

# Reproducing Natural Behaviors in Conversational Animation

Matthew Stone and Insuk Oh

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

**Abstract.** Building animated conversational agents requires developing a fine-grained analysis of the motions and meanings available to interlocutors in face-to-face conversation and implementing strategies for using these motions and meanings to communicate effectively. In this paper, we describe our research on signaling uncertainty on an animated face as an end-to-end case study of this process. We sketch our efforts to characterize people’s facial displays of uncertainty in face-to-face conversation in ways that allow us to simulate those behaviors in an animated agent. Our work has led to new insights into the structure, timing, expressive content and communicative function of facial actions that we must take into account to explain our empirical findings and to build agents that reproduce people’s effective use of the face in managing the dynamics of conversation.

## 1 Introduction

Our cooperative conversations are amazingly complex activities. As we talk, we must get the right content across, we must give each contribution the right emphasis to show why we are making it, and we must support the interaction by collaborating to ensure that it is understood. With all this to do, it’s no wonder that people produce and attend to a wide range of nonverbal behaviors, as well as the words they use in conversation.

Research has shown that people can use these behaviors as an integral part of their face to face conversations. For example, coverbal gesture can help convey content in dialogue [1]. Eyebrow flashes can help to mark emphasis in the accompanying speech [2]. Even how one holds one’s body [3] and directs one’s head [4] can signal how the discourse fits together and whether interlocutors understand this.

Indeed, the lesson of this research is that we have a rich repertoire of expressive coverbal behaviors—much more than any one study can investigate. How do we get a grip on these expressive capabilities? Concretely, just focusing on the face, what specific expressions do people have at their disposal in conversation? And how do those movements help them to successfully contribute to the ongoing communication? These are the questions we aim to open up in this chapter.

We describe here how we have extended the capabilities of our face animation system RUTH (Rutgers University Talking Head) [5] to support the study of specific, naturalistic facial movements as performed in synchrony with coverbal speech.<sup>1</sup> The revised implementation of RUTH synthesizes a range of insights from across the science of communication. For example, as described in Section 3, we have drawn on the Facial Action Coding System (FACS [6]) in extending the inventory of basic movements of our agent. We have drawn on previous animated agents, and the literature on speech prosody, such as [7], to describe how those movements align with speech. These principles are realized in our analyses of the behaviors of people’s spontaneous utterances; see Section 2.1. They are also formalized operationally within our animation engine, which brings its own assumptions and limitations; see Section 4. Finally, our methodology embraces psychological techniques for human-subjects investigations of the interpretation of embodied utterances. In particular, we make sure we can not only integrate our new animations into conversational systems, but also take our new animations into the lab and evaluate them. We begin in Section 2 by motivating the technical work on RUTH by presenting an example of this methodology—characterizing expressions of uncertainty—and highlight the challenges it presented for the original version of RUTH.

## 2 Context

As a test case for enriching the expressive repertoire of animated agents, we chose to focus on expressions that may help to convey one interlocutor’s uncertainty about a contribution to conversation. There are three reasons for this choice. First, we expect such expressions to be important. Uncertainty is crucial to the dynamics of grounding and reaching mutual understanding in dialogue [8, 9]. On the theoretical side, characterizing interlocutors’ uncertainty is necessary to explain how people avoid, detect, and resolve misunderstandings. Practically, a system that can signal its own uncertainty and recognize its interlocutors’ may be able to interact more robustly with people. Uncertainty has therefore received particular attention for human–computer interaction, for example in tutorial dialogue [10, 11].

Second, we expect signals of uncertainty to be diverse and frequent. We constantly show our level of understanding and perceive others’ level of understanding through a variety of audiovisual channels in order to make an appropriate conversation continue [4, 12–14]. In the case of uncertainty in particular, a recent survey [15] suggested that nonverbal cues could include frowns, side-ward eye movements, lip-pouting, lip-pursing, tensed mouths, side-to-side head-shakes, head tilts, self-touch gestures, hand-behind-head cues, palm-up gestures, and the shoulder-shrug—elaborating a common-sense inventory that goes back at least to Darwin [16]. Informal investigation has identified such expressions of

---

<sup>1</sup> The updated version of RUTH will be released on the web for research use in Spring 2007. The final version of this chapter will therefore serve as an archival description of this new release, as well as documenting the process by which it was developed.

uncertainty, including related states such as confusion, as some of the most frequent displays in American conversation [17]. Swerts and Kraemer [13, 14] took a more systematic approach to study the expression of speaker’s uncertainty by means of audiovisual prosody. They found that eyebrow movements, smiling, diverted gaze, and marked facial expressions were reliable visual cues that speakers produced and addressees used to distinguish responses where speakers felt confident in the information they were providing from those where speakers did not.

