Analysis and Predictive Modeling of Body Language Behavior in Dyadic Interactions From Multimodal Interlocutor Cues

Zhaojun Yang, Student Member, IEEE, Angeliki Metallinou, and Shrikanth Narayanan, Fellow, IEEE

Abstract—During dyadic interactions, participants adjust their behavior and give feedback continuously in response to the behavior of their interlocutors and the interaction context. In this paper, we study how a participant in a dyadic interaction adapts his/her body language to the behavior of the interlocutor, given the interaction goals and context. We apply a variety of psychology-inspired body language features to describe body motion and posture. We first examine the coordination between the dyad’s behavior for two interaction stances: friendly and conflictive. The analysis empirically reveals the dyad’s behavior coordination, and helps identify informative interlocutor features with respect to the participant’s target body language features. The coordination patterns between the dyad’s behavior are found to depend on the interaction stances assumed. We apply a Gaussian-Mixture-Model-based (GMM) statistical mapping in combination with a Fisher kernel framework for automatically predicting the body language of an interacting participant from the speech and gesture behavior of an interlocutor. The experimental results show that the Fisher kernel-based approach outperforms methods using only the GMM-based mapping, and using the support vector regression, in terms of correlation coefficient \( RMSE \). These results suggest a significant level of predictability of body language behavior from interlocutor cues.

Index Terms—Behavior coordination, body language, dyadic interactions, interaction goals, motion capture.

I. INTRODUCTION

The coordination of verbal and nonverbal human behavior during interpersonal interactions has been studied in many diverse areas including neuroscience, engineering, psychology and behavioral sciences [1]. Interaction participants generally adjust their behavior and give feedback continuously based on the behavior of their interlocutors whom they are having conversations with, as well as based on their own interaction goals and context. The adaptation of behavior, such as facial expressions or body movements, facilitates the communication to move smoothly, efficiently and coherently. In addition to understanding the fine details of human interaction mechanisms, modeling behavior coordination is important for developing intelligent conversational agents which can naturally respond non-verbally in real-time while human interlocutors are speaking.

Body language is an essential element of nonverbal behavior in human communication, which includes facial gestures, head, limb and other body postures and movements. It conveys attitudes and emotions of a person towards his/her interlocutor. The adaptation of an interacting participant’s body language to his/her interlocutor’s behavior depends on their interaction goals and their nature, including the stance or attitude assumed such as being friendly or conflictive. For example, two participants with friendly attitudes may tend to approach each other, while those with conflictive attitudes may avoid each other and try to terminate the conversation. Hence, the coordination pattern of a dyad’s behavior is influenced by the nature of their interaction goals.

This work focuses on studying how an interacting participant adapts his/her body language to the multimodal behavior, i.e., gesture and speech cues, of the interlocutor, based on their communication goals in a dyadic interaction. Our objective is two-fold: 1) to examine the coordination between an interaction participant’s body language with the interlocutor’s multimodal behavior given the nature of interaction stances (friendliness or conflict); 2) to computationally model the dyadic coordination and automatically predict an interaction participant’s body language from the multimodal information of his/her interlocutor. The estimation of a subject’s body language based only on the interlocutor’s behavioral cues is a novel research direction. This could lead to the development of expressive and goal-driven virtual agents that would display appropriate body language in response to the human interlocutor’s behavior such as to increase the human-agent rapport.

In this work, we use the multimodal USC CreativeIT database that consists of goal-driven improvised interactions [2]. It contains detailed full body Motion Capture (MoCap) data of both interacting participants in addition to audio data, providing a rich resource for studying, modeling and predicting body language patterns in dyadic interactions. We investigate various features inspired by the psychology literature on nonverbal communication [3] to describe body language behavior, including of head and body position, hand and body motion, and relative motion and orientation. We consider eight distinct target...
body language features for modeling and predicting the interacting participant’s behavior with respect to the interlocutor. The eight target features are selected because they capture key aspects of human body language and expressiveness, viz., body posture, velocity, orientation and relative motion toward the interlocutor. The dyad’s coordination is investigated using canonical correlation analysis (CCA). We find statistically significant correlations that empirically verify the dyad’s behavior coordination. We also rank the interlocutor’s body language and speech features for each interaction stance, according to their importance to the target participant body language feature, using a correlation-based criterion. The coordination pattern of the dyad’s behavior for each interaction type is represented by a list of ranked body language and speech features. We observe that coordination patterns depend on the interaction goals; in our study, we find that the types of highly ranked interlocutor features vary depending on whether the interaction is friendly or conflictive.

Motivated by these analyses, we propose a method to model the coordination of dyadic behavior for each interaction stance and to automatically predict the target participant body language features using the multimodal information derived from the interlocutor. For this purpose, we adopt a Fisher kernel based methodology with Gaussian Mixture Model (GMM) to map from the interlocutor’s behavior to the body language of the target participant. Experimental results show that the Fisher kernel based approach outperforms the generative method of MLE-based mapping and the support vector regression (SVR) model, in terms of both correlation coefficient and Root Mean Squared Error ($RMSE$) with respect to the ground truth data. In addition, the inclusion of the interlocutor’s speech information improves the prediction performance, compared to using only the interlocutor’s body language. These results indicate a significant level of predictability of body language from the interlocutor behavior, which could be exploited towards interaction-driven and goal-driven body language synthesis for intelligent virtual agents.

The rest of the paper is organized as follows. Section II describes related work on studying the correlation between human behavior during communication. Section III introduces the multimodal database of dyadic interactions, followed by the description of behavior feature (body language and speech) extraction in Section IV. In Section V, we define the categories of behaviors for interlocutor pairs in an interaction, which are used for analysis and modeling. Based on the behavior pairs, we analyze how the behavior is coordinated between two interaction subjects given communication goals in Section VI. Section VII presents the predictive framework for modeling the behavior coordination according to the interaction goals. Experimental results are presented in Section VIII, and Section IX provides the conclusion and future work along this research direction.

