2018 13th IEEE International Conference on Automatic Face & Gesture Recognition

Generative Multimodal Models of Nonverbal Synchrony in Close Relationships

Joseph F. Grafsgaard, Nicholas D. Duran*, Ashley K. Randall*, Chun Tao*, Sidney D'Mello
University of Colorado Boulder, Boulder, CO, USA
*Arizona State University, Tempe, AZ, USA
Joseph.Grafsgaard@colorado.edu, {nduran4, Ashley.K.Randall, ctao5}@asu.edu, Sidney.DMello@colorado.edu

***Abstract*— Positive interpersonal relationships require shared understanding along with a sense of rapport. A key facet of rapport is mirroring and convergence of facial expression and body language, known as nonverbal synchrony. We examined nonverbal synchrony in a study of 29 heterosexual romantic couples, in which audio, video, and bracelet accelerometer were recorded during three conversations. We extracted facial expression, body movement, and acoustic-prosodic features to train neural network models that predicted the nonverbal behaviors of one partner from those of the other. Recurrent models (LSTMs) outperformed feed-forward neural networks and other chance baselines. The models learned behaviors encompassing facial responses, speech-related facial movements, and head movement. However, they did not capture fleeting or periodic behaviors, such as nodding, head turning, and hand gestures. Notably, a preliminary analysis of clinical measures showed greater association with our model outputs than correlation of raw signals. We discuss potential uses of these generative models as a research tool to complement current analytical methods along with real-world applications (e.g., as a tool in therapy).**

Keywords-close relationships; couples therapy; facial expression; LSTM; neural networks; nonverbal synchrony

I. INTRODUCTION

Close relationships require shared understanding and rapport to achieve positive outcomes. A key aspect of rapport is interpersonal synchrony, in which interlocutors share behaviors, mannerisms, and common ground, which aids social bonding [1], [2]. Interpersonal communication operates over channels of speech and nonverbal behavior, with the spoken channel conveying content and meaning as well as prosodic qualities (i.e., *how* you've expressed yourself) [3]. Nonverbal behavior provides insight into momentary reactions, cognitive states, and feelings. We can categorize outward manifestations of nonverbal behavior into main channels of facial expressions [4], [5], body movements [6], [7], and gestures [2], [8]. Here, we focus on nonverbal synchrony, which encompasses the moment-by-moment changes in nonverbal behavior associated with interpersonal synchrony.

Nonverbal synchrony is intuitively expressed and experienced in our daily lives. However, these outward behavioral manifestations are overlaid on a complex substrate of internal states (cognitive & affective), knowledge and beliefs, and communicative goals. Modern theories of emotion acknowledge the complex nature of momentary affective responses, with many component processes recruited to appraise a situation and arrive at a response spanning multiple levels, such as neurobiological, physiological, behavioral, cognitive, and meta-cognitive [9], [10].

This inherent complexity of nonverbal synchrony belies the need to leverage multiple data channels. Traditional analytical approaches provide a lens to focus on a particular channel of nonverbal synchrony. For instance, we may examine how head [7] or body [11] movement change throughout a conversation. A particular pair of interlocutors may experience greater or lesser synchrony from moment to moment. However, focusing on one channel at a time or one dyad at a time (and averaging across multiple dyads) loses the larger picture. Analyses of nonverbal synchrony stand to benefit from multiple data channels (multimodality) jointly analyzed across numerous interlocutors (generalizability).

One way to address the dual concerns of multimodality and generalizability is to turn to data-driven, model-based approaches. Generative models can be used to learn patterns of nonverbal behavior that drive cross-modal synchrony, predicting behavior of an interlocutor at any given point of the conversation. By incorporating data across numerous interactions, generalized patterns of nonverbal synchrony that extend beyond any one dyad may be inferred. Inspecting how and when behaviors are produced should provide insight into proximal phenomena, such as moment-to-moment cognitive and affective processes or interpersonal dynamics. In other words, we may construct models of nonverbal synchrony across diverse settings and social relationships to identify how and why interpersonal communication generalizes or diverges.

Generative models can provide novel insight into nonverbal synchrony. Rather than describing nonverbal synchrony (as in analytical approaches) or predicting discrete categories of behavior (as in classifier models), generative models output the behaviors themselves, which allows us to query the model to assess predicted behavior. This capability has obvious applied implications (e.g., to drive virtual agent expression) in addition to offering insights on how nonverbal synchrony develops, fosters, and contextualizes social interaction.

A. Related Work

A survey of prior approaches to measure nonverbal synchrony is beyond the scope of this paper. For a concise overview of prior techniques, we refer the reader to a recent article by Delaherche and colleagues [1]. However, we will contrast the present generative model-based approach to prior analytical and model-based techniques.