Third, we expect a rich and complex relationship between people’s signals of uncertainty and their underlying states. Subjectively, uncertainty involves both cognitive and affective components [18, 19], and the relationship among these components is hotly debated in the communication literature [17, 19]. So it is wide open how specific expressions might function to convey uncertainty by signalling specific elements of an interlocutor’s cognitive or affective state.

## 2.1 Visual communication of uncertainty

We designed and carried out a four-stage experiment to investigate the expression and interpretation of uncertainty in face to face conversation. We began by collecting a corpus of speech and video of subjects engaged in a face-to-face interaction designed to elicit a range of contexts and degrees of uncertainty. Subjects spoke in an informal conversation with an experimenter, and discussed a familiar topic, Rutgers University. Along the way, the experimenter asked a series of questions about the university drawn from the University’s frequently-asked questions website. At key questions, the experimenter also asked subjects how certain they were of their answer on a scale of zero to one hundred.

For further analysis, we focused on two questions that elicited a wide range of different certainty responses: *Who is the current Rutgers University president?* and *Do you know in what year the University was chartered?*<sup>2</sup> With this analysis, we first aimed to discover whether subjects’ responses showed reliable cues to their level of certainty. In fact, as we describe elsewhere [20], we found empirically that judges’ ratings of a speaker’s apparent certainty do correlate quite closely with the certainty the speaker themselves reported—as long as judges get at least the video or the audio from a recording of the subject’s original delivery. People cannot recover uncertainty merely from the text transcript of what was said.

We then wanted to investigate whether we could reproduce the visual cues to uncertainty on an animated agent, and use those cues to communicate the agent’s uncertainty. We therefore created animations with RUTH that reproduced subjects’ deliveries in our test items as closely as possible. We showed (silent) videos of these deliveries to subjects again, and found that viewers’ ratings of RUTH’s

---

<sup>2</sup> The latter question might seem somewhat obscure to outsiders. The Rutgers campus is festooned with flags and monuments commemorating the founding of the University in 1766. So students definitely feel this is something they *should* know, even if they can’t always come up with the exact year.

apparent certainty also correlate quite closely with the subject’s reported certainty. Moreover, by manipulating the elements of the animated performance, we were able to show that both facial movements (displays of affect) and head and eye movements (manifestations of cognitive state) seem to contribute to viewers’ judgments. We describe these experimental methods and results in full in [21].

Our focus in this paper is on the process of encoding human behaviors onto RUTH, and the particular lessons we learned by developing a representation of the performance of specific embodied utterances that bridges the fields of communication, linguistics, and computer animation. Although we have always aimed at such a synthesis in developing RUTH [5], we had never before attempted to realize animations that faithfully reproduced such complex signals of interactional and expressive state from qualitative specifications. So our initial starting points—our initial prototype of RUTH, and a FACS coding of people’s behaviors in specific utterances—were further apart than we anticipated.

## 2.2 Starting Point: RUTH

RUTH [5] is a real-time facial animation system that animates conversational facial displays and head movements in synchrony with speech and lip movements. RUTH’s animation primitives include deformations of an underlying polygonal mesh, which are summed linearly when applied in combination. RUTH handles head movements by applying rotations and translations to part of the model; the effect fades out across the neck to maintain smooth geometry. The eyes also rotate in place. For speech, RUTH uses a coarticulation model based on dominance functions [22–24] to capture the fact that different speech sounds are visible to different degrees and for different durations during articulation.

