Modeling Social Cues: Effective Features for Predicting Listener Nods

Faisal Khan, Bilge Mutlu, Xiaojin Zhu
Department of Computer Sciences
University of Wisconsin–Madison, Madison, WI 53706
{faisal, bilge, jerryzhu}@cs.wisc.edu

Abstract

Human communication involves a number of verbal, vocal, and non-verbal cues that facilitate face-to-face communication. Building a computational understanding of how these cues facilitate social and cognitive processes enables new research directions in social and cognitive sciences and the design of socially interactive systems such as agents and robots. In this paper, we present preliminary work in modeling a particular communicative mechanism—listener nods—toward designing humanlike agents and robots that show appropriate social behavior.[*]

1 Introduction

A computational understanding of human social behavior including verbal, vocal, and nonverbal cues will enable new directions in basic research in social and cognitive sciences and the design of socially interactive systems such as humanlike agents and robots. Recent work in these areas has shown that careful studies of temporal and spatial patterns in these cues help to better understand developmental and traumatic injuries [3, 16] and design more effective communicative mechanisms for artificial systems, particularly avatars in immersive environments [1], virtual characters [8], and humanlike robots [14]. These promising results, however, are shadowed by major challenges in modeling social cues, particularly due to significant differences across individuals, social situations, and cultures [2, 4, 7] and interdependencies between people [5] and between social cues and cognitive processes [6, 11]. The lack of appropriate computational methods to model these differences and complex interdependencies further aggravate these challenges.

These challenges have led to the development of new approaches to modeling social cues that draw on machine learning techniques [17, 12]. Our research seeks to contribute to the efforts in gaining a computational understanding of human social cues with the goal of building socially adept humanlike agents and robots. Existing approaches to this modeling problem have explored the use of temporal models such as conditional random fields (CRFs) [13] and dynamic Bayesian networks (DBNs) with semi-Markov processes [15]. Morency and his colleagues [12] developed a temporal model called latent-dynamic conditional random fields (LDCRFs) that created a mapping between observations of verbal and nonverbal cues from the speaker and listener nods and a hidden *substructure* in a latent conditional model. The predictive power of this model outperforms alternative stochastic approaches [12] and rule-based models [19]. This approach, however, uses a rich set of human-coded conversational features that depend on precise speech and video analysis. In this paper, we focus on building predictive models of social behavior, particularly listener nods in a storytelling scenario, using a small set of automatically extracted multimedia features that we can later deploy in real-time human-computer interaction situations.

2 Methodology

Our exploration of how a small set of automatically extracted features might predict listener nods involves (1) collecting multimodal data from a large number of dyadic interactions, (2) processing this

[*] This research was supported by the National Science Foundation Award IIS-1017952.

data using speech and vision processing techniques to extract verbal and nonverbal "raw" features, (3) engineering a feature vector that incorporates raw features and the temporal and interactional interdependencies among them, (4) training an SVM-based learning algorithm using data from a subset of the dyadic interactions we captured, and (5) testing the trained model with data from a different subset of the interactions. This section describes these steps in detail.

2.1 Data Collection

Experimental Setup – Our experimental setup mimicked stereotypical face-to-face communication in which two unfamiliar participants sit across each other at a "social distance" [9] of five feet (Figure 1). The data collection equipment included three high-definition video cameras at 1080p resolution and 30p frame rate, two high-fidelity lapel microphones, and an omni-directional microphone. Two of the cameras captured the upper torsos of the participants from a direct frontal angle and the lapel microphones captured participants' speech. The third camera and the omni-directional microphone captured the speech and nonverbal behaviors of both participants from the side.

Experimental Task – The experimental task involved three conversational scenarios that we developed to capture a wide range of behavioral and interactional mechanisms. Nodding while listening is an example of such mechanisms. These scenarios include *storytelling*, *interviewing*, and *discussion*. The current study uses data collected *only* from the storytelling scenario in which one of the participants narrated the plot of their favorite movie to the other participant. We expected this scenario to provide us with a rich context to observe listener nods.

Procedure – The experiment started by providing participants with a brief description of the experiment and asking participants to review and sign a consent form. The experimenter than seated the participants, provided them with more detail on their conversational roles, and set up the data collection equipment. Before performing the experimental task, participants performed an acclimation task (getting acquainted). The participants then performed the three conversations in a stratified order. At the end of the experiment, the experimenter debriefed the participants. The overall experiment took a total of 45 minutes. Participants received $10 for their participation.

