Understanding RUTH: Creating Believable Behaviors for a Virtual Human Under Uncertainty

Insuk Oh and Matthew Stone

Department of Computer Science
Rutgers, The State University of New Jersey
110 Frelinghuysen Road, Piscataway NJ 08854-8019

Abstract. In pursuing the ultimate goal of enabling intelligent conversation with a virtual human, two key challenges are selecting nonverbal behaviors to implement and realizing those behaviors practically and reliably. In this paper, we explore the signals interlocutors use to display uncertainty face to face. Peoples' signals were identified and annotated through systematic coding and then implemented onto our ECA (Embodied Conversational Agent), RUTH. We investigated whether RUTH animations were as effective as videos of talking people in conveying an agent's level of uncertainty to human viewers. Our results show that people could pick up on different levels of uncertainty not only with another conversational partner, but also with the simulations on RUTH. In addition, we used animations containing different subsets of facial signals to understand in more detail how nonverbal behavior conveys uncertainty. The findings illustrate the promise of our methodology for creating specific inventories of fine-grained conversational behaviors from knowledge and observations of spontaneous human conversation.

Keywords: Uncertainty expression, uncertainty recognition, embodied conversational agent, talking head, RUTH.

1 Introduction

People can achieve natural and effective interaction with each other partly by expressing and recognizing how much they understand or do not understand each other [1]. People have a rich repertoire of nonverbal cues – on their faces, heads and bodies, and in their voices – that they can use to show how much uncertainty they have about the state of the conversation; see for example [2, 3]. Our prior work underscores that viewers can attend to these signals [3]. When asked to rate the uncertainty conveyed in videos of human speakers, judges' ratings were highly correlated with the level of uncertainty originally reported by the speaker.

Preliminary work suggests that conversational agents will also achieve more natural interaction with users when those agents signal what they have understood and signal what they remain uncertain about [4]. In fact, uncertainty communication may well prove particularly important for conversational agents, because we can expect agents to have less reliable perceptual information than people in conversation and

V.G. Duffy (Ed.): Digital Human Modeling, HCII 2007, LNCS 4561, pp. 443–452, 2007.

less complete knowledge of their partners and their contexts. Conversational agents will therefore have to take risks to optimize their behavior in the presence of uncertainty rather than to find one absolute best way that is guaranteed to work in all situations; and, to collaborate effectively with their partners, they will also need to signal that uncertainty.

As a stepping stone toward designing interactive agents that achieve this ultimate goal, we focus here on what nonverbal behaviors we should add into computer agents to allow them to express their uncertainty in face-to-face interaction and how we can implement these behaviors practically and reliably. We have identified and annotated a set of uncertainty signals through a systematic coding methodology and then implemented the simulated behaviors onto our ECA (Embodied Conversational Agent), RUTH [5]. The implementation exploits the fact that ECAs resemble the appearance of real humans and are capable of carrying out many common behaviors we saw in people, so we can create animations in RUTH by analyzing and specifying the time course of behaviors as observed in natural human-human communication. As a follow-up to our communication and animation research [3, 6], here we aim at evaluating how well people can recognize uncertainty expressions as animated with RUTH, and at using RUTH to discover the effects of different layers of nonverbal signals that we have found in uncertainty communication.

2 Believable Interaction with a Believable Agent

Making an agent believable involves addressing users' expectations for both the appearance and the behavior of the agent [7]. And the bar keeps getting higher – as users gain increasingly diverse experiences with digital entertainment and computer-mediated communication in everyday life, they may develop correspondingly heightened expectations of ECAs. System designers can't realize agents with all the naturalness users might call for. As Nass et al., [8] point out, as designers, we must identify where we can get by with endowing our agents with a transparently synthetic character, and where we must make them more closely resemble humans – on what dimensions and on what level of detail. The evaluation we report here is a check on how well we have managed these design trade-offs.