To explore the relationship between nonverbal actions and speech, RUTH offers a high-level specification of utterance realization in which qualitative nonverbal behaviors are timed in synchrony with the prosodic structure of utterances. RUTH assumes the ToBI model of English intonation [7], and uses the timestamps of accented syllables and of breaks between phrases as possible synchronization points between speech and gesture. Short behaviors (such as a nod) can be specified to synchronize with an accented syllable. Longer behaviors (such as raised eyebrows) can be specified to synchronize with a phrase as a whole; in this case, key milestones in the time-course of the animated behavior are timed to coincide with the beginning of the phrase, the first accented syllable in the phrase, the last accented syllable in the phrase, and the end of the phrase. In the resulting animations—as in people’s natural utterances—units of prosodic structure coincide with units of gesture and prominent syllables coincide with the most prominent phases in the realization of accompanying behaviors [25–28].

We prototyped RUTH using a small collection of recordings of scripted utterances, mostly from the domain of broadcast news. Our initial design captured only the most frequent behaviors from these recordings. In particular, we started with those behaviors that signal discourse structure and emphasis, rather than those that express affect or help manage interaction. The prototypical display of emphasis is a symmetrical movement of the brows, highlighting an individual

word or an extended phrase. See also [2]. Speakers may use raised eyebrows or, less commonly, a frown. Brief movements of the head offer a more general clue to the function of a word or phrase in the ongoing discourse. A downward nod on a word or phrase is the most common case, but speakers can also avail themselves of a range of other behaviors, including head raises, tilts, and turns, to mark what is said with a particular prominence or contrast.

### 2.3 Data and Analysis

We selected ten interaction segments from our data set for in-depth analysis. We transcribed all spoken words and obtained timings for individual phonemes in the speech by using the Praat speech analysis program ([www.praat.org](http://www.praat.org)) to hand-correct an automatic alignment of sound and transcript. Independently, two newly-trained coders used the Facial Action Coding System (FACS [6]) to characterize the movements of the face. FACS specifies criteria to assess the type, intensity, and timing of facial movement. A set inventory of types of facial action are considered, such as, for example, *eyelid tightening*, action unit 7, which achieves a squinting look that, in conversation, can convey doubt, and *chin raise*, action unit 17, which results in a pouting effect that is part of the classic “facial shrug” emblem shown in Fig. 4(b) and (c). FACS guidelines describe the degree of each appearance change as an intensity level from smallest (A) to greatest (E). Time denotes the duration of a movement, from the onset where the motion is first visible, to the apex where it reaches its maximum intensity level, to the offset where the motion is last visible. One of the most important aspects of using FACS to extend the behavioral repertoire of conversational agents is its comprehensiveness—it measures all visible facial movement, not just that presumed to be associated with emotion or cognitive states—and thus it allows for discovery of relationships between movement and psychological states [29].

We used an event-based coding, which means that we marked the durations of actions and their peak of intensity, rather than giving a frame-by-frame classification of action and intensity. We departed from the typical event-based coding in FACS, however, in allowing coders to identify an interval for the apex over which an action peaks, rather than just a single frame. This is important for conversational displays because they are often consciously held for an extended interval; it also anticipates our goals of linking the annotations to the time course of an animation and to events in the synchronous speech.

To code head pose and eye direction, we developed our own rough categorization of the appearance of the subject. (We found the FACS coding for head and eyes, which directly describe individual degrees of freedom for movement, somewhat counterintuitive and difficult to use.) We coded the apparent tilt of the subject’s head as a clock direction, so that for example 12 is straight up, 1 slightly to the left, 11 slightly to the right, and so forth. Then, we coded the intensity and direction of the subject’s eyes and nose as another clock direction (e.g. 3 is turned to the left, 9 to the right, N for straight ahead), judged with respect to the natural axis of the subject’s head, which was perhaps tilted away from the vertical. We didn’t explicitly mark forward and backward motion,

but we did attempt to reconstruct it later when realizing specific movements in animation.

The overall annotation we did can be summarized in a behavioral map [30], showing the different layers and components of a person’s aggregated behavior in one schematic table. Figure 1 offers one example. This overview of the utterance helps to suggest how speakers demonstrate their uncertainty in multimodal presentations. It also showcases the rich new repertoire of expressions and fine-grained structure of movement we were confronted with in looking at facial signals in natural conversation.

### 3 Design Rationales for Extensions

Our work led to two major changes in RUTH: a new approach to head pose (Sec. 3.1) and a new inventory of facial actions (Sec. 3.2).

