# TOWARD BODY LANGUAGE GENERATION IN DYADIC INTERACTION SETTINGS FROM INTERLOCUTOR MULTIMODAL CUES

*Zhaojun Yang, Angeliki Metallinou and Shrikanth Narayanan*

Signal Analysis and Interpretation Lab (SAIL), University of Southern California, Los Angeles, CA

zhaojuny@usc.edu, metallin@usc.edu, shri@sipi.usc.edu

**ABSTRACT**

During dyadic interactions, participants influence each other's verbal and nonverbal behaviors. In this paper, we examine the coordination between a dyad's body language behavior, such as body motion, posture and relative orientation, given the participants' communication goals, e.g., friendly or conflictive, in improvised interactions. We further describe a Gaussian Mixture Model (GMM) based statistical methodology for automatically generating body language of a listener from speech and gesture cues of a speaker. The experimental results show that automatically generated body language trajectories generally follow the trends of observed trajectories, especially for velocities of body and arms, and that the use of speech information improves prediction performance. These 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 generation.

***Index Terms***— body language generation, motion capture, speech, dyadic interactions, communication goals

## 1. INTRODUCTION

The coordination of verbal and nonverbal human behavior during interactions has been studied in many diverse areas including neuroscience, engineering and psychology. Participants generally adjust their behavior and give feedback, for example through facial expressions, body movements and speech prosody, based on the behavior of their interlocutors as well as their own communication goals. In addition to understanding the fine details of human interaction mechanisms, modeling such behavior coordination is important for developing more natural human-machine interfaces.

Body language is an important element of nonverbal behavior conveying attitudes and emotions in human communication, including gestures, body postures, etc. In interactions, participants may express body language differently depending on their communication goals and attitudes, leading to different coordination patterns of a dyad's body language. The goal of our work is two-fold: first to examine how the body language of a listener and a speaker are related in an interaction given their communication goals (friendly or conflictive); second to automatically generate a listener's body language from the body language and speech information of a speaker in dyadic interaction settings of friendly or conflictive nature.

In this work, we use the multimodal USC CreativeIT database that consists of goal-driven improvised interactions [1]. It contains detailed full body Motion Capture (MoCap) data of both participants, providing a rich resource for studying and generating body language. We extract various body language features including head and body position, hand and body motion, and relative orientation. The dyad's body language correlations are analyzed using canonical correlation analysis (CCA), for the interaction types of friendliness and conflict. We observe that correlation patterns depend on the interaction types, and also find statistically significant correlations that empirically verify the dyad's body language coordination.

Motivated by these analyses, we propose a method for automatically predicting the listener's body language using multimodal information derived from the speaker, conditioning on the dyad's communication goals. For this purpose, we use a Gaussian Mixture Model (GMM) based approach that estimates a statistical mapping from the speaker's body language and speech cues to the listener's body language. Experimental results show that generated body language trajectories generally follow the trends of observed trajectories, especially for velocities of body and arms, and that inclusion of speech information generally improves performance. These results indicate the existence of a significant level of predictability of body language in the examined improvisations, which could be exploited towards interaction-driven and goal-driven body language generation. The generation of a subject's body language based only on the interlocutor's multimodal information is a novel and unexplored research direction. This direction could lead to the development of expressive and goal-driven virtual agents that would display body language consistent with their own communication attitudes, and in response to the user's audio-visual input.

## 2. RELATED WORK

The correlation between body language and speech features, has enabled much research progress in prosody-driven body gesture synthesis. Busso et al. have proposed a prosody-driven approach for synthesizing expressive rigid head motion [2, 3]. Mariooryad et al. built a joint speech and facial gesture model to generate head and eyebrow motion for conversational agents [4]. Likewise, Sargin et al. synthesized head gestures from speech by exploring the joint correlation between head gesture and prosody patterns [5]. Frameworks for full body motion synthesis, in real time, that also use prosody information, are developed in [6, 7]. While these works mainly focus on synthesizing body language of individuals, this paper aims at interaction-driven body language generation in dyadic settings.

Chartrand et al. described that humans unconsciously mimic the behavior of their interaction partners to achieve more effective and pleasant interactions [8]. Ekman found that body language of interviewees is distinctly different between friendly and hostile job interviews [9]. Many engineering works are based on this mutual influence of interlocutors. Morency et al. predicted head nods for

virtual agents from the audio-visual information of a human speaker based on sequential probabilistic models [10]. Researchers have also used the emotional state of an interlocutor to inform that of a speaker by modeling emotional dynamics between two participants [11, 12]. The influence model proposed in [13] models participants in conversational settings as interacting Markov chains. Lee et al. proposed prosody-based computational entrainment measures to assess the coordination in married couples' interactions [14].

