Interactional Competency for Conversational Characters

Scott Prevost
Timothy Bickmore

FX Palo Alto Laboratory
3400 Hillview Avenue, Bldg. 4
Palo Alto, California 94304 USA
{prevost,bickmore}@pal.xerox.com

Justine Cassell

The MIT Media Laboratory
20 Ames Street
Cambridge, Massachusetts 02139 USA
justine@media.mit.edu

Abstract

Conversational characters are animated humanoids which interact with users using the modalities of human face-to-face conversation. Face-to-face conversation consists of information exchanges at the propositional layer–those things that contribute to the intended meaning of the speaker–and the interactional layer–those things that, among other functions, serve to regulate the flow of information among the interlocutors. We describe issues in developing such characters with multimodal communicative competencies at both the propositional and interactional levels.

Introduction

Recent work on the social nature of human-computer interactions (Reeves and Nass 1996) along with the emergence of synthetic characters and virtual worlds, has prompted research on building animated, embodied, anthropomorphic characters that can interact with human users to provide assistance in operating complex systems. One of the motivations for building these "conversational characters" is to facilitate natural communication with the system using the model of human face-to-face interaction, rather than forcing users to learn complex, and often non-intuitive, interfaces. Such interfaces may leverage users' tendency to attribute humanness to the interface, thereby making their interactions with the system easier (Cassell and Thórisson, forthcoming).

For the purposes of producing natural face-to-face conversation, several communicative modalities other than speech need to be represented. Gestures, facial expressions, intonational patterns, gaze, and body posture all provide information that may reinforce, color, or augment the interpretation of the spoken words alone. It is also these nonverbal modalities that accomplish much of the work of conversational regulation. It is therefore necessary for a system that attempts to engage in such complex interactions to have a model of how meaning is composed from observed multimodal behaviors, and how meaning is distributed among the various modalities during speech production.

In this position paper, we differentiate two layers of multimodal information in systems that generate face-to-face interactions–the propositional layer and the interactional layer–and argue that both must be modeled by a competent conversational character. The propositional layer involves those verbal and non-verbal behaviors that contribute to the intended meaning of the corresponding speech. For example, an intonational prominence on a pronoun may disambiguate the intended referent from other potential referents in a given context. An iconic gesture may supply information that is not present in the speech, such as a typing gesture in conjunction with the utterance "I'll send it right away," implying that the speaker intends to send "it" by email.

The interactional layer involves those verbal and non-verbal behaviors that regulate, coordinate and manage the flow of information between interlocutors. For the purposes of this paper, we concentrate on non-verbal interactional behaviors. For example, the direction of gaze has been shown to be correlated with turn-taking in dyadic conversation (Duncan 1974, Kendon 1974), regulating control of the speaking floor. Back-channel non-verbal behaviors, such as head nods, are performed by the listener as a means of providing limited feedback without taking a full speaking turn. Interactional information can also be communicated verbally, especially when other modalities are degraded or unavailable (e.g., "uh-huh"). Conversation initiation, termination, interruption, and filled pauses are examples of information encoded in the interactional layer. More complex verbal behaviors, such as flattery and small talk, can also serve to manage information flow, essentially greasing the wheels of interaction.

Previous Work

"Animated Conversation" (Cassell et al. 1994) was the first system to automatically produce context-appropriate gestures, facial movements and intonational patterns for animated agents. The focus of this project was on the production of speech and coordinated non-verbal behaviors for two animated characters, Gilbert and George, that engage in an automatically generated dialogue to accomplish a simple goal in the domain of banking. The generation of non-verbal behaviors was driven by an analysis of natural data based on the linguistic notion of *information structure*, which describes how information in an utterance is packaged with respect to the discourse context. For example, gestures and rise-fall intonation patterns were associated with the parts of utterances that, roughly speaking, provide new information rather than contextual links to previously mentioned or inferable information. The emphasis was on the production of non-verbal behaviors that emphasize and reinforce the content of the speech—the propositional layer. Regulatory signals, such as turn-taking cues, were derived automatically from the strict turn-taking produced by the dialogue planner rather than from observable gaze behaviors or back-channels.

The "Ymir" system, (Thórisson 1996) takes a much different approach, focusing on integrating multimodal input from a human user, including gesture, gaze, speech, and intonation, and producing multimodal output in an animated character called "Gandalf." The system allows a user to interact with Gandalf to manipulate a graphical model of the solar system and to pose simple questions about the planets.

Unlike Gilbert and George in the Animated Conversation project, Gandalf's understanding and production of non-verbal behaviors occurs mostly in the interactional layer. Although Gandalf has very limited knowledge and extremely restricted linguistic abilities, his ability to perceive and generate turn-taking cues and backchannel behaviors (e.g. head nods, direction of gaze) creates a smoothness of interaction that surpasses anthropomorphic interfaces without these interactional characteristics (Cassell and Thórisson, forthcoming).