2.1 RUTH

Our research starts from our embodied conversational agent, RUTH (Rutgers University Talking Head) [5]. RUTH is a freely available cross-platform real-time facial animation system developed by the VILLAGE Laboratory at Rutgers University. It is available at http://www.cs.rutgers.edu/~village/ruth/. A key goal of developing this animation platform has been to offer a methodological tool for developing and testing theories of functions and behaviors that occur during natural face-to-face conversation. Our work on the communication of uncertainty built on the initial design of RUTH and helped shape a number of extensions to the underlying capabilities of the system [6].

Fig. 1. RUTH, our talking head agent

RUTH displays a talking head rendered in three dimensions – the animation shows no hands or torso, just a head and neck. Figure 1 is a snapshot of RUTH smiling. RUTH's appearance is intentionally designed to be somewhat ambiguous as to gender, age, and race. RUTH is most often described as male although it has been seen as a female figure. With its large, stylized eyes, RUTH can be reasonably believed to be a range of ages, from elementary school to young adult. RUTH has the characteristics of a number of different ethnicities. The idea was to appeal to a wide range of users. However, our experience also taught us that it was also possible for certain users to find RUTH a bit discomfiting because RUTH seems to lack a specific identity.

2.2 Encoding Uncertainty Facial Signals onto RUTH

RUTH takes input for behaviors on the face and head, including the possibility of most of the facial actions inventoried in the standard Facial Action Coding System (FACS) [9, 6]. RUTH can work in conjunction with the Festival speech synthesis system to derive a sound file together with animation instructions for corresponding lip, face, and head movement. These specifications then guide a synchronized realization of the animated speech in real time. The work reported here describes silent

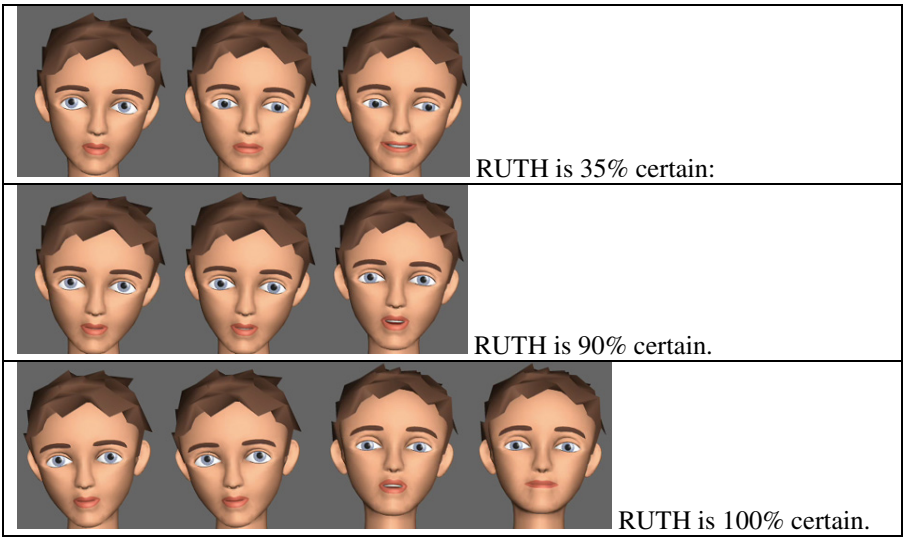

Fig. 2. RUTH's simulation of real humans' uncertainty expressions

animations that were specified directly from the time course of human behaviors; there was no speech synthesis. This illustrates a general feature of RUTH's architecture – the specification for RUTH can be done both in simple and abstract ways, depending on human analyst's coding scheme [5].