Prior analytical approaches to measure nonverbal synchrony have included techniques to measure synchrony at fixed lags (or leads), visualize synchrony dynamics, and adopt dynamical systems modeling approaches (e.g., cross-recurrence quantification analyses, coupled nonlinear oscillators). In the case of lagged correlations, it is possible

978-1-5386-2335-0/18/$31.00 ©2018 IEEE
DOI 10.1109/FG.2018.00037

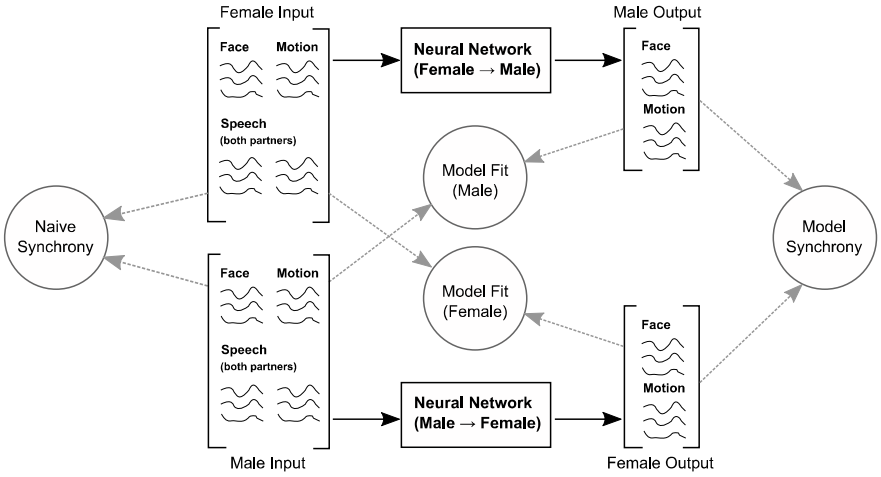

Figure 1. Model-based approach to nonverbal synchrony

to identify how a discrete behavior follows another (e.g., head movements of dyads) [7], [12]. Visualizations of synchrony may provide labeled co-occurrence of behavior across a conversation, showing how synchrony occurs at different lags across a conversation [2], [11]. Finally, dynamical systems approaches acknowledge the complexity of nonverbal behavior, quantifying how deterministic or random patterns of behavior are [13], [14]. However, these analytical techniques have usually been limited to specific modalities (vs. combined modeling of face, head, and body movement in the current work) and/or behaviors of individual dyads (vs. cross-validation across dyads as in the present approach).

Prior model-based classification approaches have focused on specific phenomena associated with nonverbal synchrony, such as discrete behaviors (e.g., turn-taking, backchannels) [15], [16] and inferring interaction states (e.g., clusters of behavior) [17], [18]. Models that predict dialogic phenomena may benefit applications such as intelligent virtual agents, robotics, and natural language interaction [15], [16]. Models of interaction states have revealed patterns of discrete synchrony behaviors across varied contexts [17], [18]. In contrast to these classification approaches that focused on discrete behaviors, we seek to produce continuous, generative models to identify and measure generalizable, multimodal patterns of nonverbal synchrony.

In very recent work, a generative model of nonverbal synchrony was created to produce facial expressions. Feng and colleagues trained feed-forward neural networks using over two hundred Skype conversations posted to YouTube [4]. The researchers focused on modeling facial keypoints (i.e., landmarks) to drive the behaviors of facial expression avatars. Their model was found to produce believable facial behaviors when presented to human raters. Whereas their model focused on expression synthesis, our present approach seeks to generatively model continuous nonverbal behaviors in order to measure and understand patterns of synchrony. In turn, our model-based measures of synchrony may be applied in clinical practices.

B. Contribution & Novelty

We constructed recurrent neural network models of nonverbal synchrony from facial expression, body movement, and speech in conversations of romantic couples (Figure 1). These models offer insight into patterns of synchrony across multiple modalities of nonverbal behavior and generalized across romantic dyads. Our analyses show that the model predictions perform significantly above chance baselines and simpler feed-forward neural networks. Our models learned behaviors encompassing facial responses, speech-related facial movements, and head movements.

To our knowledge, ours is the first attempt to produce data-driven, multimodal, generalizable models of nonverbal synchrony. As generative models of nonverbal synchrony account for the multiple modalities of interpersonal communication at ever increasing degrees of accuracy and realism, they may both offer generalizable insights into human behavior and drive humanlike interactive systems.

II. Method

A. Participants

A total of 54 heterosexual couples (108 individuals) from the Southwestern region of the United States were recruited through advertisements posted on professional and university listservs, Craigslist, and Facebook. Inclusion criteria required individuals to be at least 18 years of age, the couples to be romantically involved for at least three months, and both partners willing to participate. Compensation for participation was set at $35 for each partner. Of those initially recruited, 10 couples did not complete all portions of the study, and 15 couples had incomplete data for one or more of the recorded data channels (e.g., video, speech, or physiology). For the final set of 29 couples (N = 58 individuals), men's average age was 29.5 years (SD = 5.67), and women's average age was 30.0 years (SD = 6.33). The couples had been together for an average of 4.89 years (SD = 3.51); 15 were married and 9 had children. The majority of participants identified as White (N = 40), followed by Hispanic (N = 12), Asian American (N = 2), African American (N = 1), and three identified as other ethnicities. The majority of participants also reported completing an undergraduate or post-graduate degree (N = 20 men, 27 women).

B. Study Protocol

The study was divided into two phases. For the initial phase, each participant was sent an electronic survey that contained questions about basic demographics and relationship-focused issues. The survey was to be completed at home and couples were asked to avoid discussing answers with each other. The second phase required the couples to visit a local university campus, where they would engage in a sequence of six-minute conversations. During these conversations, their video, audio, movement, and bracelet accelerometer were recorded (see Figure 2). Video was recorded with high definition camcorders (30 frames per second at 1920x1080 resolution), audio via lapel mics, and accelerometer with Empatica E3 bracelets.

