Specifying and Animating Facial Signals for Discourse in Embodied Conversational Agents

Doug DeCarlo, Matthew Stone
Dpt. of Computer Science and Ctr. for Cognitive Science, Rutgers University
{decarlo,mdstone}@cs.rutgers.edu
http://www.cs.rutgers.edu/˜village/ruth

Corey Revilla
Entertainment Technology Center, Carnegie Mellon University
crevilla@andrew.cmu.edu

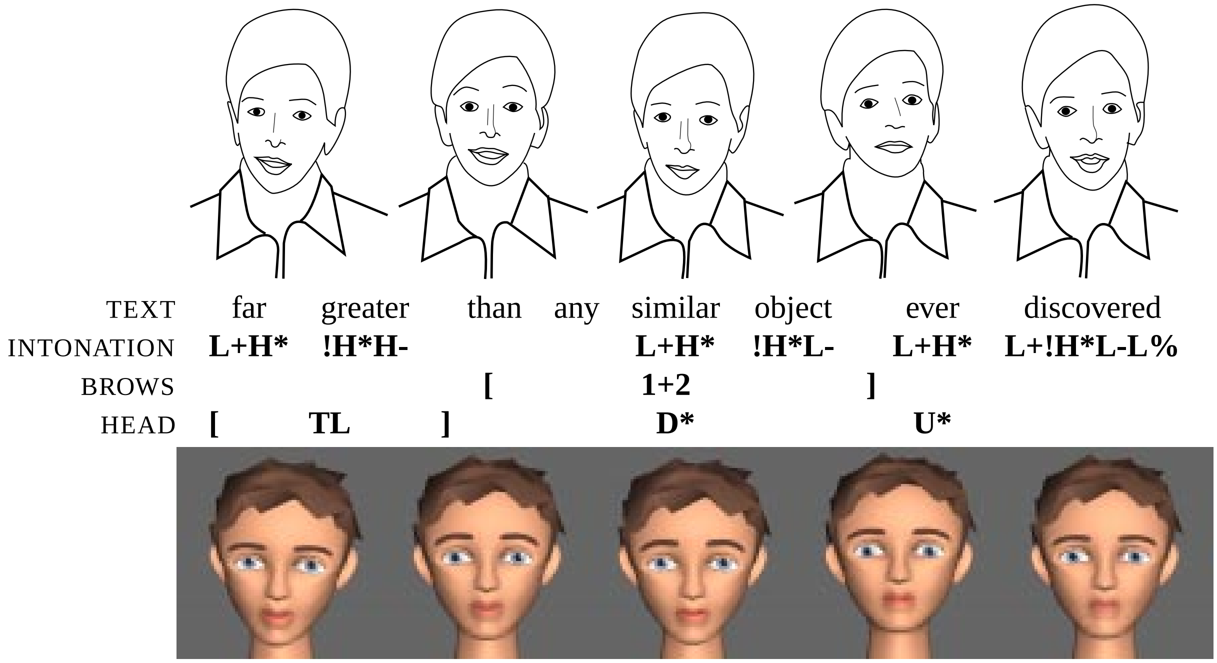
Jennifer J. Venditti
Dpt. of Computer Science, Columbia University
jjv@cs.columbia.edu

July 3, 2003

Abstract

People highlight the intended interpretation of their utterances within a larger discourse by a diverse set of nonverbal signals. These signals represent a key challenge for animated conversational agents because they are pervasive, variable, and need to be coordinated judiciously in an effective contribution to conversation. In this paper, we describe a freely-available cross-platform real-time facial animation system, RUTH, that animates such high-level signals in synchrony with speech and lip movements. RUTH adopts an open, layered architecture in which fine-grained features of the animation can be derived by rule from inferred linguistic structure, allowing us to use RUTH, in conjunction with annotation of observed discourse, to investigate the meaningful high-level elements of conversational facial movement for American English speakers.

Keywords: Facial Animation, Embodied Conversational Agents
Running head: Animating Facial Conversational Signals
This research was supported in part by NSF research instrumentation grant 9818322 and by Rutgers ISATC. Dan DeCarlo drew the original RUTH concept. Radu Gruian and Niki Shah helped with programming; Nathan Folsom-Kovarik and Chris Dymek, with data. Thanks to Scott King for discussion.



TEXT	far	greater	than	any	similar	object	ever	discovered
INTONATION	**L+H***	**!H*H-**			**L+H***	**!H*L-**	**L+H***	**L+!H*L-L%**
BROWS			**[**		**1+2**	**]**		
HEAD	**[**	**TL**	**]**		**D***		**U***	

Figure 1: Natural conversational facial displays (a, top), a high-level symbolic annotation (b, middle), and a RUTH animation synthesized automatically from the annotation (c, bottom).

Introduction

When people communicate, they systematically employ a diverse set of nonverbal cues, and highlight the intended interpretation of their utterances. Consider the example in Figure 1a, the final segment of a brief news story as read by Judy Fortin on CNN headline news in October 2000:

> NASA scientists have spotted something floating in space that's headed our way. But they're not sure if it's an asteroid or part of an old spacecraft. The odds are one in five hundred the unidentified object will collide with Earth—far greater than any similar object ever discovered.