#### 3.1 Head Pose

We did not account for posture shifts in the initial specification of RUTH. Head movements were treated as excursions from a fixed neutral position. Utterances always began in the neutral position and finished in the neutral position. This simple scheme sufficed for the monologue data which we used to guide our initial design.

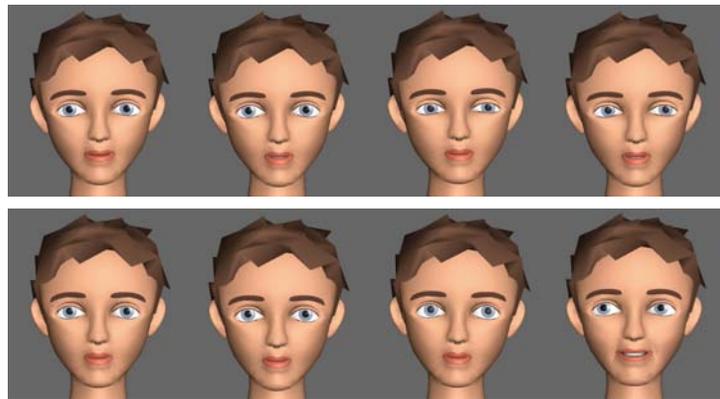
The naturalistic data we collected, however, made it clear that posture shifts are an important element of conversational behavior. Interlocutors seem to use these posture shifts to dramatize the status of their participation in the conversation. Of course, the current state of empirical research is not far enough along to support a scientific description of exactly how these behaviors function in conversation. Instead, we will try to convey our intuitions that these behaviors have relatively transparent meanings by offering the impressions that we have as viewers. For example, while not speaking or formulating an utterance, participants often adopted a characteristic “listening” pose, with the head brought forward and posed slightly lower, in a kind of pantomime of attention and deference. Participants often marked their efforts to search for information or plan an utterance by repositioning the head and eyes to look upward and away. They adopt distinctive poses for the delivery of planned utterances, and thereby embody the attitude with which the material is contributed: raising themselves and facing directly forward in confident deliveries; orienting their face slightly sideways but their eyes directly forwards, as if in anticipation of a sceptical response, to proffer controversial suggestions; or tilting their head on tentative replies as if to invite a collaborative response. And they often mark disfluencies by abrupt head movements, for example to resume the kinds of attitudes seen in utterance planning.

It is useful to describe such behaviors as posture shifts rather than as excursions for three reasons. First, successive movements adopt a distinct series of poses without any return to a neutral resting position. A speaker may move

		
phrases	(silence)	seventeen
eyes	2; then 10; then 11	12
head	1 tilt; N turn	(held)
face	38B; then R9C+R14C+17C+5D+26C 38: nostril dilator 9: nose wrinkler 14: dimpler 17: chin raiser 26: jaw drop 5: lid raise	6C+12C 6: cheek raiser 12: lip corner puller
		
phrases	forty two	maybe
eyes	N	N
head	12 tilt; 6B turn	12 tilt; 8B turn
face	4E+7C; 6C+12C continues 6: cheek raiser 12: lip corner puller 4: brow lowerer 7: lid tightener	6C+12C continues 6: cheek raiser 12: lip corner puller

**Fig. 1.** A map of one subject's behavior in a specific interaction, as we annotated it. The top row shows representative images tracing out the subject's embodied interaction. The words are broken out into phrases: we have first the initial silence while the subject planned her response; then the answer, demarcated into three short phrases that help to set off *seventeen* as a sure answer and mark *forty-two* as one possibility, open for further discussion. The head and eye movements and facial actions are synchronized with these phases in the presentation of the utterance and, as the detailed coding suggests, integrate a complex range of behaviors.

directly from a listening pose, to a planning pose, to a sceptical delivery, without ever adopting or even passing through a neutral rest position facing the interlocutor with the head held high. Second, there seems to be no intrinsic limit to the duration of a posture shift, and accordingly no expectation that an interlocutor who adopts a distinctive pose must soon end it. A listening attitude, for example, can last indefinitely. It’s not a delimited excursion—it is in effect a change in what counts as the agent’s neutral state. Finally, the meanings of the behaviors follow from the position adopted, and not the motion or excursion that the speaker effects. Our examples illustrate this. Our discussion narrates the import of these behaviors by interpreting where a speaker has placed their head as they deliver their contribution, without reference to where their head has been or where they might put it next.