The three conversations focused on: 1) a source of stress specific to one partner outside the relationship (external stressor), 2) mutual source of stress within the relationship (internal stressor), and 3) a mutual topic of enjoyment. The most prominent external stressors were work, finances, and school. The decision of whose external stressor to select (male or female) was counterbalanced based on the order in which they were recruited. The most prominent internal stressors were disturbing habits, difference of opinion, and insufficient behavior. The most popular enjoyment topics were good past experiences, children, and family pet. As in prior research, the order of topics was fixed to ensure consistent experiences across couples, ending with a positive conversation [19].

The couples also self-reported on multiple clinical measures. These included relationship duration (taken before the conversations) and multi-item post-conversation rating scales which encompassed aspects of rapport, conversational style, and affective outcomes focusing on oneself and one's partner (TABLE I).

III. Model Design and Training

A. Feature Processing

Facial expressions were extracted from the video recordings (separate videos of female and male partners) using Emotient SDK, a commercial computer vision tool that identifies specific facial muscle movements. These facial movements correspond to a subset of those described in the Facial Action Coding System [21], which enumerates all possible facial muscle movements of the human face as facial action units (AUs). Additionally, the Emotient SDK provides information on head pitch (nodding), yaw (shaking side-to-side), and roll (tilting to the side). The video recordings were also processed for gross body movements by extracting binary frame to frame pixel differences, which were averaged to obtain a percentage of change per frame.

We used the openSMILE toolkit [22] (version 2.3.0 with eGeMAPSv01a.conf) to measure fundamental characteristics of speech in the separate audio channels of female and male partners. These included fundamental frequency, loudness, formant frequencies F1-F3, jitter and shimmer. Using this minimal set of speech features, we accounted for the acoustic-prosodic features of speech without representing the spoken content.

Physiological recordings in natural interactions can be problematic as interlocutors make gestures and move

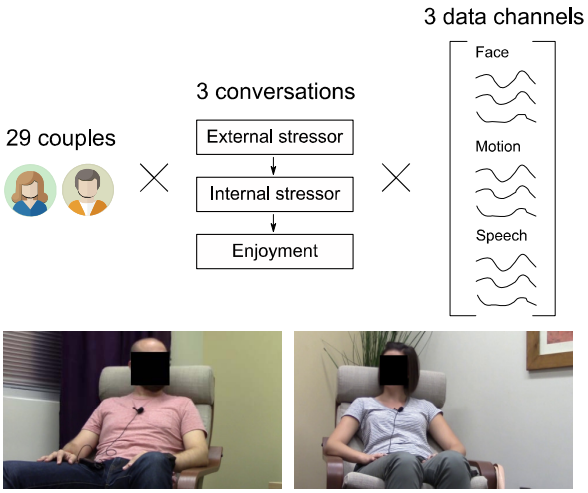

Figure 2. Overview and example of study recordings

TABLE I. Relationship Variables and Ratings

When taken	Self-report	Description
Pre-conversation	Relationship duration	Single item. Relationship duration in months.
Pre-conv.	Rel. quality	Multi-item survey [20]
Post-conversation	Partner-induced affect	Multi-item survey. Specific positive and negative emotions caused by partner.
Post-conversation	Conversation style	Multi-item survey. Ratings of partner's perceived conversation style (e.g., warmth, in-depth responses).
Post-conversation	General rapport	Multi-item survey. Ratings of closeness, understanding, and self-agency in the conversation.
Post-conversation	Adjective pairs	Multi-item survey. Dichotomous adjective pairs rating partner (e.g., dishonesty vs. honesty, coldness vs. warmth).

during conversations. Although the Empatica E3 enabled recording of electrodermal activity and heart rate, we observed significant physiological signal disruptions corresponding to these movements, so we did not use the physiological recordings in the present analyses. However, the three-axis accelerometer in the E3 physiological bracelet provided information regarding the couples' wrist and arm movements. We used the accelerometer to measure gross wrist and arm movement (i.e., gestural motion), computed as the Euclidean distance of accelerometer values (x,y,z) from second to second.

The motion features of head movement, gross body movement, and gestural motion are relative measures that depend on where a partner was seated, how much of the video frame a partner occupied, and how much a partner gestured during a conversation. We applied an individual z-score per partner per conversation to bring them to standard units that accounted for idiosyncratic variations in body movement.

Because the face, speech, and motion features were recorded at different time scales, we averaged the values of each feature at each second of conversation. If a feature had no data across an entire second (i.e., no average could be calculated), it was treated as a missing value. Only facial expression and head movement had missing data, though most video frames were tracked (M = 87% frames tracked, SD = 16%). We removed the first 10 seconds of each conversation to ignore the first moments as the discussion was being initiated.

B. Neural Network Modeling

We constructed neural networks to model nonverbal synchrony using the Keras toolkit with TensorFlow [23]. Our models were designed to use speech (both partners), facial expression (one partner), and motion features (one partner) as input in order to predict facial expression and motion of the other partner. By including both partners' speech as input, the design accounts for variations in nonverbal behavior due to speech production and turn taking dynamics (e.g., mouth movements occur during speech and facial expressions often accentuate utterances).

