MODELING MUTUAL INFLUENCE OF MULTIMODAL BEHAVIOR IN AFFECTIVE DYADIC INTERACTIONS

Zhaojun Yang and Shrikanth Narayanan

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

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

ABSTRACT

To accomplish effective communication, interaction partners generally adapt their verbal and non-verbal behavior to that of their interlocutors. This behavior adaptation is often modulated by the underlying emotional states of partners. Modeling such mutual behavioral influence is critical for emotion characterization in an interaction. In this paper, we focus on explicitly modeling the mutual influence of multimodal behavior (speech and hand gesture) in affective dyadic interactions. In our framework, the behavior adaptation in each interaction is modeled by an *interaction matrix* which assembles all behavioral information on the path between the dyad's behavior. Experimental results show that our modeling approach can significantly improve the performance of emotion recognition. We further investigate the properties of the interaction model. Analysis results reveal that the entrainment effect of dyad's behavior can be better embodied by interaction modeling, and that the interaction patterns captured by interaction matrices are dependent on the emotional states of interaction partners. These observations corroborate the validity of our interaction model for capturing emotion-dependent mutual influence of dyad's interaction behavior.

Index Terms— Multimodal behavior, emotion recognition, behavior entrainment, interaction modeling, mutual influence

1. INTRODUCTION

During social interactions, interpersonal influence is naturally induced along aspects of spoken words, speech prosody, body gestures and emotional states. To accomplish effective communication, individuals generally adapt their verbal and non-verbal behavior to that of their interlocutors. Such behavior adaptation, also known as entrainment or coordination, includes synchronizing in time or displaying similar or dissimilar behavior [1]. The mutual behavior effect controls the dynamic flow of a conversation and describes the overall interaction patterns. Understanding human interaction mechanisms and computationally modeling interaction dynamics of human behavior can bring insights into automating emotion recognition as well as the design of human-machine interfaces.

The entrainment phenomenon in human communication in terms of vocal patterns, head motion, or body gestures has been well-established in both psychology and engineering domains. Levitan et al. have found that interacting partners tend to utilize similar sets of backchannel-preceding cues which are a combination of speech cues of an individual in response to one's interlocutor [2]. The work in [3] has demonstrated a high degree of unintentional coordination between rhythmic limb movements of two partners. Robotics research has shown that human subjects use the robot's cues to regulate conversations and to convey affective intentions, resulting in a smoother interaction with fewer interruptions [4].

Since emotion is one of the major elements influencing multimodal channels of human speech, body gestures, or facial expressions, the interaction patterns of a dyad's behavior are accordingly shaped by the underlying emotional states [1]. For example, two participants with friendly attitudes may tend to approach each other, while those with conflictive attitudes may try to fight with or avoid each other. The analysis in our previous work [5] has empirically revealed that the coordination patterns of a dyad's behavior depend on the interaction stances assumed (e.g., friendly *vs.* conflictive). Lee et al. also investigated the relationship between affective states (positive *vs*. negative) and the vocal entrainment strength in married couples' interactions. A higher degree of vocal entrainment was found for couples with positive attitudes [6]. They further proposed a PCA-based scheme to quantify the turn-wise vocal entrainment for more complex interaction scenarios, and demonstrated its effectiveness for differentiating positive or negative affect [7].

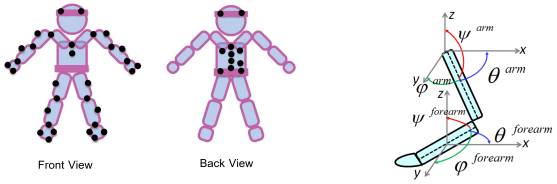
Motivated by these findings on the interrelation between emotions and interpersonal influence, researchers have attempted to model such mutual effect of emotions for assessing the emotional states of individuals [8] [9]. The benefits of incorporating mutual influence into the emotion recognition framework have been validated in these studies. There has also been much research attention on modeling the behavioral interaction for action recognition [10] [11]. Most of these studies have relied on training with statistical models, e.g., coupled HMMs, which usually require significant amounts of data. Mariooryad et al. exploited emotion-related patterns in behavioral interaction [12], which is similar to this work. However, they simply concatenated behavioral information of two partners for assessing the emotional state of an individual without modeling the latent interaction structure of the dyad's behavior.

