# Perceptive Animated Interfaces: First Steps Toward a New Paradigm for Human–Computer Interaction

RONALD COLE, SAREL VAN VUUREN, BRYAN PELLOM, MEMBER, IEEE, KADRI HACIOGLU, JIYONG MA, MEMBER, IEEE, JAVIER MOVELLAN, SCOTT SCHWARTZ, DAVID WADE-STEIN, WAYNE WARD, AND JIE YAN, MEMBER, IEEE

*Invited Paper*

*This paper presents a vision of the near future in which computer interaction is characterized by natural face-to-face conversations with lifelike characters that speak, emote, and gesture. These animated agents will converse with people much like people converse effectively with assistants in a variety of focused applications. Despite the research advances required to realize this vision, and the lack of strong experimental evidence that animated agents improve human–computer interaction, we argue that initial prototypes of perceptive animated interfaces can be developed today, and that the resulting systems will provide more effective and engaging communication experiences than existing systems. In support of this hypothesis, we first describe initial experiments using an animated character to teach speech and language skills to children with hearing problems, and classroom subjects and social skills to children with autistic spectrum disorder. We then show how existing dialogue system architectures can be transformed into perceptive animated interfaces by integrating computer vision and animation capabilities. We conclude by describing the Colorado Literacy Tutor, a computer-based literacy program that provides an ideal testbed for research and development of perceptive animated interfaces, and consider next steps required to realize the vision.*

Manuscript received November 12, 2002; revised March 31, 2003. This work was supported in part by the National Science Foundation's Information Technology Research Program under Grants REC-0115419 and IIS-0086107; in part by the NSF's Interagency Educational Research Initiative under Grants EIA-0121201 and 1R01HD-44276.01; in part by California Digital Media Innovation under Grant 01-10130; and in part by the Coleman Institute for Cognitive Disabilities.

R. Cole, S. van Vuuren, B. Pellom, K. Hacioglu, J. Ma, S. Schwartz, D. Wade-Stein, W. Ward, and J. Yan are with the Center for Spoken Language Research, Institute of Cognitive Science, University of Colorado, Boulder, CO 80309 USA (e-mail: cole@cslr.colorado.edu; sarel@cslr.colorado.edu; pellom@cslr.colorado.edu; hacioglu@cslr.colorado.edu; jiyong@cslr.colorado.edu; schwarts@cslr.colorado.edu; steind@cslr.colorado.edu; whw@cslr.colorado.edu; jie@cslr.colorado.edu).

J. Movellan is with the Machine Perception Laboratory, Institute for Neural Computation, University of San Diego, La Jolla, CA 92093 USA (e-mail: movellan@mplab.ucsd.edu).

Digital Object Identifier 10.1109/JPROC.2003.817143

***Keywords*—*Animation, conversational agents, dialogue system, electronic books, human–computer interaction, interactive books, literacy teaching, machine perception, spoken language, tools for cognitive disabilities, virtual humans, vocabulary tutors.***

## I. THE VISION

We envision a new generation of human–computer interfaces that interact with people like people interact with each other. These interfaces will use intelligent and embodied animated agents to engage users in natural face-to-face conversational interaction to accomplish a wide variety of tasks. An intelligent agent is one that mimics the behaviors of real persons and behaves intelligently in the context of a specific application or task domain. An embodied agent is one that resembles a real person. We call these interfaces of the future *perceptive animated interfaces*. The systems we envision extend the scope of conversational interfaces beyond audio processing of spoken language to include face-to-face conversational interaction with animated computer characters. The thesis of this paper is that we can begin to develop perceptive animated interfaces today that will produce more effective and desirable communication experiences than existing systems.

Perceptive animated interfaces will be populated with one or more lifelike three-dimensional (3-D) computer characters, also known as animated conversational agents or avatars, which combine human language, computer vision, and character animation technologies to engage users in natural face-to-face conversations. Animated agents will interact with people like two people interact with each other when conversing—through speech, head nods, eye contact, facial expressions, and hand and body gestures. These lifelike computer characters will orient to the user, interpret the speaker's auditory and visual behaviors to infer his or her intentions and

cognitive state, provide real-time feedback while the user is speaking (e.g., to indicate agreement, puzzlement, desire to speak, etc.), engage in interactive turn-taking behaviors, and communicate both linguistic and emotional content using speech, facial expressions, and gestures.

The advent of perceptive animated interfaces will revolutionize human–computer interaction by enabling users to communicate with machines using their natural communication skills, and by enabling system developers to design interactive experiences that are more personal, emotional, meaningful, enjoyable, and effective. Natural communication with intelligent animated agents will present new and unprecedented opportunities to individuals, including those who cannot read or type, to learn new skills, communicate more effectively, and increase their participation in the emerging information society.

Animated characters can enhance the user experience in several important ways. They bring a remarkable communication instrument—the human face—to the human–computer interface. During speech production, faces are informative linguistically. When producing speech, the lips, tongue, and jaw provide visual cues that complement auditory cues; for example, the difference between /ba/ vs. /da/ and /ma/ vs. /na/ can be distinguished by watching the speaker's lips. Animated faces today can synthesize visible speech with sufficient accuracy to improve speech recognition in noisy environments (relative to sound-alone conditions) and to improve speech recognition accuracy for individuals with hearing problems [1]–[3]. Whereas human faces are informative linguistically during everyday interaction, in language training tasks, animated characters can become "superinformative" through special effects, such as becoming semitransparent to show the movements of the tongue within the mouth from different visual perspectives.

Although enhancing the acoustic signal is an important benefit of animated characters, their greatest impact is likely to be their potential to change fundamentally the *experience* of communicating with machines by fully engaging the user's senses and emotions. An animated agent that behaves as if it perceives the user, understands the user's speech, accurately interprets the user's emotions, and responds in an appropriate and sensitive manner has the capability to produce intense, immersive, and emotional interpersonal experiences.

Artists at Disney Entertainment understand well the importance of communicating emotions through animated characters. They have developed detailed procedures and languages incorporating characters' facial expressions and gestures, colors, lighting effects, and other features for designing storyboards for animated productions to achieve emotional milestones [4]. Spoken dialogue is added to the production only after emotional milestones are achieved. As emotion is virtually ignored in current conversational systems, language technology researchers can learn much from Disney's philosophy—that emotion is a fundamental dimension of human communication, and that good communication experiences speak to the heart as well as to the mind. The development of perceptive animated agents provides a new and powerful opportunity for researchers to investigate the visual synthesis of emotions and gestures and their effects in human–computer interaction.

