Computational Study of Human Communication Dynamic

Louis-Philippe Morency
Institute for Creative Technologies
University of Southern California
Los Angeles, CA 90094
morency@ict.usc.edu

ABSTRACT
Face-to-face communication is a highly dynamic process where participants mutually exchange and interpret linguistic and gestural signals. Even when only one person speaks at the time, other participants exchange information continuously amongst themselves and with the speaker through gesture, gaze, posture and facial expressions. To correctly interpret the high-level communicative signals, an observer needs to jointly integrate all spoken words, subtle prosodic changes and simultaneous gestures from all participants. In this paper, we present our ongoing research effort at USC MultiComp Lab to create models of human communication dynamic that explicitly take into consideration the multimodal and interpersonal aspects of human face-to-face interactions. The computational framework presented in this paper has wide applicability, including the recognition of human social behaviors, the synthesis of natural animations for robots and virtual humans, improved multimedia content analysis, and the diagnosis of social and behavioral disorders (e.g., autism spectrum disorder).

Categories and Subject Descriptors
I.2.7 **[Artificial Intelligence]**: Natural Language Processing - *Discourse*; I.2.11 **[Artificial Intelligence]**: Distributed Artificial Intelligence - *Intelligent agents.*

General Terms
Algorithms, Experimentation, Theory

Keywords
Human communication dynamics, context-based recognition, backchannel feedback prediction

1. INTRODUCTION
Face-to-face communication is a highly interactive process where participants mutually exchange and interpret verbal and nonverbal messages. Communication dynamics represent the temporal relationship between these communicative messages. Even when only one person speaks at the time, other participants exchange information continuously amongst themselves and with the speaker through gesture, gaze, posture and facial expressions. The transactional view of human communication shows an important dynamic between communicative behaviors where each person serves simultaneously as speaker and listener [15]. At the same time you send a message, you also receive messages from your own communications (individual dynamics) as well as from the reactions of the other person(s) (interpersonal dynamics) [2].

Individual and interpersonal dynamics play a key role when a teacher automatically adjusts his/her explanations based on the student nonverbal behaviors, when a doctor diagnoses a social disorder such as autism, or when a negotiator detects deception in the opposite team. An important challenge for artificial intelligence researchers in the 21st century is in creating socially intelligent robots and computers, able to recognize, predict and analyze verbal and nonverbal dynamics during face-to-face communication. This will not only open up new avenues for human-computer interactions but create new computational tools for social and behavior researchers --software able to automatically analyze human social and nonverbal behaviors, and extract important interaction patterns.

In this paper, we present recent results from USC Multimodal Communication and Machine Learning Laboratory (Multicomp Lab) to build computational models of human communication dynamics. We use the example of listener backchannel feedback to illustrate the importance of integrating all level of human communication dynamics. We present latent variable probabilistic models that were specifically created to learn the multimodal aspect of individual dynamic. Then we present predictive models that learn the interpersonal dynamic between listener and speaker. We show that integrating opinions from multiple listeners (known as wisdom of crowds) significantly improve the predictive power of our probabilistic models. Finally, we present an approach to integrate explicitly the individual and interpersonal dynamics.

2. HUMAN COMMUNICATION DYNAMICS
Human face-to-face communication is a little like a dance, in that participants continuously adjust their behaviors based on verbal and nonverbal behaviors from other participants. We identify four important types of dynamics during social interactions:

- **Behavioral dynamic** A first relevant dynamic in human communication is the dynamic of each specific behavior. For example, a smile has its own dynamic in the sense that the speed of the onset and offset can change its meaning (e.g., fake smile versus real smile). This is also true about words when pronounced to emphasize their importance.

J-HGBU'11, December 1, 2011, Scottsdale, Arizona, USA.