In this work, we propose an unsupervised subspace-based approach for quantitatively modeling the mutual influence of multimodal behavior (speech and hand gesture) in affective dyadic interactions. Our interaction model is inspired by a geodesic flow-based methodology for the problem of domain adaptation which explores the sharing characteristics of two distinct datasets [13]. In our framework, the behavior adaptation from the interlocutor to the target participant in an interaction is parameterized by a geodesic flow connecting the behavior subspaces of the two interaction partners. The association of the dyad's behavior is then modeled by an *interaction matrix* that is obtained by aggregating all behavior information on the geodesic path. We evaluate the effectiveness of the interaction model in multimodal emotion recognition. Experimental results

show that the recognition performance can be significantly improved by interaction modeling. We further study the properties of the interaction model in the aspects of expressing behavior entrainment and capturing emotion-dependent interaction patterns. Analysis results reveal that the turn-wise entrainment effect of speech or hand gesture becomes more pronounced by interaction modeling, and that interaction patterns embedded in interaction matrices are dependent on the emotional states of interaction partners. These observations corroborate the validity of the interaction model for capturing emotion-related mutual influence of dyad's interaction behavior.

2. DATABASE DESCRIPTION

In this work, we use the USC CreativeIT database for dyadic interaction modeling [14] [15]. It is a multimodal database of dyadic theatrical improvisations performed by pairs of actors. Interactions are goal-driven; actors have predefined goals, e.g., *to comfort* or *to avoid*, which can elicit natural realization of emotions as well as expressive multimodal behavior. There are 50 interactions in total performed by 16 actors (9 female). The audio data of each actor was collected through close-up microphones at 48 kHZ. A Vicon motion capture system with 12 cameras captured the detailed full body Motion Capture (MoCap) data at 60 fps, i.e., the (x, y, z) positions of the 45 markers of each actor, as shown in Fig. 1(a).



(a) Motion Capture Markers. (b) Angles for hand joints.

Fig. 1. (a) The positions of the Motion Capture markers; (b) The illustration of Euler angles for hand joints.

2.1. Gesture and Acoustic Features

This work focuses on multimodal behavior of speech and hand gesture which are highly expressive forms in human communication. We manually mapped the motion data, i.e., the 3D locations of markers, to the angles of different human body joints using MotionBuilder [16]. The joint angles are popular for motion animation [17] [18] and have also been applied for exploring attitude-related gesture dynamics in our previous work [19]. Fig. 1(b) illustrates the Euler angles (θ, ϕ, ψ) of hand joints (arm and forearm) in x, y, z directions. The angles of both right and left hand joints are used as hand gesture features. In addition, we extracted acoustic features of pitch and the rms energy, as well as 12 Mel Frequency Cepstral Coefficients (MFCCs) for each actor. These features were extracted every 16.67 ms (60 fps) with an analysis window length of 30 ms, in order to match with the MoCap frame rate. The pitch features were smoothed and interpolated over the unvoiced/silence regions. We further augment both hand gesture and acoustic features with their $1st$ derivatives to incorporate the temporal dynamics.

2.2. Emotion Labels

The emotional state of each actor was annotated in terms of activation (excited *vs*. calm) and valence (positive *vs*. negative) by three or four annotators. To preserve the continuous flow of body gestures during an improvisation, we annotated time-continuous emotion for each actor throughout the interaction. Annotators used the Feeltrace instrument [20] to time-continuously indicate the emotion attribute value from -1 to 1 for each actor while watching the video recording. More details of the annotation process can be found in [21].