Both everyday observation (young people transfixed by cartoons and immersed in video games) and scientific research shows the propensity of people to become immersed in media, including their computer systems. In *The Media Equation*, Reeves and Nass [5] summarize a set of experiments demonstrating that people interact with computer programs like they interact with other people; their experiments replicate many of the major findings in social psychology by replacing the traditional "stooge" in a social psychology experiment with a software application that behaves in a like manner. For example, it has been shown that we like people who compliment our performance in collaborative tasks more than we like people who do not compliment our performance; we also like software more that gives us compliments. When an animated agent is involved, this effect of personalizing the interaction between the user and computer system can be intensified greatly. We observed this phenomenon in classrooms at the Tucker-Maxon Oral School (TMOS), Portland, OR, where educators and profoundly deaf students have interacted daily with the animated agent Baldi for the past five school years. Teachers and students alike personalized Baldi and perceive it as an entity with speech perception and production abilities, rather than a computer program of integrated language technologies. When the recognition system makes errors, teachers and students say, "Baldi did not understand" or "I did not speak well enough to make Baldi understand."

While perceptive animated interfaces are still science fiction, we argue here that human communication technologies have matured to the point where it is now possible to conceptualize, develop, and test initial system prototypes. This claim is made with full knowledge of present realities—that there are few (if any) solved problems in human language technology; that conversational interfaces work well only in specific task domains, and even in these domains do not approach human performance; that underlying speech and language recognition and generation technologies are fragile and inaccurate relative to human performance; and that research breakthroughs are needed in nearly all areas of language technology before human–computer interaction can mimic conversational interaction among people. Despite these limitations, we believe that perceptive animated interfaces can be developed today that will provide great benefits to individuals in specific task domains by combining existing technologies in novel and creative ways. We also believe it is critically important to undertake development of these futuristic systems today to determine their feasibility, to provide testbeds for research and development of research architectures and technology components, to identify missing knowledge, and to assess the benefits that perceptive animated interfaces may have to help individuals acquire new knowledge and skills.

## II. State of the Field

What is the current state of research and development of perceptive animated agents, and how effective are these

agents in improving human–computer interaction? A vital and growing multidisciplinary community of scientists worldwide is addressing these questions, and significant efforts are underway to develop and evaluate virtual humans in various application scenarios. To date, researchers have generated powerful conceptual frameworks, architectures, and systems for representing and controlling behaviors of animated characters to make them believable, personable, and emotional [6]–[11]. Gratch *et al.* [12] and Johnson *et al.* [13] present excellent overviews of the scope of enquiry and the theoretical, cognitive, and computational models underlying current research aimed at developing believable virtual humans capable of natural face-to-face conversations with people.

Animated conversational agents have been deployed in a variety of application domains. Some researchers have embedded animated conversational agents in information kiosks in public places (e.g., [14], [15]). In pioneering work conducted over the past ten years at KTH, Stockholm, Sweden, Gustafson [15] and his colleagues developed a series of multimodal dialogue systems of increasing complexity incorporating animated conversational agents: 1) *Waxholm*, a travel-planning system for ferryboats in the Stockholm archipelago [16], [17]; 2) *August*, an information system deployed for several months at the Culture Center in Stockholm [18], [19], in which the animated character moved its head and eyes to track the movements of persons walking by the exhibit and produced facial expressions such as listening gestures and thinking gestures during conversational interaction; and 3) *AdApt*, a mixed-initiative spoken dialogue system incorporating multimodal inputs and outputs, in which users conversed with a virtual real estate agent to locate apartments in Stockholm [20]. AdApt produced accurate visible speech, used several facial expressions to signal different cognitive states and turn-taking behaviors, and used direction of gaze to indicate turn taking and to direct the user to a map indicating apartment locations satisfying expressed constraints. These systems produced important insights into the challenges of developing and deploying multimodal spoken dialogue systems incorporating talking heads in public places.

Learning is an excellent task domain for investigating perceptive animated agents, and much work has been conducted in this area [13], [21]–[28]. First, face-to-face tutoring is known to be the most effective method of instruction [29], [30], and much is known about the strategies that effective tutors use [31]; thus, research can inform the design of animated agents intended to model good, effective tutoring behaviors. Second, development of intelligent tutoring systems is an active field of research, with several highly successful systems; thus, interaction with animated characters can be incorporated into these systems and potential benefits evaluated. Third, there are published national standards and standardized tests for assessing learning in many domains (e.g., reading, language proficiency, science, and math), so performance of conversational interfaces with and without animated agents can be evaluated on established and well-accepted measures. Finally, there is great need for computer-based learning systems that improve student achievement while reducing teachers' workloads.

Are animated agents effective? Does incorporating an animated agent into human–computer interfaces make these interfaces more effective? Research to date does not produce a clear answer to this question. Dehn and van Mulken's [32] review of experiments investigating the effectiveness of animated agents in a variety of tasks showed that most studies failed to reveal improvement in user performance. They note, however, that most studies to date have compared animated agent versus no animated agent conditions in a single short session, and that benefits of animated interfaces might emerge if longer studies with multiple sessions were used. Experiments by Graesser and colleagues are illustrative. Graesser *et al.* [33] examined the effectiveness of an animated conversational agent in AutoTutor, an intelligent tutoring system that "helps students construct answers to deep-reasoning questions by holding a conversation in natural language." During dialogue interaction with AutoTutor, students typed in their responses, and the system presented information via different media conditions—print only, speech only, talking head, or talking head and print. Although significant learning gains were observed using AutoTutor relative to other learning conditions (e.g., presentation of relevant text rather than dialogue interaction), there were no differences among the four media conditions. They conclude: "Something about the dialog capabilities of AutoTutor facilitates learning, particularly at deeper levels of comprehension. In contrast, the effects of the media are nonexistent. Simply put, it is the message that is the message—the media is not the message." Graesser *et al.* [34] replicated this result in a study using 155 college students who used a Web facility using one of these same four navigational guide conditions. But in this latter study, they also note that animated conversational agents have proven to be effective when they deliver learning material in monologues and tutorial dialogues [35], [36]. Given these and other results, it is clear that evaluating the effectiveness of conversational agents in human–computer interfaces is extremely complex, and can be influenced by the nature of the task, the user's personality characteristics [37], and the believability (quality) of the animated agent.

