Changes in Verbal and Nonverbal Conversational Behavior in Long-Term Interaction

Daniel Schulman
College of Computer and Information Science
Northeastern University
Boston, MA
schulman@ccs.neu.edu

Timothy Bickmore
College of Computer and Information Science
Northeastern University
Boston, MA
bickmore@ccs.neu.edu

ABSTRACT

We present an empirical investigation of conversational behavior in dyadic interaction spanning multiple conversations, in the context of a developing interpersonal relationship between a health counselor and her clients. Using a longitudinal video corpus of behavior change counseling conversations, we show systematic changes in verbal and nonverbal behavior during greetings (within the first minute of conversations). Both the number of prior conversations and self-reported assessments of the strength of the interpersonal relationship are predictive of changes in verbal and nonverbal behavior.

We present a model and implementation of nonverbal behavior generation for conversational agents that incorporates these findings, and discuss how the results can be applied to multimodal recognition of conversational behavior over time.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentations**]: User Interfaces—*Graphical user interfaces, Interaction styles, Natural language, Theory and methods*

General Terms

Design, Experimentation, Human Factors

Keywords

conversational agent, long-term interaction, multimodal interaction

1. INTRODUCTION

As conversational agents spend more time with people and work in roles in which multiple interactions are required (such as in education or counseling), it is important that the multimodal interactions they have with users are as natural as possible — not only within a given conversation but between conversations over time. This is not only important to provide users with an experience that is as natural and similar to their interactions with other people as possible, but may be crucial if agents are to succeed in those roles that require long-term interactions. Longitudinal adaptations in multimodal conversational behavior may be key for maintaining user engagement over time, such as retention in a year-long health behavior change intervention, or cognitive involvement in the fiftieth lesson that a pedagogical agent delivers.

To inform the design of such agents, we are interested in studying verbal and nonverbal behavior in face-to-face conversation between people — including such things as hand gestures, gaze cues, and posture shifts — focusing on how this behavior changes over time as a function of interaction history and the nature of the evolving interpersonal relationship between the interactants. Interaction history includes the number, pattern, and purpose of the series of conversations two interactants have had. Interpersonal relationship includes such longitudinal constructs as trust, intimacy, and working relationship (such as therapeutic alliance in healthcare). These two variables — history and relationship — are related, but often separate factors in influencing the behavior of dyads over time [31].

An understanding of longitudinal changes in human conversational behavior is also important for building multimodal interfaces that can recognize subtle variations in user behavior that signal changes in cognitive engagement, trust, or therapeutic alliance, or which indicate that a user is about to discontinue use of a system (e.g., withdraw from an online course or drop out of a weight loss program).

In this paper, we present an observational study of conversational behavior in a longitudinal video corpus of dyadic interaction between a health counselor and her clients. We focus on behavior occurring within conversation openings, defined here as the first minute of conversations. Openings are a particularly important segment of conversation, in which effects of relationship status may be most pronounced. At the beginning of a conversation, participants' beliefs about their interpersonal relationship may be communicated and/or negotiated [11]. Prior studies that have examined differences in behavior across multiple conversation showed that such differences were larger at the beginning of conversations [25, 26].

Our analyses yield models that predict certain patterns of nonverbal behavior based on interaction history and relationship status. We then describe a model and implementation of behavior generation for conversational agents which incorporates these results, and discuss future work and im-

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.
Copyright 2012 ACM 978-1-4503-1467-1/12/10 ...$15.00.

plications for multimodal recognition of nonverbal conversational behavior over time.

2. BACKGROUND AND RELATED WORK

2.1 Conversational Behavior in Long-Term Interaction

Several prior studies have reported differences in behavior in dialogue between interactants who have had a substantial number of prior conversations, and those who have had few or none. Most such work is cross-sectional and compares, for example, dyads of friends with (different) dyads of strangers. A cross-sectional study may show differences between friends and strangers, but has limited ability to show a pattern of change over time. Cross-sectional studies also cannot separate changes over time from differences between dyads. For example, a difference in nonverbal behavior between friends and strangers may be explained as either a change that occurs as people become friends, or as a baseline difference that predicts whether a dyad will become friends.

Planalp and Benson showed that observers are able to discriminate, with 79% accuracy, audiotaped conversations between friends from conversations between acquaintances. When asked what cues they used, observers reported that friends referred more often to mutual knowledge, showed higher content intimacy, sounded more relaxed, interrupted each other more often, and had more equal distribution of floor time, compared to acquaintances [21].

Cassell et al. compared direction-giving dialogues between friends and between strangers in a cross-sectional study [7]. Strangers used more explicit acknowledgments than friends when giving or receiving information. Strangers also used more nonverbal behavior related to coordination: head nods and mutual gaze were more likely to occur during acknowledgments.

