Automatic Prediction of Friendship via Multi-model Dyadic Features

Zhou Yu, David Gerritsen, Amy Ogan, Alan W Black, Justine Cassell
School of Computer Science, Carnegie Mellon University
{zhouyu, dgerrits, aeo, awb, justine }@cs.cmu.edu

Abstract

In this paper we focus on modeling friendships between humans as a way of working towards technology that can initiate and sustain a lifelong relationship with users. We do this by predicting friendship status in a dyad using a set of automatically harvested verbal and nonverbal features from videos of the interaction of students in a peer tutoring study. We propose a new computational model used to model friendship status in our data, based on a group sparse model (GSM) with L2,1 norm which is designed to accommodate the sparse and noisy properties of the multi-channel features. Our GSM model achieved the best overall performance compared to a non-sparse linear model (NLM) and a regular sparse linear model (SLM), as well as outperforming human raters. Dyadic features, such as number and length of conversational turns and mutual gaze, in addition to low level features such as F0 and gaze at task, were found to be good predictors of friendship status.

1 Introduction and Related Work

While significant advances have been made in detecting the speech and nonverbal social signals emitted by individuals (see Vinciarelli, Pantic & Bourlard, 2009, for a review), and research has addressed the social roles and states of individuals in groups (see Gatica-Perez, 2009, for a review), considerably less computational work has focused on the automatic detection of speech or nonverbal correlates of specifically dyadic states, such as rapport. And yet rapport has been shown to have important effects on interactions as diverse as survey interviewing (Berg, 1989), sales (Brooks, 1989), and health (Harrigan et al., 1985). If we are to build interactive systems that are successful, then, we believe that the ability to build rapport with a human user will be essential.

Rapport can be instantaneous and can also build over time. Granovetter (1973) describes the strength of an interpersonal "tie" as a function of the time, emotional intensity, and reciprocity that accumulates between people. These ties mediate effects in myriad domains such as learning (Azmitia & Montgomery, 1993) and healthcare (Harrigan & Rosenthal, 1983).

Accordingly, analysis of initial exchanges and those after many years of interaction suggests that the behavioral signals that indicate rapport change over time. For example, in Tickle-Degnen and Rosenthal's highly cited model (1990), rapport consists of mutual attention, positivity, and coordination. High levels of positivity between conversational partners are common in the initial phases of a relationship, but positivity has been shown to decline, without a loss in rapport, as the number of interactions increases. In fact, Ogan et al. (2012) gave evidence that the use of playful rudeness between friends during peer tutoring correlates to greater learning. This leads to an associated challenge of spoken dialogue system development: creating systems that can develop social ties, and increase rapport with the user over repeated interactions to maximize beneficial outcomes.

While little work has addressed automatic detection, some prior work has addressed the problem of emitting signals to build rapport in dialogue and agent systems (Stronks et al., 2002; Bickmore & Picard, 2005; Gratch et al., 2006; Cassell et al., 2007; Bickmore et al., 2011), and we turn to this research for what cues might be important in rapport. The majority of this prior work, however, has addressed harmony – or instant rapport – rather than rapport over time. For those systems that have addressed friendship or the growth of rapport, most commonly the *number of interactions* has been used as a meter of relationship progression, instigating changes in the dialogue system as the social odometer scrolls onward (Cassell & Bickmore, 2003; Vardoulakis et al., 2012). Counting the times a dyad has interacted is a crude approximation of a relationship state, however; being able to detect the behavioral signals that people actually use to indicate relationship status would be superior.

In our own prior work (Cassell et al.,2007) we looked at particular hand-annotated nonverbal signals (such as nodding and mutual gaze) as operationalizations of rapport, and found that friends and non-friends indeed show differing distributions of each signal as a function of relationship state. In the current study, we move to the next step and automatically harvest a set of multimodal dyadic and time contingent features to identify those features that play a significant role in predicting friendship state. A major

Proceedings of the SIGDIAL 2013 Conference, pages 51–60,
Metz, France, 22-24 August 2013.

challenge for predicting relational states such as these is to construct a compact feature space that captures only reliable rapport signals and also generalizes across different users. To provide strength to our model (as well as to fit the multimodal nature of embodied conversational agents), we look at both acoustic and visual features. Such an approach takes advantage of the fact that multimodal aspects of communication are not redundant, but often complementary (Cassell, 2000).