The poor quality of animated conversational agents today is a major stumbling block to progress. Johnson *et al.* [13] argue that it is premature to draw conclusions about the effectiveness of animated agents because they are still in their infancy, and "…nearly every major facet of their communicative abilities needs considerable research. For this reason, it is much too early in their development to conduct comprehensive, definitive empirical studies that demonstrate their effectiveness in learning environments. Because their communicative abilities are still very limited compared to what we expect they will be in the near future, the results of such studies will be skewed by the limitations of the technology."

To summarize, development of virtual humans is still in its infancy. In the past ten years, a small but emerging community of researchers has made great progress toward identifying the scope of multidisciplinary research required and the key research challenges that need to be addressed, and by offering strong theoretical, conceptual,

and computational frameworks that provide a foundation for multidisciplinary research among computer scientists, cognitive scientists, psychologists, and researchers in other disciplines. Although much innovative research has been conducted, we conclude that experiments investigating the efficacy of animated agents are limited today by constraints imposed by the state of the art of human communication technologies, including speech and language technologies, computer vision, and real-time character animation. A grand challenge is to develop new architectures and technologies that will enable experimentation with perceptive animated agents that are more natural, believable, and graceful than those available today.

In the remainder of this paper, we describe our initial experiences using an animated talking head that produced accurate visible speech in learning tasks using the CSLU Toolkit, a publicly available platform for research and development of multimodal dialogue systems [38]–[40]. We then describe a multilaboratory research effort that aims to produce a new generation of engaging and effective perceptive animated agents by combining emerging speech and language, computer vision, and character animation technologies, and by evaluating these animated agents in learning tools in public school classrooms within a literacy program called the Colorado Literacy Tutor (CLT). We conclude with a brief discussion of coordinated efforts by the research community required to accelerate progress in development of virtual humans.

## III. Initial Experiences With a Talking Head

### A. Language Training at the Tucker-Maxon Oral School

Between 1997 and 2000, a team of researchers at the University of Colorado (CU), Boulder; the Oregon Graduate Institute, Beaverton; and the University of California, Santa Cruz, collaborated with educators at TMOS to develop computer-based learning tools that used an animated 3-D talking head, called Baldi, to teach speech and language skills to profoundly deaf children [23], [24]. Authoring tools were developed within the CSLU Toolkit [40], [41] to enable project staff and interested teachers to build vocabulary tutors quickly and to integrate these applications into daily classroom learning activities. The most useful authoring tool, proposed by the TMOS educators and developed in close collaboration with them, was a step-by-step Vocabulary Wizard [42]. This tool enabled authors to import images into an application (which were often photos taken by the students), highlight objects or regions within the images, and type the names of objects. These simple steps were used to create hundreds of Vocabulary Tutors, each of which followed a four-stage sequence: 1) a pretest to collect baseline data on vocabulary knowledge (e.g., Baldi says, "Click on the cup" or "Show me the cup"); 2) introduction to the vocabulary items (e.g., "Here is the cup"; "Click on the cup"); 3) practice identifying and saying the words (e.g., "Show me the cup"; "No, that was the plate, here is the cup," "All right!" "Now say cup."); and 4) a final test to assess learning. Results showed that: 1) students learned vocabulary words quickly and retained over 50% of the words months later; 2) their speech perception and production skills improved dramatically; 3) the students thought of Baldi as a personal coach that could be relied on to patiently help them learn; and 4) teachers reported that the learning tools made their job easier and their teaching more effective, since half of the students in the class could work independently on the computers while teachers provided individualized attention to the others [3].

The outcomes of the TMOS project showed that an animated talking head that produces accurate visible speech synchronized with synthetic speech can be a powerful tool in teaching individuals who are profoundly deaf to recognize and produce new words. Not only did learning gains occur, but also the students, teachers, and administrators who worked with Baldi on a daily basis were uniformly positive in their experiences and evaluations. The director of the school, a distinguished researcher in oral deaf education, offered his opinion that learning tools incorporating animated agents would revolutionize oral deaf education [28].

We also learned that animated learning tools can capture the public imagination. The project was featured on the National Science Foundation home page during March and April 2001, and a national television network produced a segment that showcased the project. Over a period of several months, ABC TV's *Prime Time Thursday* television crew interviewed TMOS students and staff, filmed students using the vocabulary tutors, and conducted independent tests that revealed dramatic improvements in vocabulary acquisition and the students' speech production that were attributed to the learning tools [43]. They introduced the segment with a video of a student using a vocabulary tutor and the words "This is what a small miracle looks like."

### B. Learning Tools for Children With Autism Spectrum Disorders

Following the TMOS project, a two-month pilot project funded by the Coleman Institute for Cognitive Disabilities, Boulder, CO, was conducted at CU during the summer of 2001 to assess the feasibility and applicability of using animated characters to teach vocabulary and concepts to children with autism spectrum disorders (ASD). Children with ASD have difficulty with social situations and with interpreting the emotion, intention, and perspective of others, and exhibit a strong need for predictability, consistency, and routines. They also often have auditory processing disorders and a need for auditory information to be repeated or paired with text. These characteristics of autism make it difficult for these children to learn vocabulary and other information in traditional settings.

On the positive side, many children with ASD have a natural affinity and comfort level with computers. When they are using computers, there are no social expectations, and there are consistent routines that are presented in a predictable environment. This pilot project with children with ASD attempted to create a learning environment that bridged the gap between the nonsocial computer and the humanlike, consistent interactions of the animated talking head, Baldi.


**Fig. 1.** CSLU vocabulary tutor, a computer-based learning tool that uses an animated 3-D talking head, called Baldi. (a) Screen shot showing one of the tasks: "Learning about planets." (b) Use as a learning tool for children with ASD.

The six children with ASD involved in the project ranged in age from 3 to 13 years, with developmental levels that ranged from cognitively disabled to gifted. The research team, teachers, and parents built hundreds of vocabulary tutors to meet the specific needs of the participating children. Content areas included maps, counting, letter and word identification, ancient Egypt, learning about the planets, recognizing emotions, more than/less than relations, insects, telling time, and many others. Fig. 1 shows one of the tutors in use.

Each student attended two sessions per week for eight weeks. Within one session, five of the six children were able to work on the vocabulary tutor independently. The one student who was not independent was still working on mouse skills, which proved to be the minimum requirement for accessing the software.

Parent reports and observations indicated that the children were engaged by Baldi, talked about him at home, and were eager to come to the sessions. Some sessions had to be rescheduled earlier in the day than parents had originally planned because the children knew it was their day to see Baldi and were relentless in their questioning of "Is it time yet?"

