Modeling Human Communication Dynamics

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 a time, other participants exchange information continuously among 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 [9]. 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 that are 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 interaction 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.

Challenges with Individual and Interpersonal Dynamics

Human face-to-face communication is a little like a dance in that participants continuously adjust their behaviors based on verbal and nonverbal displays and signals. 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 represents this influence and relationship between the different channels of information such as speech and gestures.

Face-to-face communication is a highly interactive process where participants mutually exchange and interpret verbal and nonverbal messages.

Example: Backchannel Feedback

A great example of individual and interpersonal dynamics is backchannel feedback, the nods and para-verbs such as “uh-huh” and “mm-hmm” that listeners produce as someone is speaking [10]. They can express a certain degree of connection between listener and speaker (e.g., rapport), a way to show acknowledgment (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 participants’ ability to communicate [1]. Accurately recognizing the backchannel feedback from one individual is challenging since these conversational cues are subtle and vary between people.

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 acknowledgment 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. Another example is illustrated in Figure 1. Interpersonal dynamics represent this influence and relationship between multiple sources (e.g., participants). Modeling the individual and interpersonal dynamics becomes a multi-signals, multichannels, and multisource problem. Both individual and interpersonal dynamics need to be taken into account when modeling human communication.
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 conditional random field (LDCRF) that incorporates hidden state variables that model the substructure of a class sequence and learn dynamics between class labels [4]. It is a significant change from previous approaches that only examined individual modalities, ignoring the synergy between speech and gestures.

The task of the LDCRF model is to learn a mapping between a sequence of observations $x = \{x_1, x_2, \ldots, x_m\}$ and a sequence of labels $y = \{y_1, y_2, \ldots, y_m\}$. Each $y_j$ is a class label for the $j$th frame of a video sequence and is a member of a set $Y$ of possible class labels, for example, $Y = \{\text{backchannel, other-gesture}\}$. Each observation $x_j$ is represented by a feature vector $\varphi(x_j)$ in $\mathbb{R}^d$, for example, the head velocities at each frame. For each sequence, we also assume a vector of “substructure” variables $h = \{h_1, h_2, \ldots, h_m\}$. 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

$$P(y \mid x, \theta) = \sum_h P(y \mid h, x, \theta)P(h \mid x, \theta),$$

where $\theta$ are the parameters of the LDCRF model. These are learned automatically during training using a gradient ascent approach to search for the optimal parameter values. Inference can be easily computed in $O(m)$ using belief propagation [pearl-belief-book], where $m$ is the length of the sequence.

We first applied the LDCRF model to the problem of learning individual dynamics of backchannel feedback. Figure 2 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 Web site (sourceforge.net/projects/hcrf).

**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 [9]. This punctuation process implies that human communication should not only be represented by signals but also with communicative acts that represent 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 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 that 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 s (observed by Ward and Tsukahara [8]). In our framework,
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 integrate the relevant speaker actions, and 3) outputs probabilistic measurements describing how likely listener nonverbal behavior are.

<table>
<thead>
<tr>
<th>Predictive Score</th>
<th>Rules</th>
<th>Random</th>
<th>Random Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Test (p-Value)</td>
</tr>
<tr>
<td>Our Prediction Model</td>
<td>0.2236</td>
<td>0.1862</td>
<td>0.4106</td>
</tr>
<tr>
<td>Rule-Based Approach</td>
<td>0.1457</td>
<td>0.1381</td>
<td>0.2195</td>
</tr>
<tr>
<td>Random</td>
<td>0.1018</td>
<td>0.1042</td>
<td>0.1250</td>
</tr>
</tbody>
</table>

**Binary encoding**: This encoding is designed for speaker features which influence on listener backchannel is constraint to the duration of the feature. The three main types of encoding templates are the following:

- **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.

**Random 0.1018 0.1042 0.1250 – –**

**Rule-Based Approach**

0.1457 0.1381 0.2195 0.0571 –

**Random**

0.1018 0.1042 0.1250 – –

**Results**

**T-Test (p-Value)**

**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 [6]. 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 (see Figure 4).
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., acknowledgment) 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 [5]. Figure 5 shows the statistically significant improvement (p < 0.0183) when integrating the interpersonal dynamic (contextual prediction) with individual dynamic (vision-based recognition).

BEHAVIOR ANALYSIS TOOL

As we have 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 goal of feature selection is to find the most relevant subset of contextual features for predicting a specific nonverbal behavior. 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, this software enables psychologists and linguists to find hidden behavioral patterns that may be too subtle for the human eye to detect, or may be just too rare during human interactions. The goal of feature selection is to find the most relevant subset of contextual features for predicting a specific nonverbal behavior. By reducing the dimensionality of the data, we allow a probabilistic model to operate faster and more effectively. The outcome of feature selection is two-fold: improved accuracy of our prediction model; and a more compact, easily interpreted representation of the relevant contextual features.

A concrete example is our recent work that studied engagement and rapport between speakers and listeners, specifically examining a person’s backchannel feedback during conversation [6]. This research revealed new predictive cues related to gaze shifts and specific spoken words that 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.

CONVERSATIONAL, EMOTIONAL, AND SOCIAL SIGNALS

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

MICRO, MESO, AND MACRO DYNAMICS

The individual and interpersonal dynamics discussed in this article are...
categorized by sociologist as microdynamics, in contrast to the mesodynamics which represents the organizational or institutional factors and the macrodynamics that drives our society and culture. The computational study of microdynamics enables a bottom-up approach to sociological research, where microbehaviors are used to define large-scale behaviors (e.g., organizational behavior analysis based on audio microdynamics [7]). As important is the top-down influence of society and culture on individual and interpersonal dynamics. The joint analysis of micro-, meso-, and macrodynamics will enable a better understanding of cultural differences in human communicative behaviors.

ACKNOWLEDGMENT

Code, data, and papers related to this work are available at http://projects.ict.usc.edu/multicomp/.

REFERENCES


REFERENCES