We used two model structures: feed-forward and recurrent. The feed-forward neural network (FFNN) transforms the input through a single fully-connected activation layer (i.e., dense hidden layer). The FFNN model is trained on individual one-second inputs, with no notion of history across these data instances. Our recurrent neural network (RNN) contained a single long short-term memory (LSTM) activation layer [24]. The LSTM layer implements mechanisms of forgetting and retaining information across input sequences. The RNN model was trained on sequences of 10 one-second inputs (i.e., the prior nine seconds and the current second), with the notion that momentary responses associated with nonverbal synchrony occur within a time period of seconds [1]. Our models used rectified linear units (ReLU) as activation function, which enabled computationally efficient training.

The three conversation types were independently modeled so as to minimize cross-over and other confounding effects. We trained our models via 10-fold couple-independent cross-validation (using scikit-learn [25]). Within each fold, data from 60% of couples were used as the training set, 30% as the validation set, and 10% as the test set. To ensure that no individual feature dominated the model, we z-scored each feature per conversation to bring all features to standard units. We further normalized the values to within a [-3, 3] range. Both z-scoring and normalization were performed using only the training data. That is, descriptive statistics (mean, standard deviation, max, min) needed for the transformations were computed from the training data and then applied to validation and testing sets. This prevents 'peeking' into the validation or testing sets which would result in overly optimistic model performance. Finally, we replaced missing values with zeroes – this enabled multimodal machine learning across all data points.

Neural networks are trained via model updates through successive complete passes on the training data (known as an epoch). At each epoch, mean squared error was used to compute the loss of the model on both the training set (training loss) and validation set (validation loss). We trained until 20 epochs passed without improvement on validation loss – there was otherwise no upper limit on epochs. We explored different numbers of neurons at the activation layer (8, 32, and 128) and found 32 neurons achieved sufficient fit with diminishing returns at greater complexity (based on training and validation loss).

The weight updates of the neurons were guided by an adaptive learning rate algorithm: Adam [26] with Nesterov momentum (Nadam) [27]. Adaptive learning rate algorithms change the magnitude of training updates on the fly, which removes the burden of fine-tuning learning rates. Finally, we used dropout (a recent regularization technique that removes neurons from a layer of the neural network at random) to prevent overfitting of our models [28]. We used dropout at both the input and activation layer with identical settings. We evaluated the effects of dropout on training and validation loss at 0% (no dropout), 20%, 50%, 80%, and 95%. Ideally, training and validation loss should converge across training epochs – this was achieved in our data at 80% dropout.

Figure 3 illustrates input data and model predictions with facial expression and body movement from the conversation on internal stress.

C. Fit and Synchrony Measures

Our primary metric to measure model performance was bivariate correlation (i.e., Pearson product-moment correlation coefficient). This metric measures covariance in time series data (i.e., whether our model predictions capture patterns of change in the data). We computed correlations where the raw data were not missing (as opposed to the model predictions, which were never missing). For instance, if the male partner had missing facial expression data during a 15-second time period, these 15 seconds were removed when computing the correlation for facial expression data. This ensures that the results were not biased by missing values, while making use of all available data during training. Correlations were performed using the pandas library in Python [29] and statistical comparisons were conducted in R [30].

We refer to correlations of model predictions with target partner's data as *model fit*. For example, we correlated the model predictions of female partners' smiling (Lip Corner Puller – AU12) with the actual sensor data of female partners across one-second intervals. We describe correlations of female/male (e.g., male AU12 with female AU12) raw data as *naïve synchrony* and correlations of female/male model predictions as *model synchrony*. We calculated all three measures for each feature, couple, and conversation.

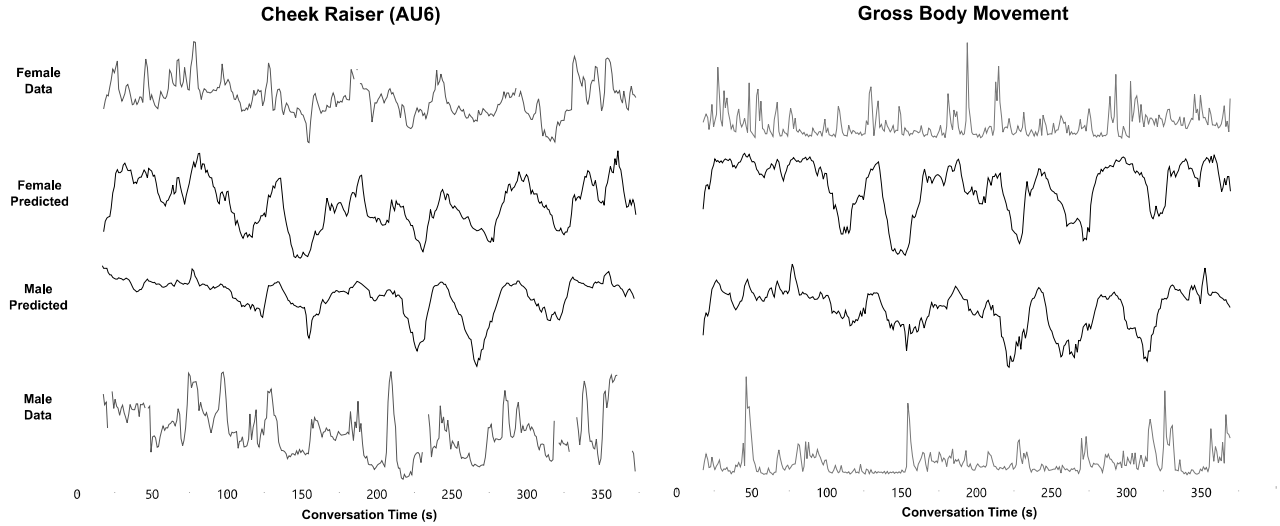

Figure 3. Example of data and model predictions from the internal stress conversation.

