

Crafting the Illusion of Meaning: Template-based Specification of Embodied Conversational Behavior

Matthew Stone and Doug DeCarlo
Computer Science and Cognitive Science, Rutgers, The State University of New Jersey
{mdstone, decarlo}@cs.rutgers.edu

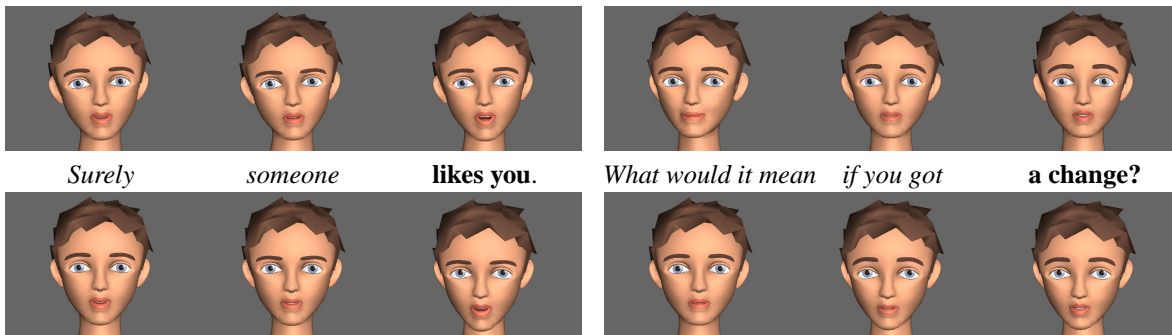


Figure 1. Instantiating fixed templates (italics) by additional words (bold). Marking-up templates and fillers for synchronized nonverbal signals can result in meaningfully different animations.

Abstract

Templates are a widespread natural language technology that achieves believability within a narrow range of interaction and coverage. We consider templates for embodied conversational behavior. Such templates combine a specific pattern of marked-up text, specifying prosody and conversational signals as well as words, with similarly-annotated gaps that can be filled in by rule to yield a coherent contribution to a dialogue with a user. In this paper we argue that templates can give a designer substantial freedom to realize specific combinations of behaviors in interactions with users and thereby to explore the relationships among such factors as emotion, personality, individuality and social role.

1 Introduction

Embodied conversational agents bring unique opportunities for delivering users new kinds of interaction with computer systems. These agents can support the conversational functions and behaviors that characterize human face-to-face communication [5]. They can enrich interaction, by giving systems visible agency in virtual

worlds [27] and even in shared spaces [6]. And by engaging meaningfully with users' social sense and expectations, they endow interaction with a new dimension—one that not only may bring advantages for usability [7], but that brings new possibilities for entertainment and perhaps even art [2].

Despite the wide-ranging research in embodied conversational agents currently underway [9], we are still accumulating tools and methodologies that allow us to prototype new embodied conversational agents. This paper aims to add a new element to our repertoire: template-based specifications of embodied conversational behavior.

Templates are the most common form of natural language generation technology, particularly in text applications. Templates are so successful because it is often much easier simply to make lists of the kinds of sentences a system needs, instead of constructing a general theory of application language. Typically, a template combines a static sequence of words with gaps that are filled in by rule. Different sentences can be constructed from the template by filling in these gaps conditionally based on available data. In more sophisticated systems, text can be constructed from templates recursively, and the system can choose which templates to expand condi-

tionally based on the context [34]. Reiter [25, 26] offers a good introduction to the motivation and functionality of templates in applied systems for language interaction; Theune [32] specifically considers template generation in the context of spoken dialogue agents.

A coding-based animation system makes it possible to extend such templates to produce specifications for embodied conversational behavior. A template can indicate the prosody with which an utterance should be delivered, and a template can also spell out concurrent actions such as facial displays or coverbal gestures. We have used templates in a number of simple conversational programs to construct utterance specifications for RUTH (Rutgers University Talking Head), a publicly-available system that animates symbolic specifications for facial displays and head movements in synchrony with visible speech [13]. Many animation systems accept such symbolic specifications, including [16] for gesture; we expect that our techniques would apply to these systems as well.

In this paper, we describe our methods and experience in constructing such systems, and assess some of the advantages and disadvantages we have found in specifying embodied conversational behaviors using templates. We argue that templates give a human designer substantial freedom to explore the relationships among such factors as emotion, personality, individuality and social role, and to achieve specific combinations of behaviors that realize the designer’s intentions for system action, while still combining productive language use with genuine interaction with the user. Thus, although templates can achieve the coverage required of a broad system only with prohibitive effort, and although templates limit the kinds of roles and personalities that agents can achieve believably, templates nevertheless have a unique role in prototyping embodied conversational agents, in framing models of social agency, and in classroom instruction.



