Predicting Video-Conferencing Conversation Outcomes Based on Modeling Facial Expression Synchronization

Rui Li[1], Jared Curhan[2] and Mohammed (Ehsan) Hoque[1]
[1] ROC HCI, Department of Computer Science, University of Rochester, New York, USA
[2] Sloan College of Management, Massachusetts Institute of Technology, Massachusetts, USA

***Abstract*— Effective video-conferencing conversations are heavily influenced by each speaker's facial expression. In this study, we propose a novel probabilistic model to represent interactional synchrony of conversation partners' facial expressions in video-conferencing communication. In particular, we use a hidden Markov model (HMM) to capture temporal properties of each speaker's facial expression sequence. Based on the assumption of mutual influence between conversation partners, we couple their HMMs as two interacting processes. Furthermore, we summarize the multiple coupled HMMs with a stochastic process prior to discover a set of facial synchronization templates shared among the multiple conversation pairs. We validate the model, by utilizing the exhibition of these facial synchronization templates to predict the outcomes of video-conferencing conversations. The dataset includes 75 video-conferencing conversations from 150 Amazon Mechanical Turkers in the context of a new recruit negotiation. The results show that our proposed model achieves higher accuracy in predicting negotiation winners than support vector machine and canonical HMMs. Further analysis indicates that some synchronized nonverbal templates contribute more in predicting the negotiation outcomes.**

I. INTRODUCTION

Video-conferencing (VC) becomes a popular platform for people to interact in professional and personal capacities [5]. On the other hand, some technical issues still exist, for example the limited view of the person, disengaged eye contact, and occasional interruptions resulting from network latency. These issues disrupt social presence, and thus lead to poor VC communication [3]. Nonverbal behavior plays a significant role to enhance social presence. It provides a source of rich information about the speaker's intentions, goals, and values [1][3][4]. This motivates us to investigate facial expressions in VC communication in order to gain insight into effective communicative skills that will improve productivity and conversational satisfaction.

In this study, we investigate interactional synchrony of facial expressions in VC-mediated conversations, as shown in Figure 1. Interactional synchrony refers to patterned and aligned interactions occurring over time [11]. In a synchronic interaction, nonverbal behaviors (e.g., facial expressions, posture, gesture) of the individuals are coordinated to the rhythms and forms of verbal expressions. As a key indicator of interactional involvement, rapport, and mutuality, it has been used in deception detection, online learning, interpersonal trust evaluation, and a variety of other fields [1][7][8][10]. However, the quantification of interactional synchrony is challenging, and it depends on the specific social context. We address this challenge by modeling facial expression synchronization of VC conversation partners given the social context of negotation.

This work was partially supported by DARPA

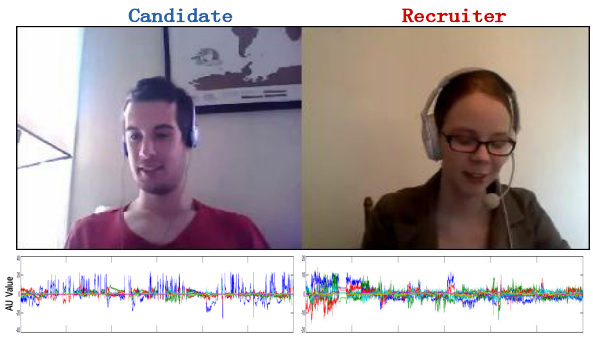

Fig. 1: An illustration of one VC conversation pair. The two participants communicate via our web-based VC platform in their own natural environments. The lower panels show the first six principal components of their facial expression action units (AUs) evolving over time.