In this study, we used animations that were constructed as follows. A FACS coder analyzed a recording of a person's natural utterance, and specified the type of facial *action unit* (AU) that occurred at each moment (AU), as well as its intensity level (on a scale of A-E, from minimum to maximum), and four timing values: onset, apex-start, apex-end, and offset. The intensity levels were converted to numerical parameters for applying deformations in the animation engine so as to mirror the appearance changes seen in the original video. The timing of onset, apex and offset was used to adjust RUTH's generic temporal profile for animating behaviors to match speakers' particular motions. Figure 2 contains snapshot images of three utterances from our data set, as animated with RUTH. More details about the animation methodology and its implications for the technology of conversational animation can be found in [6].

3 Recognition Experiment: Understanding RUTH

In using RUTH to explore signals of uncertainty, we developed a set of expressive resources that allow RUTH to closely mirror an original human performance in terms of the available behaviors and the temporal coordination among them. Nevertheless, the details of RUTH's appearance and motions are highly simplified and stylized, leaving substantial differences with a video of a person talking. A key question is therefore whether RUTH's architecture and the animations we generate with it can succeed in conveying different levels of certainty to a conversational partner. To investigate this question, we carried out a recognition experiment that corresponded as closely as possible the video only condition from our earlier judgment study [3]. The major difference was that, instead of presenting recordings of real human subjects, the animations by RUTH served as the test items.

3.1 Stimuli

We started with ten videos of human subjects giving short answers in conversation. These videos were collected in a conversation experiment designed to elicit utterances with varying levels of certainty; after giving their responses the original subjects were asked how certain they were, giving a self-reported certainty level for each clip. The questions to which subjects responded were "Do you know in what year the University was chartered?" and "Who is the current Rutgers University president?", and arose in a dialogue covering questions a student might ask about the University [3]. The videos were then analyzed and recoded to provide input to RUTH to synthesize a corresponding delivery [6].

We used the resulting specifications to create three sets of RUTH animations as stimuli for the present experiment. The first set, the *full facial behavior condition,*

included all the observable nonverbal behaviors in our analysis, including the face, head and eye movements. The second set, the *facial gyration only condition,* included just the deformations that play out on the face, but did not include eye or head movements. The third set, *the head/eye movement only condition,* included movements of the head and eyes, as well as the movements of the lower face driven by the production of visual speech (mostly movements of the lips, tongue and jaw), but did not include the other, expressive movements of the face.

3.2 Participants

All participants were current Rutgers University students. The majority of students were recruited from the Psychology department subject pool system with the exception of a few volunteers. There were total of sixty participants. Fifty five subjects were undergraduate students while five others were graduate students. Each participant was randomly assigned to one of the three conditions. There were nine males and eleven females for full facial behavior condition, fifteen males and five females each for both the facial gyration only condition and the head/eye movement only condition.

3.3 Procedure

Subjects performed the experiment using a web based interface which allowed them to view stimulus videos and to record their judgments about what they saw. Before making each judgment, they could view the video up to three times. For each video, subjects were asked to rate how uncertain the speaker looked on a scale of 0 (completely uncertain) to 100 (completely certain). Subjects were also asked to specify their own level of confidence in making the judgment, as either very confident, somewhat confident, or not confident at all. Subjects saw two practice items and, then, rated ten test items associated with the condition to which they had been assigned. Items were presented in random order to avoid order effects. Upon each rating, a response was automatically stored into a preconfigured MySQL database through a web interface with other information such as subject number, condition, test item, and experiment date. At the debriefing and an exit interview, participants were informally asked about their overall impression on the experiment and the difficulty in accomplishing the tasks.

4 Overall Results

The two primary goals of the recognition experiment were: first, to investigate how well people could recognize the uncertainty expressions that were coded onto ECA and, second, to learn the effects of different layers of the nonverbal signals. Table 1 lists the mean values of each certainty level by the subjects' original self reports and the judges' ratings of the three different conditions of RUTH animation. In addition, Spearman's rho correlation coefficients were computed for these five variables and are presented in Table 2.