2 Motivating Templates

In proposing template generation, we aim to make it easier to construct conversational agents that use their animated bodies to express themselves in a way that adds to what they say. The challenge is that these agents must always present consistent contributions to conversation, in which all actions fit together into a coherent whole. In natural conversation, this coherence derives from our ability to recognize our partners’ beliefs and intentions indirectly from the defeasible evidence that their actions provide. Unfortunately, it is not easy to formalize this inference in a productive way.

We illustrate the problem with (1). In this dialogue

fragment, we imagine that *A* is a guest at a dinner for *B*, where the host has just served dessert. With (1a), *A* reports finding the dessert a bit overwhelming. *B* reacts, either as in (1b), or as in (1c):

(1) a *A*: This dessert has too much chocolate.

b <i>B</i> :		I like chocolate. [eyebrows raised, head turned].
c <i>B</i> :		I like chocolate. [frowning, head down]

Even though *B* uses the same speech in both, *B*’s two responses carry quite different implications for the dialogue—and even for dinner. *B*’s facial signals make (1b) a polite explanation for the choice of dessert (with the possible suggestion that, given *B*’s preferences and appetite, none of the dessert will go to waste). In contrast, *B*’s facial signals make (1c) a stern rebuke: at *B*’s party, guests must accept chocolate—indeed, perhaps they must eat it and enjoy it.

Of course, the difference in interpretation makes sense. *B*’s conversational signals in the two utterances show whether *B*’s utterance reports a surprising contrast (to which a reply is expected), or whether it reports an obstacle to *B*’s goals (where some change is necessary). This fits the general functions of raised eyebrows and frowns in conversation [14, 11]. But note how these functions follow from linking *B*’s conversational signals to the specific utterances they accompany. Were *B* to display simply that something unexpected was happening, it need not convey the favorable spin we see in (1b): we learn from the meaningful alignment between display and speech that the surprise *is B*’s like for chocolate. Conversely, had *B* simply displayed some unspecified disappointment in (1c), we might have taken *B* to have agreed that chocolate was the wrong choice; instead, we learn from the meaningful alignment between display and speech that *B* is in fact disappointed by *A*, not by the dessert.

We believe that (1) is representative of the embodied signals by which we infer interlocutor’s emotions and personalities—the signals by which we understand who a character is and how they mean to relate to us at any point in a given interaction. The specific functions of conversational signals follow in part from our inference in resolving ambiguities in interpretation. The contribution of one signal may determine the intended disambiguation—with radical effects for meaning.

There are many options to realize such utterances on

an individual basis. Performance-based animation is one possibility. Many conversational systems include the capability to realize completely prescribed, or “canned”, specifications for animation [31], sometimes alongside selective generative synthesis, as in [8]. However, there are few generative ways in existing architectures to explore the subtleties of behavior illustrated in (1). For example, Pelachaud et al. [20] and Cassell et al. [10] animate raised eyebrows for emphasis, by predicting them by rule from a representation of linguistic form. This does not apply for (1) because here the linguistic form is the same. Randomizing the behaviors, as in Perlin’s seminal work [22, 23], will not distinguish *B*’s alternative replies either. While successful applications can be built using only text (and see also [30]), their conversational behaviors can be, at best, redundant to the text.

What we would hope for eventually is to generate (1) using a strategy like that of Poggi et al. [24], which allows for actions that may add meaning to an utterance by using rules to generate eyebrow raises from an underlying representation of an agent’s intentions for discourse. But in such an architecture, implementation is only possible *after* you have good abstractions of agents’ communicative intentions and the behaviors they might use to achieve them. Initially we may just have specific intuitions about what the agent should do.

Templates offer a compromise for reconciling content with agent affect, making system-building easier (even practicable) at the expense of generality. A template abstracts away from a specific utterance, so it can be reused across a productive range of circumstances. However, a template builds off of a specific utterance that we understand intuitively; the template can retain a substantial part of the overall linguistic structure of that utterance, and consequently can preserve the implicit semantic connections that give the utterance a coherent and unified interpretation in context. For example, suppose we schematize the dialogue forms in (1) by abstracting away the nouns *dessert* and *chocolate*. *A*’s utterance reduces to *This N has too much Q*. *B*’s reply is either *I like Q*, with a synchronized brow raise and head turn; or *I like Q*, with a synchronized frown and nod. When we represent utterances with limited abstraction this way, we can expect that all instances will get parallel interpretations, without spelling out the general-purpose mechanisms that we or other conversational agents might use to recognize or produce these interpretations.

