# Face-to-Face Interfaces

**Scott Prevost, Peter Hodgson, Linda Cook, Elizabeth Churchill**
FX Palo Alto Laboratory
3400 Hillview Avenue, Bldg. 4
Palo Alto, California 94304 USA
{prevost,hodgson,cook, churchill}@pal.xerox.com

## ABSTRACT

Recent work on the social nature of human-computer interactions [3] has prompted research on animated, anthropomorphic characters in user interfaces. Such interfaces may simplify user interactions by allowing them to use and interpret natural face-to-face communication techniques such as speech, gestures and facial expressions. We describe our initial implementation, a character that controls the A/V facilities in a state-of-the-art conference room, and outline the goals of our ongoing project.

### Keywords

Conversational Characters, Multimodal Systems

## INTRODUCTION

In this paper, we describe an emerging research program on embodied, conversational interfaces. The overarching goal is to determine the efficacy of animated character interfaces in various kinds of applications with various user communities. We hypothesize that, just as other HCI metaphors (e.g. the virtual desktop) have limitations that impose constraints on their applicability to certain tasks and users, so to will the face-to-face conversation metaphor that we propose. In light of this, we need to formulate a better understanding of how people accomplish tasks through face-to-face interactions with each other, and apply that understanding to virtual characters. Therefore our secondary goal is to formulate rules for face-to-face interactions based on the existing literature, implement the rules in character interfaces, and analyze interactions with real users.

Our third central goal is to study the effects of personality on modulating face-to-face conversational behaviors. Although not specifically addressed in this paper, this work has implications not only for determining the relevant rules for face-to-face behaviors, but also for selecting appropriate applications and users. We hypothesize that by modulating the behaviors in various ways, users will attribute different personalities to characters, and that certain personalities will be better suited for some applications and users than others. We have taken great care in designing the visual appearance of our character to avoid caricatures, allowing a range of personality types to be displayed.

This paper, and indeed our research so far, focuses mainly on the second goal articulated above—building a character that understands and employs the social rules of face-to-face interactions. We leave the evaluation of those rules with respect to various applications, personalities and user communities to future publications. In the sections that follow we describe our approach to modeling face-to-face interactions, briefly review the underlying system architecture, and provide a short description of our initial application, a character that controls our conference room.

## FACE-TO-FACE INTERACTIONS

Face-to-face interactions are perhaps the most natural way for people to communicate with each other. Speaking with someone face-to-face has several advantages over written text and disembodied speech. In face-to-face situations, information can be conveyed through multiple modalities, including the spoken words, the intonation (roughly, the melody) of speech, facial expressions, gaze movements, gestures, and body posture. Face-to-face interaction also allows information to be transmitted continuously and bi-directionally between interlocutors, so that for example, one participant in a conversation may be signalling agreement by nodding his head, while at the same time the other participant is speaking and gesturing with her hands.

One of the key problems in understanding the various types of behaviors involved in face-to-face interactions are that some behaviors can be “understood” only in the context of other behaviors. For example, the utterance “it's over there” makes sense only when “there” is specified, either by a previous utterance or a pointing gesture. To confuse matters more, behaviors which appear similar on the surface don't always have the same function. For example, while many speakers unwittingly raise their eyebrows on emphasized (intonationally stressed) words, this has a very different function from raising eyebrows to show surprise.

In our system, we differentiate two layers of behaviors for face-to-face interactions—the propositional layer, and the interactional layer. The propositional layer involves those verbal and non-verbal behaviors that contribute to the intended meaning of the corresponding speech. For example, 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 example, the direction of gaze has been shown to be correlated with turn-taking in dyadic conversation [2], regulating control of the speaking floor. Other interactive behaviors include back-channels, such as head nods, which are performed by the listener as a means of providing limited feedback without taking a full speaking turn.

## THE ARCHITECTURE

In order to support these two layers of behaviors, we have developed an architecture jointly with the Gesture and Narrative Language Group at the MIT Media Laboratory. The architecture (Figure 1) is composed of various modules operating in parallel with differing reaction times.

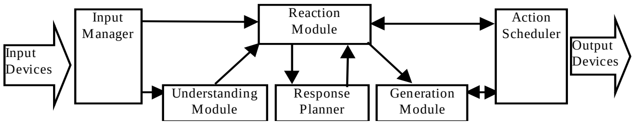


*Figure 1: Conversational Character Architecture*

The input manager is responsible for collecting data from the various input devices, normalizing it, time-stamping it, and routing it to the appropriate modules. Inputs for our current implementation include speech recognition, a vision-based tracking system for determining the user's location in the room, and a head-mounted electronic compass for determining the user's head orientation and approximating the direction of gaze. We hope to improve our vision component to provide gross gesture information in the near future.

The action scheduler is responsible for executing output behaviors in the animated character. It receives requests for the character to perform certain behaviors or collections of synchronized behaviors (e.g. an utterance and a gesture), and initiates the proper routines in the animation renderer and the speech synthesizer.

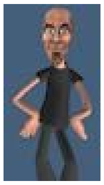
The other modules constitute the "brain" of the system. The reactive module is responsible at each computation cycle for selecting what the character should be doing. This involves arbitrating among various demands to perform interactional and propositional behaviors, striking a balance so that turn-taking and back-channel (interactional) protocols are appropriately intertwined with content-bearing behaviors.

The understanding, response planning, and generation modules operate in parallel to perform the more deliberative tasks. The response planner is responsible for formulating a dialogue plan, re-planning when necessary and informing the other modules of the current plan state. The understanding module is responsible for fusing together multiple inputs to determine a unified meaning. The generation module has the reverse task—taking a plan to convey some "meaning," and determining appropriate speech, gestures and facial expressions to perform.

This architecture differs from previous architectures for animated characters in several significant ways. The coordination of planning and language generation differentiates the current approach from Ymir [4], an embodied character architecture that focuses mostly on the interactional layer of behavior, and does not have an explicit mechanism for intertwining dialogue planning and natural language generation. Our architecture also differs from "Animated Conversation" [1], which focuses on the propositional layer of behaviors, in that ours handles real-time multimodal input at the interactional level.

## CONCLUSIONS

Our first prototype application, to be demonstrated in a brief video clip, is a character who controls our state-of-the-art conference facility at FXPAL. Users of the conference room are frequently confused and frustrated by the existing touch-screen menu interface for controlling A/V equipment. For example, a user who just wants to show a slide on the projection screen doesn't really need to navigate menus that deal with setting up conference calls. Our new character interface gives users the ability to pose a question or command, to perform a task immediately or get right to the crucial information.



The character lives in a large screen display at the front of the conference room. Users can approach the character, wander about the space in front of him, ask questions and give commands. The character is able to make sense of the multiple inputs and display the appropriate behaviors in response. For example, if the user is looking in the general direction of the character and the user begins speaking, the character will orient his gaze at the user and continue to display the *listening-gaze* behavior (fixed on the user with occasional glances to the side) as long as the user is speaking.

We believe that embodied characters in the interface show great promise for making HCI more intuitive. In our future work we intend to refine the rules sets that define character behaviors, and evaluate interactions with different user communities in a variety of applications.

## REFERENCES

1. Cassell, J., Pelachaud, Badler, Steedman, Achorn, Tripp, Douville, Prevost, & Stone. Generation of Facial Expression, Gesture & Intonation for Multiple Conversational Agents, SigGraph'94. 1994.

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

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

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