Appears in Computer Animation 2002

Making Discourse Visible: Coding and Animating Conversational Facial Displays

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

Jennifer J. Venditti
Institute for Research in Cognitive Science, University of Pennsylvania
jjv@unagi.cis.upenn.edu

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.*

1. 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:

(1) 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 toward the right of the frame in synchrony with the prosodic unit *far greater*; raised eyebrows on the prosodic unit *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 [12, 8] have created agents that use animated head and gaze direction to manage speaking turns in face-to-face conversation. Nagao and Takeuchi [27] and Poggi and Pelachaud and colleagues [33] 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 [1, 2, 21, 24, 7]. 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, be-

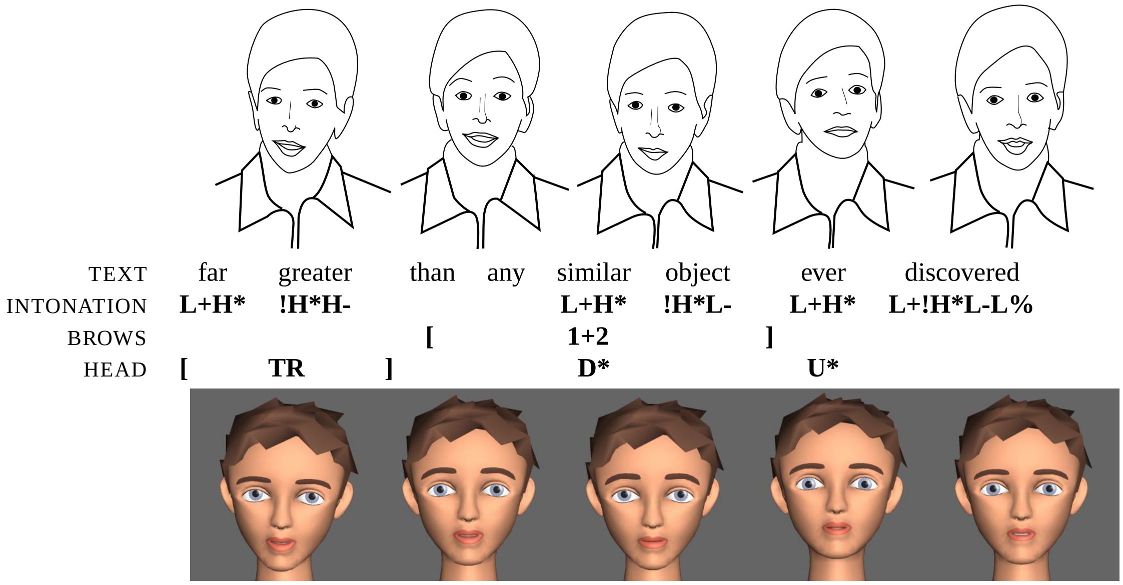
TEXT far greater than any similar object ever discovered
INTONATION L+H* !H*H- L+H* !H*L- L+H* L+!H*L-L%
BROWS [1+2]
HEAD [TR] D* U*

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

cause 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*; human judgments remain necessary where meaning provides an important aid in classifying behavior.

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 [34, 3]; **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 [16]; **1+2** is the action unit for the neutral brow raise. Finally, the head movements are labeled by categories that we observed frequently in our data: **TR** for the 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 prosodic phrasing and peaks of movement coincide with prominent syllables. 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.

We return to such issues of action and meaning in animated conversational agents in Section 5, but first we describe the design and implementation of RUTH more fully. RUTH implements a pipeline architecture with well-defined interfaces, described in Section 2, which supports flexible deployment by allowing for information at higher stages to be provided by internal modules or external applications. At the lowest level (Section 3), RUTH animates a schedule of animation instructions for our lifelike character (though not an anatomically realis-

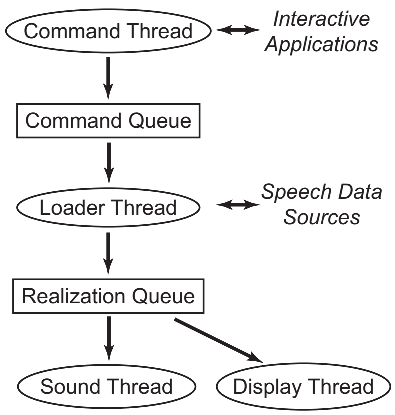

Figure 2. The architecture of RUTH.

tic one), by applying deformations to a polygonal mesh, in part using a coarticulation model in the tradition of [22, 15, 18]. A higher level (Section 4) derives a schedule of animation instructions from annotated text, by instrumenting the internal representations of the public-domain speech synthesizer Festival [4] 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 achieves frame rates of 30 per second or better on Solaris Ultra 10s with Elite 3D graphics, or Pentium III PCs with good graphics cards.