We propose a novel probabilistic model to learn an effective representation of facial interactional synchrony. This representation contains a set of facial synchronization templates displayed by multiple conversation pairs, as shown in Fig 2. In particular, we utilize a hidden Markov model (HMM) to describe the temporal properties of each speaker's facial expression. The Markovian property assumes that if a speaker smiles at previous time step, it is likely that he/she maintains the smile at the current time step, for instance. We further assume that there exists mutual influence between a pair of conversation partners. Namely, if a speaker's conversation partner displays a smile, it is likely that the speaker responds with a smile. To capture the mutual influence between a pair of conversation partners, we couple their two HMMs together as interacting processes. We thus model the multiple conversation pairs with the corresponding multiple coupled HMMs. Furthermore, we summarize the multiple coupled HMMs by introducing a stochastic process as a prior. This prior allows us to uncover the shared facial synchronization templates among the multiple conversation pairs. In this representation, a couple of conversation partners' facial expressions can be decomposed into instantiations of a particular subset of the globally shared synchronization templates. This novel representation of facial expression synchronization enables

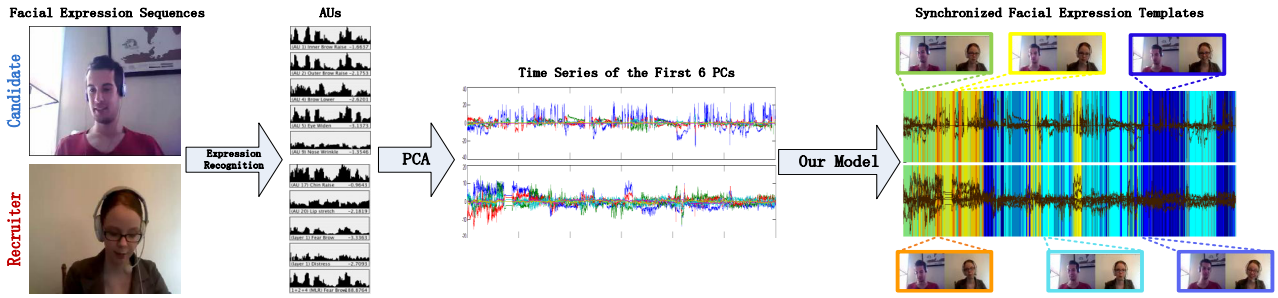

Fig. 2: Diagram of our approach illustrated on one conversation pair. From left to right, 28 facial expression action units (AUs) are extracted from the conversation partners' videos using CERT toolbox [9]; the time series of the first 6 principal components are transformed from both the AUs with the principal component analysis (PCA), and these six-dimensional PC time series are the input to our model; the model automatically decomposes the facial expression time series into the salient segments (color coded) which correspond to a subset of globally shared facial synchronization templates displayed by this pair.

us to not only interpret effective VC communication skills but also predict the outcomes of the conversations.

To conduct this study, we develop a VC system that works via a web browser without any additional download or plugin support. The platform is designed to enable auto audio and video upload in a remote server every 30 seconds as two people engage in a video-conference. This functionality allows the framework to be deployed in Amazon Mechanical Turk, where remote human workers with access to a web browser and webcam communicate with each other. To the best of our knowledge, this is the first study to investigate VC-mediated facial expression synchrony. The contributions of our study include:

- We build a novel probabilistic model to learn an effective representation of facial interactional synchrony in VC communication. This novel representation decomposes multiple pairs of conversation partners' facial expression sequences into a set of globally shared synchronization templates.
- We further represent a conversation by the frequencies of occurrence of its facial synchronization templates, and achieve superior accuracy (78% on average) than support vector machine (SVM) and canonical HMMs to predict conversation outcomes.

II. METHOD OF COMPUTER-MEDIATED NEGOTIATION STUDY

We validate our model using the dataset collected from a study engaging Mechanical Turkers in a recruitment negotiation [2]. In this study, the conversational speech and facial expressions are recorded. The outcomes are the number of points earned by each participant and a post-negotiation questionnaire regarding the participants' evaluation of their counterparts and the negotiation process.

A. Participants

242 Mechanical Turkers participate in the study. Participants are informed that their negotiations would be recorded and that the study's purpose is to investigate negotiation skills. The data collected from 150 of the Turkers is available for further analysis. Among them, 43 participants ($\%29$) are female. The remaining Turkers either had damaged videos or lacked post-questionnaire data.