IV. RESULTS

RQ1. How accurate are our generative models of nonverbal behaviors and do results vary by gender?

Our first analysis examined whether the recurrent neural network performed better at predicting nonverbal behavior than the feed-forward neural network. For this analysis, we averaged across gender (M→F, F→M), features, and conversations, to obtain one model fit value per couple for each network type. A two-tailed (for this and all subsequent analyses) paired-samples *t*-test indicated that the recurrent (M = .09, SD = .03) significantly outperformed the feed-forward net (M = .06, SD = .02), $t(28) = 6.45$, $p < .0001$. Hence, further analyses focused exclusively on the recurrent models.

Prior research on interpersonal synchrony suggests that females drive interaction in dyadic conversations [6], [31]. Accordingly, a paired *t*-test indicated that models predicting the female partner's data (M = .11, SD = .04) had significantly higher model fit than models predicting the male's behavior (M = .07, SD = .04), $t(28) = 5.00$, $p <$ .0001.

Whereas the previous analyses averaged over features, the third analysis of RQ1 compared model fit to chance (i.e., zero correlation) for individual features averaged across conversations and gender (conversation effects were analyzed in RQ3). We performed two-tailed one-sample *t*-tests to identify which features had significantly non-zero model fit. Bonferroni correction was applied on these tests (and all other feature-level analyses) to correct for multiple comparisons. The correction resulted in a significance threshold of .002 (.05/25 for 25 output features). The recurrent models predicted 15 out of 25 features significantly above chance. These features are shown under "Model Fit" in TABLE III. We found that the models best predicted facial movements of the nasolabial region, eyes, mouth & chin, Cheek Raiser (AU6), and head roll. Altogether, these nonverbal behaviors encompassed a broad range of momentary facial responses, speech-related facial movements, smiling, and head movement.

We also examined whether model fit was robust to random shuffling of partners. Each partner was paired with a random surrogate from another couple and then model fit was calculated. A paired *t*-test showed that model fit (M = .09, SD = .03) was indeed greater than random shuffling (M < .01, SD = .03), $t(28) = 10.57$, $p < .0001$, which produced chance-level estimates.

RQ2. Is model-based nonverbal synchrony comparable or better than naïve synchrony?

We calculated *model synchrony* for each feature, couple, and conversation, and then averaged across conversations. In effect, this represents the generalized patterns of behavior learned by our models, since the female/male model predictions are correlated to measure model synchrony. We also computed naïve synchrony by correlating the raw female/male signals.

We performed one-sample *t*-tests to identify which features had significantly non-zero correlations. Our models produced nonverbal synchrony across 10 features - see "Model Synchrony" in TABLE III, whereas naïve synchrony was significantly non-zero for 12 features (these are shown under "Naïve Synchrony" in TABLE III).

Both model synchrony and naïve synchrony were absent across body motion and head movements. Similarly, seven facial movements did not yield synchrony when examined via model or naïve synchrony: eye closing, jaw dropping, frowning (AU15), lip stretching, dimpling, brow lowering, and outer brow raising.

Four facial movements showed patterns of model synchrony but not naïve synchrony: chin raising, lip pressing, lip parting, and inner brow raising. In contrast, three facial movements only showed naïve synchrony: lip tightening, lip sucking, and smiling (AU12). Head roll and head yaw showed patterns of anti-correlation for naïve synchrony.

The remaining set of six facial movements were sources of both model synchrony and naïve synchrony. We compared these using paired two-tailed *t*-tests with a Bonferroni threshold of .0083 (.05/6). Upper Lip Raiser (AU10; p = .01), Lid Tightener (AU7; p = .11), and Lip Pucker (AU18; p = .97) were not significantly different. However, model synchrony was greater for Nose Wrinkler (AU9), $t(28) = 2.88$, p = .007, and Upper Lid Raiser (AU5), $t(28) = 5.28$, $p < .001$. Naïve synchrony was greater for Cheek Raiser (AU6), $t(28) = -3.44$, p = .002.

We also compared model synchrony to random shuffling. A paired *t*-test showed that model synchrony (M = .11, SD = .06) was greater than this random baseline (M < .01, SD = .05), $t(28) = 7.83$, $p < .0001$.

RQ3. Are there differences in nonverbal synchrony across conversations?

Given the substantial differences in conversation topics, we compared conversations on average model fit and model synchrony in separate analyses. One-way repeated measures ANOVA showed no significant differences in model fit ($F(2, 56) = 2.05$, $p = .14$) or model synchrony ($F(2, 56) = 2.20$, $p = .12$). Model fit and synchrony of each conversation are shown in TABLE II.

TABLE II. MODEL FIT & SYNCHRONY BY CONVERSATION

Conversation	Model fit		Model synchrony	
	M	*SD*	*M*	*SD*
External Stressor	.07	.05	.12	.11
Internal Stressor	.10	.06	.12	.09
Enjoyment	.09	.05	.08	.09

RQ4. Are model fit and synchrony associated with clinical measures?

We performed a preliminary analysis of how our measures of synchrony associated with clinical measures reported by the couples (TABLE I). Each post-conversation rating included multiple items so we computed means across items and partners to produce a single value for each clinical measure per couple and conversation (e.g., mean rapport of a couple for the internal stress topic).