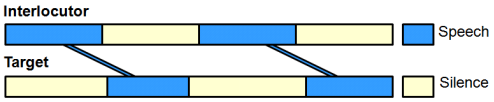
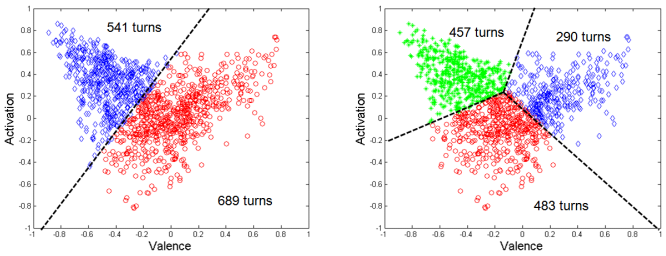


Fig. 2. Illustration of setting up dialog turn pairs. The target dialog turns are with emotion annotations.

As described in [21], we define the inter-rater agreement for the continuous emotion annotations as the linear correlation between two annotators. For each actor recording, we compute the correlation between every pair of annotators and only keep the annotator pairs with correlations greater than 0.5. We further partition each actor recording into dialog turns according to speech regions. As a result, we have 1230 annotated dialog turns (referred to as the target turn hereafter) in total. Each target turn is paired with the corresponding interlocutor's previous turn, as illustrated in Fig. 2. Our work focuses on modeling interaction behavior between the paired dialog turns. The values of activation and valence of each target turn are calculated by averaging the annotations among frames and across annotators. We jointly consider activation and valence by creating K emotional clusters in the valence-activation space using k-means algorithm. Such K-class recognition scheme has also been adopted in [12] [22]. We consider clusters with $K = 2$ and $K = 3$, and Fig. 3 shows the corresponding clustering results.



(a) $K = 2$. (b) $K = 3$.

Fig. 3. Resulting emotion classes in the valence-activation space for $K = 2$ and $K = 3$.

3. GEODESIC FLOW-BASED INTERACTION MODELING

Our objective is to model mutual influence of a dyad's multimodal behavior in an interaction. The main idea of our interaction model is inspired by a geodesic flow-based methodology for the problem of domain adaptation that explores the sharing characteristics of two distinct datasets [13]. The two datasets are the behavior characteristics of two interaction partners in our case. This method explicitly constructs an infinite-dimensional feature space $\mathcal{H}^\infty$ assembling all geometric and statistical information on the path between two datasets, i.e., the dyad's behavior. Thereby, $\mathcal{H}^\infty$ is assumed to capture the behavior connection of the two partners.

We assume the behavior characteristics of each participant in an interaction can be embedded in a low-dimensional linear subspace

$\mathbf{P} \in \mathcal{R}^{D\times d}$, where D is the feature vector dimensionality and d is the subspace dimensionality. The collection of all d-dimensional subspaces of D-dimensional vectors constitute the Grassmannian $\mathbf{Gr}(d, D)$. We compute subspaces respectively for the target and interlocutor behavior in an interaction, both of which can be seen as two points on $\mathbf{Gr}(d, D)$. To model the interaction of a dyad's multimodal behavior, we focus on establishing connection of the two behavior subspaces on $\mathbf{Gr}(d, D)$. We believe that such subspace connection can accordingly capture the dyadic interaction in the behavioral feature space.

Let $\mathbf{P}_T$, $\mathbf{P}_I \in \mathcal{R}^{D\times d}$ denote the two sets of bases of the subspaces for the target and interlocutor behavior in an interaction, respectively. The incremental changes between $\mathbf{P}_T$ and $\mathbf{P}_I$ can be parameterized by the geodesic flow that defines the shortest path between two points on $\mathbf{Gr}(d, D)$. The geodesic flow between $\mathbf{P}_T$ and $\mathbf{P}_I$ is expressed as $\mathbf{\Phi}(t) \in \mathbf{Gr}(d, D)$:

$$\mathbf{\Phi}(t) = \begin{pmatrix} \mathbf{P}_I & \mathbf{R}_I \end{pmatrix} \begin{pmatrix} \mathbf{U}_1 & 0 \\ 0 & -\mathbf{U}_2 \end{pmatrix} \begin{pmatrix} \mathbf{\Gamma}(t) \\ \mathbf{\Sigma}(t) \end{pmatrix},$$