B. Apparatus

The negotiators interact with each other through a computer-mediated platform based on a browser-based VC system. The existing freely available video software (e.g., Skype, Google+ Hangouts) often requires users to download their application or install a plugin. In addition, Skype's current API and Google+ do not allow us to capture and record audio or video streams. To handle these hurdles, we develop our own browser-based VC system that is capable of capturing and analyzing the video stream in the cloud. We implement the functionality to transfer audio and video data every 30 seconds to prevent data loss and dynamically adapt to variant network latency.

C. Task

The task of this experiment is defined as a recruitment case. A recruitment case involves a scenario in which a candidate, who already has an offer, negotiate the compensation package with the recruiter. The candidates and the recruiters need to reach an agreement on eight issues related to salary, job assignment, location, vacation time, bonus, moving expense reimbursement, starting date, and health insurance. Each negotiation issue offers 5 possible options for resolution. Each option is associated with a specific number of points for each party. The goal of the negotiators is to maximize the total points they can possibly earn (e.g., the 5 optional offers on salary issue range from 65K to 45K. Candidate receives maximum points if he/she could settle with salary of 65k whereas recruiter loses maximum points, and vice versa).

D. Procedure

As Amazon mechanical Turkers take the HIT, they are formed into 75 conversation pairs sequentially. The social

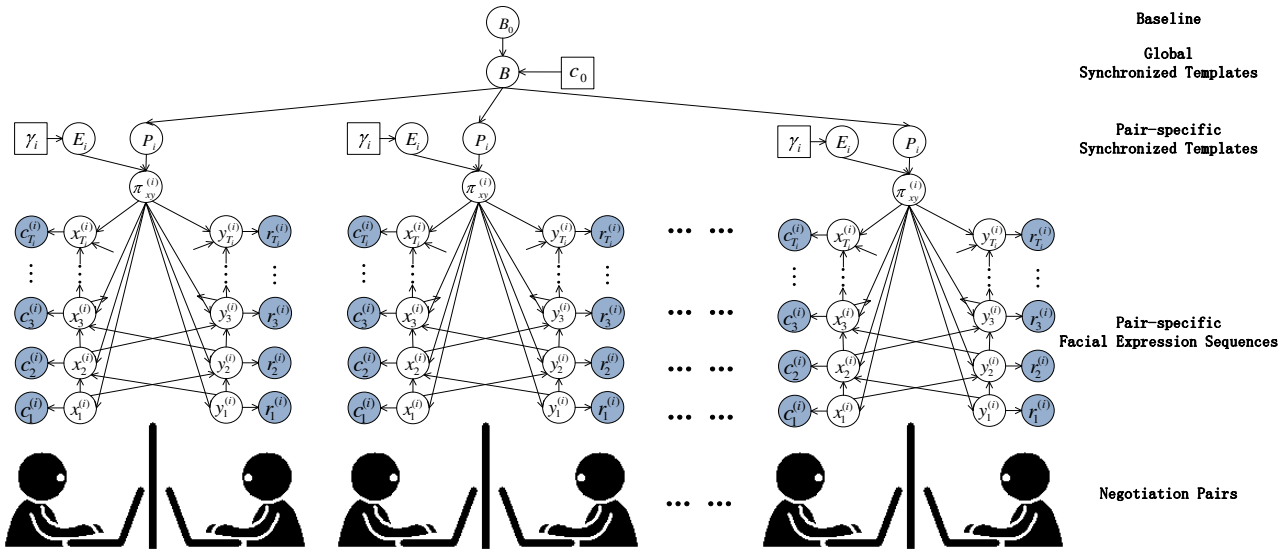

Fig. 3: Our proposed probabilistic graphical model. For each speaker's facial expression sequence, we use a HMM to describe its dynamic process during the conversation. We further couple the two HMMs of the conversation partners to describe their mutual influence, while allowing each of them to maintain his/her own dynamic process. At the top level, we use a beta process prior to summarize the facial synchronization templates shared across multiple conversation pairs. In this hierarchical structure, each conversation pair exhibits a particular subset of the globally shared facial synchronization templates. Shaded disks represent the observed facial expressions in the video frames.

roles are randomly assigned to the conversation partners. The participants coordinate with their partners to choose the locations and times for the VC-based negotiation, so they may interact in convenient and comfortable circumstances.