The clinical measures were used to predict model fit, model synchrony, and naïve synchrony for each feature using linear regression (i.e., 75 (3 outputs x 25 features) regression models in all). We computed goodness of fit (R^2) for the individual models and averaged them across features. We found that clinical measures were marginally more strongly associated with model synchrony (M = .14, SD = .06) compared to naïve synchrony (M = .11, SD = .05), $t(24) = 1.93$, $p = .066$. Model fit had a similar level of association with clinical measures (M = .13; SD = .05) compared to model synchrony ($p = .49$).

We also compared each measure with random shuffling, which showed significantly greater association for model synchrony (M = .14, SD = .06) vs. shuffled model synchrony (M = .07, SD = .03), $t(24) = 6.00$, $p <$ .0001, and model fit (M = .13, SD = .05) vs. shuffled model fit (M = .08, SD = .03), $t(24) = 4.52$, $p = .0001$. However, naïve synchrony (M = .11, SD = .05) was only marginally significantly different from shuffled naïve synchrony (M = .08, SD = .05), $t(24) = 1.82$, $p = .081$. These initial results demonstrate that our models produced clinically relevant outputs that were notably better than naïve synchrony.

A. Impact of Speech Features and Sequence Length

We investigated the impact of excluding speech features of the target partner in the model input. We averaged model fit across couples for models with target speech (M = .09, SD = .03) and without target speech (M = .08, SD = .03) and found a significant difference, $t(28) = 3.06$, $p = .005$. Given that the magnitude of the difference seemed negligible, we performed *t*-tests to identify which nonverbal behavior regions underlie this effect. The main drivers were greater model fit of nasolabial region with target speech (M = .24, SD = .10), $t(28) = 5.98$, $p < .0001$, vs. without (M = .17, SD = .09); and the eyes with target speech (M = .13, SD = .07), $t(28) = 4.54$, $p < .0001$, vs. without (M = .07, SD = .04). However, model fit of the smile region was significantly lower with target speech (M = .10, SD = .05), $t(28) = -6.95$, $p < .0001$, vs. without (M = .19, SD = .08). Thus, including or excluding target speech has a tradeoff in performance across facial regions.

We also compared further models on 3-, 5-, 20-, and 30-second sequences. An ANOVA on average model fit per couple showed significant differences ($F(4, 112) = 34.30$, $p < .0001$). Average model fit decreased from the shortest sequences (3-seconds M = .11, SD = .17; 5-seconds M = .10, SD = .17; 10-seconds M = .09, SD = .15) to longest sequences (20-seconds M = .07, SD = .14; 30-seconds M = .06, SD = .14). However, all models used 32 neurons, so greater fit may be due to lower complexity of the shorter sequences. Further work may find that optimal sequence length differs across nonverbal behaviors.

V. DISCUSSION

In the present work, we created recurrent models of nonverbal synchrony from a study of romantic couples engaging in multiple conversations. Our recurrent models (LSTMs) significantly predicted behavior above baselines of feed-forward neural networks and random chance, with models that predicted female data outperforming male (RQ1). Examinations of model synchrony vs. naïve synchrony revealed patterns of nonverbal synchrony that were uniquely learned by our models (RQ2). No significant differences in model fit were found across conversations, indicating that our models learned generalized synchronous behavior not specific to conversation topic (RQ3). Finally, our model-based fit and synchrony were most associated with clinically relevant measures when compared with naïve synchrony (RQ4).

A. Sources of Nonverbal Synchrony

Our models exhibited multiple strong sources of nonverbal synchrony. These covered facial movements of different regions (nasolabial, smiling, eyes, mouth & chin, and eyebrows), but not body movements. We will consider each of these in turn.

The facial movements tracked in the nasolabial region include Upper Lip Raiser (AU10) and Nose Wrinkler (AU9). In past emotion literature, these were identified as major components of the disgust facial prototype [32], though both AU9 & 10 may occur as markers of disagreeing or assessing. However, we did not observe extreme negative affect in this study.

Smiling includes Lip Corner Puller (AU12), which brings the lips to the sides, and Cheek Raiser (AU6), which elevates the cheeks while also raising the lower eye region. Together, these form the basis of the Duchenne (or genuine) smile [33]. Both facial movements were significantly fit by our models, though the model failed to detect the AU 12 synchrony pattern. Smiling promotes positive social bonds, so modeling this pattern of nonverbal synchrony is important.

TABLE III. MEAN CORRELATIONS OF MODEL FIT (PREDICTED, ACTUAL) AND NONVERBAL SYNCHRONY (FEMALE, MALE)