Pre- and post-test scores from the Vocabulary Tutor indicated that the children were able to learn basic concepts (maps, capitals, items from ancient Egypt, insect names, etc.) quickly and retained what they learned over the course of the summer. More abstract concepts such as more than/less than and wh- question words were more difficult for the children to learn using this paradigm.

Teacher and parent reports indicated that information learned in the summer program generalized to the classroom in the fall; students were raising their hands and offering information that was taught during the summer sessions.

The Vocabulary Tutor using an animated character proved to be an effective, engaging teaching tool for children with ASD. The Vocabulary Tutor is currently being used in the schools and homes of children with ASD to teach basic vocabulary and concepts.

### C. Summary of Initial Experiences

Our initial experiences using a talking head in language training and learning tasks produced positive experiences with two groups of exceptional children. In fact, the learning gains, positive evaluations, and visibility resulting from the research exceeded all reasonable expectations. We were, thus, both encouraged and surprised by these initial results, especially in view of the limited capabilities of the animated character in these applications. Baldi was not a perceptive agent; speech perception was disabled for this population during the applications because of unacceptable accuracy, so "perception" was limited to knowledge about whether the student moved the cursor and clicked the mouse on the correct object. Although Baldi was certainly an animated agent, animation was limited to eye blinks, raised eyebrows, and accurate movements of the lips, tongue tip, and jaws. Viewed objectively, our initial applications incorporated a disembodied head producing synthetic speech unaccompanied by natural facial expressions or emotions; a far cry from the engaging, emotional, full-bodied, lifelike characters we hope to invent in the future. Still, teachers and students perceived Baldi as helpful and effective, and the learning tools produced impressive learning gains. We thus conclude that even primitive animated agents have the potential to engage and help students learn in a supportive learning environment. These results provide excellent motivation to develop the next generation of perceptive animated interfaces.

## IV. Foundations for Research and Development of Perceptive Animated Interfaces

In this section, we attempt to show that there currently exists sufficient knowledge, infrastructure and resources to support research and development of perceptive animated interfaces. We describe a set of research tools that provide a foundation for transforming spoken dialogue systems into perceptive animated interfaces. These tools include the Galaxy architecture, the CU Conversational Agent Toolkit (CAT), CU Animate, and computer vision technologies.

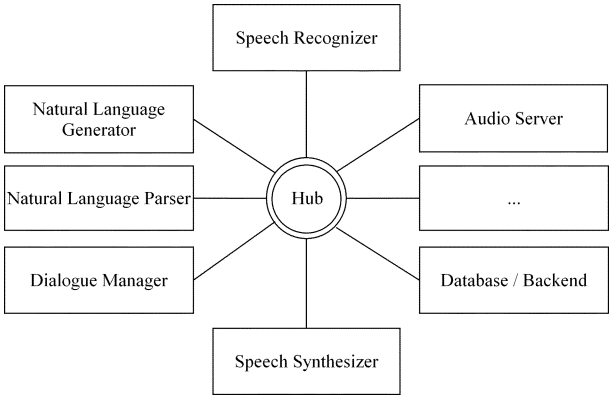



**Fig. 2.** Galaxy Hub architecture with CU CAT component modules. Communication among the component modules happens through the Hub.

### A. Galaxy

Galaxy is a public domain, plug-and-play architecture developed by the Massachusetts Institute of Technology Spoken Language Systems Group, Cambridge, to support research and development of advanced dialogue systems [44]. It is well-tested, open source, used by laboratories worldwide, and maintained by MITRE under support from the Defense Advanced Research Projects Agency [45]. Galaxy supports distributed systems and multisite collaboration—through its plug-and-play architecture, different sites can develop software-compliant technology servers that can communicate through a common architecture. Galaxy supports servers running on different computers with different operating systems, enabling widespread, inexpensive Web-based deployment of applications to different platforms.

The architecture (see Fig. 2) has a programmable Hub and application program interface that enables flexible control of interaction among servers, and a set of libraries for rapid prototyping including a graphical user interface (GUI) for controlling and monitoring the processes. The following major functions of the Hub, which have been proven to be useful for developing spoken dialogue systems, are also useful for developing perceptive animated agents.

1) *Routing*: The architecture handles message traffic among the distributed servers.
2) *State Maintenance:* The architecture provides a means of storing and accessing state information for all servers.
3) *Flow control:* The architecture manages the progress of an utterance through its processing stages, server by server.

In the next section, we describe a conversational agent toolkit that has been developed based on the Galaxy architecture. This toolkit provides a foundation for research in and the development of perceptive animated interfaces.

### B. Conversational Agent Toolkit (CU Communicator)

CAT, developed at the Center for Spoken Language Research (CSLR), Boulder, CO, provides a general-purpose platform, a set of technology modules (see Fig. 2), and tools for researching and developing advanced dialogue systems—systems that enable completely natural and unconstrained mixed-initiative conversational interaction with users in specific task domains. The toolkit provides tutorials to help users develop new systems using existing technology servers. For example, to illustrate the concepts of spoken dialogue system design, CAT includes the complete CU Communicator travel-planning system [46]. This travel-domain dialogue system enables users to say what they want, when they want, while conversing with the system to make travel plans involving planes, hotels, and rental cars [46]–[48]. The system can be accessed from the CU Communicator project home page [49].

The component technologies within CAT are described in detail in the following sections.

*1) Audio Server:* The audio server receives signals from the microphone or telephone and sends them to the speech recognizer. The server also sends synthesized or prerecorded speech to the PC speakers or telephone. The recording process is pipelined to the speech recognition server, and the play process is pipelined to the text-to-speech (TTS) server. Our telephony audio server supports barge-in using the Dialogic hardware platform [50], [51].

*2) Speech Recognizer:* Speech recognition plays an integral role in any perceptive animated agent interface. Our large vocabulary continuous speech recognition system, *Sonic*, was recently developed at CU [52], [53]. In addition to large vocabulary speech recognition, the recognizer has been developed to support both keyword/phrase spotting and constrained grammar-based speech recognition. The recognizer provides an integrated environment that incorporates voice activity detection (VAD) and speech enhancement as well as various feature and model-based speaker adaptation and normalization methods. The recognition architecture provides support for rapid portability to new languages. Sonic has been ported from English to the French, German, Italian, Japanese, Spanish, and Turkish [54] languages. It has also been trained on children's speech for use in interactive books, described later.