After both participants provide consent, a button appears that leads each individual to the correct video chat room with signals that the two can speak with each other. The participants then proceed to play out the scenario outlined in their instructions. Recording begins when the two participants build the connection, and stops when one participant hangs upon completion of the negotiation. Participants are free to offer information, arguments, and proposals, although they may not exchange their confidential instructions. The candidates won 47 conversations.

III. HIERARCHICAL COUPLED HIDDEN MARKOV MODEL

The graphical model is described as two levels. At the lower level, there are N coupled HMMs corresponding to N conversation pairs. At the top level, We use a Beta process prior to discover the facial synchronization templates shared among these distinct yet related conversation pairs in the given social context, as shown in Fig. 3.

A. Dynamic Likelihoods

We assume that the conversation partners' facial expressions are interdependent, and interact by influencing each other's emotional states or communicative strategies. Additionally, each speaker's facial expression sequence maintains its own internal dynamic. To encode these assumptions, we couple two Markov chains via a matrix of conditional probabilities between their hidden state variables. We denote the observations of the ith conversation pair's facial expression sequences as $O_i = \{c^{(i)}_{1:T_i}, r^{(i)}_{1:T_i}\}$, where $c^{(i)}_{1:T_i}$ are the observed facial expressions of the candidate, and $r^{(i)}_{1:T_i}$ are the recruiter's. The observations are The PCA components of the facial expression AUs extracted from the videos of a conversation pair.

We further define $S_i = \{x^{(i)}_{1:T_i}, y^{(i)}_{1:T_i}\}$ as the hidden state sequences, where $x^{(i)}_{1:T_i}$ are the hidden states of the candidate, and $y^{(i)}_{1:T_i}$ represent the hidden states of the recruiter. These hidden states index some patterned facial expressions of both conversation partners. The state transition probabilities are defined as

$$x^{(i)}_{t+1}|x^{(i)}_t, y^{(i)}_t \sim Mult(\pi^{(i)}_{x^{(i)}_t, y^{(i)}_t}) \quad (1)$$

$$y^{(i)}_{t+1}|y^{(i)}_t, x^{(i)}_t \sim Mult(\pi^{(i)}_{x^{(i)}_t, y^{(i)}_t}) \quad (2)$$

The emission distributions are defined as normal distributions:

$$c^{(i)}_t|x^{(i)}_t \sim N(\mu^{(i)}_{x^{(i)}_t}, \Sigma^{(i)}_{x^{(i)}_t}) \quad (3)$$

$$r^{(i)}_t|y^{(i)}_t \sim N(\mu^{(i)}_{y^{(i)}_t}, \Sigma^{(i)}_{y^{(i)}_t}) \quad (4)$$

B. Combinatorial Prior

We propose to use a beta process prior to summarize the stereotypical and idiosyncratic facial synchronization exhibited by the multiple conversation pairs. This prior not only allows flexibility in the number of facial synchronization templates, but also enables each conversation pair to exhibit a subset of the globally shared templates.