**Fig. 2.** Posture shifts on the head in the delivery of embodied utterances. Each row shows four images drawn from our realization of a specific interaction from our data set. The sequence begins with the subject’s neutral pose, followed by an attentive pose marked, among other things, by a lowering of the head. The subject turns away in utterance planning, then adopts a distinctive final pose to deliver their response. These are good examples of posture shifts, as opposed to nods, shakes, or other conversational head gestures, because these motions meaningfully target particular poses one after the other, without returning to a neutral rest.

Our conclusions align with those of Vilhjálmsón, Cassell and colleagues [31, 3], who study the posture shifts that agents manifest on their bodies as a whole. What we say about the face, in dramatizing the state of an interaction and dramatizing the attitude with which an interlocutor contributes to the discourse, mirrors, in the small, the same kinds of orienting behaviors that also play out on people’s shoulders and hips.

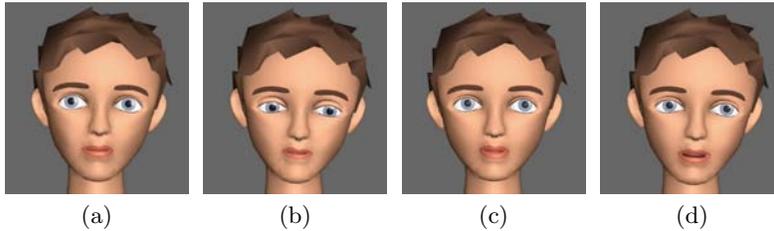
Posture shifts continue to align very closely with the prosodic structure of utterances. We do find posture shifts in periods of silence, of course. They are

part of how people signal their participation in the conversation as they listen and plan their utterances. But posture shifts that occur along with utterances seem to align in only one of two ways. Most posture shifts are timed so that the head adopts its new pose at precisely the same moment as an accompanying prosodic event. (In our analyses, precisely the same moment refers to the interval of a single frame of video, or about 0.03 seconds.) Although some such posture shifts are timed to align with the beginning or the end of an utterance, they most often coincide with the most prominent accented syllable in a phrase. For example, in returning after looking away in thought, speakers generally face back forward again in time with the delivery of the main accent of the contribution they have planned. This alignment is perhaps somewhat surprising and does add to the analytical difficulty of distinguishing between posture shifts and movements like nods that are meaningful as excursions. Finally, of course, we do find some posture shifts that begin with distinctive prosodic events. This particularly fits two cases: disfluencies, where the speaker must wait until the disfluency occurs to signal it and recover from it; and the ends of utterances, where the resumption of a listening pose is usually effected in the subsequent silence. These patterns of timing are consistent with our earlier model [5] and with much other research in nonverbal communication in assuming a close relationship between the timing of utterances and the timing of coverbal behaviors.

### 3.2 Facial Displays

We omitted a wide range of facial movements in the original specification of RUTH. We worked primarily with eyebrow movements and assumed movements were performed symmetrically on both sides of the face. This fit the data we originally worked with—factual content delivered in a dispassionate style—but probably did so only because that data offered few cases where speakers depicted the emotional content of what they said or dramatized the attitude with which they presented it. Our new experiment elicited a wide range of additional displays.

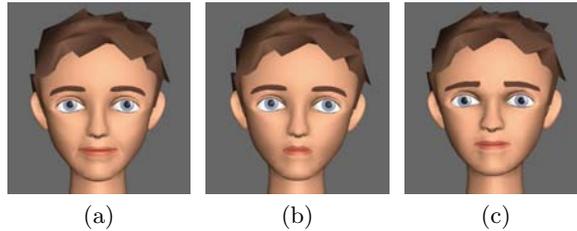
Our subjects used many of the elements of prototypical emotional expressions. However, these seemed less signs of felt emotions than icons that marked speakers' appraisals of their own contributions to conversation. This showed both in the forms of the behaviors and—at least impressionistically—in their meanings. The behaviors had a stylized form; they typically involved only part of the realization of a true emotion, and were often realized much more intensely on one side of the face than the other. And the behaviors seemed specifically tied to the ongoing communicative activity. For example, one subject displayed the characteristic nose scrunching and lip raising of disgust, just on the right side, and thereby seemed to convey not genuine revulsion at the proceedings, but rather a lighthearted indication that her contribution to the conversation was, as it were, a bit fishy. In another utterance, a different subject showed part of the signature of fear—specifically the left lip stretcher motion that pulls the lip down and to the side—midway through utterance planning, as though to signal her concern that she might be unable to provide needed information.