Judy Fortin's expressive movements in Figure 1a include a tilting nod to her left in synchrony with words *far greater* which she utters as a single speech unit; raised eyebrows on the phrase *any similar object*, along with a brief downward nod on *similar*; and an upward (and also slightly rightward) head motion on *ever*. We use the term *facial conversational signals* to refer to movements such as these. In context, these movements link the utterance with the rest of the story. They juxtapose the unidentified object with alternative space objects, emphasize the wide range of objects being considered, and highlight the unidentified object's uniqueness. They thereby call attention to the point of the story—why this possible collision with Earth, an improbable event by ordinary standards, remains newsworthy.

These movements are quite different in character from the interpersonal and affective dimensions that have been investigated in most prior research on conversational facial animation. For example, Cassell and colleagues[1;2] have created agents that use animated head and gaze direction to manage speaking turns in face-to-face conversation. Nagao and Takeuchi[3] and Poggi and Pelachaud[4;5] have created agents that produce specific emblematic displays (that is, complete expressions involving brows, mouth, eyes and head, with a single meaning) to clarify interaction with a user. Animated emotional displays (and corresponding differences in personality) have received even wider attention[6–10]. The movements of Figure 1a do not engage these interpersonal or affective dimensions; they signal internal *semantic* relationships within Judy Fortin's presentation.

Although these signals and their interpretations have not been much studied, we believe that they represent a key challenge for animated conversational agents, because they are so pervasive and so variable. In exploratory data analysis we have found that, as in Figure 1a, small head movements related to discourse structure and interpretation are among the most common nonverbal cues people provide. And Figure 1a already shows three qualitatively different head movements which each suit the synchronous speech.

In this paper, we describe a freely-available cross-platform real-time facial animation system, RUTH (for *Rutgers University Talking Head*), which animates such signals in synchrony with speech and lip movements. RUTH adopts an open, layered architecture in which fine-grained features of the animation can be derived by rule from inferred linguistic structure. RUTH therefore accepts input simply and abstractly, as a compact symbolic description of conversational behavior. Human analysts can produce such specifications for observed data, through the process we refer to as *coding* or *annotation*.

For example, Figure 1b gives a sense of RUTH's input by presenting the annotation that a group of four analysts arrived at in coding the original CNN footage from Figure 1a. The intonation is specified according the *Tones and Break Indices* (ToBI) standard[11;12]; **L+H***, **!H*** and **L+!H*** mark accents on syllables while **H-**, **L-** and **L-L%** record tones at the boundaries of prosodic units. The conversational brow movements are categorized in terms of the *facial action unit* (AU) involved, following Ekman[13]; **1+2** is the action unit for the neutral brow raise. Finally, the head movements are labeled by new categories that we observed frequently in our data: **TL** for a tilting nod on a phrase; **D*** for a downward nod accompanying a single syllable; and **U*** for an upward nod accompanying a single syllable.

The annotation of Figure 1b exhibits a typical parallel between verbal and nonverbal channels: units of motion coincide with units of speech phrasing and peaks of movement coincide with prominent syllables[13–16]. RUTH's animation retains this unity, because RUTH orchestrates the realization of nonverbal signals and speech sounds and movements as part of a single process with access to rich information about language and action. Figure 1c displays still shots from RUTH's rendition of the annotation. The comparison is not that the motions of Fortin and RUTH are identical—the symbolic input that drives RUTH is much too abstract for that—but that the motions are sufficiently alike to *mean* the same.

RUTH implements a pipeline architecture with well-defined interfaces which can link up either with internal modules or external applications. At the lowest level, RUTH animates a schedule of animation instructions for our lifelike character (though not an anatomically realistic one), by applying deformations to a polygonal mesh, in part using a dominance-based coarticulation model [17–19]. A higher level derives a schedule of animation instructions from annotated text, by instrumenting the internal representations of the public-domain speech synthesizer Festival[20] to keep track of synchronous nonverbal events and flesh them out into animation instructions using customizable rules; further utilities help support RUTH's use for dialogue research and in conversational systems. RUTH is available for use in research and education from our web site:

```
http://www.cs.rutgers.edu/˜village/ruth
```

RUTH easily achieves real-time frame rates (i.e., 30 per second or better) on any modern desktop computer with 3D graphics hardware.

RUTH requires annotated input rather than plain text because intonation, facial expressions and head movements can often add something new to the interpretation of an utterance; they are not always redundant. Bavelas and Chovil[21] offer a recent survey of the psychological evidence for such an integrated message model of face-to-face communication. On this view, the independent contribution of facial signals cannot be derived from text (automatically or otherwise); it has to be specified separately. Thus our perspective contrasts with approaches to face animation such as Perlin's[22;23] or Brand's[24], and animated agents such as Smid and Pandzic's[25], where animation is driven from generative statistical models based solely on the text. RUTH's annotated text input enables researchers to experiment with meaningful ways of selecting intonation, facial expressions and head movements to complement simultaneous speech. RUTH is also compatible with text input, of course. For example, RUTH can be used with systems that automatically annotate text for embodied delivery, such as Cassell and colleagues' BEAT system[26]. Alternatively, simple heuristics to annotate text can be quite effective in constrained domains. Nevertheless, human judgments are still necessary to vary the signals of embodied conversation meaningfully.

Implementation

Architecture

The architecture of RUTH is diagramed in Figure 2. The program consists of a tier of independent threads that use queues to coordinate and communicate. The queue implementation enforces mutual exclusion for queue operations, and allows threads waiting on the queue to suspend until the state of the queue changes. This semantics makes the multithreaded implementation of stages in the pipeline simple and elegant.