In this work, we apply a GMM-based statistical mapping methodology for body language generation, which was originally proposed for articulatory to acoustic mapping [15] and spectral conversion between speakers [16]. It has also been used for continuous emotional state estimation based on body language and speech [17].

## 3. DATABASE DESCRIPTION

We use the CreativeIT database, which is a multimodal database of dyadic theatrical improvisations [1]. It contains detailed full body Motion Capture (MoCap) data of both participants (recorded at 60 fps), as shown in Fig. 1(a), and speech information. The interactions are either improvisations of theatrical plays or theatrical exercises. Here 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. The interactions were guided by a theater expert (professor/director), and were performed following the Active Analysis improvisation technique [18]. According to this technique, the interactions are goal-driven; actors have predefined goals that they try to achieve through the appropriate use of body language and speech prosody. The goal pair of each dyad defines the attitudes of the interlocutors towards each other and the content of the interaction.

Our premise is that the dynamics of the interactions differ depending on the participant's goals. Hence, we group the interactions into 3 cases: friendliness, medium conflict and high conflict, as defined by the goal pairs. In friendly interactions both participants have friendly goals, in medium conflict interactions one participant is friendly while the other is creating conflict, and in high conflict interactions both participants' goals are conflictive. Interactions that do not fit into these categories are excluded from our analysis. This grouping is described in Table 1, along with examples of characteristic goal pairs. Friendliness, medium and high conflict groups contain 10, 30 and 8 interactions respectively, including 16 actors (8 female). Each interaction has an average length of 3 min and contains on average 35 sentences per actor.

**Table 1**. Friendly, and Medium and High Conflict Cases

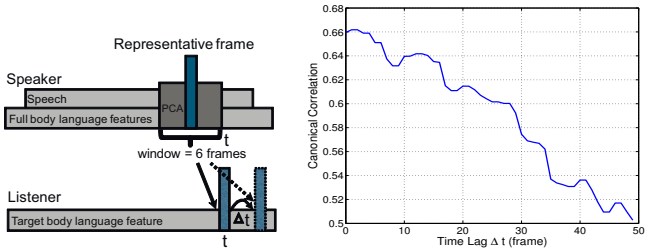
| Cases | Actors | Example goal pairs |
|---|---|---|
| Friendly | friendly - friendly | to make peace - to comfort |
| Medium Conflict | friendly - conflictive | to convince - to reject |
| High Conflict | conflictive - conflictive | to accuse - to fight back |

## 4. BODY LANGUAGE AND SPEECH FEATURES

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. Our features, presented in Fig. 1(c), are motivated from the psychology literature which indicates that body language behaviors, such as looking at the other or turning away, approaching, hand gesturing, etc, are informative of a subject's attitude towards his/her interlocutor [19]. While some of our features are informative of a subject's individual posture and motion, others describe relative behaviors towards the interlocutor, e.g., body and head orientation, approaching vs moving away, etc. The features are geometrical and are computed in a straightforward manner by defining global and local coordinate systems and by computing Euclidean distances and relative positions, angles and velocities, as illustrated in Fig. 1(b) (see [17] for more information). For example, if the cosine of the face angle ($cosFace$) shown in Fig. 1(b) is close to 1, the actor is looking towards the interlocutor, while $cosFace < 0$ indicates looking away. Features marked with * in Fig. 1(c) are the target features which will be analyzed and generated using the interlocutor's feature set. We also extract standard speech features from the speaker's sentences, i.e., pitch, energy and MFCCs.

## 5. ANALYSIS AND GENERATION SETUP

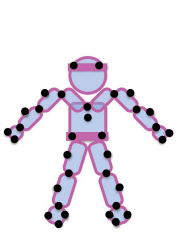
Our objective is to analyze how a listener's body language is influenced as a response to a speaker's speech and body language. In the friendliness and high conflict cases, both actors have the same goal, therefore a friendly (conflictive) listener is analyzed (or generated) with respect to a friendly (conflictive) speaker. However, the medium conflict case includes two subcategories for analysis and generation: a friendly listener with respect to a conflictive speaker, and a conflictive listener with respect to a friendly speaker.