**Fig. 3.** People adapt the expressions of emotions to provide feedback about the state of the conversation. In (a), the wide eyes and lowered jaw of surprise combine with a gesture pulling the lip to the right—perhaps emblematic of chewing over something tough—to suggest that the speaker is working through a tough surprise. In (b), the lip raiser on the right evokes disgust, and hints that the utterance being planned will be unsatisfying. In (c), the lip is pulled down on the left, drawing on the basic expression of fear, again conveying the speaker’s negative appraisal of the projected outcome of their search for the right thing to say. Example (d) gives another, more striking example of a conversational facial gesture colored by fear.

Facial behaviors that go beyond our original model show up in expressions unrelated to the emotions as well. A commonly-seen case is the “facial shrug”, formed by pressing the lips together and raising the chin, arching the upper lip and allowing the lower lip to bulge outward. The expression, like a shoulder shrug, is a conventional signal of ignorance. Another case is the emblem of “holding oneself back”, lowering the jaw but keeping the mouth closed and tight using a dimpler behavior, often with the lips sucked slightly in. Speakers who use this emblem while looking away in silence, for example, create the impression that they are worried they might blurt out something wrong—they are reconsidering their utterance, rather than thinking or planning it from scratch.

To represent these examples, we had to extend the range of motion of our character. Where a motion can play out separately on different sides of the face, we introduced separate controls for the two sides. At the same time, we added motions for widening the eye in surprise (the FACS upper lid raiser), and tightening the eyelid to squint (the FACS lid tightener); for wrinkling the nose and for raising the lip, as in disgust; for lowering the lower lip (the FACS lip corner depressor) and stretching the lips downward and outward (the FACS lip stretcher), as in fear; and for raising the chin. In a number of cases, these behaviors no longer combine together in the linear way presupposed by our underlying animation engine. To keep the animation believable, we therefore created separate primitives to describe key behaviors in combination. These cases typically involve treating motions with lips closed differently to corresponding motions with lips open, to take into account the changes in deformation as the lips start to press against one another.

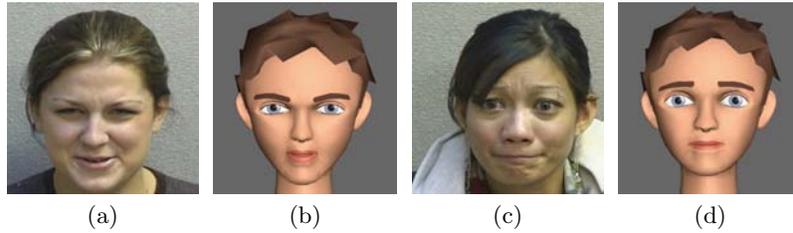


**Fig. 4.** Not all expressions get their meaning by reference to emotion. The pose in (a) is our rendering of an emblem for “holding oneself back”, with the jaw low and mouth stretched. (We cannot animate RUTH sucking the lips in, another common ingredient of this emblem.) This speaker seems to be zeroing in on the required information—in contrast to the situation in (b), where the speaker raises the chin (and the lower lip along with it), as part of an emblem of ignorance: the “facial shrug”. Image (c) shows a facial shrug on the lower face colored by elements of surprise and fear across the eyes, a pose that works as an ensemble to portray ignorance with its negative affect as well as its cognitive consequences.