Tickle-Degnen and Rosenthal propose a model of dyadic rapport that deepens over time [30, 31], and consists of three components: mutual attentiveness, positivity, and coordination. The relative importance of these components is predicted to vary throughout the course of the relationship, with coordination increasing and positivity decreasing. They suggest that components of rapport can be observed via correlates in nonverbal behavior. A meta-analysis indicated that a participants' evaluation of their partner's level of positivity was positively associated with the partners' nonverbal behaviors including forward trunk leaning, smiling, nodding, direct body orientation, and uncrossed arms.

In a series of observational studies, Schulman and Bickmore report changes in verbal and nonverbal behavior across multiple conversations. While longitudinal, these studies examine only a limited range of behaviors relative to the present study. Participants had faster articulation rates on discourse markers in later conversations [25]. Participants also used more posture shifts at the beginning of a conversation than the end, and this decrease was significantly more rapid in later conversations [26].

2.2 Conversational Agents in Long-Term Interaction

Bickmore introduced and explored the concept of "relational agents": computer agents designed to form long-term social-emotional relationships with their users [4]. As an implementation of this concept, he created a conversational agent that used various conversational behaviors intended to promote an interpersonal relationship with a user, including small talk [24], humor, empathetic messages, and reciprocal self-disclosure [2]. Bickmore's relational agent implementation was focused on identifying particular messages which, when delivered appropriately during conversation, would promote a strong user-agent relationship.

Several researchers have demonstrated (e.g., [12]) that the nonverbal behavior of agents can affect user-agent rapport, focusing on short-term rapport within a single conversation. The nonverbal behaviors associated with improved rapport may change across multiple conversations [31, 29], suggesting that this work could be meaningfully extended toward long-term interaction.

3. THE EXERCISE COUNSELING CORPUS

Our corpus for this work is a longitudinal video corpus of weekly face-to-face conversations between a human counselor (a certified exercise trainer) and clients (Figure 1). The corpus contains up to six sessions per client, resulting in 32 conversations (mean duration 15.6 minutes), comprising approximately 8.3 hours of recorded video and approximately 100,000 words of spoken dialogue.

This data was collected with the goal of modeling changes in verbal and nonverbal behavior — including both the counselor's behavior and the client's — which occur across these six sessions. Behavior change counseling is an area for which conversational agents have been repeatedly applied (e.g., [4]), and nonverbal behavior is associated with the development of the client-counselor relationship [29].

We recruited clients (N=6, 5 female) who stated they did not currently exercise regularly. The same counselor conducted all conversations, and in each conversation attempted to encourage the client to increase his or her daily physical activity. All conversations were held in the same room, with both client and counselor seated in office chairs, and were videotaped from three angles. The participants were informed the conversations would be videotaped and examined, but were not told what behaviors were being investigated.

After each conversation, both partcipants completed (separately) questionnaires to assess their interpersonal relationship. We used the short revised Working Alliance Inventory (WAI-SR) [13]. This instrument is an assessment of therapeutic alliance [5], a model of interpersonal relationship specific to counseling and psychotherapy, and including components of agreement on overall goals of counseling, specific tasks, and interpersonal bond or rapport. Strong therapeutic alliance is predictive of positive counseling outcomes [17].

There is evidence of the development of strong counselor-client interpersonal relationships over time: therapeutic alliance increases across sessions, both as reported by clients (from mean 3.7 in the first session to mean 4.7 in the last, on a 1–5 scale), and by the counselor (from 2.6 to 3.8).

4. METHODS

The goal of our analysis was to identify whether there were systematic changes in counselor and client nonverbal behavior across conversations, as a function of interaction history (the number of conversations, and whether the current conversation is the last), relationship strength (mea-