Participants – We recruited 48 participants (24 females and 24 males) from the University of Wisconsin–Madison. They studied a diverse set of majors and aged between 18 and 28. All participants were native English speakers. We assigned participants into dyads, conversational roles, and conversational scenarios following a fully stratified design to control for effects of gender composition of the dyads and the ordering of conversations.

2.2 Cue Extraction

The final dataset included 4 hours and 52 minutes of audio and video data. To extract features of social cues from this dataset, we used automatic audio and video processing techniques, particularly speech segmentation, speaker classification, pitch extraction, head movement detection, and nod detection. These processes are described below.

Speech Segmentation– To distinguish speech from silence, we calculated the maximum likelihood of two Gaussian models that we estimated from frame energy using the Audioseg toolkit, [1] labeling the audio frames with the lowest mean as silence.

Speaker Classification – We classified each speech segment as belonging to either the speaker or the listener (while the listener's role did not involve speaking, we observed a minimal amount of speaking in the form of backchannel responses) by comparing the amount of energy in recordings from the two participants' microphones.

Pitch Extraction – Using the Praat toolkit,[2] we calculated the pitch values for each speech segment of the speaker and its slope to determine whether it ended with a rising or dipping intonation.

Head Movement Detection – To calculate the speaker's of head movement, we identified the bounds of the speaker's face using the Viola-Jones algorithm [18] and calculated the horizontal and vertical displacement using a frame-by-frame block-matching.

Head Nods – We identified nods by marking significant vertical displacements as backchannel state changes following the algorithm described by Kawato and Ohya [10].