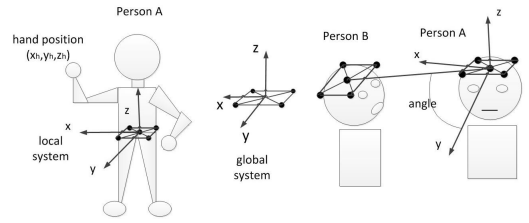

(a) (b)

**Fig. 2**. Canonical correlation analysis of the target body language feature and source features. (a) Illustration of analysis window. (b) Canonical correlation of the listener's delayed $cosLean$ with the speaker's body language features varying with time lag.

Our analysis and generation experiments are performed around the sentences in an interaction. We split each interaction into segments according to the sentences, and for each sentence we extract features for speakers and listeners (short segments are increased to 50 frames, by including frames surrounding the sentence). All the extracted features are z-normalized into the same scale. Each target feature of the listener at frame $t$ is individually analyzed/generated based on the speaker's information during a window preceding (including) frame $t$, as illustrated in Fig. 2(a). Specifically, the speaker's information consists of the speaker features at the representative frame, which is the center frame of the window, concatenated with statistical functional information of the speaker features over the window. We extract 11 statistical functionals for each speaker body language feature, such as mean, standard deviation, median, minimum, maximum, range, skewness, etc. The dimensionality is then reduced by Principal Component Analysis (PCA), by keeping only the first 10 components (preserving about 90% of the total variance). This functional information is included to provide context of the body language around the representative frame. We empirically determined a window size of 6 frames (0.1 sec), where the representative frame is the third frame of the window.



(a) Body Markers




(b) Global and Local Positions, Face orientation angle

| | |
|---|---|
| cosFace* | Cosine of subject's face angle towards Other (Fig 1(b)) |
| cosBody* | Cosine of subject's body orientation angle towards Other |
| cosLean* | Cosine of subject's body leaning angle towards Other |
| px, py, pz | Subject's (x,y,z) coord. of body position (global, Fig 1(b)) |
| absVbody* | Absolute velocity of subject's body |
| absVarmr,l* | Absolute velocity of subject's right/left arms |
| absVfeetr,l | Absolute velocity of subject's right/left feet |
| relVbody* | Relative velocity of subject's body towards Other |
| relVhandr,l | Relative velocity of subject's right/left hands towards Other |
| relVfeetr,l | Relative velocity of subject's right/left feet towards Other |
| rhandx,y,z | Subject's right hand (x,y,z) coord position (local, Fig 1(b)) |
| lhandx,y,z | Subject's left hand (x,y,z) coord position (local, Fig 1(b)) |
| dHands | Distance between subject's right and left hands |

(c) Extracted Body Language Features

**Fig. 1**. Body Language Feature Extraction from MoCap. Features marked with * will be generated using the interlocutor's feature set.

## 6. ANALYSIS OF THE DYAD'S BODY LANGUAGE

We perform Canonical Correlation Analysis (CCA) between each individual body language feature of the listener at frame $t$ and the full body language features of the speaker at the representative frame, for each of the 3 cases of goal pairs in Table 1. In this experiment, we do not include the speaker PCA information over the window, to increase analysis interpretability. CCA finds two subspaces in which the projections of two sets of data with different dimensionalities are maximally correlated [20]. By examining the magnitude of the canonical weight assigned to each normalized body language feature, we can assess their contribution to the projection onto the corresponding optimal subspace. In our case, the speaker's body language features with larger magnitudes of weights are more informative regarding the target body language feature of the listener. Hence, we select a subset of informative body language features for each target feature based on this analysis. These selected features will be later used for body language generation in Section 7.

Overall, we notice that the selected features differ for each case of goal pairs, suggesting different coordination patterns for different interaction types. Specifically, for friendly situations we select many hand position features, suggesting the expressiveness of hand gestures in these interactions. In contrast, many relative and absolute velocities, and face orientation features are found informative in medium and high conflict cases, indicating more body motion and approach-avoidance behaviors. For example, Table 2 presents the top-5 selected speaker features for the listener's target features in the medium conflict case (conflictive speaker, friendly listener).