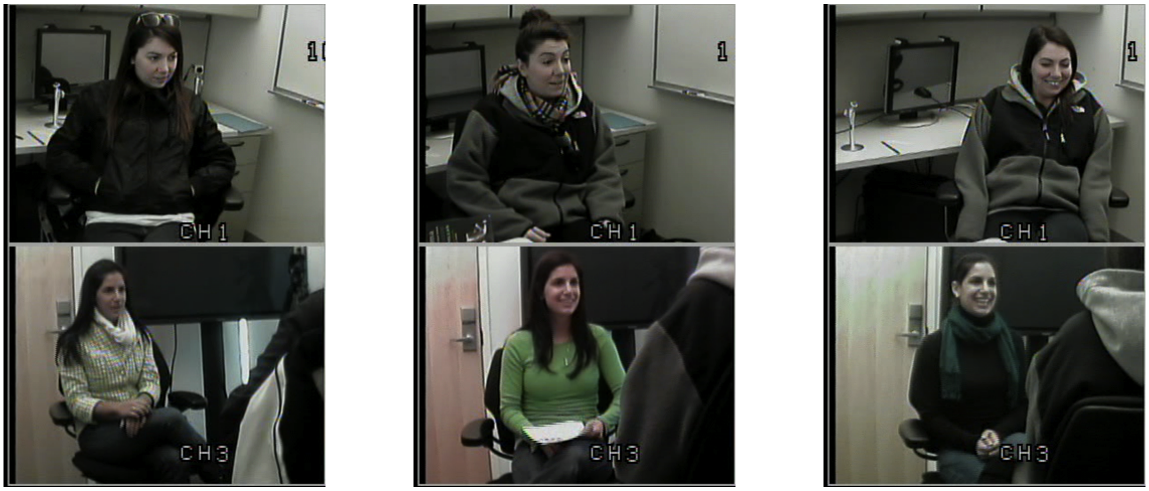

Figure 1: The Exercise Counseling Corpus: Samples from 1st, 3rd, and 6th conversations

sured by therapeutic alliance), or both. We did not attempt to account for all variability in behavior, and substantial unexplained variability remains after our analysis.

A one-minute segment of each conversation was selected, beginning from the first point at which both participants were judged to be fully seated; participants sat facing each other immediately after entering the room in all conversations. The resulting 32 minutes of video were manually annotated for various nonverbal behaviors (detailed below) using ANVIL [14]. A word-aligned orthographic transcription of the corpus, performed for previous work [25], was used to identify segments where each participant was speaking.

4.1 Outcome Variables

We chose the following set of outcome variables for analysis based on those behaviors which prior work suggested might show changes associated with varying interpersonal relationship:

- *The proportion of time spent speaking*: friends are reported to share speaking time more equally than strangers [21].
- *The number of gaze-aways during speech*: the amount of gaze-away during speech is reported to be associated with topic intimacy [1].
- *The proportion of time, when not speaking, spent nodding*: friends are reported to use less nodding for acknowledgement than strangers [7]. Restricting to time when not speaking controls for varying opportunity to show acknowledgment in different videos.
- *The proportion of time spent smiling or frowning*, or more generally with the mouth in a non-neutral position: increased facial expressivity is associated with higher immediacy [3].
- *The proportion of time spent performing self-adaptors, when not speaking*: the use of self-adaptors — self-touching gestures that do not signal meaning in conversation, and often serve to release bodily tension — is associated with perceptions of anxiety [32]. A qualitative inspection of the corpus indicated that most self-adaptors occurred when not speaking.
- *The proportion of time spent performing gestures (other than self-adaptors), during speech*: frequent and expressive gestures are associated with immediacy [3], and most hand gestures co-occur with speech.
- *The proportion of time spent with eyebrows raised or lowered, during speech*: eyebrows raises and frowns are a component of displays of affect and other facial expressivity, associated with immediacy.

A preliminary analysis indicated no significant changes in behavior within a single one-minute video. Therefore, all outcome variables are aggregates of behaviors over a video clip.

4.2 Coding of Nonverbal Behavior

The following behaviors were coded in order to determine values for the outcome variables for each of the 32 one-minute video clips.

Gaze.

An event was coded whenever a participant looked away from the partner's eyes, in any direction (up, down, or sideways).

Eyebrows.

An event was coded whenever a participant raised or lowered his or her eyebrows away from a neutral facial expression.

Head Movement.

An event was coded for any head movement which caused any part of the head to move at least two inches in any direction. However, nodding, shaking, and other rhythmic and repetitive movements were always coded. Each event was categorized as one of: nod (up-and-down movement), jerk (single quick upward movement), back (movement away from the partner), forward (toward the partner), turn (rotation either left or right), or tilt (leaning to either side).

Mouth Shape.

An event was coded whenever a participant's mouth took a shape that differed from a neutral facial expression (e.g.,

Table 1: Interrater reliability for coding of nonverbal behavior

Behavior	Cohen's κ
Gaze-away	0.71
Eyebrows	0.65
Head movement (occurrence)[a]	*0.68*
Head movement (categorized)	0.67
Mouth	0.81
Gesture (occurrence)[a]	*0.71*
Gesture (categorized)[a]	*0.57*
Gesture (self-adaptor)[b]	0.91

[a] not used in subsequent analysis
[b] a composite of several categories

corners up or down, lips protruded or retracted), other than to open during speech.

Hand Gesture.