where $t \in [0, 1]$ and $\mathbf{R}_I$ is the orthogonal complement to $\mathbf{P}_I$. Particularly, $\mathbf{\Phi}(0) = \mathbf{P}_I$ and $\mathbf{\Phi}(1) = \mathbf{P}_T$. $\mathbf{U}_1$ and $\mathbf{U}_2$ are orthogonal matrices given by generalized singular value decomposition (SVD),

$$\begin{pmatrix} \mathbf{P}_I^T\mathbf{P}_T \\ \mathbf{R}_I^T\mathbf{P}_T \end{pmatrix} = \begin{pmatrix} \mathbf{U}_1 & 0 \\ 0 & \mathbf{U}_2 \end{pmatrix} \begin{pmatrix} \mathbf{\Gamma} \\ -\mathbf{\Sigma} \end{pmatrix} \mathbf{V}^T.$$

The singular values on the diagonals of $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ are $cos(\theta_i)$ and $sin(\theta_i)$. Accordingly, $\mathbf{\Gamma}(t)$ and $\mathbf{\Sigma}(t)$ are diagonal matrices with elements $cos(t\theta_i)$ and $sin(t\theta_i)$ respectively.

Given two original D-dimensional behavior vectors $\mathbf{x}_T$ and $\mathbf{x}_I$ from the target subject and the interlocutor, their projections into a subspace $\mathbf{\Phi}(t)$ for a $t \in [0, 1]$ are $\mathbf{\Phi}(t)^T\mathbf{x}_T$ and $\mathbf{\Phi}(t)^T\mathbf{x}_I$. These projections integrate the behavioral characteristics of both interaction partners, since the geodesic flow parameterizes the gradual adaptation from the interlocutor to the target participant. To model the mutual effect of a dyad, we utilize all subspaces $\mathbf{\Phi}(t)$ on the geodesic flow. Concatenating projections into all subspaces, we obtain feature vectors $\mathbf{z}_T^\infty$ and $\mathbf{z}_I^\infty$ in an infinite-dimensional feature space $\mathcal{H}^\infty$. The inner product of $\mathbf{z}_T^\infty$ and $\mathbf{z}_I^\infty$ is given by,

$$\begin{aligned} \langle \mathbf{z}_T^\infty, \mathbf{z}_I^\infty \rangle &= \int_0^1 (\mathbf{\Phi}(t)^T\mathbf{x}_T)^T(\mathbf{\Phi}(t)^T\mathbf{x}_I)dt \\ &= \mathbf{x}_T^T(\int_0^1 \mathbf{\Phi}(t)\mathbf{\Phi}(t)^T dt)\mathbf{x}_I \\ &= \mathbf{x}_T^T\mathbf{G}\mathbf{x}_I, \end{aligned}$$

where the matrix $\mathbf{G} \in \mathcal{R}^{D\times D}$ is defined as geodesic flow kernel in [13] and can be easily obtained from the above equations. Since $\mathbf{G}$ is positive semidefinite, we further decompose it as $\mathbf{G} = \mathbf{M}\mathbf{M}^T$. We compute $\mathbf{M}$ by SVD: $\mathbf{M} = \mathbf{U}_g\mathbf{\Gamma}_g^{\frac{1}{2}}$, where $\mathbf{G} = \mathbf{U}_g\mathbf{\Gamma}_g\mathbf{V}_g^T$. $\mathbf{M} \in \mathcal{R}^{D\times D}$ embeds behavioral association of a dyad in an interaction. We refer to $\mathbf{M}$ as the *interaction matrix*.

