Dynamic Mode Decomposition with Control as a Model of Multimodal Behavioral Coordination

Lauren Klein
kleinl@usc.edu
University of Southern California
Los Angeles, California, USA

Victor Ardulov
ardulov@usc.edu
University of Southern California
Los Angeles, California, USA

Alma Gharib
agharib@chla.usc.edu
Children's Hospital Los Angeles
Los Angeles, California, USA

Barbara Thompson
thom1756@msu.edu
Michigan State University
East Lansing, Michigan, USA

Pat Levitt
plevitt@med.usc.edu
Children's Hospital Los Angeles
Los Angeles, California, USA

Maja Matarić
mataric@usc.edu
University of Southern California
Los Angeles, California, USA

ABSTRACT

Observing how infants and mothers coordinate their behaviors can highlight meaningful patterns in early communication and infant development. While dyads often differ in the modalities they use to communicate, especially in the first year of life, it remains unclear how to capture coordination across multiple types of behaviors using existing computational models of interpersonal synchrony. This paper explores Dynamic Mode Decomposition with control (DMDc) as a method of integrating multiple signals from each communicating partner into a model of multimodal behavioral coordination. We used an existing video dataset to track the head pose, arm pose, and vocal fundamental frequency of infants and mothers during the Face-to-Face Still-Face (FFSF) procedure, a validated 3-stage interaction paradigm. For each recorded interaction, we fit both unimodal and multimodal DMDc models to the extracted pose data. The resulting dynamic characteristics of the models were analyzed to evaluate trends in individual behaviors and dyadic processes across infant age and stages of the interactions. Results demonstrate that observed trends in interaction dynamics across stages of the FFSF protocol were stronger and more significant when models incorporated both head and arm pose data, rather than a single behavior modality. Model output showed significant trends across age, identifying changes in infant movement and in the relationship between infant and mother behaviors. Models that included mothers' audio data demonstrated similar results to those evaluated with pose data, confirming that DMDc can leverage different sets of behavioral signals from each interacting partner. Taken together, our results demonstrate the potential of DMDc toward integrating multiple behavioral signals into the measurement of multimodal interpersonal coordination.

ICMI '21, October 18–22, 2021, Montréal, Canada

ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479916

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing design and evaluation methods**; • **Applied computing** → **Health informatics**.

KEYWORDS

dynamical systems models, dyadic interaction, infant-mother interaction, multimodal behavioral coordination

ACM Reference Format:
Lauren Klein, Victor Ardulov, Alma Gharib, Barbara Thompson, Pat Levitt, and Maja Matarić. 2021. Dynamic Mode Decomposition with Control as a Model of Multimodal Behavioral Coordination. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3462244.3479916

1 INTRODUCTION

Supportive interaction with a primary caregiver is essential to infant well-being [15]. For decades, this has motivated researchers to observe infant-mother interaction to understand the communication dynamics which support healthy child development [1]. More recently, advances in automated video and audio feature extraction combined with computational modeling approaches have enabled more detailed analyses of the coordination between mother and infant behaviors. Typically, computational approaches have leveraged dynamical system models, correlation, or analysis of overlaps and delays between actions to evaluate the relationships between a given type of infant behavior and a given type of mother behavior. Past work modeling infant-mother dyads as dynamical systems demonstrated how Markov processes can model the transitions between expert-annotated affective states [5] or between smile onset and termination [14]. Automated approaches have analyzed how the timing of infant and mother body movements relate to expert ratings of interactions [13], or how head pose coordination changes across stages of an interaction paradigm [9]. A recent study by Song et al. [17] evaluated how real-time feedback regarding infant and parent vocalizations could promote healthy turn-taking dynamics.

While prior work has modeled infant-mother interaction using a range of modalities, it is unclear how automated methods may integrate multiple signals from different modalities. Infants and mothers may communicate with gestures, gaze, vocalizations, and shared affect. As communication skills develop, it is common for

infants and their mothers to differ from each other in the modalities they use to interact [12]. Accounting for infant and mother communication across multiple modalities without manual annotation (or training classifiers on large amounts of data to recognize higher-level behaviors) is an unsolved problem. Our past work [11] demonstrated how windowed cross-correlation with peak-picking may be used to observe coordination between infants and their mothers using different modalities for each partner. While windowed cross-correlation uses a single pair of signals, this paper explores a method for integrating multiple behavioral signals to model multimodal interpersonal coordination.

This work explores Dynamic Mode Decomposition with control (DMDc) as a model for multimodal interpersonal coordination during dyadic interactions. DMDc is a method of extracting low-order models from observations of high-dimensional systems while explicitly accounting for external input or actuation [16]; therefore, it provides two key benefits to the analysis of infant-mother interaction. First, DMDc can incorporate multiple behavioral signals, which is useful given the multimodal nature of infant-mother interaction. Second, modeling actuation enables estimation of the impact of each interaction partner's behaviors on the other partner.

We evaluated our approach using an existing longitudinal dataset [11] of recordings of infants and their mothers participating in the Face-to-Face Still-Face (FFSF) procedure [20], a validated experimental interaction paradigm, collected from 2 to 18 months of age as part of a larger longitudinal project. Pose signals and vocal fundamental frequencies were extracted at 30 Hz using OpenPose [4] and Praat [3], with speaker identification performed manually. As past work [11] indicates that the infants in this dataset modulated their head pose and arm pose differently during the FFSF procedure, (head pose is a proxy for visual attention [18] while arm pose changes during object manipulation,) we considered these features as separate modalities. We studied trends in interaction dynamics during the experimental procedure and across infants using DMDc, comparing results evaluated on unimodal and multimodal data.