Table 1. The mean values of the judges' uncertainty ratings across conditions, ordered by the self-report of the original speaker from whom the animation was modeled

Self-report: certainty level	All: face and head	Face only	Head/eyes only
100	83.2	74.5	73.3
90	75.8	65.5	85.8
80	62.3	64.5	59.1
75	62.1	64.5	62.1
65	61.8	49.9	63.5
50	27.7	74.8	23.5
40	29.4	49.9	25.8
35	25.05	56.9	28.2
10	31.75	50.0	38.9
0	23.7	26.5	36.1

Table 2. Correlation coefficients relating judges' mean ratings with other factors: the original self reports, or judges' ratings of completely expressive animations or ratings of face only

Spearman's rho		All: face/head	Face Only	Head/eyes only
Self report	Correlation Coefficient	**0.915	*0.683	*0.709
	Sig. (2-tailed)	0.000	0.030	0.022
	N	10	10	10
All: face and head	Correlation Coefficient	1	0.524	**0.830
	Sig. (2-tailed)		0.120	0.003
	N		10	10
Face only	Correlation Coefficient		1	0.207
	Sig. (2-tailed)			0.565
	N			10

The most important result here is that judges' ratings of RUTH animations containing all of the analyzed behaviors correlate very highly with the self-reports of the original speakers ($r = .915$, $p < .001$). Indeed, when we compared the correlations between judges' ratings of the original videos of human speakers [3] and judges' ratings of the RUTH animations, we also get a very high correlation (Spearman $r = .903$, $p < .001$, 2-tailed, N=10). Moreover, judges' scores track uncertainty levels fairly accurately across all conditions. When we made comparisons with the self-reported certainty levels from the conversation experiment, the ratings in the signal identification experiment and in all three conditions from the recognition experiment showed significant associations. The relationship was lower when we removed head and eye movements ($r = .683$) and pure facial gyration signals ($r = .709$). Concretely, this means that about 46% of the variance ($.683^2$) in judges' responses to pure facial gyration was accounted for by its linear relationship with speakers' self-reported certainty.

Likewise, about 50% of the variance ($.709^2$) in judges' responses to head and eye movements was accounted for by its linear relationship with speakers' self-reported certainty.

However, the different nonverbal channels seem to be providing judges with very different kinds of information. Note that we found no evidence that judges' responses to an animation with head and eye movements removed covaried with judges' responses to the same animation with facial gyration removed ($r = 0.207$, $p = 0.565$ NS). This suggests that the different modalities present complementary information that is reconciled in viewers' overall understanding of a talking speaker, as in the integrated message model of nonverbal communication [10]. Indeed, it may be that facial expressions get their precise meanings in part by how they co-occur temporally with movements of the head and eyes – since movements of the head and eyes seem to signal what an interlocutor is doing to contribute to the conversation (e.g., listening, planning an utterance, presenting information, questioning or revising previous contributions) while other displays seem to serve to appraise how well that ongoing activity is proceeding. It is striking in this connection that the correlation of judges' responses to an animation with the head and eye movements removed had such a weak correlation with judges' responses to the full animation ($r = 0.524$, $p = .12$ NS). We hope to investigate this further in future work.

4.1 Judgments for Individual Viewers

So far, we reported the effective reliability of the mean judgments. This shows that there is no systematic bias or information loss in viewers' judgments of the RUTH videos. This does not show, however, that individual viewers recognize the uncertainty of a specific video reliably or agree with one another's judgments of what they see. To explore this question, we computed correlations using all 200 individual data points for each condition. The overall tendencies remained the same as those shown in Table 2. The Spearman's rho correlation coefficient between the self-reported and the rated certainty level was the strongest when all available visual cues were presented to them ($r = .635$, $p = 0.01$) and the weakest when head/eye movements cues were all removed ($r = .405$, $p = 0.01$). The correlation was .541 when we removed all observable facial gyration signals.