II. RELATED WORK

Body language is an important indicator of emotions or attitudes during social interactions. Ekman extensively studied the relation between facial expressions and emotions, and concluded that an emotional state could be inferred from an expression with cross-cultural agreement [4]. The analysis of movement structure in [5] has demonstrated the dynamic congruency between emotion and movement. Conty et al. showed that the early binding of body gesture, gaze direction and emotion could possibly promote a more adaptive response to the interlocutor in an interaction [6]. These theoretical foundations have inspired studies on automatic emotion recognition utilizing body language. Affective content has been explored from the dynamics of body movement in [7]. Gunes et al. integrated body gestures along with facial expressions to recognize emotions [8]. Bernardt and Robinson have attempted to detect affective information in the knocking motion using derived motion primitives [9].

The correlation between body language and speech behavior has enabled much progress in prosody-driven body gesture synthesis. Busso et al. have proposed a prosody-driven approach for synthesizing expressive rigid head motion [10], [11]. Mariooryad et al. have built a joint speech and facial gesture model to generate head and eyebrow motion for conversational agents [12]. Likewise, Sargin et al. synthesized head gestures from speech by exploring the joint correlation between head gesture and prosody patterns [13]. Frameworks for full body motion synthesis in real-time, which also use prosody information, have been developed in [14], [15]. In addition to the speech-driven methods, emotion-based body language synthesis has also been studied. Beck et al. generated emotional body language for robots by investigating the effect of varying the positions of joints on the interpretation of emotions [16]. While these works mainly focus on synthesizing body language of individuals, this paper aims at interaction-driven and goal-driven body language modeling in dyadic settings, which we believe is important for the development of natural human-machine interfaces.

In social interactions, there exists interpersonal influence along aspects of speech prosody, body language, and emotional states. To accomplish effective communication, people often adapt their interaction behavior to the interlocutor and such behavior adaptation includes synchronizing in time or displaying similar or dissimilar behavior [17]. Chartrand et al. described that humans unconsciously mimic the behavior of their interaction partners to achieve more effective and pleasant interactions [18]. Ekman found that body language of interviewees is distinctly different between friendly and hostile job interviews [19]. Neumann et al. reported that the emotions in speech would induce a consistent mood state in the listeners [20]. Kendon qualitatively described detailed interrelations between movements of the speaker and the listener by analyzing sound films of social interactions [21]. He also found that the movement of the listener might be rhythmically coordinated with the speech and movement of the speaker in a social interaction.

Many engineering works have also been developed based on this mutual influence of interaction subjects. Morency et al. predicted head nods for virtual agents from the audio-visual information of a human speaker based on sequential probabilistic models [22]. In contrast, our work aims at predicting more diverse full body movements in response to the interlocutor’s behavior. Heylen et al. studied what types of appropriate responses, e.g., facial expressions, an agent should display when a human user is speaking, to increase rapport in human-agent
conversation [23]. Researchers have also used the emotional state of an interlocutor to inform that of a speaker by modeling emotional dynamics between two interaction participants [24], [25]. The influence framework proposed in [26] models participants in conversational settings as interacting Markov chains. Lee et al. proposed prosody-based computational entrainment measures to assess the coordination in the interactions of married couples [27]. Xiao et al. studied and modeled the interaction synchrony of head motion behavior using GMM [28], [29].

A Gaussian mixture model (GMM) based statistical mapping methodology has been applied in our previous work for interaction-driven body language prediction [30], i.e., predicting the listener’s body language from the speaker’s speech and gestural behavior for both friendly and conflictive interaction stances. A GMM which describes the probability distribution of an overall population is represented by the mixture of Gaussian distributions of subpopulations within an overall population. More details about GMM are described in Section VII. This GMM-based mapping was based on a method originally proposed for articulatory to acoustic mapping [31] and spectral conversion between speakers [32], and has also been used for continuous emotional state estimation based on body language and speech [33]. Our preliminary work in [30] showed that the predicted trajectories of the target body language features generally follow the trend of the observed curves. However, this generative approach produces a large difference between the exact values of predicted and observed body language. To overcome this limitation, we extend this work by applying the Fisher kernel framework [34] to model the interactive behavior coordination. The Fisher kernel technique combines the advantages of generative and discriminative approaches. It represents a data sample as a gradient vector of the generative model over the data, i.e., GMM in our case, and inputs the representation into a discriminative model. Fisher kernels on GMM have been successfully applied in classification problems, such as image categorization [35], audio classification [36] and speaker verification [37]. In this paper, we employ them to improve the prediction performance of the participant’s target body language features from the interlocutor’s behavior.

III. DATABASE DESCRIPTION

We use the USC CreativeIT database in our experiments, which is a multimodal corpus of dyadic theatrical improvisations [2]. The database is freely available for research at http://sail.usc.edu/data.php. It contains detailed full body Motion Capture (MoCap) data of the two interacting participants recorded by 12 Vicon MoCap cameras at 60 fps, and audio data obtained through close-up microphones at 48 KHZ with 24 bits. The MoCap cameras were located on the ceiling of the recording room and the motion capture process retrieved the \((x, y, z)\) positions of 45 markers for each participant. The markers were placed across an actor’s body, as shown in Fig. 1(a). Our work is based on the features extracted from the MoCap and audio data, which is described in Section IV. The process of collecting and postprocessing the MoCap data in our database requires much effort and is time-consuming. Recently, more advanced

and improved techniques have been developed and applied for motion data collection, such as with Kinect [38] or the newer Vicon system.\(^1\)

The interactions performed by pairs of actors are either improvisations of scenes from theatrical plays or theatrical exercises where actors repeat saying sentences to express specific goals featuring specific communication stances or attitudes. In this work, we examine the theatrical exercises, a simplified form of interactions with restricted lexical content (only limited phrases can be used), which encourages rich body language expressions. For example, in a theatrical exercise, one actor says *Make me smile*, and the other one replies *Why should I*. The interactions were guided by a theater expert (professor/director), and were performed following the Active Analysis improvisation technique pioneered by Stanislavsky [39]. According to this technique, the interactions are goal-driven; the pair of actors in each interaction has a pair of predefined interaction goals, e.g., *to approach* vs. *to avoid*, which they try to achieve through the appropriate use of body language and speech prosody. The goal pair of each dyad defines their attitudes towards each other and the content of the interaction.