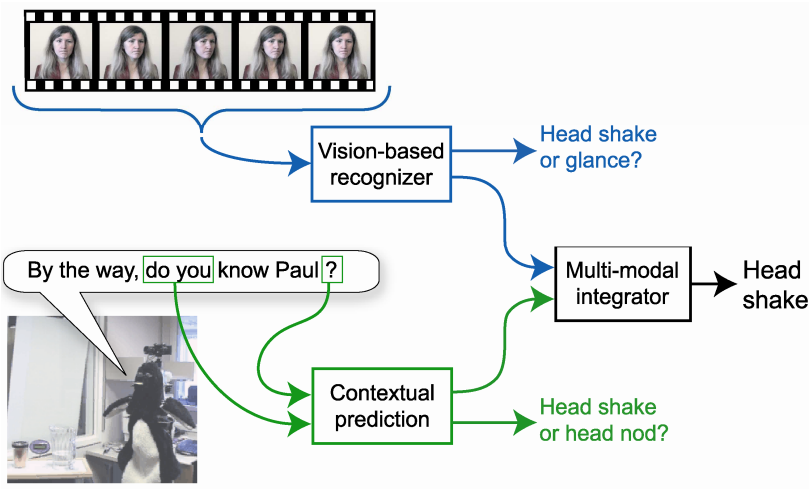

Figure 1: Example of individual and interpersonal dynamics: Context-based gesture recognition using prediction model. In this scenario, contextual information from the robot's spoken utterance (interpersonal dynamic) helps disambiguating the listener's visual gesture (individual dynamic).

The behavioral dynamic needs to be correctly represented when modeling social interactions.

- **Individual dynamic** Even when observing participants individually, the interpretation of their behaviors is a multimodal problem in that both verbal and nonverbal messages are necessary to a complete understanding of human behaviors. Individual dynamics represent this influence and relationship between the different channels of information such as language, prosody and gestures. Modeling the individual dynamics is challenging since gestures may not always be synchronized with speech and the communicative signals may have different granularity (e.g., linguistic signals are interpreted at the word level while prosodic information varies much faster).
- **Interpersonal dynamic** The verbal and nonverbal messages from one participant are better interpreted when put into context with the concurrent and previous messages from other participants. For example, a smile may be interpreted as an acknowledgement if the speaker just looked back at the listener and paused while it could be interpreted as a signal of empathy if the speaker just confessed something personal. Interpersonal dynamics represent this influence and relationship between multiple sources (e.g. participants). This dynamic is referred as micro-dynamic by sociologists [3].
- **Societal dynamic** We categorize the organizational (often referred as meso-level) and societal (often referred as macro-level) dynamics in this general category which emphasize the cultural change in a large group or society. While this paper does not focus on societal dynamics, it is important to point out the bottom-up and top-down influences. The bottom-up approach emphasizes the influence of micro-dynamics (behavioral, individual and interpersonal) on large-scale societal behaviors (e.g., organizational behavior analysis based on audio micro-dynamics[9]). As important is the top-down influence of society and culture on individual and interpersonal dynamics.

2.1 Example: Backchannel Feedback

A great example of individual and interpersonal dynamics is backchannel feedback, the nods and para-verbals such as "uh-huh" and "mm-hmm" that listeners produce as someone is speaking [15]. They can express a certain degree of connection between listener and speaker (e.g., rapport), a way to show acknowledgement (e.g., grounding) or they can also be used for signifying agreement. Backchannel feedback is an essential and predictable aspect of natural conversation and its absence can significantly disrupt participant's ability to communicate [1]. Accurately recognizing the backchannel feedback from one individual is challenging since these conversational cues are subtle and vary between people. Learning how to predict backchannel feedback is a key research problem for building immersive virtual human and robots. Finally, there are still some unanswered questions in linguistic, psychology and sociology on what triggers backchannel feedback and how it differs from different cultures. In this article we show the importance of modeling both the individual and interpersonal dynamics of backchannel feedback for recognition, prediction and analysis.

3. MODELING LATENT DYNAMIC

One of the key challenges with modeling the individual and interpersonal dynamics is to automatically learn the synchrony and complementarities in a person's verbal and nonverbal behaviors and between people. We developed a new computational model called Latent-Dynamic CRF (see **Error! Reference source not found.**) which incorporates hidden state variables that model the sub-structure of a class sequence and learn dynamics between class labels [6]. It is a significant change from previous approaches which only examined individual modalities, ignoring the synergy between speech and gestures.