[1] https://gforge.inria.fr/projects/audioseg

[2] http://www.fon.hum.uva.nl/praat/

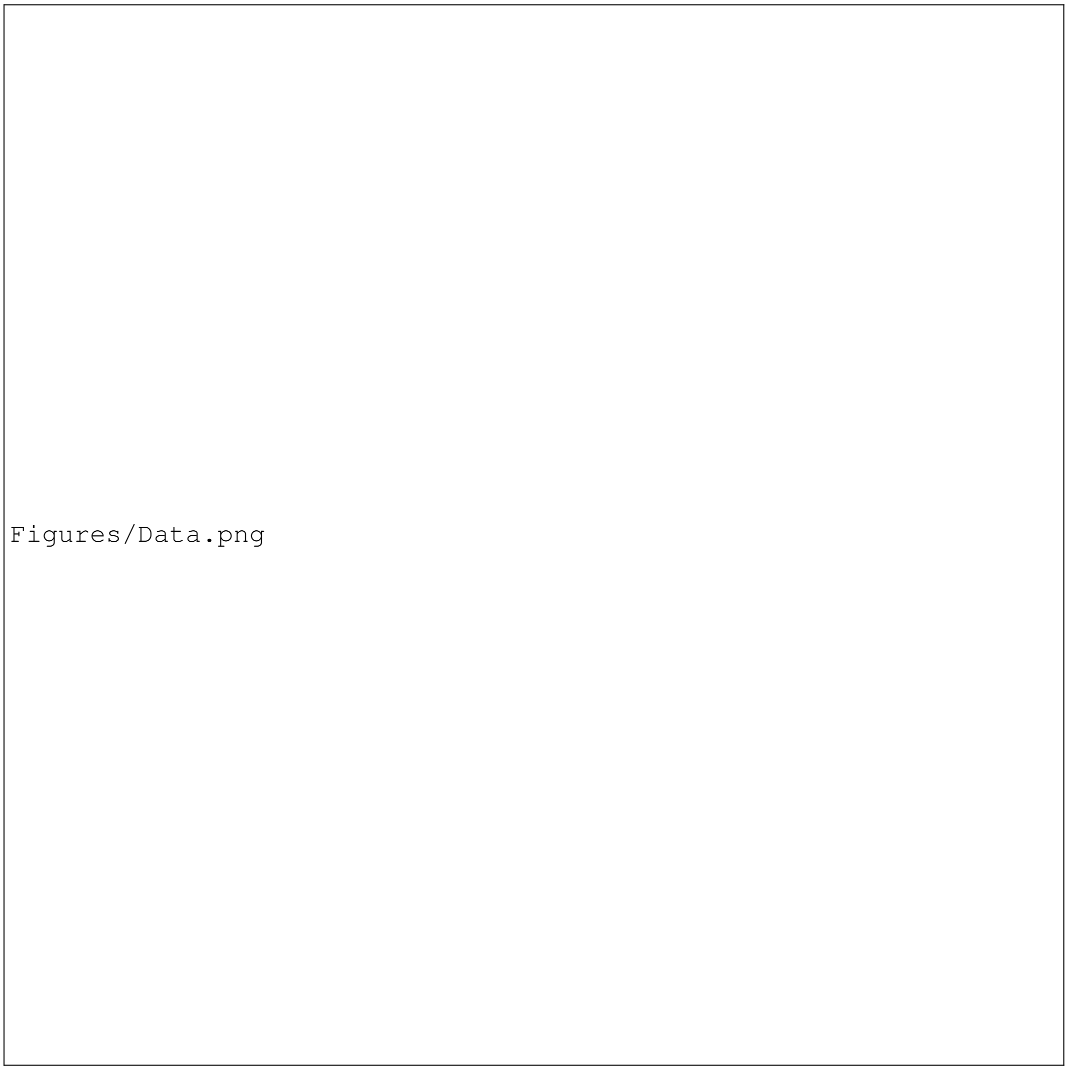

Figure 1: The spatial and equipment setup of our data collection (left), snapshots from the three video cameras (center), and automatically-extracted raw features and annotated output class (right).

2.3 Modeling

Our final feature set incorporated a set of “raw” and “derived” features for any given frame of the video and audio data. The *raw* features $\mathbf{r}_i$, where i indexes video frames, captured the following binary and normalized continuous variables.

- The absence (0) or presence (1) of speech segments, $speech = \{0, 1\}$
- Whether the designated speaker (0) or listener (1) is speaking, $speaker = \{0, 1\}$
- The speaker’s head displacement, $head_x \in (0, 1)$, $head_y \in (0, 1)$
- Whether the designated speaker is nodding (1) or not (0), $nodding = \{0, 1\}$
- Pitch value for speaker’s speech, $pitch \in (0, 1)$
- The slope of pitch values, $s \in (0, 1)$, over the window $i - 2^n$ to i, where $n = 1, \ldots, 9$

These features construct the 15×1 raw feature vector $\mathbf{r}_i$,

$$\mathbf{r}_i = [\ speech \quad speaker \quad head_x \quad head_y \quad nodding \quad pitch \quad s_1 \quad . \quad . \quad . \quad s_9\ ]'$$

The vectors of *derived* set of features, $\mathbf{g}_i^m$, captured the *average* values of these raw features over the window $i - 2^m$ to i, where $m = 1, \ldots, 7$ and, $\mathbf{h}_i^m$, captured the *change* in average values between two windows, $i - 2^m$ to i and $i - 2^{2m}$ to $i - 2^m$. Specifically,

$$\mathbf{g}_i^m = \frac{1}{2^m} \sum_{k=0}^{2^m - 1} \mathbf{r}_{i-k} \ \text{ and } \ \mathbf{h}_i^m = \mathbf{g}_i^m - \mathbf{g}_{i-2^m}^m$$

The 225×1 final feature vector $\mathbf{f}_i$ expresses the raw features and derived features for frame i,

$$\mathbf{f}_i = \left[\begin{array}{cccccccccc} \mathbf{r}_i & \mathbf{g}_i^1 & . & . & . & \mathbf{g}_i^7 & \mathbf{h}_i^1 & . & . & . & \mathbf{h}_i^7 \end{array} \right]'$$

The output class characterized whether the listener is nodding (1) or not (0), $nods = \{0, 1\}$. Ten percent of all instances had the label "nod." We established ground-truth for listener nods by manually labeling the dataset. A primary coder labeled 100% of the dataset. To evaluate coding reliability, a second coder labeled 10% of the data. The inter-rater reliability analysis showed substantial agreement between the two raters with an agreement of 94% and a Cohen's kappa of 0.72.

3 Results

We conducted four-fold cross-validation experiments to test the predictive performance of our features. In each fold, we trained an SVM learner with a randomly-selected set of 18 conversations and tested the learner with the remaining six conversations. The classification task involved using features from the speaker to classify the nodding behavior of the listener. To reduce training time, we subsampled the *training set* to include all frames with the label "nod" and the same number of frames with the label "not nods" that we observed immediately before and immediately after the nodding, producing twice as many instances of "not nods" than "nods." We used all frames of the *test set* for evaluation. We evaluated the predictions of our model using p (precision), r (recall), and F_1 scores. In the context of our experiments, p represents the proportion of the frames classified as nods that are actually nods, r denotes the proportion of the frames classified as nods out of the total number of frames that are nods in ground truth, and F_1 provides an equally weighted harmonic mean of the p and r scores. Our cross-validation tests produced a p score of 0.1083, an r score of 0.3165, and an F_1 score of 0.1605.