As described in Section I, our premise is that the dynamics of interactions differ depending on the stances of interacting participants, which may result in distinct coordination patterns of dyad’s behavior. Hence, we group the interactions into three cases according to the defined goal pairs: friendliness, medium conflict and high conflict. In friendly interactions both participants have friendly attitudes, e.g., one is to make peace and the other is to comfort; in medium conflict interactions one participant is friendly while the other is creating conflict; and in high conflict interactions the attitudes of both participants are conflictive. Interactions that do not fit into these categories are excluded from our analysis. This grouping is described in Table I, along with examples of characteristic goal pairs. Friendliness,
TABLE I

<table>
<thead>
<tr>
<th>Friendly</th>
<th>Medium Conflict</th>
<th>High Conflict</th>
<th>Example goal pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actors’ attitudes</td>
<td>friendly - friendly</td>
<td>friendly - conflictive</td>
<td>to make peace - to comfort</td>
</tr>
<tr>
<td></td>
<td>conflictive - conflictive</td>
<td>conflictive - conflictive</td>
<td>to convince - to reject</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>to accuse - to fight back</td>
</tr>
</tbody>
</table>

TABLE II

<table>
<thead>
<tr>
<th>Extracted Body Language Features. Features Marked With * Will Be Predicted Using the Interlocutor’s Feature Set (Both Marked With * and Not Marked)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cosFace*</td>
</tr>
<tr>
<td>HAngle</td>
</tr>
<tr>
<td>cosBody*</td>
</tr>
<tr>
<td>cosLearn*</td>
</tr>
<tr>
<td>px, py, pz</td>
</tr>
<tr>
<td>aVbody*</td>
</tr>
<tr>
<td>aVarmr*,a*</td>
</tr>
<tr>
<td>aVfeet,l</td>
</tr>
<tr>
<td>rVbody*</td>
</tr>
<tr>
<td>rVfeet,l</td>
</tr>
<tr>
<td>rhandx,y,z</td>
</tr>
<tr>
<td>lhandx,y,z</td>
</tr>
<tr>
<td>dHands*</td>
</tr>
</tbody>
</table>

medium and high conflict groups contain 10, 30 and 8 interactions respectively, including 16 actors (9 female). Each interaction has an average length of 3.5 minutes.

IV. Feature Extraction

A. Body Language Feature Extraction

The availability of full body MoCap information, as shown in Fig. 1(a), enables us to extract detailed descriptions of each actor’s body language expressed during an interaction. The features are motivated from the psychology literature which indicates that body language behaviors, such as looking at the other, turning away, approaching, touching or hand gesturing, are informative of a subject’s attitude and emotion towards his/her interlocutor [3]. While some of our features carry information about an individual’s posture and motion, such as hand or body positions, others describe behaviors relative to the interlocutor, e.g., body and head orientation, approaching and moving away. The features are computed in a geometrical manner by defining global and local coordinate systems illustrated in Fig. 1(b) and by computing Euclidean distances, relative positions, angles and velocities. Similar movement features have also been applied for analyzing the link between emotion and gesture dynamics [33], [40].

We extract 24 body language features in total for each actor in an interaction, as summarized in Table II. Some example features are illustrated in Fig. 1(c) and (d). The absolute velocity of a subject’s body is computed from the movement in one’s local coordinate system, while the relative velocity towards one’s interlocutor is obtained by projecting the velocity vector in the direction between the two partners, as shown in Fig. 1(c). Similarly in Fig. 1(d), the angle of a subject’s face towards the other, i.e., cosFace, is the angle between the head orientation in one’s local system and the direction between the two participants. If cosFace is close to 1, the actor is looking towards the interlocutor, while cosFace < 0 indicates looking away. Among the 24 body language features, we selected eight distinct features as the target features of the participant which we use for the analysis and prediction from the interlocutor’s full set of 24 features. These target features are marked with * in Table II. Note that all the 24 body language features including the target ones are used as the interlocutor’s features for analysis and prediction. The eight target features were selected because they capture key behaviors of human expressiveness, such as body and face orientations, body posture, hand gesture and relative motion toward the interlocutor [3]. In future work, the information conveyed by these selected target features can be used in virtual agent animation.

B. Audio Feature Extraction

We extract audio features from the speech regions of interacting participants, i.e., pitch, energy and 12 Mel Frequency Cepstral Coefficients (MFCCs), which are commonly used by the affective computing community [41]. Pitch and energy are also used for prosody-based body and head gesture synthesis [11], [14]. We extract these features every 16.67 ms with an analysis window length of 30 ms, in order to match with the MoCap frame rate.

V. The Dyad’s Behavior Pair Setup

Our objective is to investigate how an interaction participant’s body language is influenced as a response to his/her interlocutor’s speech and body language behavior in a dyadic interaction. A dyad refers to a group of two individuals and a dyadic interaction defines the communication between the pair of persons in the dyad. In the friendly and high conflictive cases, both actors have the same goal, therefore the body language of a friendly (conflictive) target participant is analyzed or predicted with respect to the behavior of a friendly (conflictive) interlocutor. The medium conflictive case includes two subcategories for analysis and modeling: a friendly target participant interacting with a conflictive interlocutor, and a conflictive target participant interacting with a friendly interlocutor.

In our work, we pair each target feature of the participant at frame $t$ with the interlocutor’s audio-visual behavioral information over a window preceding (including) frame $t$, i.e., we analyze or estimate the target participant body language from the past context information of the interlocutor. Specifically, the interlocutor’s information consists of body language or both body language and speech, as illustrated in Fig. 2. Similar to [42] that
uses utterance-level statistics of prosody features for emotion recognition, we extract 11 statistical functionals over the context window for each body language (Table II) and speech feature—mean, standard deviation, median, minimum, maximum, range, skewness, kurtosis, first quantile, third quantile and interquantile range—to represent context information. Specifically, skewness measures the asymmetry of the data around its sample mean; kurtosis describes the width of peak in the probability distribution of a random variable; and inter-quantile range is the difference between the third and first quantiles. Because of the high dimension statistical functional vector, we reduce the dimensionality using Principal Component Analysis (PCA), by preserving about 90% of the total variance.