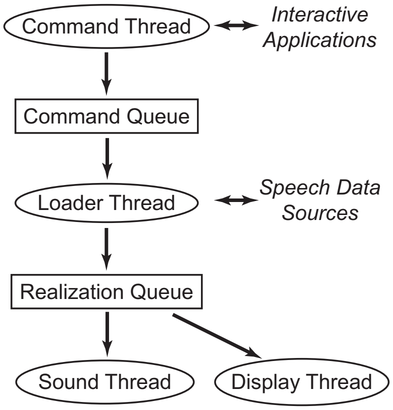


Figure 2: The architecture of RUTH.

The highest-level thread is the *command thread*, which interfaces with interactive applications. The command thread accepts and posts abstract requests for animation, such as to follow a precomputed script, to synthesize speech and control information for a new utterance, or to interrupt an ongoing animation.

Next is the *loader thread*, which supports flexible processing in linking animation with speech data. The loader thread is responsible for populating a realization queue with specific actions to animate at precise times relative to the start of speech. It implements a number of alternative strategies for marshaling the required information, including communication with the Festival speech-synthesis server[20] and access to precomputed data.

Finally, the *display thread* and the *sound thread* coordinate to realize the animation, through careful deployment of operating-systems primitives for concurrency. The display thread updates model geometry and renders frames on a real-time schedule driven by a global animation clock. The sound thread sends data to the audio device in small units (enabling graceful interruption), and monitors the results to keep the playing sound and the animation clock in agreement.

Model

RUTH supports deformable polygonal models. We combine a common underlying geometry of the model with a set of deformations, parameterized from 0 (representing no deformation) to 1, which represent independent qualitative changes to the model. Current deformations describe the mouth movements and tongue movements involved in speech, as in Figure 3; see also Cohen and Massaro [17]. There are also deformations for brow action units **1** (inner raise), **2** (outer raise), and **4** (frowning), smiling and blinking. We apply a deformation by adding offsets to the underlying geometry; the offset is interpolated from key offset values as a piecewise linear function of the deformation parameter. RUTH also permits rotations and translations over parts of the model: the eyes rotate;

Parameter	Effect
rotate jaw	opens the mouth used for low vowels
stretch mouth	tightens the lips common in many visemes
lower corners	gives the lower lip an arched look seen particularly in *p*, *b* and *m*
round upper lip	gives the upper lip a rounded shape seen for example with rounded consonant *w*
raise upper lip	raises lip with less rounding seen for example in *sh*
pout lower lip	brings lower lip forward seen for example in *sh*
lower lower lip	gives the lower lip a rounded look seen for example in *w*
tuck lower lip	draws the lower lip back under the teeth seen particularly with *f*, and *v*
raise tongue	draws the tongue up to the palate seen particularly with *t*, and *d*
stick tongue out	draws the tongue out over and past the teeth seen particularly with *th*

Figure 3: Deformations for visible speech in RUTH.

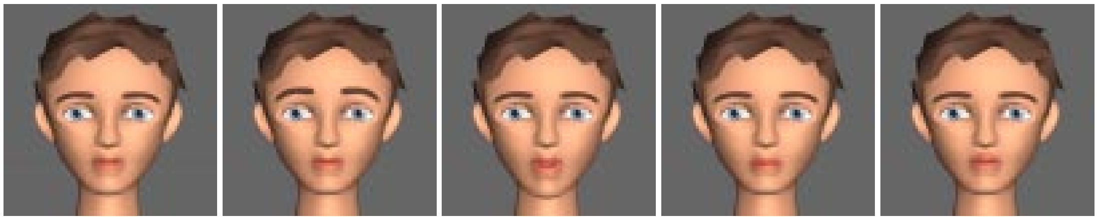


Figure 4: RUTH's underlying geometry; deformations for **1+2**, jaw opening, puckering mouth corners and raising upper lip.

the head rotates and translates, maintaining a smooth join with the neck. At the boundaries of parts, the effect of the transformation fades out gradually across a prespecified region.

Our model and some of its deformations are illustrated in Figure 4. In designing the model, we have adopted the aesthetic of illustration rather than that of photorealism, in order to obtain an attractive and believable result within reasonable computational demands. In all, the model has some 4000 polygons; appearance is determined by varying material properties rather than texture. We have moreover attempted to keep the model relatively ambiguous as to sex, race, and age (e.g. elementary school to young adult); this way, as wide a range of users as possible can regard themselves and RUTH as matched, an important aspect of usability[27].

RUTH implements mouth movements for speech using a dominance-based coarticulation model [17–19]; see King[18] for explanation and further references. The animation schedule specifies *visemes*, categories of facial appearance that correspond to particular categories of speech sounds. Visemes have *goals*, particular parameters for offset deformations at peak; and *dominance functions*, which characterize how visible these deformations are in articulation as a function of time. Deformations that affect the lips (such as smiling) also supply dominance functions which factor into the computation of speech lip-shapes. Mouth offsets in each frame are computed by applying goals for active visemes in relative proportion to their current dominance.

