Modeling Behavioral Manifestations of Coordination and Rapport over Multiple Conversations

Speaking Rate as a Relational Indicator for a Virtual Agent

Daniel Schulman and Timothy Bickmore

College of Computer and Information Science, Northeastern University
360 Huntington Ave – WVH 202, Boston MA 02115
{schulman,bickmore}@ccs.neu.edu

Abstract. Many potential applications of virtual agents require an agent to conduct multiple conversations with users. An effective and engaging agent should modify its behavior in realistic ways over these conversations. To model these changes, we gathered a longitudinal video corpus of human-human counseling conversations, and constructed a model of changes in articulation rates over multiple conversations. Articulation rates are observed to increase over time, both within a single conversation and across conversations. However, articulation rates increased mainly for words spoken separately from larger phrases. We also present a preliminary evaluation study, showing that implementing such changes in a virtual agent has a measurable effect on user attitudes toward the agent.

Keywords: Speaking Rate, Embodied Conversational Agent, Relational Agent, Rapport.

1 Introduction

Embodied Conversational Agents (ECAs) simulate face-to-face conversation with users by reproducing human verbal and nonverbal conversational behavior as much as possible [1]. ECAs have been employed in a variety of applications domains, including education, counseling, and social engagement. Many of these application domains require that an ECA have multiple conversations with each user, possibly over a long period of time. Since the conversational behavior of human dyads is known to change over time as their relationship evolves [2,3], it is important that ECAs be able to simulate this behavior as well.

Toward that goal, we discuss a preliminary investigation of one aspect of verbal behavior: changes in the articulation rate of speech, defined here in terms of the duration of words spoken, excluding any silences or pauses. Our methodology is based on the collection and analysis of a longitudinal corpus of human-human interaction, in which we examine multiple conversations over time between the

J. Allbeck et al. (Eds.): IVA 2010, LNAI 6356, pp. 132–138, 2010.

same participants, in the context of developing interpersonal relationships. In our initial analysis, we construct a model of changes in articulation rates, both within and across conversations. We then present a preliminary evaluation in which some of the changes predicted by this model are incorporated in an ECA.

2 Related Work

A number of studies have examined the effects of speaking rates on listeners, focusing particularly on how changes in speaking rate affect the listener's perceptions of the speaker's personality. Smith et. al. found that increased speaking rate was perceived as increased competence, while perceived benevolence was highest at normal speech rates, and lower otherwise [4]. Nass and Lee showed that users perceived synthesized computer speech as more extroverted when it was generated with greater speech rate, volume, pitch, and pitch variation. Users tended to prefer speech perceived as matching their own introversion/extroversion [5].

Several researchers have examined differences in speaking rates, and other features of verbal behavior, usually with a cross-sectional design (comparing friends to strangers or acquaintances). Planalp and Benson showed that observers could judge with 80% accuracy whether audiotaped conversations were between strangers or friends. Several cues that observers commonly cited (although a small percentage of all cited) were related to articulation rates, such as "pace", "tone of voice", and "smoothness" [6]. Yuan et. al. compared several large corpora of conversational telephone speech in English and Chinese. Corpora consisting primarily of conversations between friends or family members had a higher average speaking rate than those consisting primarily of conversations between strangers [7].

Cassell et. al. compared dialogue between dyads who were either friends or strangers, using a direction-giving task. Friends used significantly more turns per minute than strangers, although it is not known whether this was due to changes in speaking rate. Friends also used significantly fewer acknowledgments when receiving directions [3].

3 A Longitudinal Corpus of Counseling Conversations

We gathered a video corpus intended to allow exploratory analysis that could identify possible changes in verbal and nonverbal behavior over multiple conversations. The corpus contains multiple conversations over time between the same people, and thus can be used for longitudinal analysis. This allows us to examine changes over time and separate them from differences between individuals.