3 Implementation

To explore these ideas, we implemented a simple template-based mechanism for embodied conversational interaction. As we describe in Section 3.1, the key new

requirement of this mechanism is to ensure the characteristic alignment of nonverbal action with prosody in generated utterances. As a simple testbed illustrating this mechanism, we created two alternative animated versions of Weizenbaum’s Eliza [33] using templates for embodied utterances; we outline our philosophy and implementation in Section 3.2.

3.1 Templates: Structure and Instantiation

Our templates produce input for RUTH, the Rutgers University Talking Head [13]. This input consists of text that is marked-up to specify intonation and facial conversational signals. RUTH uses the phonetic structure of the input utterance to derive a schedule of animation instructions for lip synch and additional nonverbal actions. It then renders these instructions by applying deformations to a polygonal mesh, in part using a coarticulation model for speech [12, 15, 13]. RUTH does not yet combine conversational actions with autonomous behaviors like those of [22, 23, 1], but in general we believe such control is compatible with the template structures and methodology we have developed.

To specify prosody, RUTH uses the Tones and Break Indices (ToBI) model of English intonation [29, 3]. 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. 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. 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 end of the phrase. 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 that are somehow incomplete.

RUTH animates brow raises and frowns, smiles, and translations and rotations of the head. These conversational signals are either *batons*, which occur on a single accented syllable, or *underliners*, which span a prosodic constituent at the level of the intermediate phrase or higher; see [14, 18]. We specify head movements with

a small set of qualitative values, including rotations left (**L**) and right (**R**), nods up (**U**) and down (**D**), and tilting motions clockwise (**J**) and counterclockwise (**C**).

For an embodied conversational agent, a template will consist of text with a gap, marked up for a specific realization. In order to produce input for RUTH, the templates spell out the prosodic structure RUTH requires, and must respect RUTH's constraints on conversational signals. Given these constraints, the treatment of gaps is particularly important. In filling a gap, the system can make the following distinctions:

- Whether the material that fills the gap should be accented, or whether it can be reduced because it is not contrastive in the context. In (1), one possibility is to give the adjective filler *Q* a contrastive accent, highlighting its status as one of a number of alternative possible backgrounds for the utterance. Elsewhere, the filler for a template might not have to be accented, for example because the filler evokes an uncontroversially agreed topic, on which the utterance provides some further continuation.
- Whether the gap falls at the boundary of a prosodic phrase, or is embedded entirely within a single larger unit. When the filler marks a boundary, it must be decorated accordingly, with appropriate specifications at the onset of phrases to set pitch range and initiate underliners, and appropriate specifications at the ends of phrases to introduce boundary tones and to end ongoing nonverbal underliners. These ongoing events must be explicitly associated with the gap. The brow raise and head turn of (1b), for example, must be marked to end on the filler for *Q*.
- Whether the filler for the gap is to be realized with nonverbal batons, and which they should be. (This is possible only for accented fillers.) We do not use this for (1).

For example, we might capture (1b) in a template by combining fixed marked-up text as in (2):

```
(2) (i ((register "H") (jog "L")
        (brow "1+2")))
    (like ((accent "H*") (tone "L-")
          ))
```

with a gap as in (3)

```
(3) GAP: full phrase
    START: (register "L")
    ACCENTS: "L+H*"
    END: (tone "L-H%") (jog) (brow)
```

In sequence, this indicates that the utterance consists of an initial intermediate phrase, with an **H*** accent on *like*, followed by another phrase (at low register) filled in by rule with **L+H*** accents, and an **L-H%** boundary tone. The complete ensemble is underlined with a brow raise and a head turn. Meanwhile, we might capture (1c) by combining fixed marked-up text as in (4):

```
(4) (i ((register "H") (jog "D")
        (brow "4")))
    (like ((accent "H*") (tone "L-")
          (brow)))
```

with a gap as in (5)

```
(5) GAP: full phrase
    START: (register "L")
    ACCENTS: "L+H*"
    END: (tone "L-H%") (jog)
```

In sequence, this indicates an utterance with the same prosody as before. But now there is a frown underliner just on the initial intermediate phrase (it seems the frown should be shorter than the raise), while the whole ensemble is marked by a downward nod.