Our analysis and prediction experiments are performed using dyad’s behavior pair along the entire interaction (whether the interlocutor is speaking or not). Context information is computed using both body language and speech in the window if the speech region exceeds 50% of the window size. Otherwise, the context information only includes the body language features in the window.

VI. ANALYSIS OF THE DYAD’S COORDINATION PATTERNS

In this section, we investigate how the coordination patterns of the dyad’s behavior differ depending on the interaction goals. We approach this problem by examining to what degree the participant’s body language is correlated with the interlocutor’s behavioral context information, and what types of interlocutor features are especially informative of the participant target body language given interaction attitudes.

We first perform Canonical Correlation Analysis (CCA) between the set of target participant body language features at frame $t$ and the interlocutor information over a window, for each case of goal pair in Table I. Note that we have four cases for analysis since the medium category has two subgroups. CCA respectively finds vectors $v$ and $u$ for two sets of data $X$ and $Y$ with different dimensionalities such that the correlation of $Xv^t$ and $Yu^t$ is maximized [43]. It can sequentially seek vectors $v^t$ and $u^t$ by maximizing the correlation of $X v^t$ and $Y u^t$ subject to the constraints $v^t v^t = 0$ and $u^t u^t = 0$. In our case, $Y$ represents the target body language information and has a dimension of 8 (marked target features in Table II); $X$ contains interlocutor information that is used for prediction. CCA then returns $\min(d \dim X, \dim Y)$ canonical correlations in a descending order, and each correlation can be tested for significance through the $F$ statistic based on Rao’s approximation [44]. We set the window size in Fig. 2 to be 30, 60, 90 and 120 frames. Fig. 3 shows the first (highest) canonical correlation between the dyad’s behavior varying with the window size in each interaction case. We can see that the correlation respectively achieves the maximum with the window size of 60, 120, 30 and 60 in the case of friendliness, high conflict, medium conflict (friendly target participant) and medium conflict (conflictive target participant). In the analysis and prediction experiments that follow, we fix the window size in each case accordingly. Furthermore, we find that the maximal leading canonical correlation in each case is around 0.50 ($p = 0.000$), implying the existence of the dyad’s behavior coordination in communication. Table III presents the summary of the statistics for the first canonical correlation in each interaction case.

Next, we examine what types of interlocutor features are especially informative of the target participant body language features given interaction attitudes. The list of interlocutor features which are ranked according to their informativeness represents the coordination pattern of the dyad’s behavior in each interaction case. The minimal redundancy maximal relevance criterion ($mRMR$), introduced in [45], is helpful for us to select informative but decorrelated interlocutor features with respect to each target body language feature. The original $mRMR$ was computed based on mutual information between pairs of continuous variables. Inspired by this mutual information based metric, we compute the minimal redundancy maximal relevance criterion based on correlation instead. This correlation-based metric is denoted as $mRMR_c$, which has also been applied for continuous emotion tracking in [33]. The final feature set selected by $mRMR_c$ contains features that are maximally correlated with the target feature but minimally correlated with each other. In this way, we can avoid collinearity issues among features for later analyses and experiments.

We average the interlocutor frames over the window and obtain $N$ averaged body language ($N = 24$) or speech ($N = 14$) features of the interlocutor, denoted as $\{\bar{x}_t\}_{t=1}^N$. The target body language feature of the participant is denoted as $y_t$, and the correlation between two continuous variables is denoted as $C(\cdot)$. The correlation-based metric $mRMR_c$ is defined as

$$mRMR_c(i) = C(\bar{x}_i, y_i) - \frac{1}{N-1} \sum_{j=1,j \neq i}^N |C(\bar{x}_i, \bar{x}_j)|. \quad (1)$$

Fig. 3. The canonical correlation between the dyad’s body language in relation to the window size.

<table>
<thead>
<tr>
<th>Interaction case</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly</td>
<td>$F(192, 72260) = 422.7$</td>
</tr>
<tr>
<td>Conflicitive</td>
<td>$F(192, 457100) = 279.5$</td>
</tr>
<tr>
<td>Medium (Conflicitive)</td>
<td>$F(192, 987700) = 818.5$</td>
</tr>
<tr>
<td>Medium (Friendly)</td>
<td>$F(192, 172700) = 833.0$</td>
</tr>
</tbody>
</table>

TABLE III
SUMMARY OF STATISTICS FOR THE MAXIMAL CANONICAL CORRELATION UNDER EACH INTERACTION TYPE ($p = 0.010$ FOR ALL CASES)
A higher $mRMR_{R_e}$ indicates that the corresponding interlocutor feature is more correlated with the target feature, and less correlated with other interlocutor features. Hence, we are able to rank the interlocutor’s body language or speech features for each target participant feature in each interaction case. The absolute value of the correlation $\rho$ is used in Eq. (1), because we are interested in both highly positive and highly negative relations.

Table IV presents the top-5 ranked body language features of the interlocutor with respect to the target participant features in each case, associated with $mRMR_{R_e}$ values. Overall, we notice that the types of highly ranked features differ across the goal pairs, implying different coordination patterns for different interaction types (attitudes, stances). Specifically, in friendly situations hand position features are more informative than the others, suggesting the expressiveness of hand gestures in these interactions. In contrast, many relative and absolute velocities of arms, hands and feet, as well as face orientation features are found informative in medium and high conflict cases, indicating more body motion and approaching-avoidance behaviors in these cases. For speech features, we observe that pitch and energy of the interlocutor are more informative of the target participant features in medium and high conflict cases than in the friendly one (detailed results of speech features are omitted for the lack of space).