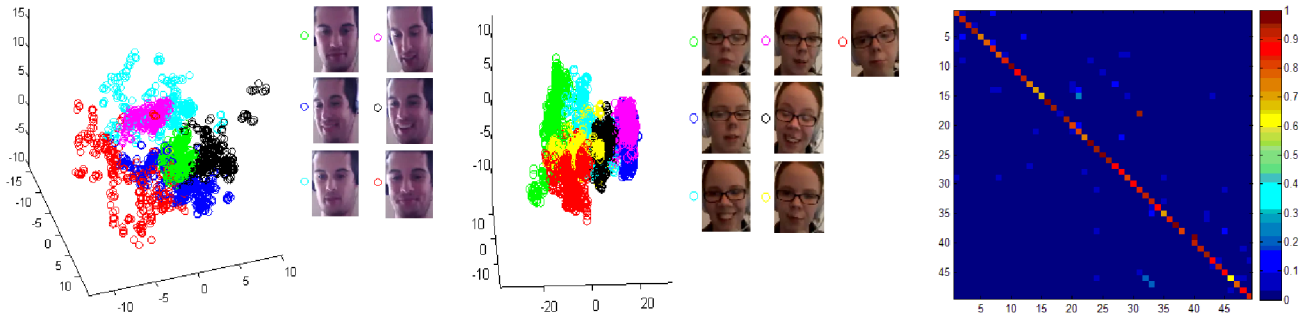

Fig. 4: Model performance illustrated on one conversation pair. From left to right, we show the candidate's and recruiter's clusters of the patterned facial expressions projected in the space of the first 3 principal components, respectively. One facial expression sample from each color-coded cluster is visualized. Each data point represent a facial expression in one video frame. In this scenario, a facial synchronization template is characterized by a particular combination between one patterned facial expression from the candidate cluster and one patterned facial expression from the recruiter's. The right panel shows the transition matrix between the synchronization templates for this conversation pair.

Let B_0 denote a fixed continuous random base measure on a space Θ which represents a space of all the potential facial synchronization templates. For multiple conversation pairs to share a set of these templates, let B denote a discrete realization of a beta process given the prior $BP(c_0, B_0)$, which is a discrete random measure on Θ following the beta process. Its elements' locations correspond to the set of facial synchronization templates shared among multiple conversation pairs. Its elements' probabilities represent how likely these templates are shared. Let P_i denote a Bernoulli measure given the beta process B. P_i is a binary vector of Bernoulli random variables representing whether a particular synchronization template displayed in the observed facial expression sequences of conversation pair i. This construction can be formulated as follows:

$$B|B_0 \sim BP(c_0, B_0) \tag{5}$$

$$P_i|B \sim BeP(B) \tag{6}$$

$$P_i = \sum_k p_{ik}\delta_{\theta_k} \tag{7}$$

where $B = \sum_k b_k \delta_{\theta_k}$. This term shows that B describes a set of countable number of synchronization templates $\{\theta_k\}$ drawn from the space Θ, and their corresponding probability masses $\{b_k\}$. The combination of these two variables characterizes how likely the synchronization templates are shared among the conversation pairs. Thus P_i is a Bernoulli process realization from the random measure B where p_{ik} as a binary random variable denotes whether conversation pair i displays the facial synchronization template k, given the probability mass b_k. Based on the above formulation, for $k = 1...K$ templates we readily define $\{(\theta_k, b_k)\}$ as a set of globally shared facial synchronization templates and their probabilities to be shared among the conversation pairs, and define $\{(\theta_k, p_{ik})\}$ as conversation pair i's subset of synchronization templates drawn from $\{(\theta_k, b_k)\}$.

The transition distribution $\pi_{xy}^{(i)} = \{\pi_{x_t^{(i)}, y_t^{(i)}}\}$ of the coupled HMMs at the bottom level governs the transitions between the i^{th} pair's subset of templates θ_k. It is determined by the element-wise multiplication between the subset $\{p_{ik}\}$ of pair i and the gamma-distributed random variables $\{e_{ik}\}$:

$$e_{ik}|\gamma_i \sim Gamma(\gamma_i, 1) \tag{8}$$

$$\pi_{xy}^{(i)} \propto E_i \bigotimes P_i \tag{9}$$

where $E_i = [e_{i1}, ...e_{iK}]$. So the effective dimensionality of π_i is determined by P_i.