However, dyadic behaviors such as conversational turns, mutual/non-mutual smile, mutual/non-mutual gaze, and mutual/non-mutual lean forward provide an additional challenge in modeling; no matter how important, they appear relatively rarely in conversational data. Thus standard non-sparse linear models, normally trained on high frequency factors, might assign too much weight to low frequency (i.e., sparse) features. In order to address issues of this sort Yuan and Lin (2007) introduced the group lasso. To address the sparse nature of our features in real-world data and the noise that occurs from different production sources, we propose an extension to this genre of technique in the form of a Group Sparse Model (GSM) which enforces sparsity with a L2,1 norm instead of the group lasso penalty (Chen, et al., 2011), due to the relatively efficient optimization process of L2,1 norms (Liu, et al., 2009). Unlike a straightforward sparse linear model (SLM) (Yang et al., 2010), which treats each feature independently, GSMs group features which share the same production source in the optimization process. In the GSM linear model, the removal of the assumption of independence between features means that the penalty is on group rather than individual features. Thus the model has general robustness to noise, since grouping features from the same production source can increase the overall confidence of the feature group.

Our contributions in this work, then, are three-fold: we (1) designed and implemented a method for automatic dyadic feature extraction which is based on low level features, and which yields strong predictive power of friendship status, (2) propose a new Group Sparse Model (GSM) with L2,1 norm, that deals with the noisy and sparse nature of the feature sets, and (3) illuminate, from this model, the nature of verbal and nonverbal behavior between friends and non-friends in a peer tutoring setting.

The remainder of the paper is organized as follows. We first describe the data set and introduce the features used in our experiments. We then describe the performance of the three computational models we evaluated. Finally, we discuss the contributions of different features to friendship prediction and provide an error analysis of our proposed model.

2 The Data Set

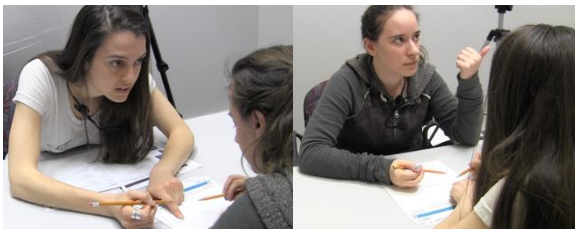

Figure 1: Camera View 1 and Camera View 2

We collected data from dyads of students engaged in a reciprocal peer tutoring task. We chose peer tutoring as it is a domain in which friendship has been shown to have a positive effect on student learning (see e.g. Ogan et al, 2012). In addition, tutoring systems that rely on dialogue are common, and peer tutoring dialogue systems are increasingly common. Thus, being able to assess friendship state in this domain is a useful step on the path to creating a peer tutoring agent that can use rapport to increase learning gains.

Each dyad consisted of two American English speakers with a mean age of 13.3 years (range = 12 – 15). We collected data from 12 dyads, of which 6 dyads were already friends. Dyads were either both girls or both boys, and each condition contained 3 boy dyads and 3 girl dyads.

Each dyad came to the lab for 3 sessions, with an average interval between visits of 4.6 days (*SD* = 3.1), totaling 36 sessions across all dyads. Each session consisted of about 90 minutes of interaction recorded from three camera views (a frontal view of each participant and a side view of the two participants). With close talk microphones, we also recorded the participants' speech in separate audio channels for the purpose of automatic dyadic acoustic feature extraction. The setting is shown in Figure 1.

Each session began with a short period of time for participants to become acquainted. After that, using a standard reciprocal tutoring procedure (see Fantuzzo et al., 1989), participants tutored each other on procedural and conceptual aspects of an algebra topic in which both participants were relatively novice. Order of seating and assignment of tutoring roles (tutor or tutee) was determined in the first session by alphabetical order of participant name. Tutoring roles alternated from that point on, such that both participants had the opportunity to take on the role of "expert" during each session. After a period of individual study time to familiarize

themselves with the material, the first tutoring period began and lasted approximately 25 minutes. This was followed by a 5 minute break, after which students' tutoring roles were reversed for a second tutoring period of 25 minutes. Finally, each student answered a survey about the interaction.

The current study examines only the tutoring sections of each session, which were divided into 30-second clips or "thin slices" (Ambady et al., 2006). In total, the data points used for modeling comprise 2259 clips from the 12 dyads.

3 Multimodal Information

In our analyses, low-level audio and visual features were automatically extracted using three off-the-shelf toolkits. Dyadic features, which are a second order derivative of the low level features, and which capture the interaction of two participants, are also automatically produced. Taken together, analysis of these features allows us to determine if the verbal and nonverbal behaviors of the participants index their friendship status in any significant way.