We also investigate the sign of the correlation between the interlocutor features $\{x_1, \ldots, x_N\}$ and the target body language feature $y$. Table IV presents the sign of the correlation between each feature pair in front of the corresponding interlocutor feature. Two feature variables with a positive correlation sign have a direct relationship with each other, while variables with a negative correlation sign have an inverse relationship. For example, for friendly interactions, a positive relationship between two subjects’ leaning angles ($\cos\text{Lean}$) is observed, indicating that they tend to lean towards and approach each other under such condition. In contrary, face orientations ($\cos\text{Face}$) or leaning angles ($\cos\text{Lean}$) of the two interacting partners are negatively related in the medium conflict cases, suggesting that a friendly participant is more likely to approach, e.g., facing/leaning towards, while the unfriendly one is more likely to avoid, e.g., facing/leaning away.

The analysis results that different coordination patterns are observed in different interaction conditions, reinforce our hypothesis that conditioning on the goal-pairs reduces some of the dyad’s coordination variability in interactions. This also helps us to focus on modeling body language coordination patterns related to the specific interaction goals of friendliness and conflict.

VII. FISHER KERNEL FRAMEWORK FOR BODY LANGUAGE MODELING

The observed correlation in the dyad’s behavior, as well as the distinct coordination patterns under different interaction types, imply certain level of predictability of body language in our goal-driven interactions of interest. We hence aim to model the dyad’s behavior coordination according to their interaction goals. We apply a Fisher kernel based methodology for estimating the participant’s body language patterns from the interlocutor’s speech and body language behavior. A generative method models the probability distribution of data, focusing on explaining the full relationships between variables. In contrast, a discriminative method focuses on learning functions that better discriminate across classes, therefore tends to yield better classification and regression results. The Fisher kernel approach allows us to combine the benefits from the two types of techniques [34].

A. Generative Model: GMM-Based Statistical Mapping

This GMM-based method estimates an optimal statistical mapping, using Maximum Likelihood Estimation (MLE), from a set of observed continuous random variables to a target continuous variable. This method was originally introduced for the problem of articulatory to acoustic mapping [31]. It can be seen as a general way to produce a mapping between two
For a large range of problems. For example, it has been applied for continuous emotional state tracking in [33]. In this work, we focus on the body language prediction problem, where the two multidimensional time series are body language and speech (if applicable) features from two interacting participants.

We denote the interlocutor feature vector (i.e., the context information within the window) at time $t$ by $\mathbf{x}_t$, and the corresponding target behavioral feature of the participant in the interaction by $y_t$. In [30], we model the joint probability distribution of the dyad’s observed behavioral information $P(\mathbf{x}_t, y_t)$ as a Gaussian Mixture Model (GMM)

$$P(\mathbf{x}_t, y_t|\theta^{(x,y)}) = \sum_{m=1}^{M} \pi_m N(\mathbf{x}_t, y_t; \mu_m^{(x,y)}, \Sigma_m^{(x,y)}),$$  \hspace{1cm} (2)

where $M$ is the number of mixtures, $\pi_m$, $\mu_m^{(x,y)}$ and $\Sigma_m^{(x,y)}$ are the weight, mean vector and covariance matrix of the $m$-th component, and $\sum_{m=1}^{M} \pi_m = 1$. The GMM-based mapping method estimates the unknown target feature $y_t$ given $\mathbf{x}_t$ using Maximum Likelihood Estimation (MLE), by maximizing the corresponding conditional distribution

$$\hat{y}_t = \arg\max_{y_t} P(y_t|\mathbf{x}_t, \theta^{(x,y)}).$$  \hspace{1cm} (3)

The conditional distribution of $y_t$ given the observation $\mathbf{x}_t$ is also represented as a GMM

$$P(y_t|\mathbf{x}_t, \theta^{(x,y)}) = \sum_{m=1}^{M} \beta_{m,t} P(y_t|\mathbf{x}_t, \theta^{(x,y)}),$$

where $\beta_{m,t}$ is the occupancy probability $P(m|\mathbf{x}_t, \theta^{(x,y)})$,

$$\beta_{m,t} = \frac{P(\mathbf{x}_t, \theta^{(x)}) P(y_t, m, \theta^{(x)})}{\sum_{m=1}^{M} P(\mathbf{x}_t, m, \theta^{(x)})} = \frac{\pi_m N(\mathbf{x}_t; \mu_m^{(x)}, \Sigma_m^{(x)})}{\sum_{m=1}^{M} \pi_m N(\mathbf{x}_t; \mu_m^{(x)}, \Sigma_m^{(x)})}$$

and according to conditional Gaussian distribution,

$$P(y_t|\mathbf{x}_t, m, \theta^{(x,y)}) = N(y_t; \mathbf{E}_{m,t}^{(y)}, \mathbf{D}_{m,t}^{(y)}),$$

$$\mathbf{E}_{m,t}^{(y)} = \mu_m^{(x)} + \Sigma_m^{(x)} \Sigma_m^{(x)^{-1}} (\mathbf{x}_t - \mu_m^{(x)}),$$

$$\mathbf{D}_{m,t}^{(y)} = \Sigma_m^{(x)} + \Sigma_m^{(x)} \Sigma_m^{(x)^{-1}} \Sigma_m^{(x)}.$$

The maximization procedure is achieved through Expectation Maximization with the minimum mean squared error (MMSE) estimate as the initial value. For more details, we refer readers to [31]. To exploit the continuous nature of the observation and target variables, we incorporate dynamic information by augmenting $\mathbf{x}_t$ and $y_t$ with their first temporal derivative estimates [31]. We denote this approach as MLE-based mapping.