The task of the Latent-Dynamic CRF model is to learn a mapping between a sequence of observations $\boldsymbol{x} = \{x_1, x_2, \ldots, x_m\}$ and a sequence of labels $\boldsymbol{y} = \{y_1, y_2, \ldots, y_m\}$. Each y_j is a class label for the j^{th} frame of a video sequence and is a member of a set $\mathbf{Y}$ of possible class labels, for example, $\mathbf{Y}$ = {*backchannel*, *other-gesture*}. Each observation x_j is represented by a feature vector $\Phi(x_j)$ in $\mathbf{R}^d$, for example, the head velocities at each frame. For

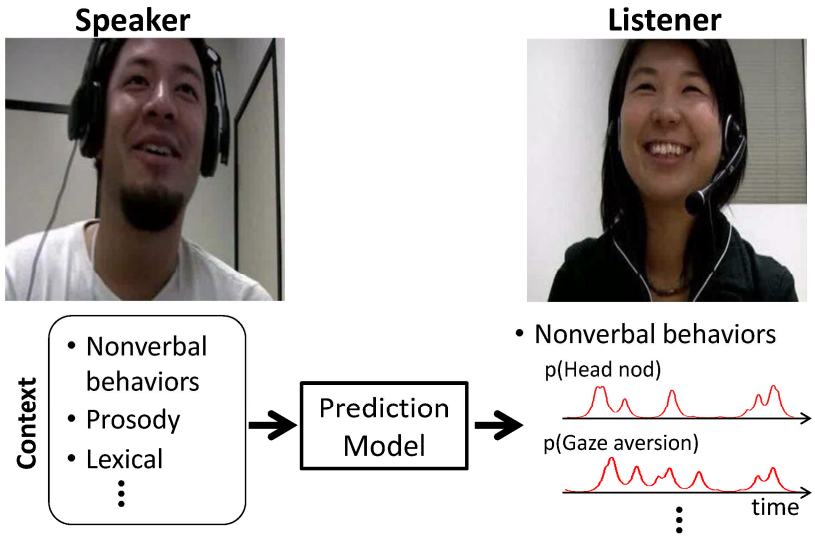

Figure 2: Prediction model of interpersonal dynamics: online prediction of the listener's backchannel based on the speaker's contextual features. In our contextual prediction framework, the prediction model automatically (1) learns which subset of the speaker's verbal and nonverbal actions influences the listener's nonverbal behaviors, (2) finds the optimal way to dynamically

each sequence, we also assume a vector of ``sub-structure" variables $\boldsymbol{h} = \{h_1,h_2,...,h_m\}$. These variables are not observed in the training examples and will therefore form a set of hidden variables in the model.

Given the above definitions, we define our latent conditional model:

$$P(\boldsymbol{y} \mid \boldsymbol{x}, \theta) = \sum_{\boldsymbol{h}} P(\boldsymbol{y} \mid \boldsymbol{h}, \boldsymbol{x}, \theta) P(\boldsymbol{h} \mid \boldsymbol{x}, \theta)$$

where θ are the parameters of the Latent-Dynamic CRF model. These are learned automatically during training using a gradient ascent approach to search for the optimal parameter values. More details can be found in [6].

We first applied the Latent-Dynamic CRF model to the problem of learning individual dynamics of backchannel feedback.

Figure 3 shows our LDCRF model compared previous approaches for probabilistic sequence labeling (e.g. Hidden Markov Model and Support Vector Machine). By modeling the hidden dynamic, the Latent-Dynamic model outperforms previous approaches. The software was made available online on an open-source website (sourceforge.net/projects/hcrf).

4. PREDICTION MODEL OF INTERPERSONAL DYNAMICS

In our contextual prediction framework, the prediction model automatically learns which subset of a speaker's verbal and nonverbal actions influences the listener's nonverbal behaviors, finds the optimal way to dynamically integrate the relevant speaker actions and outputs probabilistic measurements describing the likelihood of a listener nonverbal behavior. Figure 2 presents an example of contextual prediction for the listener's backchannel.

The goal of a prediction model is to create online predictions of human nonverbal behaviors based on external contextual information. The prediction model learns automatically which contextual feature is important and how it affects the timing of nonverbal behaviors. This goal is achieved by using a machine learning approach wherein a sequential probabilistic model is trained using a database of human interactions.