*3) Natural Language Parser:* We use the Phoenix parser [55] to map the speech recognizer outputs onto a sequence of semantic frames. Phoenix is designed for development of simple, robust natural language interfaces to applications, especially spoken language applications. Because spontaneous speech is often ill-formed and because the recognizer will make recognition errors, it is necessary that the parser be robust to errors in recognition, grammar, and fluency. This parser is designed to enable robust partial parsing of these types of input. Phoenix parses each input utterance into a sequence of one or more semantic frames.

A Phoenix frame is a named set of slots, where the slots represent related pieces of information. Each slot has an associated context-free semantic grammar that specifies word string patterns that can fill the slot. The grammars are compiled into recursive transition networks, which are matched against the recognizer output to fill slots. Each filled slot contains a semantic parse tree with the slot name as root. The developer must define a set of frames and provide grammar

rules that specify the word strings that can fill each slot in a frame.

In addition to producing a standard bracketed string parse, Phoenix also produces an extracted representation of the parse that maps directly onto the task concept structures. For example, the utterance "I want to go from Boston to Denver Tuesday morning" would produce the extracted parse

Air: [Origin].[City].Boston

Air: [Destination].[City].Denver

Air: [Date_Time].[Date].[Day_Name].tuesday

[Time_Range].[Period_Of_Day].morning

*4) Dialogue Manager:* The dialogue manager (DM) controls the system's interaction with the user and the application server. It is responsible for deciding what action the system will take at each step in the interaction. The DM is an event-driven server. It is normally in an idle state waiting for an input event. When it receives input from the Hub, it takes a set of actions, sends some frames to the Hub, and then returns to an idle state. The DM is responsible for several different functions.

1) Receiving parses from the parse server. This includes verification based on confidence assessment; ellipsis and anaphora resolution; clarification; and context update.
2) Sending natural language generation requests. This includes prompting for information; outputting information to the user; and clarification.
3) Generating database queries.
4) Receiving results from the database server.

The current context of the system is used to decide what to do next. The system does not use a dialogue network or a dialogue script, but rather a general engine operates on the semantic representations and the current context to control the interaction flow.

The basic data structures for representing domain information are frames. A frame has a name and a set of slots. Each slot is a concept hierarchy with the slot name as the root. Information is extracted from parses into frames and is stored in frames directly by the DM. Ideally, the concept structure in these frames is the same as produced in the parser extracts. In this case, the extraction from the parse to the DM frame is direct. A library of functions for manipulating frames is provided. DM frames are defined in the *task* file, which is similar to the *frames* file for the parser. The *task* file contains:

1) the definition of the system ontology (hierarchical concept structure of frames);
2) templates for prompting for information;
3) templates for confirming information;
4) templates for generating SQL queries.

This is a type of object-oriented mechanism in which the ways of prompting for and talking about information is stored in the frame with the information.

The "event driven" architecture functions similar to a production system. An incoming parse causes a set of actions, which modify the current context. After the parse has been integrated into the current context, the DM examines the context to decide what action to take next. The following actions are considered, in the order listed:

1) clarify if necessary;
2) sign off if all done;
3) retrieve data and present to user;
4) prompt user for required information.

The rules for deciding what to prompt for next are straightforward. The frame in focus is set to be the frame produced in response to the user, or to the last system prompt.

1) If there are unfilled required slots in the focus frame, then prompt for the highest priority unfilled slot in the frame.
2) If there are no unfilled required slots in the focus frame, then prompt for the missing piece of information with the highest priority in the given context.

Our mechanism does not have separate "user initiative" and "system initiative" modes. If the system has enough information to act on, then it does it. If it needs information, then it asks for it. The system does not require that the user respond with information relative to the system prompt. The user can respond with anything, and the system will parse the utterance and set the focus to the resulting frame. This allows the user to drive the dialogue, but does not require it. The system prompts are organized locally, at the frame level. The DM or user puts a frame in focus, and the system tries to fill it. This representation is easy to author; there is no separate dialogue control specification required. It is also robust in that it has a simple control that has no state to lose track of.

An additional benefit of the DM mechanism is that it is very largely declarative. The system developer creates a task file that specifies the system ontology and templates for communicating about nodes in the hierarchy. The templates are filled in from the values in the frames to generate output in the desired language.

*5) Database/Back-End:* During natural language interaction, the back-end processor receives SQL queries from the DM, interfaces to an SQL database, and retrieves data from the Web to enable learning tools to access online information. When a database request is received, the DM's SQL command is used to select records in local memory. If no records are found to match, the back-end can submit an HTTP-based request for the information via the Internet. Records returned from the Internet are then inserted as rows into the local SQL database, and the SQL statement is once again applied. Other modules may also be used to query the database (for example, a GUI may retrieve information related to a mouse click).

*6) Natural Language Generator:* The language generation module uses templates to generate words to speak back to the user based on dialogue speech acts. For example, in the CU Communicator travel-planning system, dialogue acts include "prompt" for prompting the user for needed information, "summarize" for summarization of flights, hotels, and rental cars, and "clarify" for clarifying information such as departure and arrival cities that share the same name. The natural language generator sends the resulting text to the speech synthesizer for playback to the user.

*7) Text-to-Speech Synthesizer:* The TTS synthesizer receives word strings from the natural language generator and synthesizes them into audio waveforms that can be played back to the user. Our current speech synthesizer servers make use of general-purpose TTS architectures such as the Festival speech synthesis system [56], the AT&T NextGen Synthesizer, as well as a domain-specific variable unit concatenative synthesizer currently used for the CU Communicator travel-planning system.

### C. Character Animator, Face Tracker, and Emotion Monitor

Under support from National Science Foundation Information Technology Research and Interagency Education Research Grants, additional modalities have been developed to enable conversational interaction with animated agents.

*1) Character Animator:* The character animation module receives a string of symbols (phonemes, animation control commands) with start and end times from the TTS server, and produces visible speech, facial expressions, and hand and body gestures in synchrony with the speech waveform. Our facial animation system, *CU Animate* [57], is a toolkit designed for research, development, control, and real-time rendering of 3-D animated characters. Eight engaging full-bodied characters and Marge, the dragon shown in Fig. 3, are included with the toolkit. Each character has a fully articulated skeletal structure, with sufficient polygon resolution to produce natural animation in regions where precise movements are required, such as lips, tongue, and finger joints. Characters produce lifelike visible speech, facial expressions, and gestures. CU Animate provides a GUI for designing arbitrary animation sequences. These sequences can be tagged (as icons representing the expression or movement) and inserted into text strings, so that characters will produce the desired speech and gestures while narrating text or conversing with the user.