Animation for other facial actions combines a goal with *a parameterized animation template*, which directly describes the degree to which the goal is achieved over time. Individual actions are then specified in terms of start time, end time, peak intensity, attack and decay. Figure 5 shows how we synchronize these parameters with prosodic features in speech. Actions that span prosodic units peak from the start of the first accent in a phrase to the end of the last accent in the phrase; they ramp up gradually at the start of the phrase and fall off gradually at the end. Actions that highlight individual words peak just on an accented syllable. These templates link coarse specifications for conversational actions to concrete animation parameters, and thus underlie RUTH's ability to accept qualitative, symbolic specifications. We offer a higher-level perspective on this synchrony in animation when we describe the use of RUTH later. The geometry that RUTH renders for each

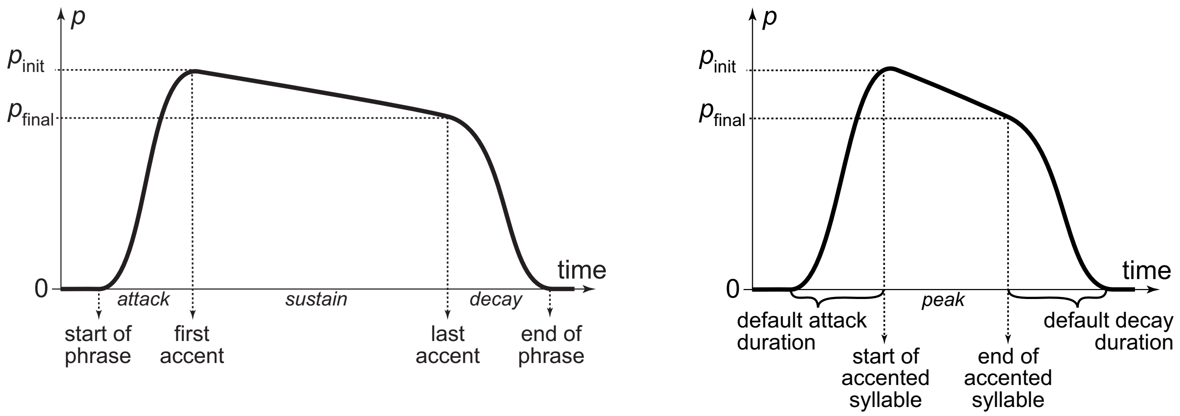


Figure 5: Action synchrony with speech. Underliner actions that synchronize with whole phrases (a, at left). Baton actions that synchronize with stressed syllables (b, at right).

frame of animation adds the computed mouth offsets and the computed action offsets for that time to the underlying geometry of the model.

Interfacing with speech

Keeping track of animation during the process of speech synthesis is a perennial problem. We have instrumented the open-source Festival speech synthesis system[20] so that it synthesizes timing data for speech and animation as an integrated whole. RUTH's loader thread includes a client for the resulting text-to-timed-animated-speech server, and RUTH's command thread accepts a `synthesize` command which instructs the loader to send specific marked-up text to Festival and to animate the results.

Festival represents linguistic structures using general graph representations. Nodes in these graphs correspond to utterance elements, including such constructs as words, phrases, phonemes and tones. A separate graph describes the relationships among elements at each linguistic level; elements can also have arbitrary features, including features that establish links between levels of linguistic analysis. Input utterances are lists of marked-up words; each list element specifies a word and (optionally) a list of attribute-value pairs which specify how the word is to be realized. For example, such attribute-value pairs can specify the prosody with which to realize the utterance. The process of text-to-speech involves repeatedly enriching the linguistic representation of this input, by adding new relationships, elements and features. This process is managed by a fully-customizable flow-of-control in interpreted Scheme. Eventually, this process determines a complete phonetic description of an utterance, including phonemes, pitch, junctures, and pauses and their timing; synthesis is completed by acoustic operations.

Festival's flexible, open architecture meshes naturally with the requirements of animation. We

```
((far       ((register "HL") (accent "L+H*")(jog "TR")))
 (greater   ((accent "!H*") (tone "H-")     (jog)))
 (than      ((register "HL-H")              (brow "1+2")))
 (any       (                               ))
 (similar   ((accent "L+H*")                (jog "D*")))
 (object    ((accent "!H*") (tone "L-")     (brow)))
 (ever      ((register "L") (accent "L+H*") (jog "U*")))
 (discovered((accent "L+!H*") (tone "L-L%"))))
```

Figure 6: Tagged speech input to Festival corresponding to Figure 1b; files use `jog` for head motions and single tags (e.g. `(jog)`) to signal ends of movements.

specify Festival input with features on words for head and brow actions as we have coded them. Figure 6 gives an example of such input. We add rules for timing these actions to Festival's text-to-speech process. Because of Festival's design, these rules can draw on structural and phonetic considerations in the utterance (as in Figure 5) by exploring its final phonetic description. We can also customize remaining quantitative parameters for specific animation actions. We add a final traversal of utterance's phonetic representation so that the server can output a series of visemes and animation commands corresponding to a synthesized waveform. For RUTH, we have also reinstrumented Festival (debugging and extending the standard release) to control pitch by annotation [28;29]; we use OGI CSLU synthesis and voices[30].

Animation schedules and speech waveforms output by Festival can be saved, reused and modified directly. This makes it easy to visualize low-level variations in timing and motion. (In the command thread, a `save` instruction constructs files for input that will reproduce the most-recently realized animation; the `canned` instruction replays the animation from a specified file.) We also support similar visualizations involving recorded speech, drawing on off-the-shelf tools to put waveforms in temporal correspondence with their transcripts and to annotate the results.

Driving RUTH with annotated text

The most abstract way to specify an animation for RUTH is to supply RUTH with text that has been marked-up to specify the head motions and other facial actions that should occur as the text is uttered. This section describes the range of delivery that RUTH supports and gives some hints about how to use RUTH's animation capabilities in the most meaningful way.