4 Discussion

In this paper, we presented the very first step of a long-term exploration in modeling human social cues, obtaining preliminary results from a model for predicting listener nods. While our prediction results are not directly comparable against those obtained using other datasets, they will serve as a baseline for our future exploration. The main contribution we sought to make in this paper is the ability to use a small set of automatically extracted features from multimodal data to predict social cues. Furthermore, we can later deploy the efficient audio and video processing techniques that we applied to extract these features to achieve real-time predictions—an important consideration in building interactive humanlike agents and robots.

Limitations and Future Work – A key limitation of the modeling approach we present here is the naive representation of temporal dependencies in social cues. In our future work, we plan to investigate how sequential approaches such as CRFs might allow us to better model these temporal dependencies. The second major limitation of our modeling approach is the use of video frames as the unit of analysis to model temporal events. While video frames provide us with a convenient way to discretize observations, it does not adequately capture the complex patterns in which social cues might co-occur. Our future work will involve exploring units of analysis and models that can capture these complex patterns and allow for varying unit lengths (e.g., modeling "events" using semi-Markov models [15]). Finally, while model predictions provide a valid measure for evaluating how well the model captures the dynamics of social behavior, high prediction accuracies do not guarantee that a humanlike agent or robot using these predictions to simulate social cues will communicate with people effectively. We plan to also conduct experiments with human subjects in which we manipulate the parameters of the predictive model that a humanlike agent or robot uses and measure the effects of these manipulations on the communicative effectiveness of a socially interactive system.

References

[1] J.N. Bailenson, N. Yee, K. Patel, and A.C. Beall. Detecting digital chameleons. *Computers in Human Behavior*, 24(1):66–87, 2008.

[2] A.P. Bayliss, G. Pellegrino, and S.P. Tipper. Sex differences in eye gaze and symbolic cueing of attention. *The Quarterly Journal of Experimental Psychology Section A*, 58(4):631–650, 2005.

[3] Z. Boraston and S.-J. Blakemore. The application of eye-tracking technology in the study of autism. *The Journal of Physiology*, 581(3):893–898, 2007.

[4] H.F. Chua, J.E. Boland, and R.E. Nisbett. Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences of the United States of America*, 102(35):12629, 2005.

[5] S. Duncan, B.G. Kanki, H. Mokros, and D.W. Fiske. Pseudounilaterality, simple-rate variables, and other ills to which interaction research is heir. *Journal of Personality and Social Psychology*, 46(6):1335, 1984.

[6] N.J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6):581–604, 2000.

[7] A. Frischen, A.P. Bayliss, and S.P. Tipper. Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological bulletin*, 133(4):694, 2007.

[8] J. Gratch, N. Wang, A. Okhmatovskaia, F. Lamothe, M. Morales, RJ Van Der Werf, and L.P. Morency. Can virtual humans be more engaging than real ones? In *Proceedings of the 12th international conference on Human-computer interaction: intelligent multimodal interaction environments*, pages 286–297. Springer-Verlag, 2007.

[9] E.T. Hall. A System for the Notation of Proxemic Behavior. *American anthropologist*, 65(5):1003–1026, 1963.

[10] S. Kawato and J. Ohya. Real-time detection of nodding and head-shaking by directly detecting and tracking the ”between-eyes”. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, page 40, Washington, DC, USA, 2000. IEEE Computer Society.

[11] S.R.H. Langton, R.J. Watt, and V. Bruce. Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2):50–59, 2000.

[12] L.P. Morency. Modeling Human Communication Dynamics [Social Sciences]. *Signal Processing Magazine, IEEE*, 27(5):112–116, 2010.

[13] L.P. Morency, I. de Kok, and J. Gratch. Predicting listener backchannels: A probabilistic multimodal approach. In *Intelligent Virtual Agents*, pages 176–190. Springer, 2008.

[14] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 61–68. ACM, 2009.

[15] K. Otsuka, H. Sawada, and J. Yamato. Automatic inference of cross-modal nonverbal interactions in multiparty conversations: who responds to whom, when, and how? from gaze, head gestures, and utterances. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 255–262. ACM, 2007.

[16] L.S. Turkstra, S.E. Brehm, and E.B. Montgomery Jr. Analysing Conversational Discourse after Traumatic Brain Injury: Isn’t It about Time? *Brain Impairment*, 7(3):234–245, 2006.

[17] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[18] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, 2004.

[19] N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in English and Japanese* 1. *Journal of Pragmatics*, 32(8):1177–1207, 2000.