3.1 Low Level Audio Features (LA)

Type	# of Features
Prosodic Features	
F0	72
Energy	38
Duration	154
Voice Quality Features	
Jitter	68
Shimmer	34
Voicing	38
Spectral Features	
MFCC	570
Total	**974**

Table 1: Acoustic Feature Groups

For acoustic feature extraction, a large set of acoustic low-level descriptors (LLD) and derivatives of LLDs combined with appropriate statistical functionals, i.e., **maxPos** (the absolute position of the maximum value in frames), **minPos** (the absolute position of the minimum value in frames), **amean** (The arithmetic mean of the contour), etc., were extracted for each of the split channel recordings. The "INTERSPEECH 2010 Paralinguistic Challenge Feature Set" in the openSMILE toolkit (Schuller et al., 2012) was used as our basic acoustic feature set. For spectral features, Mel Spectrum and LSP were excluded due to the possible overlap with MFCC. The set contained 974 features which resulted from a base of 32 low-level descriptors (LLD) with 32 corresponding delta coefficients, and 21 functionals applied to each of these 68 LLD contours. In addition, 19 functionals were applied to the 4 pitch-based LLD and their four delta coefficient contours. Finally the number of pitch onsets (pseudo syllables) and the total duration of the input were included. The dimension of each feature group is shown in Table 1.

3.2 Low Level Vision Features (LV)

Type	# of Features
Face Position Feature	10
38 Face Interest Points	114
Gaze Features	3
Face Direction Features	4
Mouth and Eye Openness	6
Smile Intensity	1
Discretized Smile	1
Total	**139**

Table 2: Vision Feature Groups

Since participants were facing the camera directly most of the time, as seen in Fig 1, current technology for facial tracking can efficiently be applied to our dataset. OMRON's OKAO Vision System was used in face detection, facial feature extraction, and basic face related features extrapolation. For each frame, the vision software returns a smile intensity (0-100) and the gaze direction, using both horizontal and vertical angles expressed in degrees. Apart from gaze direction, the software also provides information about head orientation: horizontal, vertical, and roll (in or out). 38 additional face interest points, position and confidence, were also extracted. These were normalized to pixel coordinates, which turned out to lead to quite noisy data, and hence to diminished utility of these 38 points (in the future we will consider normalizing to face coordinates). We also calculated the openness of the left eye, right eye, mouth, and the location of the face. Details are shown in Table 2. Similar to our audio feature extraction method, one static feature vector per 30 second video clip was produced. All the features were computed at the same rate as the original videos: 30 Hz. Altogether, 139 dimensions were extracted in each frame from each camera view.

3.3 Dyadic Features (DF)

All of the features discussed above are low-level acoustic and visual features, extracted with

respect to individual participants. While individual behavior may index friendship state, we posit that patterns of interaction will be more effective. For example, prior research (Baker et al., 2008) suggests that the number and length of conversational turns (Cassell et al., 2007), presence of mutual smiles and non-mutual smiles (Prepin et al., 2012), mutual gaze and non-mutual gaze (Nakano et al., 2010), as well as posture shifting (Cassell, et al., 2001; Tickle-Degnen & Rosenthal, 1990), are important features to investigate in dyadic data. While other features such as gestures and mutual pitch shift may also play a role in indexing relationship state, these are not yet a part of the dyadic features we address here.

3.3.1 Number and Average Length of Conversational Turns

We recorded individual audio channels for each participant, which makes the automatic extraction of conversational turns possible. First, we extracted intervals of silence with toolbox SoX which produced speech chunks, and then identified the speaker by comparing the speech energy (loudness) in each audio channel, as speech from each speaker is carried by the other's microphone. After that we combined the speech chunks and speaker ID to approximate conversational turns. The approximation quality is not perfect, given the variability of the audio recording, but noise can be mediated during model building.

3.3.2 Mutual Smile and Non Mutual Smile

Prepin et al. (2012) describe the role of mutual smiles (smiles that occur during the same time period) in "stance alignment" and make the point that interactional alignment of this behavior reflects synchronization of internal states. Such synchrony predicts mutual understanding and increased quality of interaction, and as such is a fundamental quality in the formation of adolescent friendships (Youniss, 1982). Cappella & Pelachaud (2002) likewise describe "mutuality" as the precondition for how smiles function in contingent ways in a dyad. Smiles are clearly therefore important to assess in data such as ours. We defined a maximum window of 500 milliseconds between the end of one participant's smile and the beginning of the next for smiles to be considered mutual.

3.3.3 Mutual Gaze and Non-mutual Gaze

Nakano & Ishii (2010) describe eye gaze as a clue to engagement, and integrate mutual gaze into their conversational agents. There is no feature for direct gaze at partner provided in the OKAO vision toolkit. Mutual gaze was therefore approximated by annotating a gaze "in front," achieved by combining the information from three directions of gaze: vertical, horizontal, and depth. Gaze "in front", or at the partner, was recorded only if the participant gaze had less than a 15 degree angle from straight forward in all of these three directions. A maximum window of 500 milliseconds for gaze to be considered mutual was also employed here.

3.3.4 Mutual Lean Forward and Non-Mutual Lean Forward

Forward leaning has been shown to be a significant predictor of the ability to establish rapport in a dyad (Harrigan et al., 1985). In fact, friends who lean in are seen as more socially competent, while strangers are seen as less socially competent when they lean in (Burgoon & Hale, 1988). For our study, lean forward was approximated by detecting the smooth trend of face enlargement within the video frame. In order to improve precision of the feature, the segments with high confidence in face detection were processed. Furthermore, posture shifting, i.e., forward leaning, is not as quickly executed as changes in gaze or smile. We therefore used a 1 second sample window for lean forward, rather than a 500 millisecond window.