We also wished to focus on a real-world task that naturally would involve multiple conversations over time, and in which changes in verbal and nonverbal behavior might plausibly have an effect on task outcomes. Therefore, we chose to study health behavior change counseling for exercise promotion, an area to which conversational agents have been applied (e.g., [8]). The rapport and relationship between counselor and client is known to affect outcomes [9], and there has been research interest in verbal and nonverbal behavior that may be related to the development of this relationship (e.g., [2]).

3.1 Procedure

Participants were asked to complete six sessions, at approximately one week intervals. The counselor was instructed to conduct a short conversation with each participant at each session, and that the conversation should encourage the participant to increase his or her daily physical activity. The conversations were video recorded.

Six participants were recruited via ads placed on craigslist.org. Participants were required to be not currently exercising regularly. Participants ranged in age from 22 to 65, although all but two were 25 or younger (median age 24). Five of the six participants were female. A single counselor (also female) interacted with all participants.

A total of 32 conversations were recorded; five of the six participants completed all six sessions. The resulting corpus comprises approximately 8.3 hours of recorded video, with approximately 100,000 words of spoken dialogue.

4 A Model of Changes in Articulation Rate

We performed a full word-aligned orthographic transcription of the corpus, producing an estimate of the duration of every spoken word. The corpus was also divided into "segments", with a segment consisting of a sequence of words by a speaker uninterrupted by silence. Note that a single turn or utterance by a speaker may contain multiple segments, if it included any intra-turn pauses. To account for differences in word lengths, the duration of each word was normalized by the number of phonemes, determined using the CMU pronouncing dictionary (version 0.7a; http://www.speech.cs.cmu.edu/cgi-bin/cmudict), with manual correction for words that did not appear in the dictionary.

We used a linear mixed-effect model to account for the longitudinal nature of the data [10], analyzed with Bayesian methods using R 2.10 and the MCMCglmm package [11]. Uninformative or very weakly-informative prior distributions were used for all effects.

To model change across conversation, we included a fixed effect of the number of previous sessions, while random effects allow for variability across subjects. Two covariates were motivated by prior work: (a) the position of a segment within a conversation [12], and (b) the length of a segment [13]. Inspection of preliminary models showed that predictions were poor for single-word segments (words bounded by silence); these had longer duration than predicted, even including segment length as a covariate. Therefore, we included, as an additional predictor ("Multiword"), whether a word was in a multi-word segment.

4.1 Results

Table 1 shows the full regression model. Word durations in later conversations tended to be shorter than word durations in earlier conversations. However, this change was observed *only* for single-word segments (shown by the fixed effects

Table 1. A Mixed-effect Regression Model Predicting Articulation Rate (average seconds per phoneme, log-transformed)

Parameter	Fixed Effect[b]		Random Effect[b]
Intercept	-2.070***	[-2.120,-2.021]	0.031 [0.017,0.087]
Session[c]	-0.015**	[-0.024,-0.005]	0.003 [0.001,0.011]
Who[d]	0.100*	[-0.002,0.176]	0.066 [0.034,0.208]
Multiword[e]	-0.592***	[-0.616,-0.564]	
Pos[f]	-0.045***	[-0.057,-0.035]	
Session × Who	0.002	[-0.001,0.005]	
Who × Multiword	-0.065***	[-0.091,-0.037]	
Session × Multiword	0.012**	[0.005,0.019]	
Multiword × Len[g]	-0.105***	[-0.108,-0.100]	
Multiword × Pos	0.038***	[0.025,0.049]	
Multiword × Len × Pos	-0.004*	[-0.008,-0.001]	

[a] *p<.05, **p<.01, ***p<.001

[b] Posterior mode and 95% credible interval.

[c] Previously completed sessions (starts at zero).

[d] 0=counselor, 1=client

[e] 1 if the word is part of a longer segment, 0 otherwise.

[f] Number of segment within a conversation, centered and standardized.

[g] Length of segment in words, log-transformed, centered, and standardized.

"Session" and "Session × Multiword"). Similarly, within conversations, words near the end of a conversation tended to be shorter, again largely for single word segments (shown by "Pos" and "Multiword × Pos").