Accurate visible speech is produced in CU Animate characters using a novel approach that uses motion capture data collected from markers attached to a person's lips and face while the person is saying words that contain all sequences of phonemes (or the visual configuration of the phonemes, called visemes) in their native language. The motion capture procedure produces a set of 8 points on the lips, each represented by an $x$, $y$, and $z$ coordinate, captured at 30 frames/sec. These sequences are stored as "diviseme" sequences, representing the transition from the middle of one visually similar phoneme class to the middle of another such class. To synthesize a new utterance, we identify the desired phoneme sequence to be produced (exactly as done in TTS synthesis systems), and then locate the corresponding sequences of viseme motion capture frames. Following procedures used to achieve audio diphone TTS synthesis, we concatenate sequences of divisemes—intervals of speech from the middle (most steady-state portion) of one phoneme to the middle of the following phoneme. By mapping the motion capture points from these concatenated sequences to the vertices of the polygons on the lips and face of the 3-D model, we can control the movements of the lips of the 3-D model to mimic the movements of the original speaker when producing the divisemes within words. This approach produces natural-looking visible speech, which we are now evaluating relative to videos of human talkers.

**Fig. 3.** Characters currently used in CU animate. Eight of the characters were designed by Sherer digital automation.

*2) Face Tracker:* A face tracking system was developed by Movellan and his colleagues at the Machine Perception Lab at the University of California, San Diego, to track faces in real time (at 30 frames/sec) under arbitrary illumination conditions and backgrounds (which may include moving objects). The system combines both color and gray-scale information. The main advantages of color-based trackers are that they are resistant to changes in pose (e.g., in-depth rotations, facial expressions) while requiring minimal computational resources. Unfortunately, color-based systems have several major problems: they are sensitive to illumination conditions and to the presence of flesh-colored backgrounds or clothing; the algorithms are local and, thus, tend to lose the face when large movement occurs; and the methods for estimating the scale of the face tend to be *ad hoc* and poorly motivated. In order to address these problems, a system was developed that integrated color-based face tracking and feature-based face detection within the same theoretical framework [58].

The color-based tracker analyzes on the order of 30 000 hypotheses per frame about the location and scale of faces. This is achieved by utilizing a bank of integral images [58]. The speed of the algorithm enables it to perform a global search over the entire image plane and to jointly estimate scale and location within a maximum-likelihood framework. The feature-based system was trained on a database of 5000 frontal, upright faces provided by Compaq, and a database of millions of background image patches taken from the Web. The system is very robust to changes in illumination and background conditions. Its main limitation at present is that it can only detect frontal views of upright faces. The color-based and feature-based systems run in parallel on different threads. The color-based system can run at 30 frames/sec, consuming very little computational resources. The frequency of operation of the feature-based system

depends on computational resources. Its main role is to update the statistics of the face and background color model, and to provide additional information about the likely locations of faces. The face detector achieved state-of-the-art performance on the standard CMU dataset [59].

The face detector communicates the location of the user's face to the animation server which, by triangulating between the user, camera, and animated agent, allows the animated agent's eyes to track the user.

*3) Emotion Monitor:* The Emotion Monitor, also developed at the Machine Perception Lab, is a prototype system that classifies facial expressions into seven emotion dimensions: neutral, angry, happy, disgusted, fearful, sad, and surprised. The system will be integrated into Galaxy/Communicator in the near future. When the feature-based face detector finds a face, the image is sent to a bank of 40 Gabor filters at eight orientations and five spatial frequencies. The filter bank representation is compressed using a bank of 21 SVM classifiers. The output of these classifiers is then converted into a probability distribution over seven emotional dimensions using a multinomial regression model. The system was trained and tested on Cohn and Kanade's DFAT-504 dataset [60]. It achieved intersubject generalization performance of 91.5% correct on the seven-category classification task. Demonstrations of the system can be found at the Machine Perception Laboratory Web site [61]. Performance in unconstrained images sent to the Web server is about 80% correct.

### D. Summary

Taken together, the Galaxy architecture, CU CAT, CU Animate, and computer vision technologies provide a set of foundational tools and technologies that can be used to develop perceptive animated interfaces. Most of these tools and technologies are either freely available now [62], [63] or will soon be available to university researchers and educators for noncommercial use. In the next section, we describe a testbed for using these tools and technologies to research and develop perceptive animated interfaces.

## V. Colorado Literacy Tutor: A Testbed for Research and Development of Perceptive Animated Interfaces

The CLT is a technology-based literacy program, based on cognitive theory and scientifically based reading research, which aims to improve literacy and student achievement in public schools. The goal of the CLT is to provide computer-based learning tools that will improve student achievement in any subject area by helping students learn to read fluently, to acquire new knowledge through deep understanding of what they read, to make connections to other knowledge and experiences, and to express their ideas concisely and creatively through writing. A second goal is to scale up the program to both state and national levels in the United States by providing accessible, inexpensive, and effective computer-based learning tools that are easy to use and require little or no learning curve by teachers or students.

A key feature of the CLT is the use of leading-edge human communication technologies in learning tasks, as described later. The program is, thus, an ideal testbed for research and development of perceptive animated agents that integrate auditory and visual behaviors during face-to-face conversational interaction with human learners. The program enables us to evaluate component technologies with real users—students in classrooms—and to evaluate how the integration of these technologies into learning tools incorporating perceptive animated agents affects learning using standardized assessment tools.

In the remainder of this section, we describe the learning tools used in the CLT, and the ways in which they use spoken dialogue system technologies, computer vision technologies, character animation technologies, and natural language understanding technologies.

### A. CLT Components

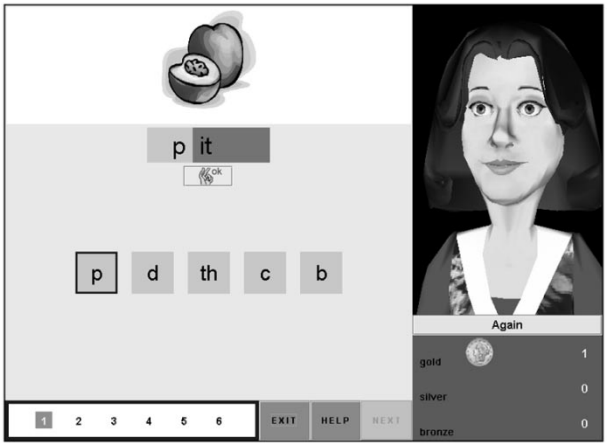
The CLT consists of five tightly integrated components: 1) managed learning environment (MLE); 2) assessment tools; 3) foundational reading skills tutors; 4) interactive books; and 5) Summary Street comprehension training. In addition, the project devotes significant effort to research on evaluating learning outcomes and designing a scalable and sustainable program.