3.3.5 Mutual Gaze followed by Mutual Smile

Mutual gaze followed by mutual smile is also approximated using a similar approach as above. It is a relatively dense feature compared to all the other possible combinations of nonverbal behaviors, thus it is the only combination that is included in the feature set in this paper. The window within which mutual gaze is considered to be followed by mutual smile is set to be within 2 seconds.

4 Computational Model

We formulate friendship prediction as a set of binary classifications. In order to have the least variance and make sure no participant appeared in both the training and testing set, a leave-one-out cross-validation setting was adopted in all of our experiments. Each session had approximately 180 30-second video clips, totaling 2259 data points. Z-score normalization by dyad was used to scale all the features into the same range. Early fusion, which is simple concatenation of feature vectors, was adopted throughout our experiments to combine different features. We evaluated our group sparse model (GSM), along with a non-sparse linear model (NLM) and sparse linear model (SLM).

4.1 Non-sparse Linear Model (NLM)

We began with a standard non-sparse linear model (NLM), which is a Support Vector Machine (SVM) (Cortes & Vapnik, 1995) with a linear kernel. The libsvm (Fan et al., 2008) package was used in our experiment, and the parameter, the slack value of SVM that controls the scale of the soft margin, was obtained by cross validation.

4.2 Sparse Linear Model (SLM)

In order to prevent over-fitting on rare dyadic features, a sparse sensitive model SLM was introduced. As well as preventing over-fitting, through weight shrinkage the sparse model can also exclude redundant features. In our experiment, an L2,1 norm sparse model with linear kernel (Yang et al., 2012) was selected as our baseline sparse model.

4.3 Group Sparse Model (GSM)

Based on the SLM, we propose a group-sparse model (GSM) with the novel use of an L2,1 norm. Instead of assuming every feature is uncorrelated to other features, the GSM groups some of the features together and utilizes their correlated information to mediate the noise of the data. For an arbitrary matrix $A \in R^{r\times p}$, its $L_{2,1-norm}$ is defined as

$$||A||_{2,1} = \sum_{i=1}^{r} \sqrt{\sum_{j=1}^{p} A_{ij}^2}$$

Suppose that we have n training data indicated by $x_1, x_2, \ldots x_n$ and sampled from c classes. In our setting, c = 2, friends or non-friends. $y_i \in \{0,1\}$ $(1 \leq i \leq n)$ is the corresponding label. The total scatter matrix S_t and between class scatter matrix S_b are defined as follows.

$$S_t = \sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T = XX^T$$

$$S_b = \sum_{i=1}^{n} n_i(\mu_i - \mu)(\mu_i - \mu)^T = XGG^TX^T$$

where μ is the mean of all samples, μ_i is the mean of samples in the i-th class. n_i is the number of samples in the i-th class, $Y = [y_1, y_2, \ldots, y_n]$.

$$G = [G_1, \ldots, G_n]^T = Y(Y^TY)^{-1/2}$$

G is the scaled label matrix. A well-known method to utilize discriminate information is to find a low dimensional subspace in which S_b is maximized while S_t is minimized (Fukunaga et al., 1990). So the object function could be easily written as follows

$$\min_W \left(W^T\left(S_t {S_b}^{-1}\right)W\right) + \gamma||W||_{2,1}$$

$$s.t. W^TW = 1$$

The optimization of the above object function was introduced in Yang et al. (2012). It is an adaptation of iterative singular value decomposition. In GSM, a block-wise constraint is imposed on the diagonal matrix (D) which is the intermediate result of the iterative single value decomposition.

$$D = diag\left(\frac{1}{||w^1||_2}I_1, \ldots, \frac{1}{||w^G||_2}I_G\right)$$

W in the equation is the weight function, w^i is the i^{th} feature group in W, and there are a total number of G sub diagonal matrices corresponding to G groups of features.

For acoustic features, Steidl et al., (2012) designed a grouping schema which consists of Prosodic Features, Voice Quality Features and Spectral features which we adopted. For visual features, based on our observation of the highly unstable performance of the 38 feature points of the face, we introduced group bondage for the entire group to prevent single face features over-fitting the classifier. Detailed group information is shown in Table 1 and Table 2.

5 Human Baseline

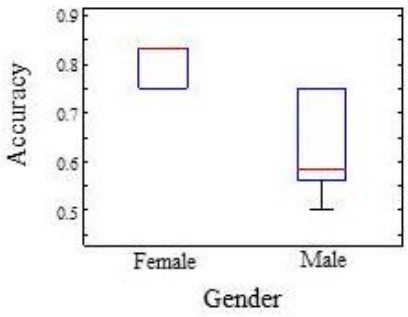

Figure 2: Boxplot of human rating accuracy with respect to gender.