Based on semiotic categories as described by McNeill [19], gestures were coded as deictic, iconic, emblematic, beat, self-adaptor, or "other".

4.3 Interrater Reliability

Three randomly-selected videos, containing sessions with three different clients, were coded separately by the primary author and by a second coder who was not involved in the development of the coding manual. The start and end times of all coded events were aligned to the nearest quarter second, and Cohen's κ was computed, treating a quarter second segment as one observation. Reliability was considered acceptable when $\kappa \geq 0.65$.

Table 1 summarizes the results. For head movement and hand gesture, Cohen's κ is reported separately for coding the occurrence of an event at the same time, and for coding the same event category at the same time. Reliability was low for categorized hand gestures: beat, iconic, and deictic gestures were all frequently confused, and emblematic gestures were rare. Combining all categories other than "adaptor" yielded good reliability, and all subsequent analysis uses only the categories of self-adaptor and non-adaptor gestures.

4.4 Analysis

For each outcome variable listed in 4.1, we fit a series of regression models. Ordinary linear regression is inappropriate here, as it assumes that observations are independent, whereas here conversations are grouped within dyads. We use generalized linear mixed-effect regression [18], an extension of ordinary linear regression that accounts for grouped data by adding "random effects", or per-group means that are assumed to be normally distributed around the population mean. We treat counselor and client behavior as separate but correlated per-conversation outcome variables, and include separate per-dyad means for the counselor and client, which may also be correlated.

We considered four variants of this model for each behavior, differing in the set of predictors included. All models include two predictors, modeling change over time: the number of prior conversations, and whether the current conversation is the last conversation for that dyad.[1] From this basic model, we consider:

A. The predictors above, with the added assumption that the effect of these predictors is the same on the counselor and the client.

B. As in A, and including self-reported therapeutic alliance from the previous conversation.

C. As in A, but with no assumption that effects on the client and counselor are the same.

D. As in B, but with no assumption that effects on the client and counselor are the same.

The models were fit to the data using Bayesian estimation with weak prior distributions: normal distributions with high variance (10^{10}) for fixed effects of parameters and inverse Wishart distributions (3 d.f.) for the dyad-level and conversation-level covariance matrices. Models were compared using the Deviance Information Criterion (DIC) [28].

The number of gaze-aways during speech was modeled as a Poisson-distributed count outcome with an added Gaussian random effect to allow for overdispersion [6]. For all other behaviors, the proportion of time during which the behavior was observed was modeled as a Gaussian-distributed outcome, following two transformations: first, by "squeezing" all values toward 0.5 slightly to avoid proportions exactly equal to 0 or 1 [27], and then by applying the inverse logit function:

$$y' = \text{logit}^{-1}\left(\frac{y * (N-1) + 0.5}{N}\right)$$

where $N = 64$ is the total number of observations.

All results are based on Markov Chain Monte Carlo simulation, performed using JAGS [22] and R 2.13.1 [23]. For each model and each behavior, 1000 samples were drawn from each of 3 different runs, and convergence was tested with Gelman-Rubin [10] diagnostics.

5. RESULTS

Table 2 gives descriptive statistics for each outcome variables, and the best-fitting regression models are summarized in Table 3. Gaze-aways, nodding, and smiling and frowning were best predicted by models in which interaction history and relationship strength have the same effect on counselor and client behavior. However, the proportion of time spent speaking, the use of both self-adaptor and non-adaptor gestures, and eyebrow raises and frowns were best predicted by models in which effects on the counselor and client differed.

Gaze.

The number of gaze-aways during speech is predicted about equally well by models with and without therapeutic alliance, although both give similar predictions: There is an increase in the rate of gaze-aways over time, which reverses in the last session (Figure 2).

[1] This predictor was added after observing that the final sessions appeared qualitatively different from others in the corpus. Omitting these sessions gives estimates similar to Table 3, although the interaction of therapeutic alliance and number of conversations on nodding (Figure 3) is only near-significant.

Table 3: Regression coefficients for best-fitting models. Bolded coefficients indicate a 95% credible interval which excludes 0.