### B. The Fisher Kernel

The Fisher kernel was originally introduced in [34] and has been applied to a variety of classification problems, such as image categorization [35] and audio classification [36]. Rather than estimating the value of the target variable directly from the generative model, the Fisher kernel technique transforms a sample data into the Fisher score from the generative model, i.e., a gradient vector of the log-likelihood. This representation (gradient vector) can be in turn used as input to a discriminative model. This technique is hence able to take advantage of both generative and discriminative approaches by embedding the generative information in the representation which can be conveniently exploited by discriminative methods. Let $P(\mathbf{x}_t, \theta^{(x)})$ be a probability model of a sequence of the interlocutor’s behavior feature vectors $X = \{\mathbf{x}_t\}_{t=1}^T$. Then the sample $\mathbf{x}_t$ can be characterized by the gradient vector

$$U_x = \nabla_{\theta^{(x)}} \log P(\mathbf{x}_t, \theta^{(x)}).$$  \hspace{1cm} (4)

This gradient vector $U_x$ describes how the parameters of the generative model contribute to the process of generating a particular sample. For the exponential family of distributions, the gradients also form sufficient statistics for the samples.

$U_x$ can be further normalized as the natural gradient $F_x$ by the Fisher information matrix $I$

$$F_x = I^{-\frac{1}{2}} U_x,$$

$$I = \mathbb{E}_{\mathbf{x}_t} \{U_x U_x^T\},$$

where $I$ defines the Riemannian manifold $M_{\theta}$ of $P(\mathbf{x}|\theta^{(x)})$ [46], and the distance metric on $M_{\theta}$ measures a distance between $P(\mathbf{x}|\theta^{(x)})$ and $P(\mathbf{x}|\theta^{(x)})$ [34].

### C. Fisher Kernels for Predicting Body Language

In our case, the probability distribution $P(\mathbf{x}_t, \theta^{(x)})$ is modeled as GMM, and parameters $\theta^{(x)}$ include the weights $\{\pi_m\}_{m=1}^M$, the component mean vectors $\{\mu_m^{(x)}\}_{m=1}^M$, and covariance matrices $\{\Sigma_m^{(x)}\}_{m=1}^M$. To save computation cost, we assume the covariance matrices are diagonal, and denote the variance vector by $\sigma_m^2 = \text{diag}(\Sigma_m^{(x)})$. The log-likelihood of the generative model is

$$L(\mathbf{x}_t; \theta^{(x)}) = \log P(\mathbf{x}_t; \theta^{(x)}).$$  \hspace{1cm} (5)

Let the index $d$ denotes the $d$-th dimension of the interlocutor feature vector $\mathbf{x}_t$. Based on the generative model (GMM) of $X = \{\mathbf{x}_t\}_{t=1}^T$, we can derive the gradient of the log-likelihood with respect to parameters $\theta^{(x)}$:

$$\frac{\partial L(\mathbf{x}_t; \theta^{(x)})}{\partial \pi_m} = \frac{\beta_{m,t}}{\pi_m} - \frac{\beta_{m,t}}{\pi_1} \left( \frac{m > 2}{\pi_1} \right),$$

$$\frac{\partial L(\mathbf{x}_t; \theta^{(x)})}{\partial \mu_m(d)} = -\frac{\beta_{m,t}}{\sigma_m(d)} (\mathbf{x}_t(d) - \mu_m(d))^2 \sigma_m(d)^{-2} - \frac{1}{\sigma_m(d)},$$

$$\frac{\partial L(\mathbf{x}_t; \theta^{(x)})}{\partial \sigma_m(d)} = \beta_{m,t} \left( \frac{\mathbf{x}_t(d) - \mu_m(d)}{\sigma_m(d)} \right)^2 \sigma_m(d)^{-3}.$$  \hspace{1cm} (9)

The Fisher information matrix $I$ can be approximated as a diagonal matrix with diagonal coefficients $I_{\pi_m}$, $I_{\mu_m(d)}$ and $I_{\sigma_m(d)}$, corresponding to the derivatives $\partial L(\mathbf{x}_t; \theta^{(x)})/\partial \pi_m$, $\partial L(\mathbf{x}_t; \theta^{(x)})/\partial \mu_m(d)$, and $\partial L(\mathbf{x}_t; \theta^{(x)})/\partial \sigma_m(d)$.
\[ \partial L(x_t; \theta^{(x)}) / \partial \mu_m(d) \text{ and } \partial L(x_t; \theta^{(x)}) / \partial \sigma_m(d) \]. Perronnin et al. approximated \( L \) in a closed form and detailed derivations can be found in the appendix of [35].

\[ I_{\pi_m} = \frac{1}{\pi_m} + \frac{1}{\pi_1} \]  
\[ I_{\mu_m(d)} = \frac{\sigma_m(d)}{\sigma_m(d)^2} \]  
\[ I_{\sigma_m(d)} = \frac{\pi_m}{\sigma_m(d)} \].

According to Eq. (5) and (11)–(13), the final normalized Fisher score vector is

\[ F_z = diag(\left[ I_{\mu_m(d)}, I_{\mu_m(d)}, I_{\sigma_m(d)} \right])^{\frac{1}{2}} \begin{bmatrix} \partial L(x_t; \theta^{(x)}) / \partial \pi_m(d); \partial L(x_t; \theta^{(x)}) / \partial \mu_m(d); \partial L(x_t; \theta^{(x)}) / \partial \sigma_m(d) \\ I_{\pi_m}^{\frac{1}{2}} \partial L(x_t; \theta^{(x)}) / \partial \pi_m(d); I_{\mu_m(d)}^{\frac{1}{2}} \partial L(x_t; \theta^{(x)}) / \partial \mu_m(d); I_{\sigma_m(d)}^{\frac{1}{2}} \partial L(x_t; \theta^{(x)}) / \partial \sigma_m(d) \end{bmatrix} \].

Once the interlocutor feature vector \( x_t \) has been transformed into \( F_z \), we apply a discriminative regression model to predict the corresponding target behavior \( y_t \). In our work, we use Support Vector Regression (SVR) as the regression model [47] where the objective is to minimize the soft margin \( L_{\gamma}\)-loss.