In order to establish a baseline of the difficulty of predicting friendship, we conducted an experiment with humans, rating whether two people in a video were friends or not, after watching a 30-second video/audio clip taken from the first session of tutoring (in which the behaviors of strangers are most likely to be distinct from friends). We recruited 14 people and screened out participants with prior theoretical knowledge of nonverbal behavior, gesture, friendship, and rapport, or who rated all 12 clips in under 8 minutes, leaving 10 participants, half male, with an average age of 23 (*SD* 4.8). Each participant was asked to watch one 30-second clip per dyad, taken from 3 minutes after tutoring began. The mean accuracy of their friendship prediction was 0.717 (*SD* 0.119), which is significantly lower than our best GSM model (trained on all three sessions) applied to those same 12 clips, with a

performance of 0.837 ($t(11) = -2.1381$ $p.<.05$). When we split the ratings by gender, we found females on average were more accurate than males (see Figure 2). According to Hall et al., (1979) females are generally better decoders of nonverbal behaviors, which may lead to better judgment of friendship.

6 Results: Models

	Human	NLM	SLM	GSM
LV		0.743	0.768	0.792*
LA		0.674	0.664	0.682*
LV+DF		0.752	0.769	0.801*
LA+DF		0.679	0.681	0.683
LV+LA		0.744	0.780	0.803*
LV+LA+DF	0.717	0.749	0.782	0.814#

Table 3: The classification accuracy of the three algorithms on different features sets. Feature sets were combined with early fusion (+). Values marked * are significantly better ($p<.05$, pairwise t-test) than other results in the same row. Values marked # are significantly better ($p<.001$, pairwise t-test) than other results in the same column.

Our group sparse model (GSM) along with the non-sparse linear model (NLM) and sparse linear model (SLM) were evaluated on different combinations of three sets of features: low-level vision features (LV), low-level audio features (LA) and dyadic features (DF), and their performance is presented in Table 3. We did not evaluate dyadic features (DF) alone due to their sparse nature.

In particular, we found that adding the automatically extracted DF to LV and LA with early fusion improved the performance ($t(2258)= -3.12, p<.001$) of the GSM model. When using fewer modalities, our newly proposed GSM outperformed NLM and SLM ($t(2258)=-1.65$, $p<0.05$). However, when the number of feature sets increased, there was no statistical difference in performance between GSM and the other two models. We suspect that when features are abundant, the information that the features provide reaches a ceiling. The advantage of the GSM was gained by mediating the noise and sparseness of the data, which resulted in better weight assignment for each feature. Alternatively, when features are abundant, even NLM can have a comparative weight assignment by performing a greedy high dimensional feature space search. Thus there is limited room for further improvement by better weight assignment among the group features which GSM assumes.

When we looked at the top features selected by NLM using the vision modality alone, two (out of 38) face features, which had an unstable nature, appeared high in rank, which suggests the possibility of NLM over-fitting the noise of these features. Surprisingly, when more modalities are added, NLM stops picking single face features as top informative features. In GSM, none of the 38 face features are listed in the top ranked features for any of the modalities, which demonstrates its ability in noise mediation.

In real world applications, data sets which produce ideal, abundant, and accurate features are rarely encountered. We often end up with data that are poor in video quality, e.g. with no split channels for each participant or no frontal face view. Our newly proposed GSM may therefore be more robust when features are noisy or certain modalities are not available.

7 Results: Contributions of Features

Feature Name	Weight
Number of Conversational Turns & Average Length of Turns	0.041
Gaze Down	-0.036
Mutual Gaze	0.014
F0	0.013
Non-mutual Gaze	-0.013
Voicing	0.014
MFCC	-0.007
Non-mutual Smile	0.004
Non-mutual Lean Forward	0.004
Mutual Gaze followed by Mutual Smile	0.001

Table 4: The top 10 informative features and their weights as trained by GSM. Positive weight is associated with friends while negative weight is associated with non-friends.

After building the model and ranking the features, we looked into the weights learned for each feature. This weight comprises not only the magnitude, which tells us if the feature is important, but also the polarity. A detailed list of the most informative features and their weights is shown in Table 4.

The strongest feature is number and length of **conversational turns** which is grouped in the table and should be interpreted as meaning that friends have more and shorter conversational turns. This is consistent with previous research on direction giving (Cassell et al., 2007), and mirrors the fact that friends are more likely to interrupt one another.

We expected that unfamiliar participants, seated about two feet across from one another, would maintain a low level of eye contact (Argyle & Dean, 1965). In fact we found that non-friends tend to **gaze down** more often. In this context, non-friends spend more time

looking down at their study materials. In turn, **mutual gaze** is higher among friends.