Our contextual prediction framework can learn to predict and generate dyadic conversational behavior from multimodal conversational data, and applied it to listener backchannel feedback [8]. Generating appropriate backchannels is a notoriously difficult problem because they happen rapidly, in the midst of speech, and seem elicited by a variety of speaker verbal, prosodic and nonverbal cues. Unlike prior approaches that use a single modality (e.g., speech), we incorporated multimodal features (e.g., speech and gesture) and devised a machine learning method that automatically selects appropriate features from multimodal data and produces sequential probabilistic models with greater predictive accuracy

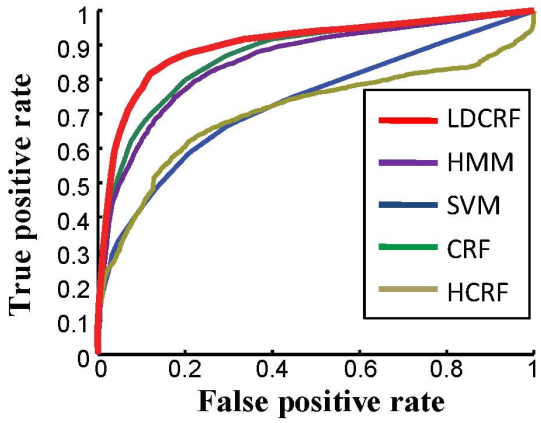

Figure 3: Recognition of backchannel feedback based on individual dynamics only. Comparison of our Latent-Dynamic CRF model with previous approaches for probabilistic sequential modeling.

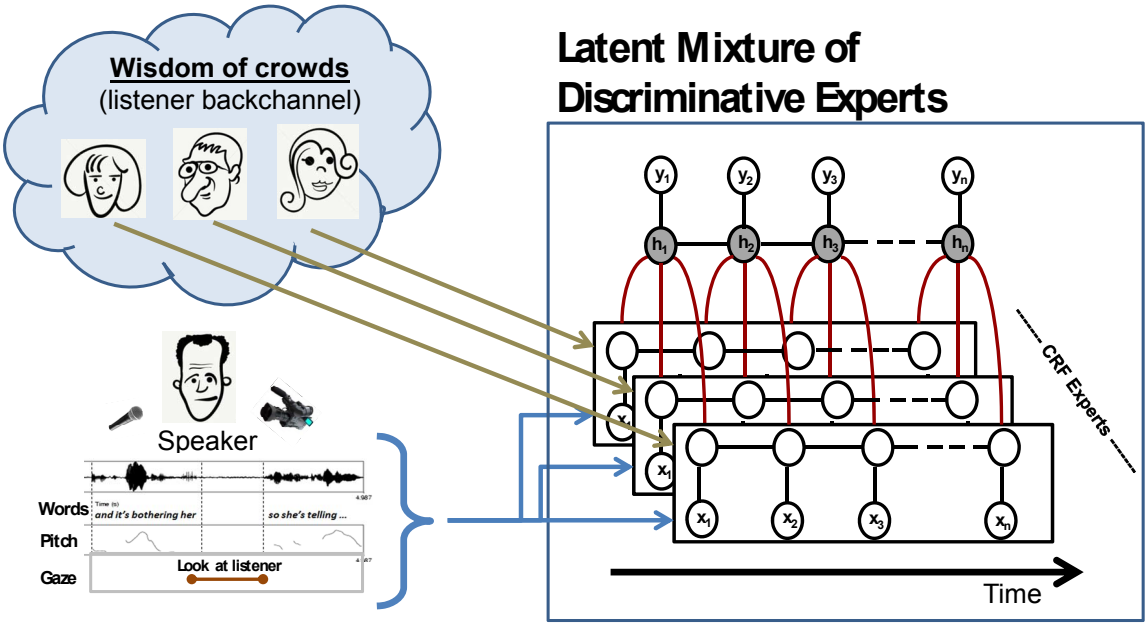

Figure 4: Our approach for modeling wisdom of crowd: (1) multiple listeners experience the same series of stimuli (pre-recorded speakers) and (2) a Wisdom-LMDE model is learned using this wisdom of crowds, associating one expert for each listener.