Our results demonstrate the ability of DMDc to integrate multiple signals into a single model of multimodal behavioral coordination, even when leveraging different sets of signals for each interacting partner. The metrics produced by this model followed known trends in mothers' behavior and infant-mother interaction across stages of the FFSF procedure, demonstrating the ability of DMDc to capture relevant social information. Effect sizes were larger when models were fit on both head pose and arm pose data, providing initial support for DMDc as a model of multimodal coordination. Additional exploration showed that for both unimodal and multimodal data, model parameters identified relationships between changes in interaction dynamics (defined in Section 3.2) from the play to still-face stages and infants' movement behavior, and trends in both individual and dyadic behaviors across infant age.

Our work makes the following contributions:

- A novel automated approach for modeling multimodal interpersonal behavioral coordination, using Dynamic Mode Decomposition with control
- Novel computational analysis of changes in individual and dyadic behaviors over the course of a developmentally relevant interaction paradigm between infants and mothers.

Section 2 of this paper provides an overview of related research that has informed this work. Section 3 describes the methods used in this work. Section 4 reports and discusses the results of our analysis. Section 5 provides conclusions based on our findings, and Section 6 describes directions for future work.

2 RELATED WORK

This section summarizes: 1) evaluation methods for monitoring infant-mother interaction; 2) computational analysis of infant-mother interaction; and 3) Dynamic Mode Decomposition with control.

2.1 Structure of Infant-mother Interaction

2.1.1 Dynamic Patterns. Infants and their mothers coordinate their behaviors across multiple time scales. Moment-to-moment, dyads engage in a "serve-and-return" pattern, an essential component of infant-mother interaction where one partner reaches out with a gesture or vocalization and the other responds with their own action [15]. However, over the course of an interaction, dyads may not remain constantly in a coordinated state. Instead, the mutual regulation model [8] asserts that alternating states of synchronous and mismatched interactions are common, as long as the dyad is able to successfully repair interactions following a mismatch [21].

2.1.2 Coding Scales. Typical approaches for analyzing infant-mother interaction rely on time-intensive manual annotation. As communication dynamics evolve over the course of an interaction, researchers or clinicians observe videos of the interactions and measure dynamic processes across multiple time scales [12]. Rating systems such as the Coding Interactive Behavior scale [6] identify global metrics that assess the entire interaction. These systems focus both on behaviors of each partner, such as the mother's gaze toward the infant, and dyadic qualities such as the fluency of the interaction. In contrast, micro-coded time series involve generating a time series of individual behaviors for both interacting partners [12]. These time series are then compared via cross-correlation [9], analysis of delays and overlaps [13], or as a dynamical system [14].

2.1.3 Face-to-Face Still-Face Procedure. To observe how dyads interact in specific situations and to control external sources of variance, researchers employ interaction paradigms, or procedures that each dyad follows. One of the most common validated experimental procedures for monitoring infant-mother interaction is the Face-to-Face Still-Face (FFSF) procedure [20]. The procedure starts with the 'play' stage, during which infants and their mothers play normally. Next, the play interaction is deliberately perturbed to evaluate how mothers and infants repair the interaction in response to stress. Mothers are instructed to maintain eye contact with their infants, but to keep an emotionless expression and refrain from responding to their infants' cues. This is termed the 'still-face' stage. Finally, during the 'reunion' stage, the dyad resumes play. Each stage lasts for 2-3 minutes. This procedure has been widely used to study early social and emotional development, emotion regulation, and differences between groups such as dyads with typically versus atypically developing infants or mothers with depression versus without depression [1]. Given the known interaction patterns that occur during this paradigm, observations of the FFSF procedure serve as useful data for evaluating models of dyadic interaction.

2.2 Computational Analysis of Infant-mother Interaction

More recently, computational methods have emerged to support automated analysis of infant-mother interaction. Past work has suggested modeling the infant and mother as a dynamical system. Markov processes have been used to model transitions between engaged and disengaged infant states in response to their mother's affective expressions [5], soothing actions of mothers in response to fussy infants [19], and smiles of the infant and mother [14]. However, identifying higher-level behaviors or states that involve multiple modalities, such as engagement and soothing actions, requires manual annotation. Typically, automated methods involve comparing a single time series representing the infant's behavior with a single time series representing the mother's behavior. Relationships between these signals are identified by cross-correlating or identifying delays and overlaps. For example, Hammal et al. [9] used windowed cross-correlation with peak picking to identify changes in head pose coordination during the FFSF procedure. Weisman et al. [22] used automatically detected fatherese to evaluate infant and father vocalizations, finding that the amount of pause by the father differed after the still-face stage of the FFSF procedure.

As infant-mother interaction is inherently multimodal, especially in the first year of life [12], our past work [11] evaluated an approach for modeling behavioral coordination across modalities. Specifically, we modeled interactions between pairs of behavioral signals from infants and their mothers during the FFSF procedure. Using windowed cross-correlation with peak-picking, we demonstrated that infant-mother behavioral coordination across modalities (for example, measuring the mother's vocalization fundamental frequency and the infant's head pose) demonstrated significant trends across both infant age and stages of the experimental procedure. Leveraging the same dataset, our current work demonstrates how multiple pairs of signals, rather than one pair of signals, can be integrated into a single model of multimodal behavioral coordination. This is a necessary step toward enabling a holistic evaluation of coordination during infant-mother interactions.