Given these results, we next examined the occurrences of single-word segments within the corpus. The most common such words ("okay", "yeah", "mm-hmm", "um", "so", "yknow", "and", "right", "but", "great"; approximately 70% of all instances), appear to consist mainly of backchannels and acknowledgements (e.g., "okay", "yeah"), and discourse markers (e.g., "so", "and").

In sum, we observed that the durations of single-word acknowledgements and discourse markers decreased over time, both within a single conversation and across multiple conversations.

5 A Preliminary Study of the Effects of Articulation Rate Changes in Conversational Agents

The changes predicated by the model are quite subtle: after five conversations, the average speaker increased their articulation rate approximately 8% (and only on specific words). We conducted a preliminary evaluation in order to test whether the model-predicted differences, when incorporated in a conversational agent's speech, were perceptible to users, and whether they had any measurable effect on attitudes toward the agent. Participants had two similar conversations, with two similar agents, which differed in articulation rates: In the SLOW condition, the articulation rate of the agent's speech was left unchanged, while in FAST, the articulation rate of acknowledgments and discourse markers was increased by the amount our model predicted would occur after five conversations (8%), and also increased at the predicated rate within a conversation (to a total of approximately 13%).

5.1 Apparatus and Measures

The two agents were chosen to have a similar appearance, and both used synthesized speech with synchronized nonverbal behavior. Participants used multiple-choice spoken input, with up to 6 utterance choices displayed by the agent at each turn. However, the agents were controlled via a Wizard-of-Oz setup [14], in order to eliminate any possible effects of speech recognition errors.

Both dialogues consisted of social dialogue only. The dialogues were designed to be approximately the same length (about 40 turns , varying slightly based on participant choices), and contained similar (but not identical) topics. Topics with a low intimacy level were used, such as weather, local sports, and features of the experiment location. Dialogues were manually tagged to identify acknowledgments and discourse markers that should increase in articulation rate when in the FAST condition.

Perceived rapport was assessed with the bond subscale of the Working Alliance Inventory [15] following each conversation. Participant introversion/extroversion was assessed using a 16-item subset of the Interpersonal Adjective Scales [16].

5.2 Procedure

The order of conditions, agents, and dialogues were randomly assigned. Following a demographics questionnaire, participants received brief instruction in how to interact with the agent. Participants were told they would be interacting with two different agents, but were not informed of differences in articulation rates, or any other specific differences. The experimenter left the room during the conversations, and returned to administer questionnaires afterward.

5.3 Results

8 participants (5 female, mean age 34.6, age range 23–63) were recruited via a contact list of potential participants who had expressed interest in previous studies but had not participated. All reported high levels of computer proficiency, and all but one were college graduates.

No significant difference was observed in perceived rapport (Working Alliance Inventory) between the SLOW and FAST speech (paired $t(7)=-0.296$, $p=0.78$). However, given results by Nass and Lee [5], we also analyzed the effect of the participant's extroversion. A linear regression showed that extroversion predicted the difference in perceived rapport between SLOW and FAST ($R^2=0.55$,

$F(1,6)$=7.39, p=0.035). Participants who were more extroverted were more likely to report a higher perceived rapport in the FAST condition.

Only one participant reported noticing a difference in speaking rate. When participants were asked to "guess" which agent spoke faster, 5 of 8 identified the correct agent; this is not significantly different from chance ($\chi^2(1)$=0.5, p=0.48). Therefore, we cannot conclude that participants consciously distinguish this difference in speaking rates in conversation.

6 Discussion

Our corpus shows evidence of changes in articulation rates over time, both within conversations and across multiple conversations. This complements results from earlier, cross-sectional studies, providing evidence that these changes in verbal behavior are in fact changes over time rather than pre-existing differences.

We also show a previously unreported nuance: increases in articulation rates were observed mainly in words bordered by silence, and these words were often acknowledgments or discourse markers. One possible explanation is that these words are the ones most easily spoken faster; longer segments of speech already tend to have faster articulation rates [13]. Alternatively, markers such as "so" may be used by speakers to coordinate their interaction. Faster articulation rates may indicate a decrease in explicit coordination as speakers increase in familiarity.