RUTH *input and its motivation*

To specify prosody, RUTH uses the Tones and Break Indices (ToBI) model of English intonation [11;12]. In ToBI, prosodic structure is described in terms of *phrasing*, clustering of words into groups delimited by perceived disjuncture, and *accentuation*, the perceived prominence of particular syllables within a group of words. Intonational tune is specified by symbolic annotations that describe the qualitative behavior of pitch at accents and phrasal boundaries. In the ToBI labeling, each utterance is required to consist of one or more phrases. Each phrase must end with appropriate phrase or boundary markers, and each phrase must contain at least one accented word.

The English tonal inventory includes pitch accents such as high (**H***), low (**L***), or rising accents that differ in whether the rise precedes (**L+H***) or follows (**L*+H**) the stressed syllable. Accents with a high tonal component are generally realized high in the speaker's pitch range for the phrase, but can sometimes be downstepped (annotated by **!** as in **!H***) to a lower pitch value (and lower prominence). Pitch accents are specified to RUTH as values of a word's `accent` attribute.

Words are grouped into two hierarchical levels of prosodic phrasing in English: the smaller intermediate phrase and the larger intonation phrase. An intermediate phrase is marked by a high (**H-**) or low (**L-**) tone immediately after the last accented syllable in the phrase, and an intonation phrase is additionally marked by a high (**H%**) or low (**L%**) tone at the right phrase edge. Common patterns for intonation phrases thus include the fall often found in declarative statements **L-L%**, the rise often found in yes-no questions **H-H%**, and a combined fall-rise **L-H%** associated generally with contributions to discourse that are somehow incomplete. Phrase and boundary tones are specified to RUTH as values of the `tone` attribute, which accompanies the final word in a phrase.

ToBI offers sophisticated resources for characterizing the pitch contour of English utterances, in terms that correlate closely with the meanings that prosodic variation can convey in particular discourse contexts; see Pierrehumbert and Hirschberg[31]. Researchers can call upon these resources in deciding to realize embodied utterances with suitable intonation. However, rich variation is not always necessary; for example, it works quite well to just put an **H*** on content words that have not been used before in the discourse[32], and to put an **L-** or **L-L%** at natural boundaries, after every few content words. These strategies offer a simple alternative for preparing specifications for RUTH by hand, or for writing algorithms that construct them automatically.

Another important aspect of English prosody is *pitch range*, the extremes of high and low that are attained over a whole phrase. This is also known as the register of speech. ToBI labels describe the qualitative changes in pitch with respect to whatever pitch range happens to be in effect. But changes in overall pitch range help to signal the organization of discourse: at the beginnings of discourse segments, pitch range is expanded and at the ends of discourse segments pitch range is contracted and generally lowered[33]. In addition, the general level of pitch is a signal of a speaker's involvement in what they say: more important contributions are delivered with higher pitch[34]. Varying pitch range is thus essential to give the variability and organization of natural speech.

RUTH's convention is that a `register` attribute on the first word of a phrase sets the pitch range for the whole phrase to one of a few qualitative values. (The convention applies for all intermediate phrases, not just intonation phrases.) RUTH's qualitative values, as given in Figure 7, are derived from the work of Moehler and Mayer[29;35].

`"H"`	primary high register (default)
`"H-H"`	expanded high register
`"H-L"`	compressed high register
`"L"`	primary low register
`"L-L"`	compressed low register
`"HL"`	expanded register including lows and highs
`"HL-H"`	full pitch range

Figure 7: Possible specifications of pitch range for RUTH[29;35].

RUTH's models of facial conversational signals build on this specification of prosody. Our new movements may function as *underliners* that accompany several successive words, or as *batons* that highlight a single word[13]. In calculating the temporal dynamics of underliners and batons, RUTH builds from the close synchrony that researchers[13–16] have found between embodied action and simultaneous speech in conversation. We anticipated this already in discussing Figure 5. RUTH assumes that underliners span complete intermediate or intonation phrases. This allows RUTH to ensure automatically that the movement appears to peak in synchrony with the first prosodic emphasis in a phrase and to be released after the last prosodic emphasis in a phrase. Similarly, RUTH assumes that batons only occur on words that are specified for accent, and times the peak of the baton to synchronize with the stressed vowel.

Aligning conversational facial signals with speech this way can help to settle difficult annotation decisions in a principled way. It is quite difficult to annotate beginnings and ends of brow movements, for example by looking at a video record of a conversation. The typical difficulty is judging where a movement starts or ends within a series of short unaccented words. Figure 1 is representative: the phrase *than any similar object* begins and ends with unstressed syllables. Coders who have to choose separately whether to include *than* or *any* as marked with a brow raise face a difficult and probably meaningless judgment.

In RUTH's input, a separate attribute of words controls each independent dimension of facial movement. For each attribute, RUTH permits at most one underliner and at most one baton at a time; a labeled word either marks the beginning or the end of an underliner or carries a baton. RUTH adopts the convention that baton labels end in `*`, while corresponding underliner labels omit the `*`.

RUTH follows Ekman in classifying brow movements in terms of the *facial action unit* (AU) involved; AUs are patterns of change in the face that trained experts can code and sometimes even perform reliably[13]. Brow movements are made up of AU **1**, which raises the inside of the brow; AU **2**, which raises the outside of the brow; and AU **4**, which narrows and depresses the brow. RUTH currently implements a neutral raise, specified as values `"1+2"` or `"1+2*"` for the attribute brow, and a neutral frown, specified as values "4" or "4*". RUTH's smile is specified with an attribute smile, and may be used as an underliner "S" or baton "S*".