4.1 Signal Punctuation and Encoding Dictionary

While human communication is a continuous process, people naturally segment these continuous streams in small pieces when describing a social interaction. This tendency to divide communication sequences of stimuli and responses is referred to as punctuation [15]. This punctuation process implies that human communication should not only be represented by signals but also with communicative acts that represents the intuitive segmentation of human communication. Communicative acts can range from a spoken word to a segmented gesture (e.g., start and end time of a pointing) or a prosodic act (e.g., region of low pitch).

To improve the expressiveness of these communicative acts we propose the idea of encoding dictionary. Since communicative acts are not always synchronous, we allow them to be represented with various delay and length. In our experiments with backchannel feedback, we identified 13 encoding templates which represent a wide range of ways that speaker actions can influence the listener backchannel feedback. These encoding templates will help to represent long-range dependencies that are otherwise hard to learn using directly a sequential probabilistic model (e.g., when the influence of an input feature decay slowly over time, possibly with a delay). An example of a long-range dependency will be the effect of low-pitch regions on backchannel feedback with an average delay of 0.7 seconds (observed by Ward and Tsukahara [12]). In our prediction framework (see [8] for details), the prediction model will pick an encoding template with a 0.5 seconds delay and the exact alignment will be learned by the sequential probabilistic model (e.g., Latent-Dynamic CRF) which will also take into account the influence of other input features. The three main types of encoding templates are:

- **Binary encoding:** This encoding is designed for speaker features which influence on listener backchannel is constraint to the duration of the speaker feature.
- **Step function:** This encoding is a generalization of binary encoding by adding two parameters: width of the encoded feature and delay between the start of the feature and its encoded version. This encoding is useful if the feature influence on backchannel is constant but with a certain delay and duration.
- **Ramp function:** This encoding linearly decreases for a set period of time (i.e., width parameter). This encoding is useful if the feature influence on backchannel is changing over time.

It is important to note that a feature can have an *individual* influence on backchannel and/or a *joint* influence. An *individual* influence means the input feature directly influences listener backchannel. For example, a long pause can by itself trigger backchannel feedback from the listener. A *joint* influence means that more than one feature is involved in triggering the feedback. For example, saying the word ``and'' followed by a look back at the listener can trigger listener feedback. This also means that a feature may need to be encoded more than one way since it may have an *individual* influence as well as one or more *joint* influences.

4.2 Wisdom of Crowds

In many real life scenarios, it is hard to collect the actual labels for training, because it is expensive or the labeling is subjective. To address this issue, a new direction of research appeared in the last decade, taking full advantage of the “wisdom of crowds” [12]. In simple words, wisdom of crowds enables the fast acquisition of opinions from multiple annotators/experts.

Based on this intuition, we model wisdom of crowds using Parasocial Consensus Sampling paradigm [4] for data acquisition, which allows quided crowd members to experience the same situation. Parasocial Consensus Sampling (PCS) paradigm is

Table 1: Comparison of our prediction model with previously published approaches. By integrating the knowledge from multiple listener, our Wisdom—LMDE is able to identify prototypical patterns in interpersonal dynamic.

Model	Wisdom of Crowds	Precision	Recall	F1-Score
Wisdom LMDE	Yes	0.2473	0.7349	**0.3701**
Consensus Classifier (Huang et al., 2010)	Yes	0.2217	0.3773	0.2793
CRF Mixture of Experts (Smith et al., 2005)	Yes	0.2696	0.4407	0.3345
AL Classifier(CRF)	No	0.2997	0.2819	0.2906
AL Classifier(LDCRF) (Morency et al., 2007)	No	0.1619	0.2996	0.2102
Multimodal LMDE (Ozkan et al., 2010)	No	0.2548	0.3752	0.3035
Random Classifier	No	0.1042	0.1250	0.1018
Rule Based Classifier(Ward et al.,2000)	No	0.1381	0.2195	0.1457

based on the theory that people behave similarly when interacting through a media (e.g., video conference).