We show some preliminary evidence that changes in articulation rates of a speaker may have a measurable effect, even though listeners may not necessarily be able to consciously perceive the changes. However, the characteristics of the listener may be equally important: Extroverted listeners may prefer a speaker that "jumps right in" with a speaking style that indicates greater familiarity.

Both studies are limited by a small number of participants, and the corpus includes only a single counselor. Additional research is needed to determine whether these results will generalize across people, languages or dialects, or cultural backgrounds. To address some of these limitations, we plan a longitudinal evaluation study, in which participants have multiple conversations with an ECA designed according to the model developed here.

This work discusses how only a single aspect of verbal behavior changes over time. However, the observation that there are changes in verbal and/or nonverbal behavior both across and within conversations has implications for researchers working with virtual agents. We believe that the ability of virtual agents to change their behavior in realistic ways over time is important for making them more lifelike, engaging, and effective, especially as people work with them for longer periods of time.

Acknowledgments. Thanks to Jenna Zaffini for her work in collecting the video corpus, Connor Westfall for assistance with the evaluation study, and to the other members of the Relational Agents Group for their help and many useful discussions.

References

1. Cassell, J.: Embodied conversational agents. MIT Press, Cambridge (2000)
2. Tickle-Degnen, L., Gavett, E.: Changes in nonverbal behavior during the development of therapeutic relationships. In: Philippot, P., Feldman, R.S., Coats, E.J. (eds.) Nonverbal behavior in clinical settings, pp. 75–110. Oxford University Press, New York (2003)
3. Cassell, J., Gill, A.J., Tepper, P.A.: Coordination in conversation and rapport. In: Workshop on Embodied Language Processing, Association for Computational Linguistics, pp. 41–50 (2007)
4. Smith, B.L., Brown, B.L., Strong, W.J., Rencher, A.C.: Effects of speech rate on personality perception. Language and Speech 18(2), 145–152 (1975)
5. Nass, C., Lee, K.M.: Does computer-generated speech manifest personality? an experimental test of similarity-attraction. In: CHI 2000: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 329–336. ACM, New York (2000)
6. Planalp, S., Benson, A.: Friends' and acquaintances' conversations I: Perceived differences. Journal of Social and Personal Relationships 9(4), 483–506 (1992)
7. Yuan, J., Liberman, M., Cieri, C.: Towards an integrated understanding of speaking rate in conversation. In: International Conference on Spoken Language Processing, INTERSPEECH-2006 (2006)
8. Bickmore, T.: Relational agents: Effecting change through human-computer relationships (2003)
9. Horvath, A.O., Symonds, D.B.: Relation between working alliance and outcome in psychotherapy: A meta-analysis. Journal of Counseling Psychology 38(2), 139–149 (1991)
10. Verbeke, G., Molenberghs, G.: Linear Mixed Models for Longitudinal Data. Springer, Heidelberg (2001)
11. Hadfield, J.: MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. Journal of Statistical Software 33(2), 1–22 (2009)
12. Quené, H.: Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. The Journal of the Acoustical Society of America 123(2), 1104–1113 (2008)
13. Nakatani, L.H., O'Connor, K.D., Aston, C.H.: Prosodic aspects of american english speech rhythm. The Journal of the Acoustical Society of America 69(S1), 82 (1981)
14. Dahlbäck, N., Jönsson, A., Ahrenberg, L.: Wizard of oz studies: why and how. In: IUI 1993: Proceedings of the 1st international conference on Intelligent user interfaces, pp. 193–200. ACM, New York (1993)
15. Horvath, A.O., Greenberg, L.S.: Development and validation of the working alliance inventory. Journal of Counseling Psychology 36(2), 223–233 (1989)
16. Wiggins, J.S.: A psychological taxonomy of trait-descriptive terms: The interpersonal domain. Journal of Personality and Social Psychology 37(3), 395–412 (1979)