2. Architecture

The architecture of RUTH is diagrammed 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.

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 from scratch, 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 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.

3. 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 six mouth movements and two tongue movements involved in speech [15], 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. We also permit support-mapped rotations and translations over parts of the model: the eyes rotate; the head rotates and translates, maintaining a smooth join with the neck.

Our model and some of its deformations are illustrated in Figure 3. In designing the model, we have adopted the esthetic 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 [28].

RUTH implements mouth movements for speech using a coarticulation model in the tradition of [22, 15, 18]; see explanation and references in [18]. 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

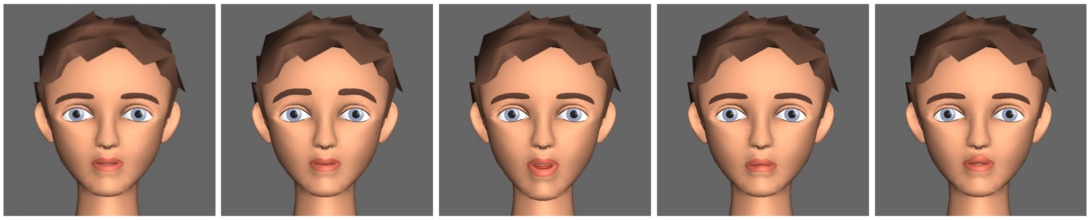

Figure 3. RUTH**'s underlying geometry; deformations for 1+2, jaw opening, puckering mouth corners and raising upper lip.**

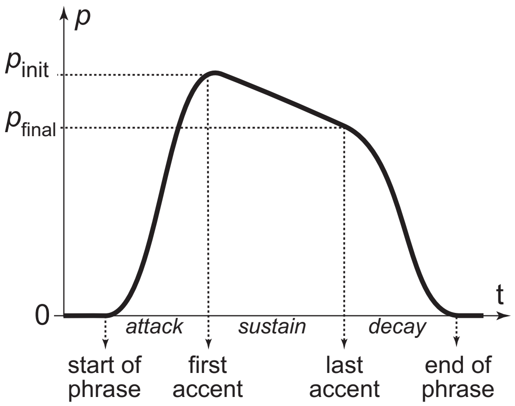

Figure 4. Action synchrony with speech

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 4 shows how we synchronize these parameters with speech. The final geometry adds the action offsets and computed mouth offsets to the underlying geometry.

4. 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 [4] 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.

Festival represents linguistic structures using general graph representations. 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. The process of text-to-speech involves repeatedly enriching the linguistic representation of input, by adding new relationships, elements and features. This process is managed by a fully-customizable flow-of-control in scheme. Eventually, this process determines a complete phonetic description of an utterance, including phones, 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 specify Festival input with features on words for head and brow actions as we have coded them. Figure 5 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 4) 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 [17, 26]; we use OGI CSLU synthesis and voices [23].

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. 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.

5. Discussion

Conversation brings motions and requirements beyond the the lip-synch and emotional expression em-

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

Figure 5. 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.

phasized in such prior models as Cohen and Massaro's [15] and King's [18]. But more general models, defined in terms of musculature [31, 36] or simulation, [35] introduce complications that 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. This also contrasts with approaches such as Perlin's [30] or Brand's [5] that animate speech merely by applying generative statistical models. 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 on generating meaningful hand gestures [9] and coordinating them with other communicative actions [10].) In fact, the facial signals of prior agents [29, 32, 13] are just eyebrow movements and are planned independently of other communicative decisions; RUTH should make it easier to take the next steps.

Likewise, in developing and testing *psycholinguistic theories* of conversation, predictable, rule-governed realization of abstract descriptions makes computer animation an important methodological tool [9, 29, 25]. 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 [25]. Krahmer and colleagues are conducting psycholinguistic studies of conversational brow movements using coding-based animation [19].