## 4 Future Extensions and Issues

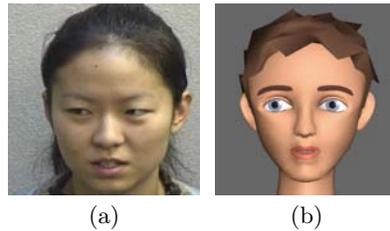
Our detailed analysis of particular utterances has left us with a more concrete sense of the gaps between human behaviors and the animations of RUTH, even as we have extended it. For one thing, there are a range of miscellaneous movements that we saw in human conversation but have not implemented in RUTH, including among others sucking in the lips, licking the lips, and blowing out air through tense lips in a sigh of frustration. These are difficult to animate with RUTH’s current architecture of prespecified deformations in linear combination; more so than the other behaviors we have animated these motions depend closely on the physical dynamics of force and contact across the face. However, these motions do fall into the important category of *manipulators* [32] some of which should probably be included in any agent’s behavioral repertoire. An agent that does not fidget when distracted or uncomfortable will lack one of the most common cues we saw to difficulty in conversation.

Even for those FACS action units which we included in RUTH’s repertoire, there are significant differences between how they look on human faces and how they look on RUTH. RUTH does not yet show changes in skin marks such as lines, wrinkles, and furrows. Of course, most human faces have permanent infraorbital furrows (below the lower eyelid) or nasolabial furrows (adjacent to the nostril wings) that become more prominent in many actions and can be crucial in allowing viewers to recognize certain of them [6]. Again, animating such changes goes beyond the resources of RUTH’s current design, which handles smooth deformations but not these other appearance changes. However, by leaving out these movements, we may have made RUTH’s facial expressions much harder to recognize.



**Fig. 5.** The absence of wrinkles and other appearance changes in RUTH animations may color the meaning of RUTH’s displays and make them more difficult for viewers to recognize.

Let us look at the example of Fig. 5. When we make side-by-side comparisons between actual human expression and RUTH’s implementation, we can see that missing skin marks can be a critical issue. The first image shows a subject expressing some confusion or uncertainty she was experiencing. The second is the corresponding RUTH implementation, which gives a viewer an impression that is closer to anger. In the same manner, the third image portrays the subject as apologetic, which is close to sadness; surprise and fear are the primary impressions from the fourth image. The difference between original human facial expression and RUTH implementation in still images can be subjective and subtle. However, it illustrates the possibility that the missing cues and marks can actually alter the meaning.



**Fig. 6.** Animation should lead to the same visual impression as the original video. That may mean exaggerating the human motions on RUTH. In this figure, the reorientation of RUTH’s eyes involves a comparable rotation in space as the original example but does not give as strong an impression of looking away.

One way to compensate may be to exaggerate RUTH’s expressions over the original human data. In fact, when an ECA coder reads the annotations without watching original video, it is still very difficult to know exactly how much movement responds to each intensity level. After all, FACS is designed to annotate observable movements from the face, not to provide specifications for computer

implementation. Therefore, ECA researchers have to make an inference as to how that FACS code should be coded into RUTH. Figure 6 shows RUTH with eyes turned away to the same degree as the humans subject. Nevertheless, the expressiveness from the human subject appeared to be relatively stronger. The slight and yet perceivable difference could be in part due to the simple physical difference in its appearance. For example, RUTH's eyes are relatively bigger than those of humans, thus diluting the effect of the same amount of movement on RUTH.

## 5 Conclusion

We believe that this methodology provides a general way to extend the effective behavioral repertoire of conversational characters. However, pursuing this work is challenging. It involves reconciling methodologies from communication, linguistics, and computer science to develop a specification for the behaviors of the agent. This paper documents the process. We first describe the data analysis we did in Sec 2.1 and some of the surprising elements we found there. In Sec 3, we then highlight how the empirical data we wanted to model led to substantial revisions in the capabilities of the animation system and the way the animation interfaces with utterance representations and dialogue architecture. Finally, as we discuss in Sec 4, the data also gives us reason to reconsider more general design principles for future animated agents.

## 6 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 thank Autodesk Alias for the use of Maya modeling and animation software.