The goals of our computational model are to automatically discover the prototypical patterns of backchannel feedback and learn the dynamic between these patterns. This will allow the computational model to accurately predict the responses of a new listener even if he/she changes her backchannel patterns in the middle of the interaction. It will also improve generalization by allowing mixtures of these prototypical pattern.

To achieve these goals, we propose a variant of the Latent Mixture of Discriminative Experts [9] which takes full advantage of the wisdom of crowds. Our Wisdom-LMDE model is based on a two step process: a Conditional Random Field (CRF) is learned first for each expert, and the outputs of these models are used as an input to an Latent Dynamic Conditional Random Field (LDCRF, see Figure 3) model, which is capable of learning the hidden structure within the input. In our Wisdom-LMDE, each expert corresponds to a different listener from the wisdom of crowds. Figure 4 show an overview of our approach.

Table 1 summarizes our experiments comparing our Wisdom-LMDE model with state-of-the-art approaches for behavior prediction (see [10] for more details). Our Wisdom-LMDE model achieves the best f-1 score. The second best f-1 score is achieved by CRF Mixture of experts, which is the only model among other baseline models that combines the different listener labels in a late fusion manner. This result supports our claim that wisdom of clouds improves learning of prediction models.

5. CONTEXT-BASED RECOGNITION: COMBINING INDIVIDUAL AND INTERPERSONAL DYNAMICS

Modeling human communication dynamics implies being able to model both the individual multimodal dynamics and the interpersonal dynamics. A concrete example where both types of dynamics are taken into account is context-based recognition (see Figure 1). When recognizing and interpreting human behaviors, people use more than their visual perception; knowledge about the current topic and expectations from previous utterances help guide recognition of nonverbal cues. In this framework, the interpersonal dynamic is interpreted as contextual prediction since an individual can be influenced by the conversational context but at the end he or she is the one deciding to give feedback or not.

Figure 1 shows an example of context-based recognition where the dialogue information from the robot is used to disambiguate the individual behavior of the human participant. When a gesture occurs, the recognition and meaning of the gesture is enhanced due to this dialogue context prediction. Thus recognition is enabled by the meaningfulness of a gesture in dialogue. However, because the contextual dialogue information is subject to variability when modeled by a computational entity, it cannot be taken as ground truth. Instead features from the dialogue that predict a certain meaning (e.g., acknowledgement) are also subject to recognition prediction. Hence in the work reported here, recognition of dialogue features (interpersonal dynamic) and recognition of feedback features (individual dynamic) are interdependent processes.

We showed that our contextual prediction framework can significantly improve performance of individual-only recognition when interacting with a robot, a virtual character or another human [7]. Figure 5 shows the statistically significant improvement ($p<0.0183$) when integrating the interpersonal dynamic (contextual prediction) with individual dynamic (vision-based recognition).

6. DISCUSSION

Modeling human communication dynamics enables the computational study of different aspect of human behaviors.

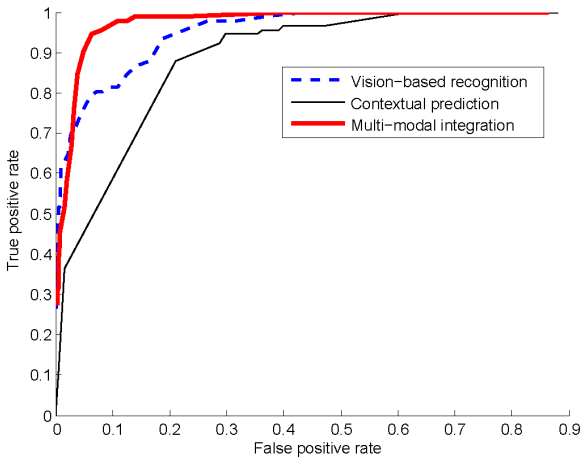

Figure 5: Backchannel feedback recognition curves when varying the detection threshold. For a fixed false positive rate of 0.0409 (operating point), the context-based approach improves head nod recognition from 72.5% (vision only) to 90.4%.