4. MULTIMODAL EMOTION RECOGNITION

In this section, we assess the effectiveness of the interaction model in multimodal emotion recognition, i.e., evaluating the emotional state in a target turn using multimodal behavior of a dialog turn pair. To this end, we conduct the mapping for both target and interlocutor features using the interaction matrix $\mathbf{M}$ per interaction: $\mathbf{f} = \mathbf{M}^T\mathbf{x}$, where $\mathbf{x}$ represents a behavioral feature vector described by speech (audio) or hand gesture (visual) features. The new feature vector $\mathbf{f}$ contains the interaction information induced by $\mathbf{M}$. For comparison, we also evaluate the recognition performance using two baselines: 1) only the behavioral features of the target turn, i.e., $\mathbf{x}_T$; 2) both behavioral features of the target and interlocutor turns, i.e., $\mathbf{x}_T$ and $\mathbf{x}_I$, which have also been applied in [12]. For each target or interlocutor turn in a dialog turn pair, eight high level statistical functionals are extracted from either mapped or baseline behavioral features: mean, median, standard deviation, range, lower quartile, upper quartile, minimum and maximum. We adopt the leave-one-interaction-out scheme and linear SVM in the experiment. In the interaction model, we apply principal component analysis (PCA) to identify subspaces $\mathbf{P}_T$ and $\mathbf{P}_I \in \mathcal{R}^{D\times d}$ for the target and interlocutor behavior in each interaction. The subspace dimension d is determined using cross-validation on the training set.

Table 1 presents the results for recognizing 2-Class and 3-Class emotions in the valence-activation space (see Fig. 3) in different conditions. Firstly, we can observe that speech cues generally show a higher discriminative power for distinguishing emotional dimensions in contrast to hand gesture behavior. One possible reason could be that the activation dimension can be better perceived from audio cues [23]. Furthermore, the inclusion of interlocutor information generally improves the recognition performance, compared to the performance with the target only information. These results are consistent with the findings in [12], validating that the interlocutor's multimodal behavior provides complementary information about the emotional state of the target subject during a dyadic interaction. More importantly, our interaction model significantly outperforms the two baselines in all cases. For example, the recognition accuracy is 58.7% and 61.8% respectively using the audio-visual information from the target turn and from both dyadic turns. The performance improves to 71.8% with interaction modeling. This observation corroborates that the dyad's behavior under such modeling can serve as improved indicators of the displayed emotion in the target turn, namely the interaction matrix $\mathbf{M}$ effectively embeds the dyadic behavior coordination.

Table 1. Accuracies ($\%$) for recognizing 2-Class and 3-Class emotions in the valence-activation space from information of the target turn (T), information from both target and interlocutor turns ($T + I$) [12], and using interaction modeling (Interaction).

	$\mathbf{K = 2}$ [Chance level: 50%]			$\mathbf{K = 3}$ [Chance level: 33%]		
Features	**T**	**T + I**	**Interaction**	**T**	**T + I**	**Interaction**
Audio	58.5	59.7	**69.0**	47.0	45.0	**54.5**
Visual	57.8	61.9	**65.5**	39.6	44.0	**52.4**
Audio-Visual	58.7	61.8	**71.8**	45.6	46.2	**55.9**

5. ANALYSIS OF MODELING RESULTS

In this section, we further study the properties of the mutual influence model to better understand its benefits for emotion characterization in an interaction.

5.1. Analysis of Behavior Entrainment

Section 4 empirically validated that the interaction model is effective in capturing the dyadic behavior cohesiveness for multimodal emotion recognition. Herein, we focus on explicitly measuring such

cohesiveness and examining whether the dyadic entrainment can be better embodied with interaction modeling.

We apply the PCA-based entrainment measure proposed in [7] to quantify the turn-wise behavioral coordination. Given two sets of behavior observations $\mathbf{X}_T$ from the target turn and $\mathbf{X}_I$ from the interlocutor turn, we respectively compute the subspaces $\mathbf{W}_T$ and $\mathbf{W}_I \in \mathcal{R}^{D\times d_\mathbf{W}}$ from each set. The subspace dimensionality $d_\mathbf{W}$ is the minimum number of eigenvectors covering 90% total variance of each behavior set. The turn-wise entrainment metric is defined as,

$$En(\mathbf{X}_T, \mathbf{X}_I) = \frac{1}{d_\mathbf{W}} trace(\mathbf{W}_T^T \mathbf{W}_I \mathbf{W}_I^T \mathbf{W}_T). \quad (1)$$

A larger En value indicates a higher entrainment level.