## References

1. Cassell, J., McNeill, D., McCullough, K.E.: Speech-gesture mismatches: evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics and Cognition* **6**(2) (1999)
2. Krahmer, E., Ruttkay, Z., Swerts, M., Wesselink, W.: Pitch, eyebrows and the perception of focus. In: *Symposium on Speech Prosody*. (2002)
3. Cassell, J., Nakano, Y., Bickmore, T.W., Sidner, C.L., Rich, C.: Non-verbal cues for discourse structure. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2001)*. (2001) 106–115
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)*. (2003) 553–561

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) (2004) 27–38
6. 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)
7. Silverman, K.E.A., Beckman, M., Pitrelli, J.F., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J.: ToBI: a standard for labeling English prosody. In: *Proceedings of the International Conference on Spoken Language Processing*. (1992) 867–870
8. Paek, T., Horvitz, E.: Conversation as action under uncertainty. In: *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI)*. (2000) 455–464
9. DeVault, D., Stone, M.: Scorekeeping in an uncertain language game. In: *BRANDIAL: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*. (2006) 64–71
10. Liscombe, J., Hirschberg, J., Venditti, J.J.: Detecting certainness in spoken tutorial dialogues. In: *Proceedings of Interspeech*. (2005)
11. Forbes-Riley, K., Litman, D.J.: Analyzing dependencies between student certainness states and tutor responses in a spoken dialogue corpus. In Dybkjaer, L., Minker, W., eds.: *Recent Trends in Discourse and Dialogue*. Springer (2007)
12. Clark, H.H., Krych, M.A.: Speaking while monitoring addressees for understanding. *Journal of Memory and Language* (50) (2004) 62–81
13. Swerts, M., Kraemer, E.: Audiovisual prosody and feeling of knowing. *Journal of Memory and Language* **53**(1) (2005) 81–94
14. Kraemer, E., Swerts, M.: How children and adults produce and perceive uncertainty in audiovisual speech. *Language and Speech* **48**(1) (2005) 29–53
15. Givens, D.B.: *The Nonverbal Dictionary of Gestures, Signs and Body Language Cues*. Center for Nonverbal Studies Press (2001)
16. Darwin, C.: *The Expression of the Emotions in Man and Animals*. John Murray, London (1872)
17. Rozin, P., Cohen, A.B.: High frequency of facial expressions corresponding to confusion, concentration and worry in an analysis of naturally occurring facial expressions in americans. *Emotion* **3** (2003) 68–75
18. Brashers, D.E.: Communication and uncertainty management. *Journal of Communication* **51** (2001) 477–497
19. Ellsworth, P.C.: Confusion, concentration and other emotions of interest. *Emotion* **3** (2003) 81–85
20. 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)
21. Oh, I., Stone, M.: Understanding RUTH: Creating believable behaviors for a virtual human under uncertainty. In: *HCHI: Human-Computer Interaction International Proceedings*. (2007)
22. Löfqvist, A.: Speech as audible gestures. In Hardcastle, W.J., Marchal, A., eds.: *Speech Production and Speech Modeling*. Kluwer (1990) 289–322
23. Cohen, M.M., Massaro, D.W.: Modeling coarticulation in synthetic visual speech. In Thalmann, N.M., Thalmann, D., eds.: *Models and techniques in computer animation*. Springer (1993) 139–156
24. King, S.A.: *A facial model and animation techniques for animated speech*. PhD thesis, The Ohio State University (2001)

25. Ekman, P.: About brows: Emotional and conversational signals. In von Cranach, M., Foppa, K., Lepenies, W., Ploog, D., eds.: *Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium*. Cambridge University Press, Cambridge (1979) 169–202
26. Bull, P., Connelly, G.: Body movement and emphasis in speech. *Journal of Nonverbal Behavior* **9**(3) (1985) 169–187
27. McNeill, D.: *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago (1992)
28. Engle, R.A.: *Toward a Theory of Multimodal Communication: Combining Speech, Gestures, Diagrams and Demonstrations in Instructional Explanations*. PhD thesis, Stanford University (2000)
29. Cohn, J.F., Ekman, P.: Measuring facial action. In Harrigan, J.A., Rosenthal, R., Scherer, K.R., eds.: *The New Handbook of Nonverbal Behavior Research*. Oxford (2005) 9–64
30. Frank, M.: Research methods in detecting deception research. In Harrigan, J.A., Rosenthal, R., Scherer, K.R., eds.: *The New Handbook of Methods in Nonverbal Behavior*. Oxford (2006)
31. Vilhjálmsón, H., Cassell, J.: BodyChat: autonomous communicative behaviors in avatars. In: *Proceedings of the International Conference on Autonomous Agents (Agents 1998)*. (1998) 269–276
32. Ekman, P., Friesen, W.V.: The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica* **1**(1) (1969) 49–98