We use the Markov chain Monte Carlo method to do the posterior inference. Based on the sampling algorithm proposed in [12], we developed a Gibbs sampling solution to sample the marginalized hierarchical beta processes part of the model. Given transition distributions $\pi_{xy}^{(i)}$, shared templates $\{\theta_k\}$, and observed facial expression sequences $c_{1:T_i}^{(i)}$ and $r_{1:T_i}^{(i)}$, within a massage passing algorithm, we compute the backward messages:

$$m_{t+1,t}(S_t^{(i)}) \propto p(O_{t+1:T_i}^{(i)}|S_t^{(i)}, \pi_{xy}^{(i)}, \{\theta_k\}) \tag{10}$$

to update the hidden state sequences $S_{1:T_i}^{(i)}$ by sampling from:

$$\begin{aligned} &p(S_t^{(i)}|S_{t-1}^{(i)}, O_{1:T_i}^{(i)}, \pi_{xy}^{i}, \{\theta_k\}) \propto \\ &\pi_{S_{t-1}^{(i)}}^{(i)}(S_t^{(i)}) N(O_t^{(i)}; \mu_{S_t^{(i)}}^{(i)}, \Sigma_{S_t^{(i)}}^{(i)}) m_{t+1,t}(S_t^{(i)}) \end{aligned} \tag{11}$$

IV. RESULTS AND DISCUSSION

The facial representation method we used is the facial expression AUs defined in facial action coding system coding (FACS) [6]. In particular, we extracted the AUs from the videos using the Computer Expression Recognition Toolbox (CERT) [10]. After converting the data with the PCA technique to reduce their dimensions, we adopted the first 6 principal components, which accounted for about 97% of the data variance, to represent the data for our further analysis.

A. Salient Segments Estimation

In Fig. 2, we illustrate the process of our approach on one conversation pair's facial expression time series. Our model decomposes the pair's facial expression sequences into a subset of salient segments which correspond to a subset of synchronization templates. Each template is characterized by