Category	Feature	Model Fit *M* (*SD*)	Model Synchrony *M* (*SD*)	Naïve Synchrony *M* (*SD*)	Synchrony Pattern
Nasolabial	Upper Lip Raiser (AU10)	**.28 (.11)**	**.28 (.17)**	**.19 (.12)**	**M = N**
	Nose Wrinkler (AU9)	**.20 (.11)**	**.24 (.18)**	**.12 (.11)**	**M > N**
Eyes	Upper Lid Raiser (AU5)	**.24 (.13)**	**.26 (.20)**	**.07 (.08)**	**M > N**
	Lid Tightener (AU7)	**.14 (.09)**	**.21 (.18)**	**.16 (.11)**	**M = N**
	Eyes Closed (AU43)	.01 (.06)	−.01 (.15)	.06 (.10)	-
Mouth & Chin	Chin Raiser (AU17)	**.19 (.12)**	**.29 (.19)**	−.01 (.09)	**M**
	Lip Pressor (AU24)	**.15 (.10)**	**.26 (.18)**	.01 (.10)	**M**
	Lip Pucker (AU18)	**.12 (.11)**	**.25 (.15)**	**.24 (.14)**	**M = N**
	Lips Part (AU25)	**.11 (.10)**	**.23 (.16)**	.07 (.15)	**M**
	Jaw Drop (AU26)	**.09 (.08)**	.05 (.21)	−.03 (.14)	-
	Lip Tightener (AU23)	**.09 (.06)**	−.06 (.17)	**.07 (.09)**	**N**
	Lip Corner Depressor (AU15)	**.06 (.06)**	−.03 (.15)	−.01 (.09)	-
	Lip Stretcher (AU20)	**.06 (.07)**	0 (.18)	.07 (.11)	-
	Lip Suck (AU28)	.02 (.08)	.10 (.20)	**.06 (.09)**	**N**
Smile	Cheek Raiser (AU6)	**.17 (.09)**	**.23 (.18)**	**.32 (.13)**	**M < N**
	Lip Corner Puller (AU12)	**.12 (.08)**	−.05 (.17)	**.38 (.14)**	**N**
	Dimpler (AU14)	.03 (.08)	0 (.18)	.03 (.11)	-
Eyebrows	Inner Brow Raiser (AU1)	.04 (.09)	**.13 (.17)**	.02 (.07)	**M**
	Brow Lowerer (AU4)	.01 (.08)	.06 (.15)	.05 (.09)	-
	Outer Brow Raiser (AU2)	−.02 (.10)	.11 (.21)	.03 (.08)	-
Head	Head Roll	**.08 (.06)**	.06 (.16)	**−.05 (.08)**	**N**
	Head Pitch	.02 (.07)	−.01 (.20)	.02 (.10)	-
	Head Yaw	0 (.07)	.02 (.20)	**−.18 (.14)**	**N**
Body	Gestural Motion	.02 (.06)	.01 (.10)	−.01 (.05)	-
	Gross Body Movement	.01 (.07)	.03 (.14)	.11 (.14)	-

Gray highlights indicate significantly non-zero correlations (after Bonferroni). For rightmost column, M = model synchrony was significantly greater than zero but naive synchrony was not; vice versa for N. M > N = model synchrony was significantly greater than naïve synchrony; vice versa for M < N. M = N denotes cases where both were significantly greater than zero but statistically equivalent. Cases where neither were significant are denoted with -.

A variety of mouth movements were implicated in patterns of nonverbal synchrony as detected by our models, including Lip Pressor (AU24), Lip Pucker (AU18), and Lips Part (AU25). These are related to speech-related mouth opening and lip movements coinciding with particular phonemes (e.g., an 'o' sound). These may also express momentary responses, such as pressing lips together in frustration.

There were a few upper face facial movements learned by our models. Lid Tightener (AU7) is a squinting facial movement that may co-occur with skepticism or thought. Inner Brow Raiser (AU1) has been linked to prototypical sadness in past literature [34]. However, eyebrow raising often emphasizes what is concurrently said. In more confrontational contexts, one may expect these facial movements to arise out of anger, fear, or disgust. Further work must reconcile these different contextual interpretations.

Body movements were not identified as strong sources of synchrony. Posture shifting is a key component of social mirroring [6], so this is an important pattern that our models do not yet capture. Both head yaw and head roll produced naïve anti-correlations, which indicate behavior occurring out of sync or periodically. For instance, head yaw mainly measures turning toward or away from one's partner in our study environment, as the chairs were placed at an angle. In a scenario with partners directly in front of each other, head yaw would more likely signal shaking of the head in disagreement [7]. These asynchronous behaviors highlight important phenomena in interpersonal communication that are not well represented by our present features.

B. Potential Application

We envision use of these models of nonverbal synchrony as a research tool as well as a clinical tool in couples therapy. A therapist may consult real-time models of nonverbal synchrony to assess a couple's style of interaction and provide targeted intervention. Such technologies may eventually surpass our innate capabilities (e.g., we may include face and body microexpressions [5]).

VI. CONCLUSION

Generative models of nonverbal synchrony yield a new approach to understand human behavior. In contrast with prior analytical approaches, generative modeling

enables learning of patterns across modalities and generalized across dyads. We have presented recurrent neural network models of nonverbal synchrony constructed from a study of romantic couples engaging in conversations. We found that facial movements, such as nasolabial, eye, and mouth movements were strong sources of nonverbal synchrony learned by our models. Further work in this vein offers a new platform to understand human behavior, complementary to existing approaches. As generative models of nonverbal synchrony account for multiple modalities of interpersonal communication at ever increasing degrees of accuracy, they may enable us to bridge across diverse social contexts to understand the patterns of behavior that bring us closer together.

VII. Acknowledgments

This research was supported by the National Science Foundation (NSF DUE-1745442) and the Institute of Educational Sciences (IES R305A170432). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. This research was also supported in part by the Institute for Social Science Research at Arizona State University.

VIII. References

[1] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *IEEE Trans. Affect. Comput.*, vol. 3, no. 3, pp. 349–365, 2012.