3.2 Animating Eliza

We used such templates to implement two versions of Eliza, which we will refer to as Nice-Eliza and Tough-Eliza; these implementations are now available with distributions of RUTH. With just a few exceptions, the two implementations use the same words and intonation in their utterances. Their animated deliveries are quite different, however. We mapped out the realization of each agent's utterances by keeping in mind a specific take on the relationship that such an agent might reasonably achieve with its partner, a putative psychoanalytic "patient". Nice-Eliza reflects a view of the interaction that casts difficulties with communication as the key challenge. We aimed for realizations of utterances that encouraged communication and explicitly drew the "patient" in for discussion about themselves, while at the same time setting strict bounds to keep the conversation on track. Nice-Eliza responds to *Nobody likes me* and *I want a change* at the top of Figure 1.

By contrast, our goal for Tough-Eliza was a more direct, confrontational style, focusing on the resolution of genuine and clear conflicts in the "patient's" situation. Here, we aimed for realizations of utterances that carried an expectation that the "patient" would naturally be forthcoming and open; and we aimed for realizations of utterances that call attention to and crystallize possible conflicts behind the "patient's" utterances. Tough-Eliza's responses appear at the bottom of Figure 1.

Nice-Eliza and Tough-Eliza each have their characteristic behaviors. Nice-Eliza often uses a circular head tilt as if to welcome further utterances, whereas Tough-Eliza requests information with a downward tilt that seems to call for franker talk. Nice-Eliza smiles more; Tough-Eliza frowns more. Both of the programs have the same repertoire of actions, though; Nice-Eliza also frowns and Tough-Eliza also smiles. The simultaneous speech allows different signals in different utterances to respond to the agents' take on the interaction.

Ultimately, the roles these programs apparently project are illusions that we as programmers imagined and designed into the interface. In other people, we would regard such roles as grounded not just in the general cultural repertoire but in some specific emotional competencies and specific personal values that an individual must put into practice to succeed in the role [21]. But the specific constructs we developed in Eliza were not based on a predefined system or organization; they emerged iteratively through our efforts to refine the agents' visible behavior. The specific advantage of templates over generative methods is precisely that they allow an artist or designer to freely craft the specific meaning of an interaction (cf. [35, 28]), and so provide an application for coding-based animation whose use in practice may go on to inform more general model-building.

4 Assessment

Our experience with the template method is positive, but guarded. Agents' delivery of animated utterances seems felicitous and natural in context, and even to reflect some broader consistency of interaction. While making the wrong move in embodied conversation can be quite jarring, our agent's delivery doesn't call attention to itself. Casual observers accept the intonation and expressions without comment. This apparent coherence should not be too surprising, however. Creating templates is not so different from coding specific utterances as people might naturally perform them, and speech synthesizers and animation engines can realize such specifications increasingly convincingly.

The difference between the two Elizas is perceptible, and suggests—if vaguely—the kinds of distinctions we aimed for. Nice-Eliza is indeed more encouraging to talk to. But Eliza does wear thin quickly. In some sense, the quality of the interaction provides the heaviest constraint on the impressions that an Eliza agent can get across, or on the kinds of intentions for interaction that a designer can realize with templates in this setting. Imagination notwithstanding, communication really is an issue with Eliza, and users really do need encouragement. With these constraints, adopting another perspective may not

be able to portray a meaningful alternative for the interaction. Of course, such limitations are familiar with any kind of interactive technology (see for example [19]).

5 Conclusions

In this paper, we have described the use of templates in embodied conversational agents. Templates are utterance specifications that combine static descriptions of words and synchronized behaviors with gaps that are filled in by rule. To implement templates effectively, we control the combined realization of text and fillers in a way that allows flexible delivery of the filler but that respects the natural synchrony between speech and nonverbal behavior. To use templates effectively, we consider specific embodied utterances that fit our design and have a clear overall interpretation; we abstract away from these utterances in a restricted way that preserves the connections between language and action that frame and disambiguate this interpretation.

We see template generation techniques as an important tool for developing embodied conversational agents, one that complements the use of canned utterances and full generative synthesis. Some applications may suffice with templates, while others may be most readily implemented by combining canned responses, templates and deep generation techniques. Templates can also be merged with behavior-based animation techniques, provided templates contain additional specifications of which actions can be altered in which ways.