Among the audio features, **F0**, which captures pitch related information such as range and mean, has been shown to differ between conversational and non-conversational speech (Bolinger, 1986). Here, friends show that more conversational style in their speech, despite the tutoring nature of the interaction.

In order to further examine the lessons to be learned from this GSM model about verbal and nonverbal behavior in friends and strangers, we also ran a repeated measures ANOVA, including both gender and friendship status as factors. There were no significant effects for gender, however, and so that factor was collapsed for further analysis. The four features described above were all significantly different between friends and strangers (although gaze down was simply a trend, at $p<.08$).

The following features were also important to the model, but did not show significance in the ANOVA, perhaps because of their sparse nature in our data. **MFCC** (Mel-Frequency Cepstral Coefficients) was associated with strangers and the similar audio feature of **voicing** was associated with friends. Both of these features have been described as approximating speech style – voicing, for example, may indicate more backchannels, such as "uh huh" and "hmm" (Ward, 2006).

In Nakano et al. (2003), listener gaze at the speaker is interpreted as evidence of non-understanding. We found similar results whereby non-friends were more likely to engage in **non-mutual** gaze – looking at one another when the other person was not looking back. Mutual smile did not distinguish between friends and non-friends, while **non-mutual smile**, on the other hand, provided indicative strength, in spite of its sparse nature, for friendship. This may relate to our prior work (Ogan et al., 2012) which found significant teasing and other behavior whereby friends appear comfortable enjoying themselves at the expense of the other.

Mutual lean forward lacked predictive power in our model, while **non-mutual lean forward** was more salient between friends. We often found, for example, that friends maintained very different postures, with a tutor leaning back much of the time, leaning forward only to answer a direct question from the tutee. Non-friends, on the other hand, tended to remain fixed on the study material. This may have been a display of formality, where a casual attitude would have been perceived as either impolite or inappropriate. In either relationship state, the tutee tended to sit hunched over the worksheet, and since we did not enter tutor state into the model, this may have washed out some tutor-specific results.

For the time contingent feature**, mutual gaze followed by mutual smile** is informative and predictive of friends.

8 Error Analysis and Discussion

Dyad ID	LA+DF	LV+DF	LA+LV+DF
1	0.732	0.809	0.819
2*	0.703	0.793	0.804
3*	0.574	0.771	0.778
4*	0.713	0.708	0.762
5	0.653	0.879	0.880
6	0.728	0.827	0.835
7	0.624	0.873	0.882
8*	0.712	0.861	0.852
9*	0.698	0.820	0.830
10	0.606	0.834	0.854
11*	0.700	0.682	0.743
12	0.749	0.780	0.785

Table 5: The average accuracy of classification in each dyad using the group sparse model (GSM) with different combination of feature sets. Dyads marked with * are friends

We performed an error analysis to understand the contexts under which our model failed to accurately predict friendship states, and here we discuss the implications of these examples for our work. Table 5 shows the average accuracy of each dyad using audio, visual, and dyadic features to predict friendship. Dyads 2, 3, 4, 8, 9 and 11 are friend dyads, and the rest are strangers.

Dyad 3 (friends) showed very low accuracy in audio and dyadic features alone, which might be explained by the fact that in one early session for this dyad, most of the 30-second clips contain very sparse numbers of low-level audio features (LA). An examination of the audio recording reveals that one of the participants was more aggressive than in the other sessions. The student told his friend, *"Just be quiet—I am trying to work,"* and *"Shh, you don't understand, so I basically have to teach you how to work that, but I'm trying to work."* At this point in the interaction, his partner stopped participating in the task and said virtually nothing for the rest of the session. This lack of speech led to a lower number of turns – a pattern with a closer resemblance to strangers than friends.

It seems that such rude behavior would be more likely between friends than strangers, meaning that ultimately our model will need to

be sensitive to this kind of variance. With more pairs of friends, different styles of friendship can be further distinguished. However, this specific phenomenon signals that in the future, lexical information which could be obtained by automatic speech recognition could further improve performance.

Dyad 11 also showed low relative accuracy in predication, particularly when the model used vision features. We found that one of the participants often tilted her head, which partially blocked the frontal camera view of the other participant, thus resulting in less confidence in automatically extracted visual features. In the future we will set our cameras in a better position in order to reach higher feature extraction accuracy.

When we combined all our features, the prediction accuracy of Dyad 3 and 11 improved, further demonstrating that multimodal information improves friendship modeling.

9 Conclusion and Future Work

As a first step towards predicting the state of friendship between two interlocutors, we analyzed a set of automatically harvested low-level and dyadic features from dialogues in a peer-tutoring task. Both low level features and dyadic features were shown to be useful in discriminating between those who are friends and those who are not.