[2] O. Oullier, G. C. De Guzman, K. J. Jantzen, J. Lagarde, and J. A. Scott Kelso, "Social coordination dynamics: Measuring human bonding," *Soc. Neurosci.*, vol. 3, no. 2, pp. 178–192, 2008.

[3] S. E. Brennan, A. Galati, and A. K. Kuhlen, "Two minds, one dialog: Coordinating speaking and understanding," *Psychol. Learn. Motiv.*, vol. 53, pp. 301–344, 2010.

[4] Y. Feng, A. Kannan, G. Gkioxari, and L. Zitnick, "Learn2Smile: Learning Non-Verbal Interaction through Observation," in *Proceedings of the International Conference on Intelligent Robots and Systems*, 2017.

[5] Y. Song, L.-P. Morency, and R. Davis, "Learning a Sparse Codebook of Facial and Body Microexpressions for Emotion Recognition," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, 2013, pp. 237–244.

[6] W. Tschacher, G. M. Rees, and F. Ramseyer, "Nonverbal synchrony and affect in dyadic interactions," *Front. Psychol.*, vol. 5, p. 1323, 2014.

[7] Z. Hammal, J. F. Cohn, and D. T. George, "Interpersonal Coordination of HeadMotion in Distressed Couples," *IEEE Trans. Affect. Comput.*, vol. 5, no. 2, pp. 155–167, Apr. 2014.

[8] D. McNeill, *Gesture & Thought*. Chicago: The University of Chicago Press, 2005.

[9] J. A. Russell, "Emotion, core affect, and psychological construction," *Cogn. Emot.*, vol. 23, no. 7, pp. 1259–1283, Nov. 2009.

[10] K. R. Scherer, "The dynamic architecture of emotion: Evidence for the component process model," *Cogn. Emot.*, vol. 23, no. 7, pp. 1307–1351, 2009.

[11] F. Ramseyer and W. Tschacher, "Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome.," *J. Consult. Clin. Psychol.*, vol. 79, no. 3, p. 284, 2011.

[12] Z. Hammal, J. F. Cohn, D. S. Messinger, W. I. Mattson, and M. H. Mahoor, "Head movement dynamics during normal and perturbed parent-infant interaction," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 276–282.

[13] N. D. Duran and R. Fusaroli, "Conversing with a devil's advocate: Interpersonal coordination in deception and disagreement," *PLoS One*, vol. 12, no. 6, p. e0178140, 2017.

[14] A. Main, A. Paxton, and R. Dale, "An exploratory analysis of emotion dynamics between mothers and adolescents during conflict discussions.," *Emotion*, vol. 16, no. 6, pp. 913–928, 2016.

[15] C.-C. Lee and S. Narayanan, "Predicting interruptions in dyadic spoken interactions," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, 2010, pp. 5250–5253.

[16] D. Ozkan, K. Sagae, and L.-P. Morency, "Latent mixture of discriminative experts for multimodal prediction modeling," in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 860–868.

[17] C. Saint-Georges *et al.*, "Do Parents Recognize Autistic Deviant Behavior Long before Diagnosis? Taking into Account Interaction Using Computational Methods," *PLoS One*, vol. 6, no. 7, p. e22393, 2011.

[18] D. M. Messinger, P. Ruvolo, N. V Ekas, and A. Fogel, "Applying machine learning to infant interaction: The development is in the details," *Neural Networks*, vol. 23, no. 8–9, pp. 1004–1016, 2010.

[19] J. M. Gottman *et al.*, "Correlates of gay and lesbian couples' relationship satisfaction and relationship dissolution," *J. Homosex.*, vol. 45, no. 1, pp. 23–43, 2003.

[20] S. S. Hendrick, "A generic measure of relationship satisfaction," *J. Marriage Fam.*, vol. 50, no. 1, pp. 93–98, 1988.

[21] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System*. Salt Lake City, USA: A Human Face, 2002.

[22] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proceedings of ACM Multimedia*, 2013, pp. 835–838.

[23] F. Chollet and others, "Keras (https://github.com/fchollet/keras)." GitHub, 2015.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv Prepr. arXiv1412.6980*, 2014.

[27] T. Dozat, "Incorporating Nesterov Momentum into Adam," 2016.

[28] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[29] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the Python in Science Conference*, 2010, pp. 51–56.

[30] R Core Team, "R: A Language and Environment for Statistical Computing." Vienna, Austria, 2013.

[31] G. Bodenmann, N. Meuwly, J. Germann, F. W. Nussbeck, M. Heinrichs, and T. N. Bradbury, "Effects of stress on the social support provided by men and women in intimate relationships," *Psychol. Sci.*, vol. 26, no. 10, pp. 1584–1594, 2015.

[32] S. C. Widen and J. A. Russell, "Children's recognition of disgust in others.," *Psychological Bulletin*, vol. 139, no. 2., pp. 271–299, 2013.

[33] Z. Ambadar, J. F. Cohn, and L. I. Reed, "All Smiles are Not Created Equal: Morphology and Timing of Smiles Perceived as Amused, Polite, and Embarrassed/Nervous," *J. Nonverbal Behav.*, vol. 33, no. 1, pp. 17–34, Mar. 2009.

[34] E. G. Krumhuber and K. R. Scherer, "Affect bursts: Dynamic patterns of facial expression.," *Emotion*, vol. 11, no. 4, pp. 825–841, 2011.