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 [20] presents evidence (and [6] provides an implementation) linking perceived emotion to pitch range and voice quality of speech; and for manual gesture, where Chi and colleagues model the emotional variables that quantitatively modulate symbolic action [14]. Combining a symbolic specification of discourse with complementary specifications of affect and personality that are realized across modalities remains important future work for facial animation. To this end, we are extending RUTH so that planned motions can undergo probabilistic transformations (as in [30]), 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 for example in [11], we expect that RUTH will provide an important resource for the scientific community. In particular, most of the end-to-end systems described in [11] 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.

6. Acknowledgments

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.

References

[1] E. André, T. Rist, S. van Mulken, M. Klesen, and S. Baldes. The automated design of believable dialogues for animated presentation teams. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, pages 220–255. MIT, 2000.

[2] G. Ball and J. Breese. Emotion and personality in a conversational agent. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, pages 189–219. MIT, 2000.

[3] M. Beckman and G. A. Elam. Guidelines for ToBI labelling, version 3.0. Technical report, Ohio State University, 1997. http://ling.ohio-state.edu/Phonetics/etobi_homepage.html.

[4] A. Black and P. Taylor. Festival speech synthesis system. Technical Report HCRC/TR-83, Human Communication Research Center, 1997.

[5] M. Brand. Voice puppetry. In *SIGGRAPH*, pages 21–28, 1999.

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

[7] B. D. Carolis, C. Pelachaud, I. Poggi, and F. de Rosis. Behavior planning for a reflexive agent. In *IJCAI*, 2001.

[8] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. Embodiment in conversational characters: Rea. In *CHI 99*, pages 520–527, 1999.

[9] J. Cassell, M. Stone, B. Douville, S. Prevost, B. Achorn, M. Steedman, N. Badler, and C. Pelachaud. Modeling the interaction between speech and gesture. In *Proceedings of the Cognitive Science Society*, 1994.

[10] J. Cassell, M. Stone, and H. Yan. Coordination and context-dependence in the generation of embodied conversation. In *First International Confernence on Natural Language Generation*, pages 171–178, 2000.

[11] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors. *Embodied Conversational Agents*. MIT, 2000.

[12] 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.

[13] J. Cassell, H. Vilhjálmsson, and T. Bickmore. BEAT: the behavioral expression animation toolkit. In *SIGGRAPH*, pages 477–486, 2001.

[14] D. Chi, M. Costa, L. Zhao, and N. Badler. The EMOTE model for effort and shape. In *SIGGRAPH*, pages 173–182, 2000.

[15] M. M. Cohen and D. 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.

[16] P. 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 University Press, Cambridge, 1979.

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

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

[19] E. Krahmer, Z. Ruttkay, M. Swerts, and W. Wesselink. Pitch, eyebrows and the perception of focus. In *Symposium on Speech Prosody*, 2002.

[20] 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.

[21] J. C. Lester, S. G. Towns, C. B. Callaway, J. L. Voerman, and P. J. FitzGerald. Deictic and emotive communication in animated pedagogical agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, pages 123–154. MIT, 2000.

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

[23] M. Macon, A. Cronk, A. Kain, and J. Wouters. OGIresLPC: diphone synthesiser using residual linear prediction coding. Technical Report CSE-97-007, Oregon Graduate Institute, 1997.

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

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

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

[27] K. Nagao and A. Takeuchi. Speech dialogue with facial displays: Multimodal human-computer conversation. In *Proceedings of ACL 32*, pages 102–109, 1994.

[28] C. Nass, K. Isbister, and E.-J. Lee. Truth is beauty: Researching embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, pages 374–402. MIT, 2000.

[29] C. Pelachaud, N. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1):1–46, 1996.

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

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

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

[33] I. Poggi and C. Pelachaud. Performative facial expressions in animated faces. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, pages 155–188. MIT, 2000.

[34] K. E. A. Silverman, M. Beckman, J. F. Pitrelli, M. Ostendorf, C. Wightman, P. Price, and J. Pierrehumbert. ToBI: a standard for labeling English prosody. In *Proceedings of the International Conference on Spoken Language Processing*, pages 867–870, 1992.

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

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