*1) Managed Learning Environment:* All student activities are organized and displayed within an MLE. The MLE logs all student and system behaviors within the program, and displays progress graphics that shows individual and aggregate student performance aligned to district, state, and national learning goals at each grade level.

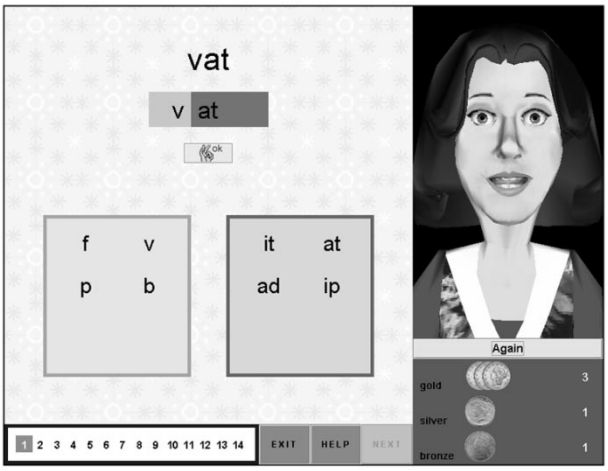
*2) Assessment Tools:* These tools provide a rapid assessment of a student's reading level, alphabet knowledge, and phonological awareness. The assessment tools are designed to identify younger students who may have learning problems and, thus, require focused instruction to acquire skills that underlie reading.

*3) Foundational Reading Skills Tutors:* These tutors provide a sequence of focused exercises in which the animated agent interacts with students to learn and practice foundational reading skills in several domains: alphabet knowledge, phonological awareness, letter-to-sound decoding, recognizing common sight words, understanding syllable structure, spelling, and vocabulary training. In a variety of engaging exercises in each domain, the tutor presents instructions, provides hints, and gives feedback and encouragement to the learner in response to mouse clicks, speech, or typed input. Fig. 4 shows examples of foundational reading skills tutors. Foundational skills tutors are presented within an automated study plan that adapts to each student's performance, and which is closely integrated with interactive books.

*4) Interactive Books:* Interactive books are the main platform for research and development of perceptive animated agents. Fig. 5 shows a page of an interactive book. Interactive books incorporate all of the spoken dialogue, language processing, computer vision, and computer animation technologies described in the previous section to enable




(a)




(b)

**Fig. 4.** Examples of foundational reading skills tutors with animated teacher shown in top right-hand corner. (a) Beginning sounds. (b) Rhyme changing.

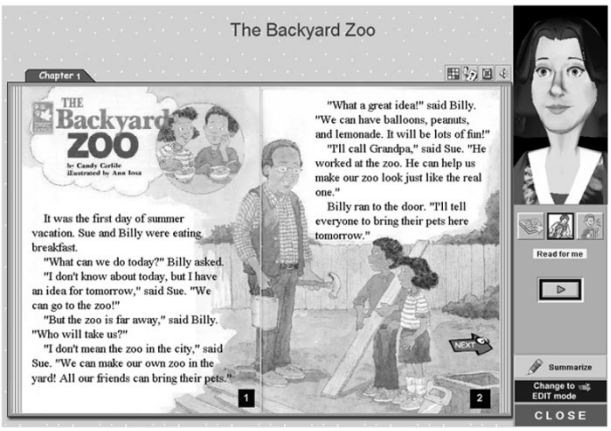



**Fig. 5.** An interactive book with animated teacher in top right corner.

natural face-to-face conversational interaction with users. Fig. 6 shows this integration within a client-server architecture that provides a platform-independent user interface for Web-based delivery of multimedia learning tools [64].

Interactive book authoring tools are designed for easy use by project staff, teachers, and students to enable authors to design and format books by combining text, images, videos, and animated characters. Once text and illustrations have been imported or input into the authoring environment, authors can orchestrate interactions between users, animated characters, and media objects. Developers can populate illustrations (digital images) with animated characters, and cause them to converse with each other, with the user, or speak or sing their parts in the stories using naturally recorded or synthetic speech. A markup language enables authors to control characters' facial expressions and gestures while speaking. The authoring tools also enable authors to prerecord sentences and/or individual words in the text as well as utterances to be produced by animated characters when narrating text or during conversations with users. This gives users the flexibility to let the animated agents speak in their voice while maintaining synchronized lip movement.

The authoring tools enable a wide range of user and system behaviors within interactive books, including having the story narrated by one or more animated characters (while controlling their facial expressions and gestures), having users converse with animated characters in structured or mixed-initiative dialogues, having the student read out loud while words are highlighted, clicking on words to have them spoken by the agent, interacting with the agent to sound out the word, having the student respond to questions posed by the agent either by clicking on objects in images or saying or typing responses, and having the student produce typed or spoken summaries which are analyzed for content using language processing techniques.

Read-aloud feedback involves following along as text is read, highlighting the read text, monitoring reading fluency and verifying pronunciation accuracy. Various feedback options are possible, such as display of just the current paragraph being read, highlighting of the read text, and a pointer that follows the current reading position. Read-aloud feedback is obtained by building a language model for the book, getting partial phrases from the speech recognizer as the user is reading, determining the current reading location using the partial phrase and an efficient Viterbi search through the book, and aligning the partial phrase with the book text using a dynamic programming search. In order to allow for skipping, the Viterbi search finds the words that when strung together minimize a weighted cost function of adjacent word proximity and distance from the reader's last active reading location. The dynamic programming search has constraints to account for boundary effects at the ends of the partial phrase.