2.3 Dynamic Mode Decomposition with Control

Dynamic Mode Decomposition with control (DMDc) was introduced by Proctor et al. [16] as a method of evaluating the underlying dynamics of high-dimensional systems with external control inputs. Zhang et al. [23] expanded DMDc to handle systems with time-varying modes by restricting the decomposition to sliding windows. The first use of DMDc to analyze social interaction was by Ardulov et al. [2], who used prosody and lexical features to estimate the impact of interviewer behavior on rapport and information disclosure during child forensic interviews. This paper discusses the ability of DMDc to analyze patterns of interaction dynamics during a validated infant-mother interaction paradigm.

3 METHODOLOGY

This section describes our approach for automated analysis of infant-mother coordination. Section 3.1 describes the feature extraction process, and Section 3.2 discusses the application of DMDc to multimodal behavioral signals. Figure 1 details the full pipeline.

3.1 Dataset

We used a dataset of audio-video recordings of infants and their mothers participating in the Face-to-Face Still-Face (FFSF) procedure described in Section 2.1.3. All data were collected at the Children's Hospital Los Angeles under IRB protocol CHLA-15-00267; recordings were labeled with anonymized participant ID numbers and age group. In this implementation of the FFSF protocol, the infant was seated on the lap of a researcher, and the dyad was given a set of toys. Each of the play, still-face, and reunion stages lasted 2 minutes; however, if the infant was fussy for a continuous period of 30 seconds during the still-face stage, the stage was stopped early and the procedure progressed to the reunion stage. As this dataset was collected as part of a larger effort to study the effects of maternal stress on infant development, only mothers were recruited to participate with their infants (fathers were not recruited).

57 unique infant-mother dyads participated in the study. Each dyad participated in the FFSF procedure up to five times, at 2, 6, 9, 12, and 18 months of infant age. Given the emotional difficulty of the procedure, some interactions could not be completed. Interactions were excluded from analysis if the mother broke the interaction protocol for any reason, if the infant was too fussy to reach the final stage of the FFSF protocol, or if the infant was asleep at any time during the procedure. Additionally, videos were excluded if the camera was paused or replaced during the interaction or if the infant or mother were too occluded to collect reliable pose data. After applying exclusion criteria, 200 videos were included for analysis. To track behaviors of both the mothers and infants, we extracted video features including arm pose and head pose, and audio features including vocal fundamental frequency. Sections 3.1.1 and 3.1.2 describe our method for tracking each partner's arm and head pose throughout each recording.

Klein et al. [11] demonstrated that the infants in this dataset modulated their head and arm pose differently throughout the FFSF procedure. On average, infants over 2 months of age showed increased head pose variance from the play stage to the still-face stage, indicating they were moving their heads more during the stressful still-face period. Meanwhile, arm pose variance was typically greater than head pose variance across all stages, but did not change significantly during the still-face stage at any infant age. This was likely due to the need for infants to move their arms in order to grasp and manipulate toys, resulting in greater arm pose variance overall. In contrast, head pose is a proxy for visual attention [18], and infants are known to change their gaze patterns during the FFSF procedure, often showing increased gaze aversion from the mother, from the play to still-face stages [1]. Given the different implications of head and arm posture during the interactions in the dataset used in this work, we consider them as separate modalities.

We used the vocal fundamental frequency (F0) values extracted from this dataset by Klein et al. [11]. F0 is similar to pitch, and is commonly used as a proxy for emotional arousal [10], which mothers and infants are known to coordinate [7]. This dataset includes 67 interactions from 39 dyads; interactions with musical toys precluded accurate collection of F0 values and were excluded. As infants in this dataset vocalized infrequently [11], only the mothers' F0 values were used for further analysis. The recordings comprising

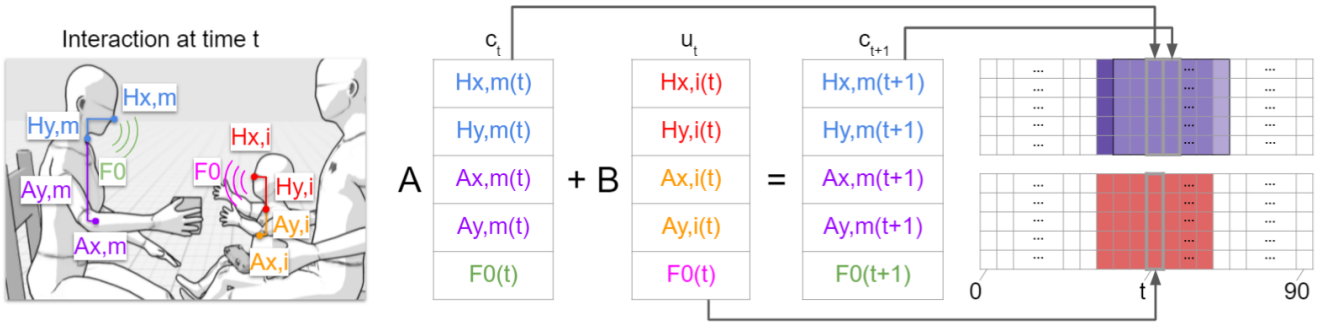

Figure 1: Computational modeling pipeline. Left: pose and audio features are extracted from a video frame; middle: dynamical system model with infant and mother features from two consecutive frames; right: matrices of infant and mother features. In this example, DMDc is applied to multimodal data and the infant's features are used as the control input.

both datasets are detailed in Figure 2. The dataset of video features is labeled as D_{video} and the dataset of audio features as D_{audio}.