**Table 2**. Top-5 selected features in medium conflict case (conflictive speaker, friendly listener). Minus sign indicates a negative relation.

| Target Listener Feat. | Selected Speaker Features (top-5) |
|---|---|
| cosBody | relVbody (−), cosFace, absVfeetl (−), absVfeetr (−), rhandz |
| cosLean | cosLean (−), dHands, pz, rhandx (−), absVbody (−) |
| cosFace | cosFace (−), absVarmr, relVhandl, rhandz (−), absVbody (−) |
| absVbody | absVbody, lhandx (−), absVarmr, dHands (−), lhandy |
| absVarmr | absVarmr, absVbody, lhandx (−), absVarml, relVhandl (−) |
| relVbody | relVbody, cosBody, rhandz, lhandx (−), cosFace (−) |

In addition, we conduct CCA between each individual target body language feature of the listener and the full body language features of the speaker, including the speaker PCA information this time. CCA is performed both separately for the 3 cases of goal pairs, and by considering all cases together. Analysis results show that canonical correlations for each case separately are higher than those over all cases, which reinforces our hypothesis that conditioning on the goal-pairs reduces some of the dyad's coordination variability and helps us focus on body language coordination patterns that are related to the dyad's communication goals. Furthermore, canonical correlation for each target body language feature in each case is greater than $0.50$ ($p < 0.01$), implying statistically significant coordination of the dyad's body language in communication.

During an interaction, one person's movement in the recent past may still influence the interlocutor's movement at current time. To investigate the effect of this reaction lag, we delay the target frame by certain time lag $\Delta t$ (see Fig. 2(a)) and examine how the correlation of a dyad's body language varies in terms of $\Delta t$. Results show that the correlation generally decreases with increasing time lag. For instance, Fig. 2(b) shows the correlation between a listener's delayed cosine of leaning angle ($cosLean$) and a speaker's body language features with varying $\Delta t$ in the friendliness case.

## 7. BODY LANGUAGE GENERATION

The coordination of a dyad's body language implies a certain level of predictability of body language in our goal-driven interactions, hence it is possible to generate a listener's body language based on the speaker's audio-visual information. For body language generation, we apply a GMM-based approach that estimates an optimal statistical mapping, using Maximum Likelihood Estimation (MLE), from a set of observed continuous random variables, here a speaker's audio-visual features, to a target continuous variable, here a listener's body language feature. This method was originally presented for the problem of articulatory to acoustic mapping [15].

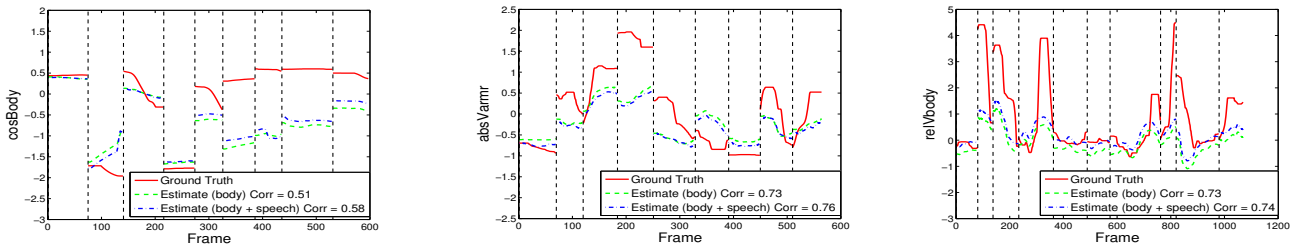
Let $x_t$ be the listener's target body language feature and $\mathbf{y_t}$ be the speaker's feature vector at time $t$. We train a GMM to model the joint distribution of $x_t$ and $\mathbf{y_t}$: $P(x_t, \mathbf{y_t}|\lambda^{(x,y)}) = \sum_{m=1}^{M} a_m N(x_t, \mathbf{y_t}; \mu_m^{(x,y)}, \Sigma_m^{(x,y)})$, where $a_m$, $\mu_m^{(x,y)}$ and $\Sigma_m^{(x,y)}$ are the weights, means and covariance matrices of the $m$-th component. The conditional distribution of the target feature $x_t$ given the observation $\mathbf{y_t}$ is also represented as a GMM:

$$P(x_t|\mathbf{y_t}, \lambda^{(x,y)}) = \sum_{m=1}^{M} P(m|\mathbf{y_t}, \lambda^{(x,y)}) P(x_t|\mathbf{y_t}, m, \lambda^{(x,y)}),$$

where $P(x_t|\mathbf{y_t}, m, \lambda^{(x,y)})$ is the conditional distribution of the $m$-th component and $P(m|\mathbf{y_t}, \lambda^{(x,y)})$ is the so-called occupancy probability. For each test recording, we estimate the target feature given the observed features by maximizing the conditional probability model: $\hat{x}_t = \arg\max_{x_t} P(x_t|\mathbf{y_t}, \lambda^{(x,y)})$. The maximization is done through the EM procedure with minimum mean squared error (MMSE) estimate as the initial value [15]. To incorporate dynamic