RUTH allows general head movements as facial conversational signals. The head can nod up and down, rotate horizontally left and right and tilt at the neck from side to side; it can also be translated front-to-back and side-to-side through motion at the neck. Like brow movements, these actions may get their meanings individually or in combination; they may synchronize with individual words, giving Ekman's batons or Hadar et al.'s *rapid movements*[36], or they may synchronize with larger phrases, giving Ekman's underliners or Hadar et al.'s *ordinary movements*. Head movements are specified using values of the attribute jog.

No standard symbolic coding of head movements exists. We have developed our own, drawing on our preliminary analysis of videotaped embodied utterances and informal observations of everyday conversation. The labels for head movements that we currently support are given in Figure 8, together with some rough speculations about the functions that these different movements might carry. We emphasize that this inventory is provisional; categorizing the movements that accompany conversational speech and accounting for their function remains an important problem for future research. At least two further steps are required to validate a system like that suggested in Figure 8. Empirical research must show that the categories fit observed conversation across a range of individuals across a range of contexts. And empirical research must confirm that interlocutors also are sensitive to the differences among categories. Such effort is proceeding; see Krahmer and colleagues[37;38] for example.

Finally, RUTH will synchronize a blink just at the end of an accented vowel when the word carries the simple attribute (blink).

Using RUTH *in applications*

As simple illustrations of the use of RUTH, we have implemented two applications: a version of Weizenbaum's famous Eliza program[39] which outputs specifications for animated speech; and a demonstration of conversational feedback that animates RUTH performing an indefinite sequence of randomized acknowledgment behaviors: nods, brow raises, and noises like "mm-hmm" and "uh-huh". Both programs are available as part of the standard RUTH release; see also Stone and DeCarlo[40]. The programs share a convenient overall architecture that a system-builder can use to add animated output to an existing application—piping the output of an ordinary interactive system as input to a RUTH process running in parallel. (The Eliza program also prints out each command before sending it to RUTH so you can see exactly what the input is to the animation.)

Our Eliza illustrates some convenient heuristics for annotating plain text to send it to RUTH. Like all Eliza systems, the meat of the program is a series of condition-response rules that describe possible responses that the system could give. (Our animated version of Eliza extends a text implementation realized as a Perl script by Jon Fernquist but modeled on a Lisp version of Eliza

Value	Effect and possible use
`"D"`	nods downward general indicator of emphasis
`"U"`	nods upward perhaps indicates a “wider perspective”
`"F"`	brings the whole head forward perhaps indicates need for “a closer look”
`"B"`	brings the whole head backward perhaps emblem of being “taken aback”
`"R"`	turns to model’s right perhaps indicates availability of more information
`"L"`	turns to model’s left perhaps indicates availability of more information
`"J"`	tilts whole head clockwise (around nose) perhaps indicates expectation of engagement from partner
`"C"`	tilts whole head counterclockwise perhaps indicates expectation of engagement from partner
`"DR"`	nods downward with some rightward movement meaning seems to combine that of D and R
`"UR"`	nods upward with some rightward movement meaning seems to combine that of U and R
`"DL"`	nods downward with some leftward movement meaning seems to combine that of D and L
`"UL"`	nods upward with some leftward movement meaning seems to combine that of U and L
`"TL"`	tilts clockwise with downward nodding perhaps indicates contrast of related topics
`"TR"`	tilts counterclockwise with downward nodding perhaps indicates contrast of related topics

Figure 8: Possible head movement (jog) codes in RUTH.

described by Norvig[41].) The condition looks for a specified sequence of words in the user's utterance, and records all the words following the matched sequence. The response is a text template for an animated utterance and can include a position where the recorded words from the user's utterance can be copied and presented back to the user, perhaps with new intonation or facial displays. To mark-up the user's utterances for prosody, templates can invoke procedures that realize it without accents, realize it just with a single accent on the final content word, or realize it with accents on all content words.

Our feedback demonstration illustrates low-level interaction with RUTH; it creates instructions for animations on-the-fly. To create the `feedback` application, we recorded and digitized a number of samples of acknowledgment sounds, and logged when the sound started, when the sound reached its peak intensity, and when the sound finished. We also took note whether the sound should be animated with the mouth closed (like "mm-hmm") or with the mouth open (like "uh-huh"), and whether the sound offers positive feedback, expressing understanding, or negative feedback, expressing confusion. Every few seconds, the feedback program wakes up and instructs RUTH to play one of the sound files and a new animation timing file that goes with it, including a randomized selection of actions—blinking, the right mouth shapes to go with whatever sound file is being played, perhaps a head jog, and perhaps a brow action.

Discussion

Conversation brings motions and requirements beyond the the lip-synch and emotional expression emphasized in such prior models as Cohen and Massaro's[17] and King's[18]. But more general models, defined in terms of musculature[42;43] or simulation[44], introduce complications that can stand in the way of real-time performance and easy customization. We have constructed a new alternative, RUTH, by organizing the design and implementation of a face animation system around the investigation of conversational signals.

In particular, RUTH is designed with *coding* in mind; RUTH accepts text with open-ended annotations specifying head motions and other facial actions, and permits the flexible realization of these schedules. Many applications demand coding. In autonomous conversational agents, for example, a rich intermediate language between the utterance generation system and the animation system helps organize decisions about what meaning to convey and how to realize meaning in animation. (See the work of Cassell and colleagues[45] on generating meaningful hand gestures and coordinating them with other communicative actions[46].) RUTH still lacks many meaningful expressions, including emblems of emotion such as disgust and emblems of thought such as pursing the lips. However, the facial signals of prior agents[26;47;48] are just eyebrow movements and are planned independently of other communicative decisions; so RUTH already makes it easier to take the next steps.