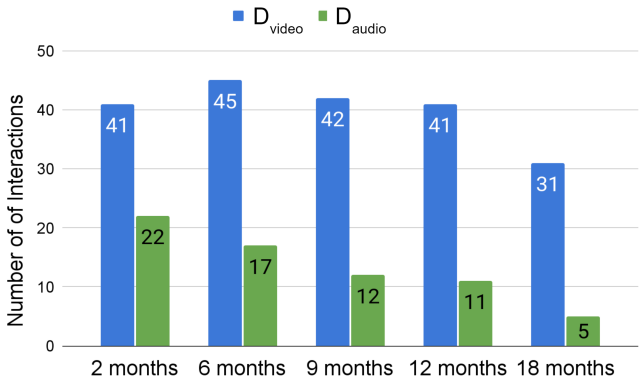

Figure 2: Interaction datasets. The set of video features is labeled as D_{video}, and audio features as D_{audio}. As music precluded the extraction of vocal fundamental frequency in a portion of the videos, the interactions with audio data are a subset of the total set of interactions analyzed.

3.1.1 Person Tracking. The pose of each participant was identified in each frame using the open-source OpenPose software [4]. Leveraging the relatively constant, seated positions of participants within the frame throughout each recording, a clustering method was used to assign each detected set of pose landmarks to the correct person and avoid tracking erroneously detected pose landmarks. First, a cluster centroid was initiated using the head position of the first three most confidently detected people, representing the infant, child, and researcher. Only head landmarks such as the nose or neck were used for tracking, as these had the most consistent positions within each video frame. Participants moved their heads while interacting; however, the overall location of each person's nose and neck landmarks changed less frequently than landmarks on the limbs, resulting in more reliable discrimination between people. If an additional visitor accompanying the dyad was present throughout the entire video, a fourth cluster centroid was initiated. In each consecutive frame, the head position of each detected person was calculated and assigned to a cluster so as to minimize the total distance between each cluster centroid and its new assigned point. The centroid of each cluster was updated after each addition. After the clustering algorithm terminated, each cluster was assigned as the mother, infant, researcher, or visitor based on the known seating positions of each individual. Only the mothers' and infants' data were used for further analysis.

3.1.2 Pose Evaluation. After the pose landmarks for both mother and infant were identified, head and arm pose were monitored within each frame. Head and arm pose were measured using the detected pose landmarks from each frame, as illustrated in Figure 1. Head pose was monitored using the horizontal and vertical distances between the nose and neck landmarks. Arm pose was monitored using the horizontal and vertical distances between the neck and elbow landmarks. As the videos were filmed in profile view, only the arm which was more consistently detected throughout the recording was considered. Since DMDc does not restrict model input to a single signal from each participant, both the horizontal and vertical pose measurements were used as inputs. This removed the need to model pose using angles, which can introduce erroneous variance during fluctuations between 0° and 359°. To normalize by the size of participants or closeness to the camera, distances were divided by the size of each participant's head, measured as the average distance between the nose and ear landmarks.

3.2 Dynamical System Modeling

For each recorded interaction, we fit windowed DMDc models to the extracted data. Separate models were fit for each modality (arm pose, head pose, F0) and for each combination of modalities. This allowed direct analysis of the ability of DMDc to identify trends by comparing results across models. The code used to implement this approach can be found at https://github.com/LaurenKlein/dmdc_analysis.

3.2.1 Window Selection. As described in Section 2.1, infant-mother interactions are characterized by time-varying dynamics; throughout typical play, the individual leading the interaction can change over time. Additionally, even for infants with typical development,

dyads will likely shift between states of coordinated communication and mismatched or asynchronous communication. To capture these changing dynamics, data from each interaction were split into non-overlapping intervals of 3 seconds. This interval was selected based on past work [9, 11] that used windowed-cross correlation to explore infant-mother pose coordination during the FFSF protocol. As the participants' arms were occasionally out of frame, windows with more than 10% of missing arm pose data were excluded from further analysis. For audio analysis, windows without vocalizations, and therefore without F0 data, were excluded.

3.2.2 Control Parameters. To appropriately capture interaction dynamics such as the "serve-and-return" pattern, it is necessary for the model to incorporate changes in the leader of the interaction. Various moments may be characterized by the infant reaching out and the mother responding or vice versa, or by an uncoordinated, mismatched state. Windowed cross-correlation addresses this by varying the lag between correlated signals. In this work, we address changes in leader-follower dynamics by fitting two types of models: one DMDc model describes the continuous evolution of the mother's behavioral signals and integrates the infant's behavior as the control signals, while the other models the infant's behavior over time and uses the mother's behavior as the control.