Behavior Model	Mouth A	Gaze A	Nod B	Speech C	Adaptor C	Gesture C	Brows C
Intercept (counselor)	-0.43	**-3.74**	**-1.48**	**-1.00**	**-3.24**	**-2.85**	**-3.48**
Intercept (client)	-0.37	**-2.84**	**-2.30**	0.04	**-2.78**	**-2.75**	**-2.86**
Session (counselor)	-0.16	**0.20**	0.06	**-0.27**	**0.29**	**-0.37**	**-0.34**
Session (client)				**0.13**	-0.06	0.14	0.14
Last Session (counselor)	**0.97**	**-0.80**	-0.44	**0.93**	-0.31	1.20	0.65
Last Session (client)				-0.36	-1.15	0.28	0.28
Alliance			**-0.28**				
Alliance×Session			**0.06**				
σ_{dyad} (counselor)	0.78	0.62	0.53	0.55	0.56	0.64	0.62
σ_{dyad} (client)	0.95	0.78	0.91	0.54	1.21	0.82	0.70
ρ_{dyad}	**0.60**	-0.02	0.23	-0.00	0.03	-0.03	-0.10
σ_{conv} (counselor)	0.56	0.56	0.45	0.65	0.93	0.92	1.06
σ_{conv} (client)	0.96	0.40	0.87	0.39	1.35	1.11	1.21
ρ_{conv}	**0.66**	0.03	0.07	**-0.62**	0.10	-0.19	**-0.33**

Table 2: Descriptive statistics for all behaviors

	Counselor		Client		Overall	
	Mean	SD	Mean	SD	Mean	SD
Mouth[a]	0.35	0.18	0.37	0.24	0.36	0.21
Gaze-Away[b]	0.04	0.06	0.11	0.08	0.08	0.08
Nod[c]	0.24	0.08	0.14	0.10	0.19	0.10
Speech[a]	0.20	0.10	0.57	0.12	0.39	0.21
Adaptor[c]	0.09	0.07	0.11	0.18	0.10	0.14
Gesture[d]	0.04	0.04	0.12	0.11	0.08	0.09
Brows[d]	0.02	0.05	0.10	0.09	0.06	0.08

[a] Proportion of time
[b] Count of events during speech
[c] Proportion of time not speaking
[d] Proportion of time speaking

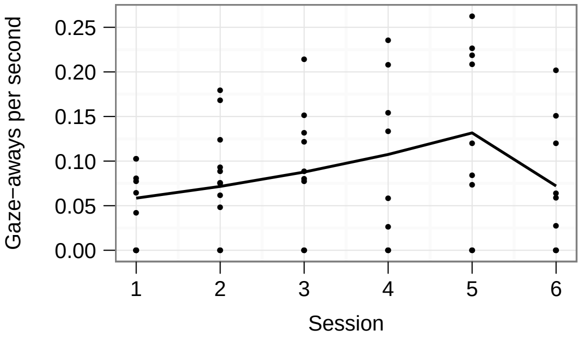

Figure 2: Gaze-aways during speech, by session. There were multiple observations of no gaze-aways by a participant: 6 in first sessions, and 3–4 in each of second through last sessions. The line is the model-based prediction for the average participant.

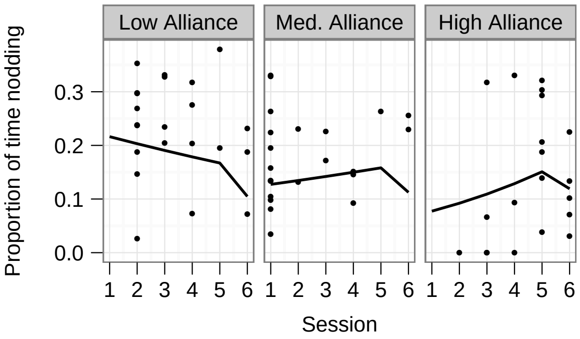

Figure 3: Nodding when not speaking, by session and alliance. The lines are model-based predictions for the average participant, at 25th, 50th, and 75th percentile therapeutic alliance.

Nodding.

The proportion of time spent nodding was best predicted by a model which included therapeutic alliance. Participants nodded less when they reported higher therapeutic alliance, but this effect was moderated by the number of sessions: in later sessions, all participants tended to nod more. There was a non-significant trend toward less nodding in the last session (Figure 3).

Mouth.

Participants used fewer non-neutral mouth positions in later sessions, but reversed this trend in the last session (Figure 4).

Speech.

The counselor spoke less in later sessions, and the clients spoke more, but this reversed in the last session for the counselor and there was a non-significant trend for it to reverse for the client.

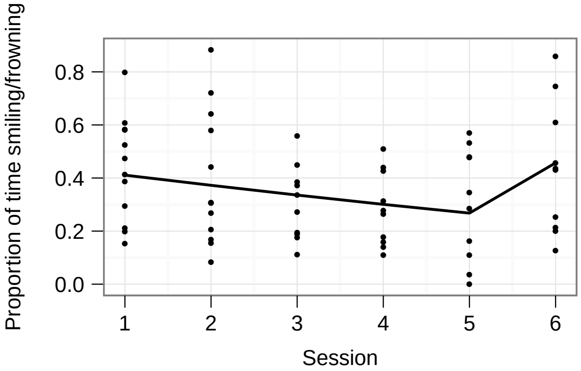

Figure 4: Occurrence of non-neutral mouth positions, by session. The line is the model-based prediction for the average participant.