VIII. EXPERIMENTS AND RESULTS

In this section, we present our experimental results of predicting body language of an interaction participant from the interlocutor’s body language and speech (if applicable). The prediction performance of each target body language feature marked with * in Table II is examined in all interaction conditions. In our experiments, we use 4-fold cross-validation by randomly leaving interactions out, and the cross-validation performance is averaged over all the 4 folds. On average, we use about 5040000 frames for training and 1680000 frames for testing. For the MLE-based mapping, we use full covariance GMMs which perform better compared to the diagonal ones. We apply the diagonal covariance GMMs to the Fisher kernel based approach to simplify the gradient derivation procedure, as described in Section VII-C.

To examine the effectiveness of the interlocutor’s speech features in predicting the participant’s body language, we compare the performance with visual (the interlocutor’s body language behavior) only information, and with audio-visual (the interlocutor’s body language and speech behavior) information. For each target body language feature and for each goal pair, we compute the interlocutor’s context information by selecting the top 6, 12, 18, 24 visual and the top 6, 12 audio interlocutor features respectively, according to the list of ranked features in Section VI. When we use only visual features for prediction, the set of 12 visual features performs the best. When we use audio-visual information for prediction, the combination of 12 visual features and 6 audio features provides the best performance, where the audio and visual features are fused by concatenation at the feature level.

We use two measures to evaluate the prediction performance: correlation coefficient \( \rho \) to assess how well the trend of body language trajectories is captured, and \( RMSE \) to examine how well the actual values are estimated.

\[ \rho = \frac{\sum_{t=1}^{T} (x_t - \overline{x})(\hat{x}_t - \overline{x})}{\sqrt{\sum_{t=1}^{T} (x_t - \overline{x})^2} \sqrt{\sum_{t=1}^{T} (\hat{x}_t - \overline{x})^2}}, \]

\[ RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (x_t - \hat{x}_t)^2}, \]

where \( T \) is the number of frames, \( x_t \) is the observed body language value at time \( t \), \( \hat{x}_t \) is the estimated value, and \( \overline{x}, \overline{x} \) are the average of the observed and estimated values, respectively. A higher \( \rho \) value indicates that the predicted body language curve follows better the trend of the observed trajectory, while a lower \( RMSE \) value tells that the estimated values are closer to the observed ones.

a) Prediction Performance: Fig. 4 and 5 present the prediction performance evaluated by correlation and \( RMSE \), respectively, under different interaction conditions. In each interaction case, we compare the performance of the Fisher kernel based, MLE-based and the basic SVR approaches using the audio-visual interlocutor features. The x-axis in these two figures represents the eight target features which have been selected in Section IV-A, i.e., cosBody, cosFace, cosLean, aVarml, rVbody, rHands, aVbody and aVarmr. The y-axis represents the correlation value in Fig. 4 and the \( RMSE \) value in Fig. 5 respectively.

In majority of interaction cases, the Fisher kernel based method generally achieves better performance than the other two approaches. For example, the correlation coefficient \( \rho \) and \( RMSE \) for \( rVbody \) in the friendly case are about 0.54 and 0.70 respectively by employing the MLE-based method, and are about 0.41 and 0.52 with the SVR model. The performance improves to 0.72 (increase in \( \rho \)) and 0.44 (decrease in \( RMSE \)) with the Fisher kernel based approach. We can further observe that MLE-based approach has a higher correlation performance than the SVR model, while the SVR model generally outperforms the MLE-based approach in terms of \( RMSE \). The results show that the Fisher kernel based method is helpful in both tracking the trends while simultaneously reducing the difference of actual values between the predicted and observed curves. This results from the fact that it leverages the advantages of both the generative and discriminative approaches.

In general, the velocity features of arms and body, i.e., aVarml, rVbody, aVbody and aVarmr, are better predicted than the other features in all the conditions. A potential application of this result to future animation work would be to use the predicted velocity values as constraints for body motion animation of a virtual agent. This could make the animation more consistent with the desired interaction type. We can also observe that the friendly case has a better performance of predicting the orientation features, i.e., cosBody, cosLean and cosFace, in terms of both correlation and \( RMSE \). The relatively higher predictability of orientation body language in the friendly case indicates that people with friendly attitudes tend to adapt more to the behavior of their interlocutors.
Fig. 4. Correlation $\rho$, the higher the better, from Fisher-kernel-based, MLE-based, and SVR prediction using audio-visual information.

also agrees with the observation in [48] that the degree of vocal entrainment is higher for interaction participants with a positive affective state.

In Table V, we compare the performance with visual only and with audio-visual information respectively using the Fisher kernel based method. It is interesting to observe that the fusion of audio and visual information generally improves the performance, compared to the visual only information. For instance, the inclusion of speech information improves the coefficient of determination for collisions from 0.42 to 0.59, while lowering $RMSE$ from 0.6 to 0.42. This observation suggests an interplay between multiple behavioral cues in interactions, such as the coordination between speech and body language. Similar observations were obtained for the other two approaches, which are not presented here for the lack of space.

b) The Effectiveness of GMM Parameters: Next, we investigate the contribution of each type of GMM parameters, i.e., weights, component means and standard deviations, to the final prediction performance. Audio-visual information is used for analysis. We find that the performance is the lowest when using the gradient only with respect to the weights compared to the performance when using gradients with respect to the means or standard deviations. The combination of gradients with respect to the means and standard deviations provides similar performance to using all the parameters. The small benefit from the weights is probably due to the fact that they capture much less information regarding the data distribution than the other two types of parameters.

C) Qualitative Results: Example trajectories of some body language features ($rVbody$, $cosBody$, $aVarml$, and $dHands$) estimated by different methods are shown in Fig. 6. Compared to the curves estimated by the SVR model, the trajectories from the other two approaches generally better follow the trend of the directly observed ones. For example, in Fig. 6(a), both MLE-based and Fisher kernel based methods capture the valley of $rVbody$ well in the first 200 frames. Moreover, the estimated feature values from the Fisher kernel based mapping are closer to the observed ones. As observed in Fig. 6(d), the Fisher kernel based estimations are almost at the same level with the directly measured values throughout, while the predictions from the other two methods mostly shift a lot from the directly observed values. In the MLE-based mapping, the estimated feature values tend to approach the mean vector of the Gaussian mixture components [31], which might be the reason for the greater difference between the exact values of the estimated and observed trajectories. Furthermore, the Fisher kernel based approach can also better capture the dynamic changes of the body language trajectories.