**Table 3**. Median correlation between generated and observed values of the target features of the listener.

| | **cosBody** | **cosLean** | **cosFace** | **absVbody** | **absVarmr** | **relVbody** | **cosBody** | **cosLean** | **cosFace** | **absVbody** | **absVarmr** | **relVbody** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Friendly (Friendly Speaker, Friendly Listener) | | | | | | Medium Conflict (Friendly Speaker, Conflictive Listener) | | | | | |
| Body | 0.27 | 0.26 | 0.18 | 0.44 | 0.43 | 0.45 | 0.29 | 0.20 | 0.30 | 0.49 | 0.52 | 0.41 |
| Speech + Body | 0.42 | 0.28 | 0.24 | 0.49 | 0.48 | 0.54 | 0.41 | 0.24 | 0.32 | 0.58 | 0.56 | 0.48 |
| | High conflict (Conflictive Speaker, Conflictive Listener) | | | | | | Medium Conflict (Conflictive Speaker, Friendly Listener) | | | | | |
| Body | 0.09 | 0.07 | 0.07 | 0.48 | 0.37 | 0.28 | 0.10 | 0.21 | 0.13 | 0.55 | 0.53 | 0.53 |
| Speech + Body | 0.18 | 0.15 | 0.18 | 0.51 | 0.54 | 0.35 | 0.24 | 0.23 | 0.20 | 0.57 | 0.56 | 0.51 |




(a) Med. conflict interaction (conflictive listener). (b) Med. conflict interaction (friendly listener). (c) Friendly interaction.

**Fig. 3**. Examples of generated body language curves using MLE mapping from visual and from audio-visual information.

information, we augment $x_t$ and $\mathbf{y}_t$ with corresponding derivative estimates [15]. By using the information from the neighborhood, this approach allows us to exploit the continuous nature of our observation and target features, and provides us with relatively smooth trajectory estimates of the listener's body language.

We use an implementation of this mapping developed for [17] and available in [21]. Our experiment setup follows Section 5. GMMs are trained using the listener's body language feature $x_t$ and the speaker's feature vector $\mathbf{y}_t$. Specifically $\mathbf{y}_t$ contains the top-10 selected body language features of the representative frame, plus PCA information extracted from the window, as described in Section 5. We first train a visual GMM based on $x_t$ and the visual vector $\mathbf{y}_t$. When including speech information, $\mathbf{y}_t$ contains the top-10 body language and top-5 speech features, along with the PCA coefficients (the top speech features are derived through CCA, similarly to the body language features). We then train an audio-visual GMM based on $x_t$ and the audio-visual vector $\mathbf{y}_t$. In the generation stage, body language of the listener is generated either from body language of the speaker, using the visual only GMM, or from body language and speech of the speaker, using the audio-visual GMM. For the increased body language frames surrounding a sentence, which do not contain speech (see Section 5), we use the visual GMM.

## 8. EXPERIMENTS AND RESULTS

In our experiments, we use 4-fold cross validation by randomly leaving interactions out for each target body language feature in each interaction condition respectively. On average, we use about 25000 frames for training and 8300 frames for testing. We are mostly interested in capturing the shape/trends of body language trajectories rather than their exact values. Hence, to evaluate the quality of body language generation, we use Pearson's correlation between predicted and observed values of the body language feature over an interaction.

Table 3 presents the median correlations of generated and observed values of target body language features, when using just body language features and when using both body language and speech features. We can observe that generated body language is positively correlated with the observed values ($p < 0.01$). In particular, velocities of body and arms are better generated compared to orientations of body and face in each case. Moreover, the performance on orientations of body and face is better in the cases of friendly speakers. In addition, inclusion of speech information generally improves generation performance of body language features. For example, speech cues improve the correlation between the MLE estimate and ground truth of the right arm absolute velocity ($absVarmr$) in the high conflict case. These results suggest that there is a significant level of predictability of body language in our goal-driven improvisations.