Self-Adaptors.

The counselor used more self-adaptors in later sessions, but there was no significant trend for clients

Hand Gestures.

The counselor performed fewer non-adaptor gestures in later conversations, but reversed this trend in the last session. There was no significant trend for clients.

Eyebrow Movement.

The proportion of time spent performing eyebrow raises or lowering was predicted poorly by every model. No systematic trends were observed.

6. DISCUSSION

We show significant longitudinal differences in most behaviors investigated. In later conversations, gaze-aways during speech are more common, and less time is spent smiling and frowning. Participants nod more when they report lower therapeutic alliance. The counselor decreases speaking time while the client increases. Most trends reverse in the last conversation.

The results presented here do not fully agree with previous work. Prior (cross-sectional) studies have reported that friends share speaking time more equally than strangers, whereas here the counselor speaks less initially and further decreases her speaking time in later conversations. This may be due to the nature of the conversational task, which is focused on the client's attitudes and behavior.

Our finding of less smiling and frowning over time is broadly in agreement with Tickle-Degnen's model of rapport [31]: positivity is more important in early conversations. In an informal examination of the corpus (during coding) we noted that few of the smiles coded are Duchenne smiles [9]. We conjecture that "performing" appropriately-valenced facial expressions as an indication of empathy may be important for rapport in early conversations.

Cassell et al. report that friends use fewer nods for acknowledgments than strangers [7]. We report an effect of relationship strength: stronger self-reported relationship is associated with less nodding, particularly in early conversations. We conjecture that Cassell et al.'s study, which is cross-sectional, is showing differences between dyads rather than change over time: dyads with a strong relationship (associated with less nodding) are more likely to become and remain friends.

The results we report on the use of hand gesture, including both self-adaptors and other gestures, are difficult to generalize. We see significant change for the counselor only, and these results may blend general and idiosyncratic factors: they may be useful for developing models of this particular counselor rather than more general models of human behavior.

Across nearly all behavior, we report a pattern where the observed change over time reverses in the last conversation. We note that participants were always aware that the sixth conversation was their last, and all dyads had an explicit discussion about the end of their relationship. We conjecture that a final interaction, like an initial interaction, has increased uncertainty about the participants' interpersonal relationship, and this uncertainty is associated with changes in behavior.

7. TOWARD BEHAVIOR GENERATION FOR LONG-TERM INTERACTION

Multiple approaches to behavior generation have been explored, including rule-based (BEAT [8] and NVBG [16]), probabilistic generation from a model of an annotated corpus [20], or generation as part of grammar-based natural language generation [15]. In this section, we explore the feasibility of implementing our reported findings in a generation system.

Our approach is to implement these findings as adjustments to the probability of generating a behavior event. This can be done with any underlying behavior generation system that outputs generation probabilities for each behavior event (or can be modified to do so). However, we make some additional assumptions: First, we assume that the results found here hold constant throughout a conversation. Second, we assume that these results combine additively with other predictors of behavior. This second assumption allows us to implement these findings as simple adjustments, ignoring how the baseline probability is generated.

As a proof of concept, we have implemented "Rhythm", a simple rule-based nonverbal behavior generator. Rhythm inputs the text of agent utterances, annotated with relevant contextual information (e.g., the number of previous conversations), and outputs synchronized nonverbal behavior annotations. The baseline probability of generating a behavior event (p, below) is produced by rules given in other work. Rhythm currently implements the following longitudinal changes:

- The probability of generating a smile or frown changes as a function of the interaction history, where s is the number of prior conversations, and f is 1 if the agent believes this is a final conversation and 0 otherwise:

$$p' = \text{logit}^{-1}(\text{logit}(p) - 0.16s + 0.97f)$$

- The probability of generating a gaze-away changes as a function of interaction history:

$$p' = 1 - (1-p)^{\exp(0.2s-0.8f)}$$

- The probability of generating a headnod changes as a function of interaction history and the agent's be-

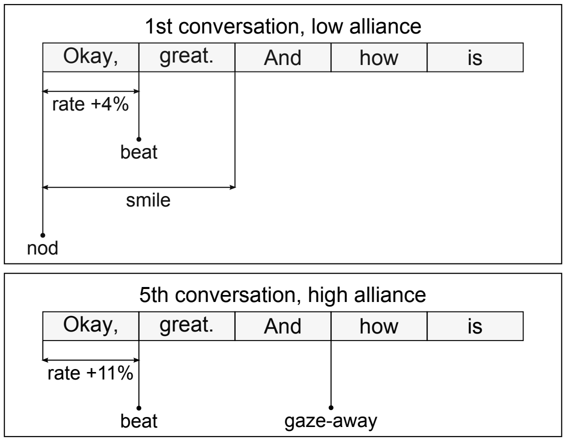

Figure 5: Sample behavior generation for the utterance "Okay, great. And how is your exercise going?"