We compute the entrainment measure for each pair of dialog turns using the original feature set $\mathbf{X}$ (audio or visual) or the mapped feature set $\mathbf{M}^T\mathbf{X}$. We also examine the computed entrainment measures w. r. t. 2-Class emotion labels (see Fig. 3(a)). Table 2 presents the average turn-wise entrainment values in each emotion class. Firstly, we can observe that the entrainment values of speech or gesture behavior have been improved by interaction modeling. This improvement in all cases is statistically significant with $p \ll 0.01$, suggesting that the interpersonal mutual influence becomes more pronounced in the interaction space induced by $\mathbf{M}$. In addition, the speech behavior exhibits a much higher entrainment level compared to the gesture behavior. This might be due to greater gesture variabilities across persons and over different temporal scales. We also find that the inter-emotion difference in entrainment values can be boosted with interaction modeling. For example, a statistical difference of gesture entrainment ($p = 0.03$) has been found between Class I and II after interaction modeling, while such difference is not observed for the original gesture features.

Table 2. The average entrainment values in each emotion class with original and mapped multimodal behavioral features.

	Class I	Class II	Class I	Class II
Feature	**Original**		**Mapped**	
Audio	0.6414	0.6374	0.6908	0.6773
Visual	0.2745	0.2892	0.3515	0.3728

5.2. Analysis of Interaction Patterns

As introduced in Section 1, the patterns of mutual behavior influence are usually shaped by the underlying emotional states of partners [5]. Since we have shown that the interaction matrix $\mathbf{M}$ of an interaction can effectively capture the mutual influence of dyadic behavior, we further investigate how the interaction patterns rooted in $\mathbf{M}$ depend on the overall emotions of two interaction partners.

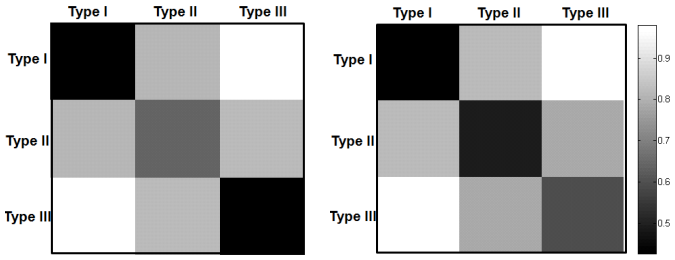
For this purpose, we have at lease three annotators rate the overall activation and valence for each actor in an interaction on a 9-point scale. The inter-rater agreement for both activation and valence is around 0.7. The global rating of each actor in an interaction is described by the average value across annotators. Similar to the emotion processing in Section 2.2, we create two emotion clusters in the valence-activation space based on the global emotional ratings. The distribution of the two clusters is similar to Fig. 3(a). Thereby, interactions can be grouped into three categories: in Type I or III, both actors have the same emotion label of Class I or Class II; in the incongruent Type II, one actor has a Class I label while the other is with a Class II label.

We examine the difference between interaction patterns within an interaction group as well as across groups. The interaction pattern difference is measured by the distance between interaction matrices of two interactions. According to Section 3, $\mathbf{M}$ is an orthogonal matrix. Therefore, the distance between two interaction matrices $\mathbf{M}_A$ and $\mathbf{M}_B$ is defined using the *Binet-Cauthy* metric [24]:

$$dist(\mathbf{M}_A, \mathbf{M}_B) = (1 - \prod cos^2(\theta_i))^{\frac{1}{2}}, \quad (2)$$

where θ_i are the principal angles between $\mathbf{M}_A$ and $\mathbf{M}_B$ [25].

Fig. 4 visualizes the average distances between pairwise interaction matrices within an interaction category as well as across categories. These distances are computed with respect to interaction matrices of speech (Fig. 4(a)) and hand gesture (Fig. 4(b)). A darker color indicates a smaller distance. As can be seen in Fig. 4, the average distance is much smaller within the same group than that computed across groups. This observation validates that the mutual influence model is capable of capturing emotion-dependent interaction patterns, which can further increase the discriminative power for emotion recognition. Moreover, we can observe the most difference between the interaction patterns in Type I and Type III, which may result from the fact that the two groups are the most distant in terms of emotions.