Likewise, in developing and testing *psycholinguistic theories* of conversation, predictable, rulegoverned realization of abstract descriptions makes computer animation an important methodological tool[45;47;49]. Coding-based animation systems allow analysts to visualize descriptions of observed events, so that analysts can obtain a more specific feel for alternative models. Coding-based systems can also generalize away from observations arbitrarily, so that analysts can, for example, explore anomalous behaviors which might be very difficult or impossible to get from people (or statistical models fit to people). The same flexibility and control makes coding-based animation a natural ingredient of empirical studies of perception; Massaro and colleagues' explorations of human speech perception that use mismatched sound and animation are the classic example[49]. Krahmer and colleagues are conducting psycholinguistic studies of conversational brow movements using coding-based animation[37].

In formulating RUTH's input as this abstract, meaningful layer, we do not discount the importance of quantitative variables in conversational agents. We simply assume that range of movement and other quantitative aspects of motion do not contribute to the symbolic interpretation of discourse. Rather, they provide quantitative evidence for speaker variables such as involvement and affect. This is already the norm for intonation, where Ladd[34] presents evidence (and Cahn[50] provides an implementation) linking perceived emotion to pitch range and voice quality of speech; and for manual gesture, where Chi and colleagues[51] model the emotional variables that quantitatively modulate symbolic action. Badler and colleagues[52;53] are exploring a similar approach to modulate facial animation. Integrating such modality-independent specifications of affect and personality with conversational signals for discourse remains important future work for facial animation. To this end, we are extending RUTH so that planned motions can undergo probabilistic transformations, as in Perlin's work[22;23], so as to achieve greater variability within RUTH's coding-based framework.

With the surge of interest in interfaces that engage in natural embodied conversation, as seen in recent surveys of embodied conversational agents[54], we expect that RUTH will provide a helpful resource for the scientific community. In particular, most embodied conversational agents create abstract schedules for animation that need to be realized; RUTH naturally fits into such an architecture and enhances its functionality. Nor is there any obstacle, at least in principle, to integrating the insights of RUTH's design and architecture into other frameworks and animation systems.

References

[1] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. Embodiment in conversational characters: Rea. In *Proceedings of CHI*, pages 520–527, 1999.

[2] J. Cassell and K. Thórisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(3), 1999.

[3] Katashi Nagao and Akikazu Takeuchi. Speech dialogue with facial displays: Multimodal human-computer conversation. In *Proceedings of ACL*, pages 102–109, 1994.

[4] Isabella Poggi and Catherine Pelachaud. Performative facial expressions in animated faces. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors, *Embodied Conversational Agents*, pages 155–188. MIT, 2000.

[5] Catherine Pelachaud and Isabella Poggi. Subtleties of facial expressions in embodied agents. *Journal of Visualization and Computer Animation*, 13(5):301–312, 2002.

[6] Elisabeth André, Thomas Rist, Susanne van Mulken, Martin Klesen, and Stephan Baldes. The automated design of believable dialogues for animated presentation teams. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors, *Embodied Conversational Agents*, pages 220–255. MIT, 2000.

[7] Gene Ball and Jack Breese. Emotion and personality in a conversational agent. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors, *Embodied Conversational Agents*, pages 189–219. MIT, 2000.

[8] Berardina De Carolis, Catherine Pelachaud, Isabella Poggi, and Fiorella de Rosis. Behavior planning for a reflexive agent. In *Proceedings of IJCAI*, 2001.

[9] James C. Lester, Stuart G. Towns, Charles B. Callaway, Jennifer L. Voerman, and Patrick J. FitzGerald. Deictic and emotive communication in animated pedagogical agents. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors, *Embodied Conversational Agents*, pages 123–154. MIT, 2000.

[10] S. C. Marsella. Sympathy for the agent: Controlling an agent's nonverbal repertoire. In *Proceedings of Agents*, 2000.

[11] Mary Beckman and Gayle Ayers Elam. Guidelines for ToBI labelling, version 3.0. Technical report, Ohio State University, 1997. http://ling.ohio-state.edu/Phonetics/etobi_homepage.html.

[12] Kim E. A. Silverman, Mary Beckman, John F. Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, and Janet Pierrehumbert. ToBI: a standard for labeling English prosody. In *Proceedings of the International Conference on Spoken Language Processing*, pages 867–870, 1992.

[13] Paul Ekman. About brows: Emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog, editors, *Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium*, pages 169–202. Cambridge, 1979.

[14] David McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992.

[15] Peter Bull and Gerry Connelly. Body movement and emphasis in speech. *Journal of Nonverbal Behavior*, 9(3):169–187, 1985.

[16] Randi A. Engle. *Toward a Theory of Multimodal Communication: Combining Speech, Gestures, Diagrams and Demonstrations in Instructional Explanations*. PhD thesis, Stanford University, 2000.

[17] Michael M. Cohen and Dominic W. Massaro. Modeling coarticulation in synthetic visual speech. In N. M. Thalmann and D. Thalmann, editors, *Models and Techniques in Computer Animation*, pages 139–156. Springer, 1993.

[18] Scott A. King. *A facial model and animation techniques for animated speech*. PhD thesis, The Ohio State University, 2001.

[19] A. Löfqvist. Speech as audible gestures. In W. J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modeling*, pages 289–322. Kluwer, 1990.