liefs about the strength of the user-agent relationship, where a is therapeutic alliance, standardized to a z-score:

$$p' = \text{logit}^{-1}(\text{logit}(p) + 0.06s - 0.44f - 0.28a + 0.06sa)$$

- Following [26], the probability of generating a posture shift changes as a function of interaction history and minutes from conversation start (m):

$$p' = \text{logit}^{-1}(\text{logit}(p) + 0.16s - 0.03m - 0.02sm)$$

In addition, following [25], the articulation rate of discourse markers increases in later conversations. Figure 5 shows an example of behavior generated by Rhythm, on identical sentences, in the context of a first conversation between a dyad with high therapeutic alliance, and a fifth conversation between a dyad with low therapeutic alliance. In the first conversation, Rhythm generates a head nod and a smile, which do not appear in the fifth. In the fifth conversation, the articulation rate on the first word ("Okay") is increased, and an additional gaze-away is generated.

8. CONCLUSIONS AND FUTURE WORK

We have described systematic changes in nonverbal behavior that occur over time, and are also associated with changes in relationship quality independent of time. These findings have been implemented in a nonverbal behavior generation system. We plan to conduct a longitudinal evaluation study which is intended to validate this generation model, to (partially) validate these findings in a larger population, and to examine the effect of these changes across multiple conversations on user-agent rapport, user engagement, and the perceived behavioral realism of an agent.

The findings raise potential research issues both for applications focused on the generation of realistic conversational behavior, and those focused on the multimodal recognition of conversational behavior in the context of long-term human-agent interaction. Models which assume that a single interaction is representative of *all* interactions may give misleading results. For example, our results indicate a greater tendency to gaze away from a conversation partner in later conversations; a model of engagement based on gaze might interpret this change (possibly erroneously) as decreased engagement.

The difference between a final conversation and earlier conversations was an unexpected finding that does not, to our knowledge, appear in prior work, and it merits further investigation. Under the assumption that these changes occur because participants know the session will end their relationship, we suggest a new research question: will other changes in the nature of a conversational task or the nature of an interpersonal relationship (for example, from a professional and impersonal relationship to a friendship) produce similar effects? As very long-term user-agent interaction is explored, such changes may become more common.

Our results show that conversational behavior in long-term interaction is a complex product of the people involved and multiple aspects of their interpersonal relationship. The findings reported could be investigated largely because of features of the exercise counseling corpus relative to those used in other work: conversations in the corpus vary on multiple dimensions, including the participants involved and the history of their interaction, and the corpus is augmented with self-report measures of interpersonal relationship. Validating these results in a larger population — and extending them to other aspects of long-term interaction — will require very large, high-quality, and well-annotated corpora of face-to-face conversation. The exercise counseling corpus, at 8 hours of video, is small enough to limit the research questions that can be investigated, but already large enough to make manual annotation a major effort. Automated and semi-automated annotation will be necessary for further research in this area.

This work illustrates that producing realistic behavior in conversational agents should take many new contextual factors into account, including characteristics of the user, the agent, their relationship, and the conversational task.

9. ACKNOWLEDGEMENTS

The authors thank Jenna Zaffini for her work in collecting the corpus, Kshitij Lohani for assistance in coding, and to the members of the Relational Agents Group for much help and useful discussion.

This material is based upon work supported by the National Science Foundation under Grant No. ISS-0545932. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

10. REFERENCES

[1] A. Abele. Functions of gaze in social interaction: Communication and monitoring. *Journal of Nonverbal Behavior*, 10(2):83–101, June 1986.

[2] I. Altman and D. Taylor. *Social penetration: The development of interpersonal relationships.* Holt, Rinehart and Winston, 1973.

[3] P. A. Andersen. Nonverbal immediacy in interpersonal communication. In A. W. Siegman and S. Feldstein, editors, *Multichannel Integrations of Nonverbal Behavior*, pages 1–36. Lawrence Erlbaum, Hillsdale, NJ, 1985.

[4] T. Bickmore. *Relational Agents: Effecting Change through Human-Computer Relationships.* PhD thesis,

Massachusetts Institute of Technology, Cambridge, MA, 2003.