We are particularly intrigued by the prospect of introducing the field of embodied conversational agents to students through template generation. In preparing templates, students can focus on developing detailed descriptions of utterances in conversation, and on gaining an informal appreciation for the general principles that give these utterances their meaning. At the same time, however, preparing templates gives students the satisfaction of creating interactive systems that they can play with, systems that may even react in unexpected ways. More generally, by highlighting the simplicity, and the challenge, of creating embodied conversational agents using templates, we hope to lower the barriers to entry to this research area, and add to its momentum.

Acknowledgments

This research was supported in part by NSF CISE CDA 9818322 and by Rutgers ISATC. Thanks for discussion to Gustavo Gallegos, Scott King, Larry LaFountain-Stokes and audiences at University of Chicago. Our speech synthesis uses Festival [4] and OGI voices [17].

References

- [1] N. Badler, J. Allbeck, L. Zhao, and M. Byun. Representing and parameterizing agent behaviors. In *Computer Animation*, pages 133–143, 2002.
- [2] J. Bates. Virtual reality, art and entertainment. *PRESENCE: Teleoperators and Virtual Environments*, 1(1):133–138, 1992.
- [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] J. Cassell. Embodied conversational interface agents. *Communications of the ACM*, 43(4):70–78, 2000.
- [6] J. Cassell. Towards a model of technology and literacy development: Story listening systems. Technical Report MIT-GNL-01-1, MIT, 2001.
- [7] J. Cassell and T. Bickmore. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and Adaptive Interfaces*, 12(1):1–44, 2002.
- [8] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsón, and H. Yan. Embodiment in conversational characters: Rea. In *CHI 99*, pages 520–527, 1999.
- [9] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors. *Embodied Conversational Agents*. MIT, 2000.
- [10] J. Cassell, H. Vilhjálmsón, and T. Bickmore. BEAT: the behavioral expression animation toolkit. In *SIGGRAPH*, pages 477–486, 2001.
- [11] N. Chovil. Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction*, 25:163–194, 1991.
- [12] 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.
- [13] D. DeCarlo, C. Revilla, M. Stone, and J. Venditti. Making discourse visible: coding and animating conversational facial displays. In *Computer Animation*, pages 11–16, 2002.
- [14] 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.
- [15] S. A. King. *A facial model and animation techniques for animated speech*. PhD thesis, The Ohio State University, 2001.
- [16] S. Kopp and I. Wachsmuth. Model-based animation of coverbal gesture. In *Computer Animation*, pages 252–257, 2002.
- [17] 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.
- [18] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, 1992.
- [19] J. Nielsen. *Usability Engineering*. Morgan Kaufmann, 1994.
- [20] C. Pelachaud, N. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1):1–46, 1996.
- [21] C. Pelachaud and I. Poggi. Subtleties of facial expressions in embodied agents. *Journal of Visualization and Computer Animation*, 13(5):301–312, 2002.
- [22] K. Perlin. Layered compositioning of facial expression. In *SIGGRAPH*, 1997. Technical Sketch.
- [23] K. 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] I. Poggi and C. Pelachaud. Eye communication in a conversational 3D synthetic agent. *AI Communications*, 13(3):169–181, 2000.
- [25] E. Reiter. NLG vs. templates. In K. de Smedt, C. Mellish, and H. J. Novak, editors, *Proceedings of the Fifth European Workshop on Natural Language Generation*, pages 95–106, Leiden, 1995.
- [26] E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Cambridge, 2000.
- [27] J. Rickel and W. L. Johnson. Animated agents for procedural training in virtual reality: Perception, cognition and motor control. *Applied Artificial Intelligence*, 13:343–382, 1999.
- [28] P. Sengers. Designing comprehensible agents. In *Proceedings of IJCAI*, pages 1227–1232, 1999.
- [29] 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.
- [30] K. Smid and I. S. Pandzic. Conversational virtual character for the web. In *Computer Animation*, pages 240–247, 2002.
- [31] T. Sowa, S. Kopp, and M. E. Latoschik. A communicative mediator in a virtual environment. In *Workshop on Information Presentation and Natural Multimodal Dialogue*, Verona, Italy, 2001.
- [32] M. Theune. *From Data to Speech: Language Generation in Context*. PhD thesis, Eindhoven Technical University, 2000.
- [33] J. Weizenbaum. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.
- [34] M. White and T. Caldwell. EXEMPLARS: A practical, extensible framework for dynamic text generation. In E. Hovy, editor, *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 266–275. Association for Computational Linguistics, 1998.
- [35] T. Winograd and F. Flores. *Understanding Computers and Cognition: A New Foundation for Design*. Addison Wesley, 1986.