To perform the analysis, we introduced a new computational group sparse model (GSM) in order to accommodate the sparse and noisy properties of multi-channel features. GSM outperformed a baseline of human raters who make these types of social judgments in everyday interactions. GSM also statistically outperformed a non-sparse linear model (NLM) and a sparse linear model (SLM) when the analysis used only a single set of low level features or single set of low level features combined with dyadic features. When all features were used, the distinctions between models decreased, since in a huge multimodal feature space, even a naïve model could greedy search for a good weight assignment. Thus our newly proposed model did not significantly outperform the others in this scenario. And in general, more features produced more accurate prediction.

Based on the outcomes of the GSM model, we investigated differences between verbal and nonverbal behavior cues as a function of different friendship states. While much research on rapport detection and building in ECAs has focused on low level features, we found that dyadic features provided some of the most distinguishing differences between friends and non-friends. For example, mutual gaze and non-mutual gaze were both indicative, as friends are comfortable looking directly at one another while non-friends may have used direct gaze only to signal non-understanding. This comfort between friends was also notable in other salient dyadic features; i.e., while non-friends often work in concert looking down at the task, friends were relaxed such that one partner could lean back, interrupt to take more conversational turns, and smile at the other without needing to reciprocate the smile each time.

In future work we will look at temporal contingency more closely, examining whether participants' actions are contingent on the behavior of their partner. We will also examine whether the behavior of friends and strangers changes over multiple sessions. In this context, we include one suggestive graph, which shows that strangers increase their mutual gaze over sessions but friends decrease it. We are currently collecting further sessions for each dyad so as to be able to further analyze the nature of these relationships over time.

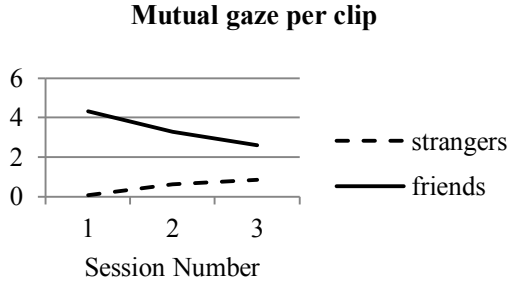

Figure 3: Weight of the mutual gaze in each session, by friendship status

To date we have found that the inclusion of automatically extracted dyadic features can lead to better prediction of friendship state. Both verbal and nonverbal behaviors were discovered that distinguish between different friendship status and that suggest how to design embodied dialogue systems that intend to spend a lifetime on the job.

Acknowledgements

Thanks to Angela Ng, Rachel Marino and Marissa Cross for data collection, Giota Stratou for visual feature extraction, Yi Yang, Louis-Philippe Morency, Shoou-I Yu, William Wang, and Eric Xing for valuable discussions, and the NSF IIS for generous funding.

References

Ambady, N., Krabbenhoft, M. A. & Hogan, D. (2006). The 30-sec sale: Using thin-slice

judgments to evaluate sales effectiveness. *Journal of Consumer Psychology*, *16*(1), 4–13.

Argyle, M. & Dean, J. (1965). Eye-contact, distance and affiliation. *Sociometry*, *28*(3), 289–304.

Azmitia, M. & Montgomery, R. (1993). Friendship, transactive dialogues, and the development of scientific reasoning. *Social Development*, *2*(3), 202–221.

Baker, R. E., Gill, A. J., & Cassell, J. (2008). Reactive redundancy and listener comprehension in direction-giving. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue* (pp. 37–45).

Brooks, M. (1989). *Instant rapport* (p. 205). New York: Warner Books.

Berg, B. L. (1989). *Qualitative research methods for the social sciences*. Boston: Allyn and Bacon.

Bernieri, F. J. (1988). Coordinated movement and rapport in teacher-student interactions. *Journal of Nonverbal Behavior*, *12*(2), 120–138.

Bickmore, T. W. & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, *12*(2), 293–327.

Bickmore, T. W., Pfeifer, L., & Schuman, D. (2011). Relational agents improve engagement and learning in science museum visitors. In *Intelligent Virtual Agents* (pp. 55–67). Reykjavik.

Bolinger, D. (1986). *Intonation and its parts: Melody in spoken English*. Stanford, CA: Stanford University Press.

Burgoon, J. K. & Hale, J. L. (1988). Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Communications Monographs*, (May 2013), 37–41.

Cappella, J. N. & Pelachaud, C. (2002). Rules for responsive robots: Using human interactions to build virtual interactions. In Reis, Firzpatrick, & Vangelisti (Eds.), *Stability and change in relationships*. New York, NY: Cambridge University Press.

Cassell, J. (2000). Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost, & E. Churchill (Eds.), *Embodied conversational agents* (pp. 1–27). MIT Press.

Cassell, J., Gill, A. J., & Tepper, P. A. (2007). Coordination in conversation and rapport. *Proceedings of the ACL Workshop on Embodied Natural Language*, 40 –50.