(a) Audio (b) Visual

Fig. 4. Average distances between pairwise interaction matrices within an interaction category and across categories.

6. CONCLUSIONS AND FUTURE WORK

In this work, we focused on explicitly modeling the mutual influence of multimodal behavior in affective dyadic interactions. In our framework, the behavior adaptation from the interlocutor to the target participant in an interaction is captured by an interaction matrix which assembles all information on the geodesic path between the dyad's behavior subspaces. The experimental results demonstrated that our interaction model can significantly improve the performance of emotion recognition over the baseline features. We further investigated the properties of the interaction model. Analysis results revealed that the entrainment effect of a dyad's behavior becomes more prominent by interaction modeling, and that the interaction patterns embedded in interaction matrices depend on the emotional states of interaction partners. These observations corroborated the validity of the interaction model for capturing emotion-related interaction dynamics of multimodal behavior.

In the future, this mutual influence model can be extended by incorporating temporal dynamics at either frame or turn level over an interaction. This study enables us to develop more natural interaction interfaces which can robustly monitor the emotional states of human users in real-time and appropriately adjust its behavior to achieve user satisfaction.

7. REFERENCES

[1] J. Burgoon, L. Stern, and L. Dillman, *Interpersonal adaptation: Dyadic interaction patterns.*, Cambridge University Press, 1995.

[2] R. Levitan, A. Gravano, and J. Hirschberg, "Entrainment in speech preceding backchannels," in *Proc. of ACL for Computational Linguistics: Human Language Technologies*, 2011.

[3] M. Richardson, K. Marsh, and R. Schmidt, "Effects of visual and verbal interaction on unintentional interpersonal coordination.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 31, no. 1, pp. 62, 2005.

[4] C. Breazeal, "Regulation and entrainment in human-robot interaction," *The International Journal of Robotics Research*, vol. 21, no. 10-11, pp. 883–902, 2002.

[5] Z. Yang, A. Metallinou, and S. Narayanan, "Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1766–1778, 2014.

[6] C-C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples.," in *INTERSPEECH*, 2010.

[7] C-C. Lee, A. Katsamanis, M. Black, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions," *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, 2014.

[8] C-C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions.," in *INTERSPEECH*, 2009.

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

[10] N. Oliver, B. Rosario, and A. Pentland, "A bayesian computer vision system for modeling human interactions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 831–843, 2000.

[11] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 3, pp. 305–317, 2005.

[12] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *Affective Computing, IEEE Transactions on*, vol. 4, no. 2, 2013.

[13] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. of CVPR*, 2012.

[14] 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.

[15] "CreativeIT Database," `http://sail.usc.edu/CreativeIT/`.

[16] Installation Guide, "Autodesk®," 2008.

[17] S. Levine, C. Theobalt, and V. Koltun, "Real-time prosody-driven synthesis of body language," vol. 28, no. 5, pp. 172, 2009.

[18] M.E. Sargin, Y. Yemez, E. Erzin, and A. Tekalp, "Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1330–1345, 2008.

[19] Z. Yang, A. Metallinou, E. Erzin, and S. Narayanan, "Analysis of interaction attitudes using data-driven hand gesture phrases," in *Proc. of ICASSP*, 2014.

[20] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop on Speech and Emotion*, 2000.

[21] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Automatic Face and Gesture Recognition (FG), IEEE International Conference and Workshops on*, 2013.

[22] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *Affective Computing, IEEE Transactions on*, vol. 3, no. 2, pp. 184–198, 2012.

[23] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *Affective Computing, IEEE Transactions on*, vol. 2, no. 2, pp. 92–105, 2011.

[24] J. Hamm and D. Lee, "Grassmann discriminant analysis: a unifying view on subspace-based learning," in *Proc of ICML*, 2008.

[25] A. Knyazev and M. Argentati, "Principal angles between subspaces in an a-based scalar product: algorithms and perturbation estimates," *SIAM Journal on Scientific Computing*, vol. 23, no. 6, pp. 2008–2040, 2002.