[20] Alan Black and Paul Taylor. Festival speech synthesis system. Technical Report HCRC/TR-83, Human Communication Research Center, 1997.

[21] Janet Beavin Bavelas and Nicole Chovil. 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.

[22] Ken Perlin. Layered compositioning of facial expression. In *SIGGRAPH*, 1997. Technical Sketch.

[23] Ken Perlin. Noise, hypertexture, antialiasing and gestures. In D. Ebert, editor, *Texturing and Modeling: A Procedural Approach, Second Edition*, pages 209–274. Academic Press, 1998.

[24] Matthew Brand. Voice puppetry. In *SIGGRAPH*, pages 21–28, 1999.

[25] Karlo Smid and Igor S. Pandzic. Conversational virtual character for the web. In *Computer Animation*, pages 240–247, 2002.

[26] Justine Cassell, Hannes Vilhjálmsson, and Tim Bickmore. BEAT: the behavioral expression animation toolkit. In *SIGGRAPH*, pages 477–486, 2001.

[27] Clifford Nass, Katherine Isbister, and Eun-Ju Lee. Truth is beauty: Resarching embodied conversational agents. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors, *Embodied Conversational Agents*, pages 374–402. MIT, 2000.

[28] Matthias Jilka, Gregor Möhler, and Grzegorz Dogil. Rules for the generation of ToBI-based American English intonation. *Speech Communication*, 28:83–108, 1999.

[29] Gregor Möhler and Jörg Mayer. A discourse model for pitch-range control. In *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, 2001.

[30] Mike Macon, Andrew Cronk, Alexander Kain, and Johan Wouters. OGIresLPC: diphone synthesiser using residual linear prediction coding. Technical Report CSE-97-007, Oregon Graduate Institute, 1997.

[31] Janet Pierrehumbert and Julia Hirschberg. The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in Communication*, pages 271–311. MIT Press, Cambridge MA, 1990.

[32] Julia Hirschberg. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63(1–2):305–340, 1993.

[33] Julia Hirschberg and Christine Nakatani. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of ACL*, 1996.

[34] D. R. Ladd, K. Silverman, F. Tolkmitt, G. Bergmann, and K. Scherer. Evidence for the independent function of intonation contour type, voice quality and F0 range in signaling speaker affect. *Journal of the Acoustical Society of America*, 78:435–444, 1985.

[35] Gregor Möhler and Jörg Mayer. A method for the analysis of prosodic registers. In *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, 1999.

[36] U. Hadar, T. J. Steiner, E. C. Grant, and F. Clifford Rose. Head movement correlates of juncture and stress at sentence level. *Language and Speech*, 26(2):117–129, 1983.

[37] Emiel Krahmer, Zsófia Ruttkay, Marc Swerts, and Wieger Wesselink. Pitch, eyebrows and the perception of focus. In *Symposium on Speech Prosody*, 2002.

[38] Emiel Krahmer, Zsófia Ruttkay, Marc Swerts, and Wieger Wesselink. Audiovisual cues to prominence. In *International Conference on Spoken Lanugage Processing*, 2002.

[39] Joseph Weizenbaum. ELIZA—a computer program for the study of natural language communication between man and machine. *Communcations of the ACM*, 9(1):36–45, 1966.

[40] Matthew Stone and Douglas DeCarlo. Crafting the illusion of meaning: template-based specification of embodied conversational behavior. In *Computer Animation and Social Agents*, 2003.

[41] Peter Norvig. *Paradigms of Artificial Intelligence: Case Studies in Common Lisp*. Morgan Kaufmann, 1992.

[42] Stephen M. Platt. *A structural model of the human face*. PhD thesis, University of Pennsylvania, 1985.

[43] Keith Waters. A muscle model for animating three-dimensional facial expressions. *Computer Graphics*, 21(4):17–24, 1987.

[44] Demetri Terzopoulos and Keith Waters. Physically-based facial modeling, analysis and animation. *Journal of Visualization and Computer Animation*, 1(2):73–80, 1990.

[45] Justine Cassell, Matthew Stone, Brett Douville, Scott Prevost, Brett Achorn, Mark Steedman, Norm Badler, and Catherine Pelachaud. Modeling the interaction between speech and gesture. In *Proceedings of the Cognitive Science Society*, 1994.

[46] Justine Cassell, Matthew Stone, and Hao Yan. Coordination and context-dependence in the generation of embodied conversation. In *First International Conference on Natural Language Generation*, pages 171–178, 2000.

[47] Catherine Pelachaud, Norm Badler, and Mark Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1):1–46, 1996.

[48] Isabella Poggi and Catherine Pelachaud. Eye communication in a conversational 3D synthetic agent. *AI Communications*, 13(3):169–181, 2000.

[49] Dominic W. Massaro. *Perceiving Talking Faces: From speech perception to a behavioral principle*. MIT, 1998.

[50] Janet E. Cahn. The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8:1–19, 1990.

[51] Diane Chi, Monica Costa, Liwei Zhao, and Norm Badler. The EMOTE model for effort and shape. In *SIGGRAPH*, pages 173–182, 2000.

[52] Norman Badler, Jan Allbeck, Liwei Zhao, and Meeran Byun. Representing and parameterizing agent behaviors. In *Computer Animation*, pages 133–143, 2002.

[53] Meeran Byun and Norman Badler. FacEMOTE: qualitative parametric modifiers for facial animations. In *ACM SIGGRAPH Symposium on Computer Animation*, pages 65–71, 2002.

[54] Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors. *Embodied Conversational Agents*. MIT, 2000.