Cassell, J., Bickmore, T. W., Campbell, L., Vilhjálmsson, H. H., & Yan, H. (2001). More than just a pretty face: Conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, *14*(1-2), 55–64.

Cassell, J. & Bickmore, T. W. (2003). Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, *13*(1), 89–132.

Chen, X., Lin, Q., Kim, S., Carbonell, J. G., & Xing, E. P. (2012). Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, *6*(2), 719–752.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Fan, R., Chang, K., Hsieh, C., Wang, X., & Lin, C. (2008). LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, *9*, 1871–1874.

Fantuzzo, J., Riggio, R., Connelly, S., & Dimeff, L. (1989). Effects of reciprocal peer tutoring on academic achievement and psychological adjustment: A component analysis. *Journal of Educational Psychology, 81(2),* 173-177.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition, Second Edition* (2nd ed., p. 592). San Diego, CA: Academic Press.

Gatica-Perez, D. (2009). Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing, 27*(12), 1775–1787.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R. J., & Morency, L.-P. (2006). Virtual rapport. In *Intelligent Virtual Agents* (pp. 14–27). Springer Berlin/Heidelberg.

Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test.* Baltimore, MD: Johns Hopkins University Press.

Harrigan, J. A., Oxman, T. E., & Rosenthal, R. (1985). Rapport expressed through nonverbal behavior. *Journal of Nonverbal Behavior*, *9*, 95–110.

Harrigan, J. A. & Rosenthal, R. (1983). Physicians' head and body positions as determinants of perceived rapport. *Applied Social Psychology*, *13*(6), 496–509.

Liu, J., Ji, S., & Ye, J. (2009). Multi-task feature learning via efficient l2, 1-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (pp. 339–348). AUAI Press.

Nakano, Y. I., Reinstein, G., Stocky, T., & Cassell, J. (2003). Towards a model of face-to-face grounding. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. ACL'03* (Vol. 1, pp. 553–561). Sapporo: Association for Computational Linguistics.

Nakano, Y. I. & Ishii, R. (2010). Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th international conference on Intelligent user*

interfaces. *IUI'10* (pp. 139–148). Hong Kong: ACM Press.

Ogan, A., Finkelstein, S., Walker, E., Carlson, R., & Cassell, J. (2012). Rudeness and rapport: Insults and learning gains in peer tutoring. In *Proceedings of the 11 International Conference on Intelligence Tutoring Systems (ITS 2012)*.

Prepin, K., Ochs, M., & Pelachaud, C. (2012). Mutual stance building in dyad of virtual agents: Smile alignment and synchronisation. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on Social Computing (SocialCom)* (pp. 938–943).

Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., … Weiss, B. (2012). The INTERSPEECH 2012 speaker trait challenge. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012)*. Portland, OR: ISCA.

Steidl, S., Polzehl, T., Bunnell, H. T., Dou, Y., Muthukumar, P. K., Perry, … Metze, F. (2012). Emotion identification for evaluation of synthesized emotional speech. In *Proceedings of the 6th International Conference on Speech Prosody 2012* (pp. 4–7). Shanghai: Tongji University Press.

Stronks, B., Nijholt, A., van Der Vet, P., Heylen, D., & Machado, A. (2002). Designing for friendship: Becoming friends with your ECA. In A. Marriott, C. Pelachaud, T. Rist, Z. M. Ruttkay, & H. Vilhjalmsson (Eds.), *Workshop on Embodied Conversational Agents - Let's specify and evaluate them!, AMAAS 2002* (pp. 91–96). Bologna: AMAAS.

Tickle-Degnen, L. & Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, *1*(4), 285–293.

Vardoulakis, L. P., Ring, L., Barry, B., Sidner, C. L., & Bickmore, T. W. (2012). Designing relational agents as long term social companions for older adults. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), *Intelligent Virtual Agents* (pp. 289–302). Santa Cruz, CA: Springer Berlin Heidelberg.

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

Ward, N. (2006). Non-Lexical Conversational Sounds in American English. *Pragmatics and Cognition, (14)*1, 113-184.

Wang, W. Y., Finkelstein, S., Ogan, A., Black, A. W., & Cassell, J. (2012). "Love ya, jerkface": Using sparse log-linear models to build positive (and impolite) relationships with teens. In *Proceedings of the 13th Annual SIGDIAL Meeting on Discourse and Dialogue* (pp. 20–29). Seoul, South Korea.

Yang, Y., Shen, H. T., Ma, Z., Huang, Z., & Zhou, X. (2010). l2,1-regularized discriminative feature selection for unsupervised learning. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence* (pp. 1589–1594). AAAI Press.

Youniss, J. (1982). *Parents and peers in social development: A Sullivan-Piaget perspective*. University of Chicago Press.

Yuan, M. & Lin, Y. (2007), Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* 68(1), 49-67.