While a backchannel feedback such as head nod may at first look like a conversational signal ("I acknowledge what you said"), it can also be interpreted as an emotional signal where the person is trying to show empathy or a social signal where the person is trying to show dominance by expressing a strong head nod. The complete study of human face-to-face communication needs to take into account these different types of signals: linguistic, conversational, emotional and social. In all four cases, the individual and interpersonal dynamics are keys to a coherent interpretation.

As we already shown in this article, modeling human communication dynamics is important for both recognition and prediction. One other important advantage of these computational models is the automatic analysis of human behaviors. Studying interactions is grueling and time-consuming work. The rule of thumb in the field is that each recorded minute of interaction takes an hour or more to analyze. Moreover, many social cues are subtle, and not easily noticed by even the most attentive psychologists.

By being able to automatically and efficiently analyze a large quantity of human interactions, and detect relevant patterns, these new tools will enable psychologists and linguists to find hidden behavioral patterns which may be too subtle for the human eye to detect, or may be just too rare during human interactions. A concrete example is our recent work which studied engagement and rapport between speakers and listeners, specifically examining a person's backchannel feedback during conversation[8]. This research revealed new predictive cues related to gaze shifts and specific spoken words which were not identified by previous psycho-linguistic studies. These results not only give an inspiration for future behavioral studies but also make possible a new generation of robots and virtual humans able to convey gestures and expressions at the appropriate times.

7. CONCLUSION

This paper presented a computational framework to analyze human social behaviors during face-to-face interactions. The framework is based on 4 levels of human communication dynamics: behavioral dynamic, individual dynamic, interpersonal dynamic and societal dynamic. We showed how behavioral, individual and interpersonal dynamics can be integrated using the example of backchannel feedback prediction. We showed that by analyzing the wisdom of crowds from multiple listener, we can identify prototypical patterns and significantly improve prediction performance. These results pave the way to new exciting research in the automatic analysis of conversational, emotional and social signals.

8. REFERENCES[1]

[1] J. B. Bavelas and L. Coates and T. Johnson, Listeners as Co-narrators, *Journal of Personality and Social Psychology*, 79(6):941-952, 2000

[2] J. DeVito, *The Interpersonal communication book*, 12th edition, Pearson/Allyn and Bacon, 2008

[3] Hawley A.H. (1950) Human ecology: A theory of community structure. Ronald Press.

[4] L. Huang, L.-P. Morency, and J. Gratch, Parasocial consensus sampling: combining multiple perspectives to learn virtual human behavior. AAMAS 2010.

[5] D. McNeill, *Hand and mind: What gestures reveal about thought*, University of Chicago Press, 1996

[6] L.-P. Morency, A. Quattoni and Trevor Darrell, Latent-Dynamic Discriminative Models for Continuous Gesture Recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June, 2007

[7] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell, Head Gestures for Perceptual Interfaces: The Role of Context in Improving Recognition. *Artificial Intelligence*. Elsevier, 171(8-9):568-585, June, 2007

[8] 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*. Springer, 20(1):70-84, January 2010

[9] D. Ozkan, K. Sagae, and L.-P. Morency. Latent mixture of discriminative experts for multimodal prediction modeling. *In International Conference on Computational Linguistics (COLING)*, 2010

[10] D. Ozkan and L.-P. Morency, Modeling Wisdom of Crowds Using Latent Mixture of Discriminative Experts, ACL 2011

[11] A. Pentland, "Social dynamics: Signals and behavior," in Proc. IEEE Int. Conf. Developmental Learning, San Diego, CA, October 2004

[12] A. Smith, T. Cohn, and M. Osborne. 2005. Logarithmic opinion pools for conditional random fields. In ACL, pages 18–25

[13] James Surowiecki. The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. *Doubleday*, 2004

[14] N. Ward and W. Tsukahara (2000), Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 23:1177-1207

[15] P. Watzlawick, J. B. Bavelas, and D. D. Jackson, Pragmatics of Human Communication A Study of Interactional Patterns, Pathologies, and Paradoxes Chapter: Psychotherapy, 1967

[16] V. H Yngve, On getting a word in edgewise, *Sixth regional Meeting of the Chicago Linguistic Society*, pp. 567-577, 1970.

[1] See our group webpage for more details about individual projects described in this paper:

http://multicomp.ict.usc.edu/