In Fig. 3, we present example curves of generated body language features from body language only and from both body language and speech. Since we generate body language for each sentence segment, the estimated trajectories for an interaction are discontinuous and the vertical dotted lines indicate boundaries between sentences. We can observe that our generated body language curves from both visual cues and audio-visual cues generally capture the trends of the observed curves, especially for velocities of body and arms, although the exact values of generated and observed body language are different in some cases. We also notice differences in the dynamics of different body language features, e.g., body orientation changes slower compared to velocities, an issue that potentially requires further investigation and modeling.

## 9. CONCLUSIONS AND FUTURE WORK

In this paper, we presented an analysis of the coordination between the body language of dyads in improvised interactions with various levels of interaction conflict, as defined by the dyad's communication goals. We also used a GMM-based statistical methodology for generating the body language of a listener in an interaction from multimodal cues of a speaker given their communication goals. The experimental results suggest a significant level of predictability of body language in the examined goal-driven improvisations, which can be exploited for body language generation.

In the future, it would be interesting to examine to what extent this methodology generalizes in less constrained interactions. We would also like to combine this work with more traditional generation approaches that are based on discrete body language units, e.g., [6]. Our long-term goal is to work towards interaction-driven and goal-driven body language synthesis, e.g., creating virtual characters with specific communication goals which would display expressive body language in response to the user's audio-visual information.

## 10. REFERENCES


[1] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database: A multimodal database of theatrical improvisation," in *Proc. of Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC)*, 2010.

[2] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1075–1086, 2007.

[3] S. Marioooryad and C. Busso, "Generating human-like behaviors using joint, speech-driven models for conversational agents," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2329–2340, 2012.

[4] C. Busso, Z. Deng, U. Neumann, and S. Narayanan, "Natural head motion synthesis driven by acoustic prosodic features," *Journal of Computer Animation and Virtual Worlds*, vol. 16, no. 3-4, pp. 283–290, 2005.

[5] M.E. Sargin, E. Erzin, Y. Yemez, A.M. Tekalp, A.T. Erdem, C. Erdem, and M. Ozkan, "Prosody-driven head-gesture animation," in *Proc. of ICASSP*, 2007.

[6] S. Levine, C. Theobalt, and V. Koltun, "Real-time prosody-driven synthesis of body language," *ACM Transactions on Graphics (TOG)*, vol. 28, no. 5, pp. 172, 2009.

[7] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun, "Gesture controllers," *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4, pp. 124, 2010.

[8] T. L. Chartrand and J. A. Bargh, "The chameleon effect: The perception-behavior link and social interaction," *Journal of Personality and Social Psychology*, vol. 76, pp. 893–910, 1999.

[9] P. Ekman, "Body position, facial expression and verbal behavior during interviews," *Journal of Abnormal and Social Psychology*, vol. 63, 1964.

[10] L.P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Journal of Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 70–84, 2010.

[11] A. Metallinou, A. Katsamanis, and S. Narayanan, "A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs," in *Proc. of ICASSP*, 2012.

[12] C.C. Lee, C. Busso, S. Lee, and S. Narayanan, "Influence of interlocutor emotion states in dyadic spoken interactions," in *Proc. of InterSpeech*, 2009.

[13] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland, "Learning human interactions with the influence model," MIT Media Lab: Cambridge, MA., 2001.

[14] C.C. Lee, A. Katsamanis, M.P. Black, B. Baucom, P.G. Georgiou, and S.S. Narayanan, "An analysis of PCA-based vocal entrainment measures in married couples' affective spoken interactions," in *Proc. of InterSpeech*, 2011.

[15] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, pp. 215–227, 2008.

[16] T. Toda, A.W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 2222–2235, 2007.

[17] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing, Special Issue on Continuous Affect Analysis, in press*, 2012.

[18] S. M. Carnicke, *Stanislavsky in Focus: An Acting Master for the Twenty-First Century*, Routledge, UK, 2008.

[19] J.A. Harrigan, R. Rosenthal, and K.R. Scherer, *The new handbook of Methods in Nonverbal Behavior Research*, Oxford Univ. Press, 2005.

[20] F.R. Bach and M.I. Jordan, "A probabilistic interpretation of canonical correlation analysis," *Technical Report 688, Department of Statistics, University of California, Berkeley*, 2005.

[21] A. Metallinou, "Emotion Tracking Code (GMM mapping)," `http://sail.usc.edu/~metallin/data_code.html`.