4.2 Judgments for Individual Items

We can also analyze responses by item, and explore the extent to which different animations are equally effective at conveying a particular level of uncertainty. To present these findings, we tabulate the *certainty recognition difference*, which is computed by subtracting the self-reported uncertainty level from a judge's ratings, for subject's ratings of each of the test items in the sample.

Figure 3 graphically displays the variance on each test item, presenting the median, interquartile range, outliers, and extreme cases within the *certainty recognition difference* variable for ten certainty levels: 0%, 10%, 35%, 40%, 50%, 65%, 75%, 80%, 90%, and 100%. In most cases, the median values were close to 0%, meaning that there was no systematic bias in how judges rated it. Judges most accurately rated the

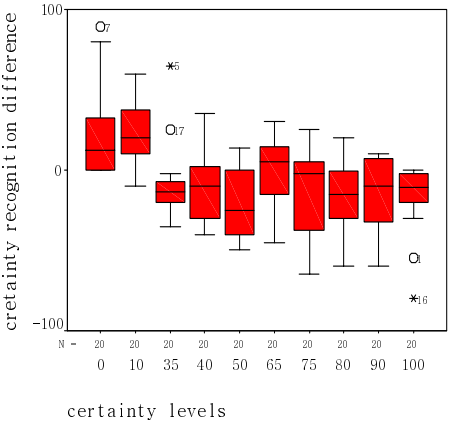

Fig. 3. Difference between the self-reported uncertainty levels and the ratings at the recognition experiment when all signals were presented

100% certainty test item. The variance was the smallest for the 35% certainty. For 0% certainty, there were three individuals who rated 0% as either 80% or 85% certainty. These judges seemed to be confused or forgetful about our operationalized definition of uncertainty – if the agent appeared not to know the answer at all, judges were instructed to give a rating of 0% certainty. During the debriefing, judges sometimes reported that they rated this test item as high certainty because they thought the agent looked "certain" of not knowing the answer. By contrast, the subject who reported 50% certainty showed lots of hesitation and confusion during the conversation experiment. To several judges, this test item offered a more intuitive illustration of high uncertainty than a speaker who quickly indicates that they cannot provide an answer.

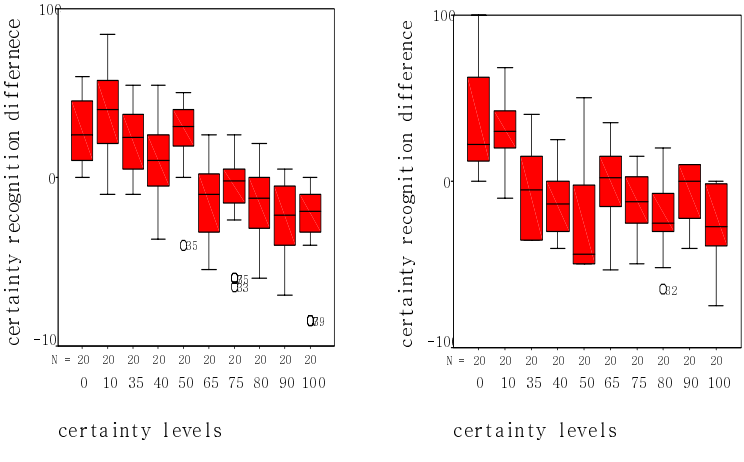

Fig. 4. Certainty recognition differences: pure facial gyration without head/eye movements (left) and head/eye movements only (right)

Figure 4 shows the box plots for the difference distributions for the simplified animations. These visualizations help to suggest the different ways viewers might combine information from facial expressions and head movements to understand an animation. In cases such as the 50% test item, it seems obvious that the head and eye movements – an extended interval of looking away and shaking the head while speaking – had more to do with the raters' very low judgments of certainty than the facial gyration – in this case, a sustained and perhaps apologetic smile. Finally, there were some cases where the signals from the two different channels seemed to be weighted together in subjects' ratings of the overall behavior. Thus, compared with the original values in the case of 35% and 40%, people gave relatively higher certainty ratings with the head/eye movement information and relatively lower ratings with the pure facial gyration signals. If this was a case where the two cues were redundant, that may explain why the full rendering of the 35% test time had the clearest signals among all test items (See Fig. 2).

