A history of a shared experimental paradigm. *Infancy* 4, 4 (2003), 451–473.

[2] P Boersma and D Weenink. 2002. Praat 4.0: a system for doing phonetics with the computer [Computer software]. *Amsterdam: Universiteit van Amsterdam* (2002).

[3] Steven M Boker, Jennifer L Rotondo, Minquan Xu, and Kadijah King. 2002. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological methods* 7, 3 (2002), 338.

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

[5] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. 2012. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing* 3, 3 (2012), 349–365.

[6] Giulio Gabrieli, Wan Qing Leck, Andrea Bizzego, and Gianluca Esposito. 2019. Are Praat's default settings optimal for Infant cry analysis. In *Proceedings of the 2019 CCRMA Linux Audio Conference, LAC, Stanford, LA, USA*. 23–26.

[7] Zakia Hammal, Jeffrey F Cohn, and Daniel S Messinger. 2015. Head movement dynamics during play and perturbed mother-infant interaction. *IEEE transactions on affective computing* 6, 4 (2015), 361–370.

[8] Patrik N Juslin and Klaus R Scherer. 2005. Vocal expression of affect. *The new handbook of methods in nonverbal behavior research* (2005), 65–135.

[9] A Kappas, U Hess, and KR Scherer. 1991. Voice and emotion: Fundamentals of Nonverbal Behavior. *Rim, B., Feldman, RS (eds.)* (1991), 200–238.

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

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

[12] Ammar Mahdhaoui, Mohamed Chetouani, Raquel S Cassel, Catherine Saint-Georges, Erika Parlato, Marie Christine Laznik, Fabio Apicella, Filippo Muratori, Sandra Maestro, and David Cohen. 2011. Computerized home video detection for motherese may help to study impaired interaction between infants who become autistic and their parents. *International Journal of Methods in Psychiatric Research* 20, 1 (2011), e6–e18.

[13] Daniel M Messinger, Paul Ruvolo, Naomi V Ekas, and Alan Fogel. 2010. Applying machine learning to infant interaction: The development is in the details. *Neural Networks* 23, 8-9 (2010), 1004–1016.

[14] Setareh Nasihati Gilani, David Traum, Arcangelo Merla, Eugenia Hee, Zoey Walker, Barbara Manini, Grady Gallagher, and Laura-Ann Petitto. 2018. Multimodal dialogue management for multiparty interaction with infants. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 5–13.

[15] Center on the Developing Child at Harvard University. 2012. The science of neglect: The persistent absence of responsive care disrupts the developing brain.

[16] Livio Provenzi, Giunia Scotto di Minico, Lorenzo Giusti, Elena Guida, and Mitho Müller. 2018. Disentangling the dyadic dance: theoretical, methodological and outcomes systematic review of mother-infant dyadic processes. *Frontiers in psychology* 9 (2018), 348.

[17] Sally J Rogers, L Vismara, AL Wagner, C McCormick, G Young, and S Ozonoff. 2014. Autism treatment in the first year of life: a pilot study of infant start, a parent-implemented intervention for symptomatic infants. *Journal of autism and developmental disorders* 44, 12 (2014), 2981–2995.

[18] Seokwoo Song, Seungho Kim, John Kim, Wonjeong Park, and Dongsun Yim. 2016. TalkLIME: mobile system intervention to improve parent-child interaction for children with language delay. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 304–315.

[19] Angela EB Stewart, Zachary A Keirn, and Sidney K D'Mello. 2018. Multimodal modeling of coordination and coregulation patterns in speech rate during triadic collaborative problem solving. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 21–30.

[20] Rainer Stiefelhagen and Jie Zhu. 2002. Head orientation and gaze direction in meetings. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*. 858–859.

[21] Chuangao Tang, Wenming Zheng, Yuan Zong, Zhen Cui, Nana Qiu, Simeng Yan, and Xiaoyan Ke. 2018. Automatic Smile Detection of Infants in Mother-Infant Interaction via CNN-based Feature Learning. In *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*. 35–40.

[22] Edward Tronick, Heidelise Als, Lauren Adamson, Susan Wise, and T Berry Brazelton. 1978. The infant's response to entrapment between contradictory messages in face-to-face interaction. *Journal of the American Academy of Child psychiatry* 17, 1 (1978), 1–13.