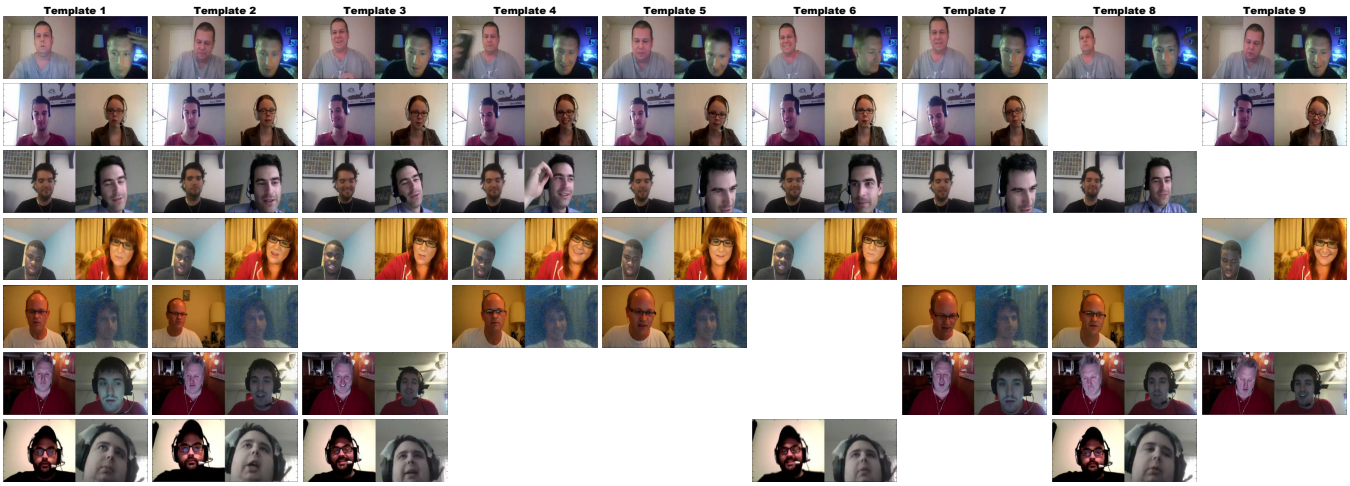

Fig. 5: Illustration of instantiations of facial synchronization templates learned by our model, with candidates on the left and recruiters on the right. Each column contains the instantiations of one particular templates shared among the negotiation pairs, and each row represents one pair of negotiators. Our model allows the pairs to exhibit different subsets of the globally shared synchronization templates. For example, Template 1 represents a case in which both parties exhibit neutral facial expressions to each other. Template 4 is a case where candidates appear distracted, while recruiters respond with a smile. In Template 8, candidates exhibit subtle polite smiles while recruiters show neutral faces.

a particular combination of two patterned facial expressions, as visualized in Fig. 4. The color-coded segments suggest that the negotiation pair periodically display some facial synchronization templates. In Fig. 4, since the data clusters are visualized in the first 3 principal component space, some separations may not be obvious. The transition probability matrix indicates that these synchronization templates are persistent with high self-transition probabilities.

B. Facial Synchronization Templates

Fig. 5 demonstrates a matrix of the shared facial synchronization templates estimated from the 75 negotiation pairs in our study. The exploratory interpretations of these templates are summarized in Table I. Each template can be quantified by the combination of the mean vector and the covariance matrix of one candidates' facial cluster and the mean vector and the covariance matrix of one recruiters' facial cluster, illustrated in Fig. 4. Note that we only illustrate a subset of global synchronized templates learned from the dataset which are shared most frequently. Among the reported templates, Since template 1, 4, 6, and 8 do not involve speaking. they are labeled as nonverbal templates.

C. Negotiation Outcome Prediction

To measure the performance of our novel representation of the negotiation processes, we try to predict the negotiation

TABLE I: Interpretation of Facial Synchronization Templates.

Template	Candidate	Recruiter
Template 1	neutral face	neutral face
Template 2	neutral face listening	speaking holding the turn
Template 3	smile	speaking holding the turn
Template 4	looking down listening	smile
Template 5	speaking holding the turn	subtle smile listening
Template 6	big smile	neutral face listening looking away
Template 7	speaking smiling revealing information	neutral face listening
Template 8	subtle polite smile	neutral face
Template 9	looking down speaking	big smile listening

outcomes based the facial synchronization templates. We randomly assign the data into training and testing sets. In particular, each negotiation pair's negotiation process is represented by the frequencies of occurrence of its subset of templates, and the ground-truth of a negotiation winner is determined by the points each party earned in the negotiation. We examine prediction performance given the training sets containing the various number of template instantiations for the 75 pairs' facial expression sequences. We have to assume each negotiation pair exhibits the same set of templates in order to implement the canonical HMMs. On the other hand, we use the corresponding segments of the facial expression time series to train a SVM. Figure 6 indicates that our model leads to significant improvement in prediction performance, particularly when fewer training instantiations are available. Canonical HMMs essentially compute a set of averaged template from the 75 pairs of facial expression sequences. This representation blurs the distinction between the conversation pairs' facial expressions as variations. A major cause to SVM's inferior performance is that it does not account for the temporal information of the sequential data. Our proposed model addresses these issues. In particular, the highest weights are assigned to Template 6 (nonverbal), 8 (nonverbal), and 9. This suggests that most predictive information is derived from these facial synchronization templates, most of which are nonverbal templates.

V. CONCLUSIONS

This paper investigate facial expression synchrony in a computer-mediated negotiation based on video-conferencing conversations. We further present a probabilistic dynamic model to automatically learn a set of facial synchronization templates. These templates are shared among negotiation pairs while they engage in a simulated negotiation task via a VC platform. The validation of these facial synchronization templates suggests that some pure nonverbal templates are strong indicators of the negotiation outcomes. This novel approach allows us to recognize the negotiation skills, and predict the negotiation outcomes. For example, in real-life scenario, professional negotiators may be trained to control their facial expressions or hide their feeling. Our approach can contribute to evaluate their performance and the effectiveness of the tactics.