3.2.3 Dynamic Mode Decomposition with control. To better understand the dynamics and relative influence that the partners have during the still-face interaction, our work leverages Dynamic Mode Decomposition with control (DMDc) which assumes the dynamical system model

$$c_{t+1} = Ac_t + Bu_t \tag{1}$$

where c_t and c_{t+1} represent the observations of the mother's behavior at times t and $t+1$ respectively, while u_t represents the child's signal. The relationship described in Eq 1 represents the underlying assumption that the observed signal for a mother is the combination of the signals from mother and infant from the previous time steps. DMDc is an algorithm for estimating the transition matrix, A, and controller matrix, B. This is accomplished by recognizing that given an observational window $C_t = [c_t, c_{t-1}, \ldots, c_{t-w}]$ and $U_t = [u_t, u_{t-1}, \ldots, u_{t-w}]$ then it holds that:

$$C_{t+1} = AC_t + BU_t = [A\,B] \begin{bmatrix} C_t \\ U_t \end{bmatrix} \tag{2}$$

From this form it is possible to solve for A and B:

$$[A\,B] = C_{t+1} \begin{bmatrix} C_t \\ U_t \end{bmatrix}^{\dagger} \tag{3}$$

where † represents the Moore-Penrose pseudo-inverse.

Since the mother's signal c always has the same dimensionality, A will always be a square matrix, implying that it is possible to study the dynamic response (or mode) of the mother's behavior by looking at the dominant eigenvalues of the transition matrix. Given that this model expresses a discrete time dynamical system, the eigenvalues correspond with the frequency responses; the eigenvalue with the largest complex magnitude represents the "dominant" dynamics. Accordingly, the magnitude of the dominant eigenvalue expresses an exponential decay while the angle to the real axis reflects an oscillatory response. To distinguish between models, we denote the eigenvalues as $\lambda_{A,I}$ or $\lambda_{A,M}$. The subscript I or M indicates the control input to the model as the behavioral signals of the infant or mother, respectively.

Furthermore, A and B give us an estimate of the relative influence each component of the previous time step has on the future. Therefore, we introduce the measure *relative influence* (R):

$$R = \frac{||B||_F}{||A||_F} \tag{4}$$

While the concept of responsiveness is multifaceted, we use relative influence R as a way to measure a component of an infant's or mother's responsiveness to their interactive partner during a given window. This measure approximates proportionally how much an observed behavior is driven by a response relative to self-driven; or how much an individual's response is influenced by their partner's behavior relative to their own behavior. To distinguish between infant-controlled models and mother-controlled models, we denote the relative influence of the infant on the mother as R_I, and the relative influence of the mother on the infant as R_M.

3.3 Analysis

A successful model of interaction dynamics must be able to recognize changes in behavior and dyadic processes, such as those caused by the stressful still-face stage or communication patterns that change as the infant develops. Therefore, we evaluated how the metrics detailed in Section 3.2.3 evolved across stages of the FFSF protocol and across infant age. The mean values of these metrics were aggregated across 3-second windows occurring in the same experimental stage. Identified trends were compared across models fit on each modality or combination of modalities to assess the ability of DMDc to evaluate multimodal behavioral coordination.

Given the larger size of D_{video} compared to D_{audio}, we conducted our analyses in two stages. First, we evaluated our approach on the dataset of head pose and arm pose features. Analyzing our approach on this larger dataset allowed for a robust analysis of DMDc's ability to incorporate multiple behavioral signals. Next, we repeated our analysis using the interactions and FFSF stages with available F0 data. A separate analysis of coordination metrics calculated using the mothers' F0 data enabled us to directly evaluate the model's ability to incorporate different sets of behavioral signals for each partner in a dyad, and to address the challenges associated with less frequent behaviors (i.e., vocalizations).

3.3.1 Model Validation. We first evaluated our model's ability to identify known trends inherent to the FFSF protocol. As mothers were instructed not to respond to their infants during the still-face stage, it follows that the relative influence of an infant on the mother should be lower during the still-face stage. Therefore, we anticipated a decrease in R_I from the play to still-face stages. This effect was tested with two-tailed Student's t-tests for each set of models. Since mothers were instructed not to interact (or vocalize) during the still-face stage, audio data were excluded from this portion of the analysis.

Representative models must capture the lead-follow or "serve-and-return" pattern of infant-mother interactions. Given the finite length of each FFSF stage, there is a trade-off between the time spent leading vs. following; if a larger part of a stage is characterized by

the infant responding to the mother's cues, less time remains for the mother to respond to her infant. Therefore, we anticipated that stages with higher mother-to-infant influence would show lower infant-to-mother influence. To test this hypothesis, we evaluated the Pearson correlation coefficient between R_I and R_M. Given instructions for mothers to be unresponsive during the still-face stage, we conducted this analysis for the play and reunion stages.

3.3.2 Exploring Trends in Interaction Dynamics. Beyond monitoring known behavior patterns, a goal of modeling infant-mother coordination is to explore trends that emerge with changes in interaction quality or as part of infant development. For example, studies have conducted the FFSF procedure with varying levels of maternal unresponsiveness, sometimes allowing different forms of touch between mother and infant during the still-face stage [1]. We explored the effect of mothers' changing behavior across FFSF stages by evaluating the relationship between the decreased influence of infants on their mothers from play to still-face stages, or decreased responsiveness from mothers, and infant behavior during the still-face stage. We measured the difference in infants' influence on their mothers across stages as R_I^{play}-$R_I^{still-face}$, capturing the difference in the relative influence metrics from play to still-face stages for models that used the infants' behavioral signals as control inputs. To explore how this value related to infant's behaviors during the still-face stage, we monitored $\lambda_{A,M}^{still-face}$, the dominant eigenvalue of the infant's transition matrix. This allowed for a direct comparison between changes in a mother's feedback across stages to the dynamics of infant behavior during the stressful still-face stage. As mothers did not vocalize during the still-face stage, only head pose and arm pose were evaluated for this analysis.