[5] E. S. Bordin. Theory and research on the therapeutic working alliance: New directions. In A. O. Horvath and L. S. Greenberg, editors, *The Working Alliance: Theory, Research, and Practice*, chapter 1, pages 13–37. Wiley, New York, NY, 1994.

[6] N. E. Breslow. Extra-Poisson variation in Log-Linear models. *Applied Statistics*, 33(1):38+, 1984.

[7] J. Cassell, A. J. Gill, and P. A. Tepper. Coordination in conversation and rapport. In *Workshop on Embodied Language Processing*, pages 41–50. Association for Computational Linguistics, June 2007.

[8] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore. BEAT: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '01, pages 477–486, New York, NY, USA, 2001. ACM.

[9] P. Ekman, R. J. Davidson, and W. V. Friesen. The Duchenne smile: Emotional expression and brain physiology II. *Journal of Personality and Social Psychology*, 58(2):342–353, 1990.

[10] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, Nov. 1992.

[11] E. Goffman. *Relations in public; microstudies of the public order.* Basic Books, 1971.

[12] J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. van der Werf, and L.-P. Morency. Virtual rapport. In J. Gratch, M. Young, R. Aylett, D. Ballin, and P. Olivier, editors, *Intelligent Virtual Agents*, volume 4133, chapter 2, pages 14–27. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[13] R. L. Hatcher and A. J. Gillaspy. Development and validation of a revised short version of the working alliance inventory. *Psychotherapy Research*, 16(1):12–25, Jan. 2006.

[14] M. Kipp. ANVIL — a generic annotation tool for multimodal dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pages 1367–1370, 2001.

[15] S. Kopp, P. Tepper, and J. Cassell. Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of the 6th international conference on Multimodal interfaces*, ICMI '04, pages 97–104, New York, NY, USA, 2004. ACM.

[16] J. Lee and S. Marsella. Nonverbal behavior generator for embodied conversational agents. In J. Gratch, M. Young, R. Aylett, D. Ballin, and P. Olivier, editors, *Intelligent Virtual Agents*, volume 4133, chapter 20, pages 243–255. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[17] D. J. Martin, J. P. Garske, and M. K. Davis. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. *Journal of consulting and clinical psychology*, 68(3):438–450, 2000.

[18] C. E. McCulloch and J. M. Neuhaus. Generalized linear mixed models. In *Encyclopedia of Biostatistics.* John Wiley & Sons, Ltd, 2005.

[19] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought.* University Of Chicago Press, Aug. 1992.

[20] M. Neff, M. Kipp, I. Albrecht, and H. P. Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Trans. Graph.*, 27(1):1–24, 2008.

[21] S. Planalp and A. Benson. Friends' and acquaintances' conversations I: Perceived differences. *Journal of Social and Personal Relationships*, 9(4):483–506, Nov. 1992.

[22] M. Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 2003.

[23] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2009.

[24] K. P. Schneider. *Small talk: Analyzing phatic discourse.* Hitzeroth, 1988.

[25] D. Schulman and T. Bickmore. Modeling behavioral manifestations of coordination and rapport over multiple conversations. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, and A. Safonova, editors, *Intelligent Virtual Agents*, volume 6356 of *Lecture Notes in Computer Science*, chapter 14, pages 132–138. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2010.

[26] D. Schulman and T. Bickmore. Posture, relationship, and discourse structure. In H. Vilhjálmsson, S. Kopp, S. Marsella, and K. Thórisson, editors, *Intelligent Virtual Agents*, volume 6895 of *Lecture Notes in Computer Science*, pages 106–112, Berlin, Heidelberg, 2011. Springer Berlin / Heidelberg.

[27] M. Smithson and J. Verkuilen. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1):54–71, Mar. 2006.

[28] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, Oct. 2002.

[29] L. Tickle-Degnen and E. Gavett. Changes in nonverbal behavior during the development of therapeutic relationships. In P. Philippot, R. S. Feldman, and E. J. Coats, editors, *Nonverbal behavior in clinical settings*, chapter 4, pages 75–110. Oxford University Press, New York, NY, USA, 2003.

[30] L. Tickle-Degnen and R. Rosenthal. Group rapport and nonverbal behavior. In C. Hendrick, editor, *Group processes and intergroup relations*, volume 9 of *Review of Personality and Social Psychology*, pages 113–136. Sage, Newbury Park, CA, 1987.

[31] L. Tickle-Degnen and R. Rosenthal. The nature of rapport and its nonverbal correlates. *Psychological Inquiry: An International Journal for the Advancement of Psychological Theory*, 1(4):285–293, 1990.

[32] P. H. Waxer. Nonverbal cues for anxiety: an examination of emotional leakage. *Journal of abnormal psychology*, 86(3):306–314, June 1977.