In addition, we investigate the correlation and $RMSE$ as a function of time for the three types of mapping given an interaction. To compute the correlation or $RMSE$ at frame $t$ of an interaction, we set a window (1 sec) centered at $t$ respectively for the observed and estimated curves, and calculate the correlation $\rho$ or $RMSE$ between the two sets of samples within the window. Fig. 7 presents the correlation and $RMSE$ curves over time for the $dHands$ example in Fig. 6(d). As can be observed in Fig. 7(a), the Fisher kernel based method generally achieves the highest correlation throughout except for a few valleys in the middle. As shown in Fig. 6(d), compared to the flatness of the observed curve in the middle region (from frame 300 to 500), the Fisher kernel based method brings more variations to the predictions, resulting in the correlation valleys in Fig. 7(a). Moreover,
we can easily observe from Fig. 7(b) that the $RMSE$ curve from the Fisher kernel-based method is significantly lower than those of the other two approaches.

**IX. Conclusion and Future Work**

In this paper, we study how an interaction participant adapts his/her body language behavior to the multimodal cues of the interlocutor, under different interaction goals. We adopt a variety of psychology-inspired features to describe the body language behavior of participants, including body motion, posture and relative orientation [3]. Our work focus is two-fold: 1) to systematically examine the coordination between the participant’s body language and the interlocutor’s multimodal speech and body language behavior for the interaction stances of friendliness and conflict; 2) to automatically predict an interaction participant’s body language from the multimodal behavior of the interlocutor for specific interaction goals. The coordination analysis empirically corroborated the correlation between the dyad’s behavior. It also showed that the interlocutor features that are especially informative of the participant’s target body language features differ across interaction types. This implies that the coordination patterns depend on the overarching interaction goals. Further, we proposed a methodology that combines Gaussian Mixture Model (GMM) based statistical mapping, and Fisher kernels, in order to automatically predict body language of an interaction participant from the multimodal speech and gesture cues of the interlocutor. The experimental results showed that the Fisher kernel-based approach generally outperforms the generative model of MLE-based mapping and the basic SVR model in terms of both correlation coefficient and $RMSE$ computed based on the ground truth. These promising results suggest that there is a significant level of predictability of body language in the examined goal-driven improvisations, which could be exploited for interaction-driven and goal-driven body language synthesis.

Experimental results also showed that the predictability of orientation body language is lower in the conflictive interactions.
than in the friendly ones. It is possible that conflictive behavior cues are inherently more self-initiated whereas friendly ones are more synchronized with the interlocutor cues. We also observed that the participant’s body language is influenced by other behavioral cues than merely the body language. This was reflected in our experimental results, where the use of both speech and body language of the interlocutor improved prediction performance over using only body language. Other cues, such as facial expressions, could also be an indicator of the interlocutor’s behavior. A further detailed study of the interplay between the various multimodal cues is a topic for future work.

Also, in the future, this work could be combined with more traditional synthesis approaches that are based on discrete body language units. For example, the interlocutor multimodal cues could be used to generate constraints on a virtual agent’s body language during data-driven, unit-based animation. In particular, the predicted velocities from the interlocutor information can be used as constraints for body motion animation for a virtual agent, which could make the animation more consistent with the desired interaction type. We are currently studying how interaction attitudes are conveyed by hand gesture dynamics based on data-driven gesture units [49]. The results in [49] demonstrate that the unit-based hand gesture dynamics can well discriminate interaction attitudes and that such derived gesture units can be used for gesture synthesis with attitudes. Our long-term goal is to work towards interaction-driven and goal-driven body language synthesis, e.g., creating intelligent virtual characters with specific interaction goals which would display expressive body language in response to the human user’s audio-visual input behavior.
Zhaojun Yang (S’13) received the B.E. degree in electrical engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2009, and the M.Phil. degree in systems engineering and engineering management from the Chinese University of Hong Kong (CUHK), Hong Kong, in 2011. She is currently working toward the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, CA, USA. Her research interests include multimodal emotion recognition and analysis, interaction modeling, and spoken dialog systems.

Angeliki Metallinou received the Diploma in electrical and computer engineering from the National Technical University of Athens, Athens, Greece, in 2007, and the M.Eng. and Ph.D. degrees in electrical engineering from the University of Southern California (USC), Los Angeles, CA, USA, in 2009 and 2013, respectively. Between 2007 and 2013, she was a member of the Signal Analysis and Interpretation Lab (SAIL) at USC, Los Angeles, CA, USA, working on spoken and multimodal emotion recognition and analysis, as well as computational approaches for healthcare applications, including autism. She also interned at Microsoft Research, Redmond, WA, USA, working on state tracking for spoken dialog systems. She is currently a Research Scientist with Pearson Knowledge Technologies, Menlo Park, CA, USA, working on automatic speech recognition, spoken language assessment, and written essay grading for education applications. Her research interests include speech and multimodal signal processing, automatic speech recognition, affective computing, machine learning, natural language processing, and dialog systems.

Shrikanth (Shri) Narayanan (S’88–M’95–SM’02–F’09) was with AT&T Labs Research, Florham Park, NJ, USA, and AT&T Bell Labs, Murray Hill, NJ, USA, from 1995 to 2000. He is currently Professor of Engineering at the University of Southern California (USC), Los Angeles, CA, USA, where he directs the Signal Analysis and Interpretation Laboratory (SAIL). He holds appointments as Professor of Electrical Engineering, Computer Science, Linguistics, and Psychology, and as the Founding Director of the Ming Hsieh Institute, Ming Hsieh Department of Electrical Engineering, USC, Los Angeles, CA, USA. He has published over 600 papers and has been granted 16 U.S. patents. His research focuses on human-centered signal and information processing and systems modeling with an interdisciplinary emphasis on speech, audio, language, multimodal and biomedical problems, and applications with direct societal relevance.