Next, we analyzed how metrics extracted by each model evolved with infant age. Motor and social skills develop with age, impacting infants' interactive capabilities and in turn influencing their mothers' responses. To evaluate trends across infant age, we conducted a linear mixed models (LMM) analysis of both individual parameters (λ_A) and dyadic parameters (R), with the id of each dyad as a random effect. LMM coefficients were evaluated to assess the direction and strength of relationships between each metric and age. Given the differences in scale (for example, infant age ranges between 2 and 18 months while eigenvalues remain close to the unit circle), values were normalized between 0 and 1 prior to statistical testing.

4 RESULTS AND DISCUSSION

We first report results with models evaluated on D_{video} to assess the ability of our approach to integrate multiple behavioral signals. These results are reported and discussed in Sections 4.1 and 4.2. Next, results from analysis including audio data, evaluated on the subset of interactions included in both D_{video} and D_{audio}, are reported in Section 4.3. We explore relationships between infant and mother behavior, trends across experimental stage, and trends across age; therefore, we consider statistical significance at $\alpha < .017$ using the Bonferroni correction for multiple comparisons.

4.1 Model Validation

4.1.1 Observing the Still-Face Instructions. Consistent with the instructions of the FFSF procedure, results showed a decrease in mothers' measured responsiveness to infants' behavior during the still-face stage. As shown in Table 1, t-tests demonstrated a significant decrease in R_I from the play to still-face stages, representing a decreased influence of infants' behavior on mothers' behavior, or decreased responsiveness of mothers to their infants. This result was strongest when measuring relative influence using both head pose and arm pose, indicating that changes in responsiveness may be best observed by integrating multiple behaviors. As mothers were instructed not to interact with their infants during the still-face stage, they did not vocalize and therefore did not produce F0 data; consequently, audio data were excluded from this analysis.

Table 1: t-statistic between R_I^{play} and $R_I^{still-face}$

Modalities	t
Head Pose	3.838**
Arm Pose	3.069*
Head Pose & Arm Pose	4.374**

*p<.017 **p<.001

4.1.2 Leading-Following Relationship. Results indicated a significant negative correlation between R_I and R_M, shown in Table 2. This finding is consistent with anticipated interaction dynamics; if a mother leads most of the interaction, there remains less time for the infant to lead. For a given stage, this result is strongest for models trained on head and arm pose, demonstrating the potential of DMDc to leverage multimodal signals in tracking the leading-following relationship inherent to infant-mother interaction.

Table 2: Correlations between R_I and R_M

Modalities	Stage	R
Head Pose	Play	-.402 **
Arm Pose	Play	-.348 **
Head Pose & Arm Pose	Play	-.430 **
Head Pose	Reunion	-.285 **
Arm Pose	Reunion	-.347 **
Head Pose & Arm Pose	Reunion	-.375 **

*p<.017 **p<.001

4.2 Trends in Interaction Dynamics

4.2.1 Infant Responses to the Still-Face Stage. Comparing metrics between play and still-face stages, results showed an inverse correlation between R_I^{play}-$R_I^{still-face}$ and the magnitude of $\lambda_{A,M}$. These results are shown in Table 3. Describing the autonomous evolution of infant motor behavior across frames, smaller eigenvalues represent a more dampened dynamic system, or an infant's faster return to baseline behavior after a change in pose. This indicates that a more drastic change across stages in a mother's responsiveness, or in an infant's ability to influence the mother's behaviors, was followed by more sporadic infant behavior during the still-face stage. This may have reflected how mothers' decreased responsiveness influenced infants' emotion regulation behaviors. These results are

shown in Table 3. The inverse correlation was strongest for models trained on both head pose and arm pose data; this result supports the potential of multimodal metrics of behavioral coordination for evaluating changes in interaction dynamics.

Table 3: Correlation between (R_I^{play} - R_I^{sf}) and $\lambda_{A,I}^{sf}$

Modalities	t
Head Pose	-.183*
Arm Pose	-.225*
Head Pose & Arm Pose	-.266**

*p<.017 **p<.001

4.2.2 Trends Across Infant Age. The linear mixed models analysis (LMM) demonstrated changes in infants' transition dynamics, and in infants' influence on their mothers' behavior, across age. The magnitudes of eigenvalues $\lambda_{A,M}$ increased with age for all models, as shown in Table 4. Increasing values of $\lambda_{A,M}$ indicate that older infants produced more stable and less sporadic movements, perhaps due to improved motor abilities. Similar results for unimodal and multimodal models indicate that not only did individual movements become more stable with age, but behaviors that involved multiple modalities (e.g., hand-eye coordination) became more stable. However, metrics from multimodal models also had larger standard errors, indicating more variance between infants. Significant trends were not found in the angle of of $\lambda_{A,M}$ to the real axis, or in the oscillatory response of infant behavior. As observed in Figure 3, variance in $\lambda_{A,M}$ between infants decreased with age, indicating more similar behaviors among older infants. This pattern was consistent across all FFSF stages. Significant trends were not found in the mothers' transition dynamics. This is unsurprising, as mothers are likely no longer developing motor skills.