Other character-based interfaces, such as Microsoft Persona (Ball et al. 1997) and PPP Persona (Andre, Muller, and Rist 1996), enforce rigid notions of alternating turns, and thus do not utilize information in the interactional layer. Speech-only mixed-initiative dialogue systems, such as TRAINS (Allen, et al. 1996), fail to process any of the non-verbal behaviors that convey interactional information in face-to-face conversation.

Research Issues

While Gilbert and George, and Gandalf represent significant advances in autonomous, embodied, animated agents, neither system is complete. Gilbert and George cannot interact with real people in real time, and fail to produce convincing turn-taking and backchannel behaviors. Gandalf, on the other hand, fails to model planning, language and propositional non-verbal behaviors at a level necessary for accomplishing non-trivial tasks. We believe that, in order to overcome these deficiencies, the next generation of animated, conversational characters must integrate the propositional and interactional layers of communication, and account for their interactions.

Consider the turn-taking problem. In "Animated Conversation," turns are allocated by a planner that has access to both agents' goals and intentions. In the real world, turns are negotiated through interactional non-verbal cues, rather than strictly imposed by the structure of the underlying task. In Gandalf, turns are negotiated by the agent and the human user in a natural way. Since Gandalf does no dialogue planning, however, each turn is artificially restricted to a single utterance. To competently handle turn-taking in non-trivial, mixed-initiative, multimodal dialogue, a system must interleave and process the propositional and interactional information in a principled way.

There are several significant research issues which must be addressed in building conversational characters with this capability. These are best described in terms of the impact of interactional competency on the components of a more traditional natural language dialogue system.

Understanding

In a multimodal conversation system, the understanding component must not only integrate information from different modalities into a coherent propositional representation of what the user is communicating, but—in order to know what function that information fills in the ongoing conversation—it must also derive the interactional information from the perceptual inputs. Moreover, it must determine when the user has communicated enough to begin analysis and be able to re-analyze input in case it misinterprets the user's turn-taking cues.

Discourse Planning

The discourse planner for conversational characters must be able to plan turn-taking sequences and easily adapt when those plans are invalidated by non-verbal cues—for example when the human refuses to give over the turn, and continued nonverbal feedback becomes more appropriate than adding new content to the conversation.

Generation

When the discourse plan calls for the generation of interactional information, the character must decide which modality to use, and must take into account interactional information from the user. For example, signaling an interruption or termination may be performed verbally or with a nod of the head depending on whether the user or the character currently has the turn.

Finally, the issues of real-time performance and synchronization are crucial factors for embodied conversational systems. Characters must respond to a user's utterance within a short period of time in order to maintain the appearance of having conversational competency. Many interactional responses, such as back-channel feedback, must be timed within a fraction of a second to convey the appropriate message. Since performing natural language understanding, discourse planning, and text generation in a principled way is computationally complex, characters may need to be constructed from a suite of communication skill processes which are operating with different response times. In fact, if the system's faster reactions are aptly timed, this may provide more time for the slower, reflective layer to come up with the correct content, either in understanding or generation. That is, a well placed "hmm, let's see" with slow thoughtful nods of the head can give users necessary information about the state of the conversation, while allowing the system a little more time to come up with a contentful response.

Conclusion

Effective conversational characters will require competency in both propositional and interactional layers of communication, and must address the issue of how these layers are integrated. We are currently developing a generic conversational character architecture which addresses these issues as a joint project between FX Palo Alto Laboratory and the MIT Media Laboratory.

References

Allen, J.; Miller, B.; Ringger, E.; and Sikorski, T. 1996. A Robust System for Natural Spoken Dialog. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, 62-70. Santa Cruz, Calif.

Andre, E.; Muller, J.; and Rist, T. 1996. The PPP Persona: A Multipurpose Animated Presentation Agent. In *the Proc. of Advanced Visual Interfaces*, ACM Press.

Ball, G.; Ling, D.; Kurlander, D.; Miller, D.; Pugh, D.; Skelly, T.; Stankosky, A.; Thiel, D.; Van Dantzich, M. and Wax, T. 1997. Lifelike computer characters: the persona project at Microsoft Research. In J. M. Bradshaw (ed.) *Software Agents*, Cambridge, MA: MIT Press.

Cassell, J.; Pelachaud, C.; Badler, N.; Steedman, M.; Achorn, B.; Tripp, B.; Douville B.; Prevost, S.; and Stone, M. 1994. Animated Conversation: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents, *In Proceedings of SigGraph 1994*, 413-420.

Cassell, J. and Thórisson, K. Forthcoming. The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence,* forthcoming.

Duncan, S. 1974. Some Signals and Rules for Taking Speaking Turns in Conversations. In Weitz (ed.), *Nonverbal Communication*: Oxford University Press.

Kendon, A. 1974. Movement Coordination in Social Interaction: Some Examples Described. In Weitz (ed.), *Nonverbal Communication*: Oxford University Press.

Reeves, B., and Nass, C. 1996. *The Media Equation*: Cambridge University Press.

Thorisson, K. 1996. Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills. PhD dissertation, MIT Media Laboratory.