The discovered facial synchronization templates can be embedded with active learning scheme as to evaluate VC communication skill and provide real-time feedback in computer-mediated communication. Our model can also be generalized to analyze other conversation scenarios such as interview, customer service, and tele-medicine.

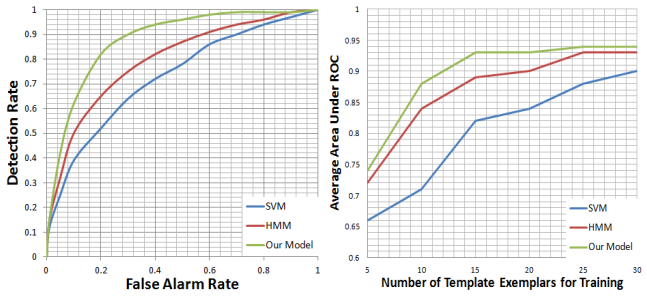

Fig. 6: ROC curve summarizing prediction performance for negotiation winners. Left: Area under average ROC curves for different numbers of template exemplars. Right: We compare our model with canonical HMM and SVM.

VI. ACKNOWLEDGMENTS

The authors acknowledge the help of Kazi Tasnif Islam, Anis Kallel and RuJie Zhao for the data collection.

REFERENCES

[1] N.E. Dunbar, M.L. Jensen, D.C. Tower and J.K. Burgoon, Synchornization of Nonverbal Behaviors in Detecting Mediated and Non-mediated Deception, *Nonverbal Behav. J.*, vol. 38, 2014, pp 355-376.

[2] J.R. Curhan, R. Li and M.E. Hoque, Predicting Negotiation Outcomes from Smiles, in preparation, 2015.

[3] C.N. Gunawardena, Social Presence Theory and Implications for Interaction and Collaborative Learning in Computer Conferences, *Educational Telecommunications Inter. J.*, vol. 1, 1995, pp 147-166.

[4] J.B. Walther, Computer-Mediated Communication Impersonal, Interpersonal, and Hyperpersonal Interaction, *Communication Research J.*, vol. 23, 1996, pp 3-43.

[5] J. Caukin, 35 Million People Concurrently Online on Skype, *Retrieved Dec. 1, 2013 from Skype*.

[6] Paul Ekman and W. V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, *Consulting Psychologists Press*, 1978.

[7] M. Gratier, Expressive Timing and Interactional Synchrony between Mothers and Infants: Cultural Similarites, Cultural Differences, and the Immigration Experience, *Cog. Dev. J.*, vol. 18, 2004, pp 533-554.

[8] J. Cassell, Embodied Conversational Interface Agents, *Comm. of the ACM J.*, vol. 43, 2000, pp 70-78.

[9] X. Yu, S. Zhang, Y. Yu and N. Dunbar, "The Computer Expression Recognition Toolbox (CERT)",*in Seventh IEEE International Conference on Automatic Face and Gesture Recognition*, Santa Barbara, CA, 2011, pp. 298-305.

[10] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan and M. Bartlett, "Automated Analysis of Interactional Synchrony using Robust Facial Tracking and Expression Recognition",*in Tenth IEEE International Conference on Automatic Face and Gesture Recognition*, Shanghai, China, 2014, pp. 1-6.

[11] M. Muhlenbrock and U. Hoppe, "Computer Supported Interaction Analysis of Group Problem Solving", *in Third International Conference on Computer Support for Collaborative Learning*, Palo Alto, CA, 1999, pp. 50.

[12] R. Thibaux and M.I. Jordan, "Hierarchical Beta processes and the Indian Buffet Process". *in Tenth International Conference on Artificial Intelligence and Statistics*, San Juan, Puerto Rico, 2007, pp. 564–571.