Table 4: Trends in $\lambda_{A,M}$ across age

Stage	Modalities	Coefficient	SE
Play	Head Pose	.043*	.016
Play	Arm Pose	.156**	.027
Play	Head Pose & Arm Pose	.133**	.025
Still-Face	Head Pose	.069	.031
Still-Face	Arm Pose	.187**	.038
Still-Face	Head Pose & Arm Pose	.142**	.035
Reunion	Head Pose	.093**	.026
Reunion	Arm Pose	.285**	.044
Reunion	Head Pose & Arm Pose	.155**	.025

*p<.017 **p<.001

Meanwhile, infants' influence on their mothers' behaviors, measured by R_I, decreased with age. Given that only two modalities are considered in this analysis, this effect may reflect different uses of individual modalities; mothers may have become less likely to respond to their infants using consistent motor movements, but remained attentive overall. It is also possible that the decrease in R_I reflects changes in the timing of responses. While this work

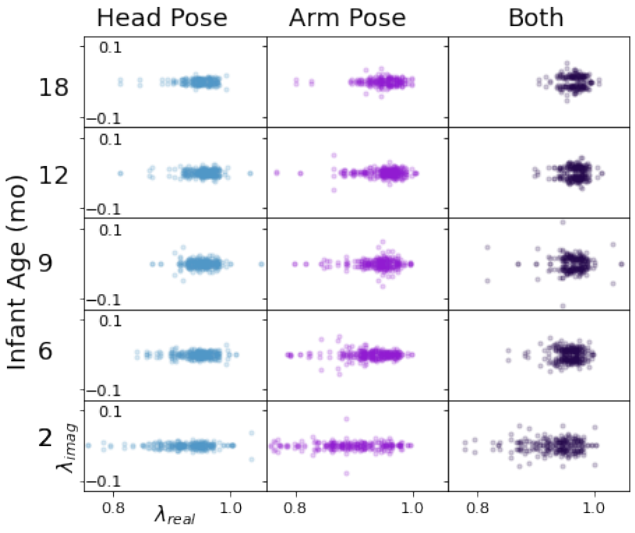

Figure 3: ($\lambda_{A,M}$) across infant age, calculated using head pose data (left), arm pose data (center), and both (right). Results include $\lambda_{A,M}$ values from each of the three stages (play, still-face, reunion) of the FFSF protocol.

evaluates frame-to-frame observations within 3-second windows, future work will explore additional time scales. Results of the LMM analysis are shown in Table 5. Across modalities for a given metric, effect sizes are smaller during the reunion stage. This may reflect between-dyad differences in interaction repair strategies or infant emotion regulation behaviors demonstrated while recovering from the still-face stage; between-subject variance in these reunion-stage behaviors may inhibit the ability to observe trends that emerge with age. Significant effects were not found for LMMs evaluated on R_M, or the infants' responsiveness to their mothers. While mothers were instructed to interact with their infants and typically remained engaged over time, infants often became distracted by toys or the environment; therefore, variance in infant behavior may have masked changes in moment-to-moment responsiveness. Evaluating multiple time scales may address this challenge.

Table 5: Trends in R_I Across Infant Age

Stage	Modalities	Coefficient	SE
Play	Head Pose	-.132**	.032
Play	Arm Pose	-.164**	.038
Play	Head Pose & Arm Pose	-.130**	.033
Reunion	Head Pose	-.064*	.019
Reunion	Arm Pose	-.057	.029
Reunion	Head Pose & Arm Pose	-.061*	.022

*p<.017 **p<.001

4.3 Incorporating Audio Data

Consistent with the results of Section 4.1.2, negative correlations were found between R_I and R_M for all combinations of modalities.

These results are shown in Table 6. For a given set of pose signals, effect sizes became smaller when F0 was included as an input. This is likely due to the sparsity of F0 data; while pose can be monitored continuously, F0 values can only be collected during vocalizations. During the play and reunion stages, mothers did not speak continuously; rather, there were often periods of silence. As a result, an average of 18 3-second windows included F0 data, and supported the fit of a dynamical model. Given the richness of mother-infant communication, exchanges that occur in just a few windows may not sufficiently reflect interaction dynamics; rather, our results suggest that longer or additional interaction recordings may be needed to evaluate coordination using observations of sparse behaviors.

Table 6: Correlations between R_I and R_M

Modalities	Stage	R
Head Pose	Play	-.436 **
Arm Pose	Play	-.378 *
Head & Arm Pose	Play	-.434 **
Head Pose & F0	Play	-.427 **
Arm Pose & F0	Play	-.345 *
Head & Arm Pose & F0	Play	-.432 **
Head Pose	Reunion	-.315*
Arm Pose	Reunion	-.304*
Head & Arm Pose	Reunion	-.342 *
Head Pose & F0	Reunion	-.119
Arm Pose & F0	Reunion	-.233
Head & Arm Pose & F0	Reunion	-.241

*p<.017 **p<.001

Similar to models evaluated on D_{video}, results showed a negative trend in R_I values, as shown in Table 7. Consistent with results reported in Section 4.2, effect sizes were smaller during the reunion stage and when including F0 data. However, similar trends for models evaluated with F0 data show the ability of the DMDc approach to leverage different behaviors for each partner. This is necessary when partners differ in their communication modalities, such as interactions between verbal and non-verbal individuals.