5 Conclusion

Our results showed that judges rate certainty for animations with RUTH in ways that correlate closely with the self-reported uncertainty of the speaker on whom the animation was based, and with judges' ratings of the original video of the speaker. These relationships strongly suggest that people could recognize different levels of uncertainty not only with another human conversational partner, but also with an embodied conversational agent that used similar cues. Our work thus promises to inform the design of agents that signal their uncertainty as effectively in conversation as people do – and, more generally, to allow us to exploit knowledge and observations of specific human behaviors to enrich human-agent interaction.

Our research also points to how viewers make the judgments they do. The association between the self-assessed uncertainty and the rated uncertainty in all three conditions (*all observable, pure facial gyration, and head/eye movements only*) proved to be highly significant. Head and eye movements were better indicators of uncertainty than pure facial gyrations. But raters made the most reliable judgments when both kinds of information were presented to them in a coordinated manner. In future work, we hope to develop a more detailed model of the way addressees combine these different signals into an integrated understanding of speakers.

The kinds of human computer interaction that ECAs can bring to us are unique and varied. What is new with ECA is that we must now enable an agent to use its embodiment in support of effective interaction. Research still has a long way to go to achieve this. However, it is increasingly practical and reliable to increase agents' believability by crafting communicative expressions like our own for them.

Acknowledgments. We are grateful for the assistance of Mark Frank and Doug DeCarlo at various stages of this research. Our work was supported in part by NSF HLC 0308121 and by the second author's Leverhulme Trust Visiting Fellowship at the University of Edinburgh 2005-2006. We are grateful to Alias for the use of Maya modeling and animation software.

References

1. Clark, H.H.: Using Language. Cambridge University Press, Cambridge (1996)
2. Swerts, M., Krahmer, E.: Audiovisual prosody and feeling of knowing. Journal of Memory and Language 53(1), 81–94 (2005)
3. Oh, I., Frank, M., Stone, M.: Face-to-face communication of uncertainty: expression and recognition of uncertainty signals by different levels across modalities. In: ICA International Communication Association (2007)
4. Nakano, Y.I., Reinstein, G., Stocky, T., Cassell, J.: Towards a model of face-to-face grounding. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2003), pp. 553–561 (2003)
5. DeCarlo, D., Revilla, C., Stone, M., Venditti, J.: Specifying and animating facial signals for discourse in embodied conversational agents. Computer Animation and Virtual Worlds 15(1), 27–38 (2004)
6. Stone, M., Oh, I.: Interdisciplinary challenges of generating animated utterances. In: Wachsmuth, I., Knoblich, G. (eds.) Modeling Communication with Robots and Virtual Humans, Springer, Heidelberg (2007)
7. Bailenson, J.N., Blascovich, J.: Avatars. In: Bainbridge's, W.S. (ed.) Encyclopedia of Human-Computer Interaction, pp. 64–68. Berkshire Publishing Group (2004)
8. Nass, C., Robles, E., Wang, Q.: 'User as Assessor' Approach to Embodied Conversational Agents. In: Ruttkay, Z., Pelachaud, C. (eds.) From Brows to Trust: Evaluating Embodied Conversational Agents (Human-Computer Interaction Series), Kluwer Academic Publisher, Netherlands (2004)
9. Ekman, P., Friesen, W.V., Hager, J.C.: Facial Action Coding System (FACS): Manual and Investigator's Guide. A Human Face, Salt Lake City, UT (2002)
10. Bavelas, J.B., Chovil, N.: Visible acts of meaning: an integrated message model of language in face-to-face dialogue. Journal of Language and Social Psychology 19(2), 163–194 (2000)