5 CONCLUSION

This paper presents DMDc as a method for evaluating multimodal behavioral coordination, specifically during infant-mother interaction. Significant changes in model output from the play to still-face stages of the FFSF protocol validated the ability of our approach to capture known trends in a developmentally relevant interaction paradigm. Stronger effect sizes across FFSF stages for models evaluated on multimodal rather than unimodal data supported the value of DMDc as a method for integrating multiple behavioral signals to construct a cohesive measure of behavioral coordination. The ability of this approach to integrate separate sets of behavioral signals for each participant makes DMDc appropriate for evaluating early communication, as infants are still developing many of the communication modalities available to adults.

This paper also demonstrates how DMDc may be used to explore developmental phenomena through multimodal computational methods. Results indicated a relationship between changes from play to still-face in observations of the infants' influence on their mothers and the dynamics of infant behaviors during still-face. Again, larger effect sizes for multimodal data analysis show the value of our approach. Transition dynamics in infant behavior, and the observed influence of infants' behaviors on their mothers' behaviors, both showed significant changes across age, showing the ability of DMDc parameters to capture developmentally relevant changes in interaction dynamics. This research can serve as an argument for the value of multimodal metrics of dyadic coordination, and the ability of DMDc to provide such metrics.

Table 7: Trends in R_I across Infant Age

Stage	Modalities	Coefficient	SE
Play	Head Pose	-.272*	.085
Play	Arm Pose	-.316*	.099
Play	Head & Arm Pose	-.254*	.083
Play	Head Pose & F0	-.170*	.062
Play	Arm Pose & F0	-.227*	.077
Play	Head & Arm Pose & F0	-.182*	.069
Reunion	Head Pose	-.140	.059
Reunion	Arm Pose	-.183*	.073
Reunion	Head & Arm Pose	-.149*	.06
Reunion	Head Pose & F0	-.105	.052
Reunion	Arm Pose & F0	-.063	.033
Reunion	Head & Arm Pose & F0	-.105	.053

*p<.017 **p<.001

6 FUTURE WORK

This work serves as a proof-of-concept for an approach to modeling multimodal dyadic coordination. The modalities used in this research are not extensive, and our future work will also incorporate physiological signals and facial expressions. Additionally, we will explore how varying windowing parameters can enable our approach to capture patterns that occur over longer or shorter periods of time.

Our future research will also investigate how the model output from each window of time relates to moment-to-moment changes in the interaction. Correlational analysis between automated metrics and expert-annotated micro-coded time series discussed in Section 2.1.2 would provide additional understanding as to the specific dyadic processes that are captured by this model. Elucidating the relationship between automatically extracted coordination metrics and expert-annotated labels will also inform the feasibility of automating aspects of the annotation process to enable significantly scaling up analysis of infant-mother interactions.

ACKNOWLEDGMENTS

This research was supported 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.

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] Victor Ardulov, Madelyn Mendlen, Manoj Kumar, Neha Anand, Shanna Williams, Thomas Lyon, and Shrikanth Narayanan. 2018. Multimodal interaction modeling of child forensic interviewing. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 179–185.

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

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 172–186.

[5] Jeffrey F Cohn and Edward Z Tronick. 1987. Mother–infant face-to-face interaction: The sequence of dyadic states at 3, 6, and 9 months. *Developmental psychology* 23, 1 (1987), 68.

[6] R Feldman. 1998. Coding interactive behavior (CIB). *Unpublished manual, Department of Psychology, Bar-Ilan University, Ramat-Gan, Israel* (1998).

[7] Ruth Feldman. 2003. Infant–mother and infant–father synchrony: The coregulation of positive arousal. *Infant Mental Health Journal: Official Publication of The World Association for Infant Mental Health* 24, 1 (2003), 1–23.

[8] Andrew Gianino and Edward Z Tronick. 1988. The mutual regulation model: The infant's self and interactive regulation and coping and defensive capacities. (1988).

[9] 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.

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

[11] 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*. 287–295.

[12] 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), e113571.

[13] Chloe Leclere, Marie Avril, S Viaux-Savelon, N Bodeau, Catherine Achard, Sylvain Missonnier, Miri 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.

[14] 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.

[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] Joshua L Proctor, Steven L Brunton, and J Nathan Kutz. 2016. Dynamic mode decomposition with control. *SIAM Journal on Applied Dynamical Systems* 15, 1 (2016), 142–161.

[17] 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.

[18] 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.

[19] Cynthia A Stifter and Michael Rovine. 2015. Modeling dyadic processes using hidden Markov models: A time series approach to mother–infant interactions during infant immunization. *Infant and child development* 24, 3 (2015), 298–321.

[20] 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.

[21] Edward Z Tronick and Andrew Gianino. 1986. Interactive mismatch and repair: challenges to the coping infant. *Zero to Three* (1986).

[22] Omri Weisman, Mohamed Chetouani, Catherine Saint-Georges, Nadege Bourvis, Emilie Delaherche, Orna Zagoory-Sharon, David Cohen, and Ruth Feldman. 2016. Dynamics of Non-Verbal Vocalizations and Hormones during Father-Infant Interaction. *IEEE Trans. Affect. Comput.* 7, 4 (2016), 337–345.

[23] Hao Zhang, Clarence W Rowley, Eric A Deem, and Louis N Cattafesta. 2019. Online dynamic mode decomposition for time-varying systems. *SIAM Journal on Applied Dynamical Systems* 18, 3 (2019), 1586–1609.