*5) Comprehension Training:* Comprehension training uses Summary Street [65], a program developed at CU by W. Kintsch and his associates to train students to achieve deep comprehension of text (e.g., the author's intent, inferences, cause and effect, relation to world knowledge). The program applies latent semantic analysis (LSA), a text processing technique, developed by Deerwester *et al.* [66] and Landauer and Dumais [67], to grade student essays by comparing them directly to text, or to graded essays, or to

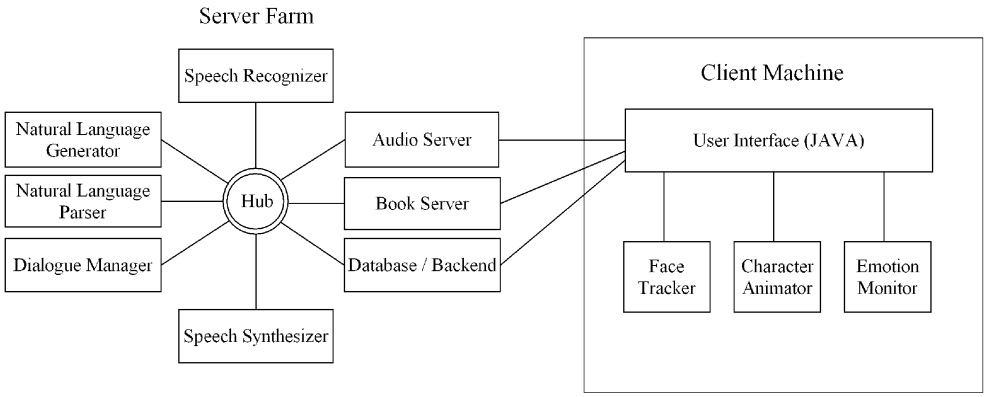



**Fig. 6.** Interactive book architecture as a client/server implementation. The client and server can reside on separate machines or on the same machine. Interactive books reside on the client side within a Java-based user interface (UI). Most of the modules run in a server farm, with the exception of the face tracker, character animator (CU animate), and emotion monitor, which run on the client side, making possible a tight and efficient integration with the UI. The client-side modules communicate directly with the Java layer, which in turn communicates with the Hub via a book server. The UI provides multimodal input and output (text, speech, graphics, video, movement, and animation), platform independence, and an easy-to-use authoring environment.

one or more "golden essays." After reading an exposition or story, students are instructed to type a summary of the text. Summary Street grades the essay and provides the student with meaningful visual feedback about the presence or absence of critical information (e.g., subthemes) in the summary. The program can then help the student acquire missing information by directing them to relevant sections of the text, and interact with students to combine redundant sections of prose and eliminate irrelevant sentences, in order to make their summaries more concise. Students who used Summary Street in high school classes doubled the time they spent on their assignments, and improved their achievement by a full letter grade [64], [68]. A Summary Street tutorial can be found online[1] [69]. Interactive books provide a digital environment for incorporating Summary Street, and for extending comprehension training to children who cannot read well or type. In this scenario, the animated character narrates a story, and then asks the child to summarize the story in her own words. The child will then produce a spoken summary, which is transcribed using automatic continuous speech recognition and then graded using LSA. The animated character will then interact with the student to revisit parts of the story containing information missing from the summary, and interact with the student to revise the summary.

### *B. Summary*

Foundational skills reading tutors and interactive books were introduced into a dozen kindergarten and first grade classrooms in Boulder, CO, during the spring of 2003. The initial deployment was limited to interaction with animated characters that speak and emote through mouse clicks by the students. At the time of this writing, the speech recognition, face recognition and emotion classification systems have been integrated into the interactive books architecture, and our new generation of perceptive animated interfaces will be introduced into classrooms in the fall of 2003.

[1]http://www.colit.org/demos.html

## VI. Research Challenges

Building systems that enable face-to-face communication with intelligent animated agents requires a deep understanding of the auditory and visual behaviors that individuals produce and respond to while communicating with each other. Face-to-face conversation is a virtual ballet of auditory and visual behaviors—a *pas de deux* of signals and cues—with the speaker and behavior simultaneously producing and reacting to each other's sounds and movements. While talking, the speaker produces speech annotated by smiles, head nods, and other gestures. At the same time, the listener provides simultaneous auditory and visual feedback to the speaker (e.g., "I agree"; "I'm puzzled"; "I want to speak"). For example, the listener may signal the speaker that she desires to speak; the speaker continues to talk, but acknowledges the nonverbal communication by raising his hand and smiling in a "wait just a moment" gesture. Face-to-face conversation is often characterized by such simultaneous auditory and visual exchanges, in which the sounds of our voices, the visible movements of our articulators, direction of gaze, facial expressions, and head and body movements present linguistic information, paralinguistic information (emotions, sarcasm, spatial referents, etc.), and communication about the conversation itself (agreement, turn taking, etc.).

Inventing systems that engage users in accurate and graceful face-to-face conversational interaction is a challenging task. The system must simultaneously interpret and produce auditory and visual signals in real time while preserving the timing relationships between perception and production appropriate to conversational interaction. The

system must interpret the user's auditory and visible speech, eye movements, facial expressions, and gestures, since these cues combine to signal the speaker's intent—e.g., a head nod can clarify reference, whereas a shift of gaze can indicate that a response is expected. Paralinguistic information is also critical, since the prosodic contour of the auditory signal or a visual cue such as rolling the eyes may signal that the user is being sarcastic. The animated agent must also produce accurate, natural, and expressive auditory and visible speech with facial expressions and gestures appropriate to the physical nature of language production, the context of the dialogue, and the goals of the task. Most important, the animated interface must combine perception and production to interact conversationally in real time—while the animated agent is speaking, the system must interpret the user's auditory and visual behaviors to detect agreement, confusion, desire to interrupt, etc., and while the user is speaking, the system must both interpret the user's speech and simultaneously provide auditory and/or visual feedback via the animated character.

To develop lifelike computer characters imbued with unique and credible personalities and capable of natural and graceful face-to-face dialogues with users, new research is needed to gain a deeper understanding of the signals and cues exchanged during face-to-face communication, and research is needed to use this knowledge to develop human communication systems that model these behaviors. By applying this knowledge to improved machine perception and generation technologies, research will lead to a new generation of perceptive animated interfaces.

## VII. Conclusion: It Takes a Community

We have argued that perceptive animated interfaces can be realized in the near future by leveraging existing infrastructure and expertise in areas of human language technologies and by extending spoken dialogue systems to incorporate computer vision (face tracking, eye tracking, expression recognition, gesture recognition, etc.) and computer animation systems. But developing perceptive animated interfaces requires far more than advancing and integrating technologies within systems in these areas. Developing perceptive animated interfaces requires a diverse community of researchers working together, motivated by strong competing theories of cognition, communication, and learning, sharing common infrastructure and measuring progress on well-defined tasks.

The importance of a coherent and focused research community, with shared research goals and infrastructure, is illustrated by progress in spoken language systems. In the past 25 years, spoken language systems have progressed from recognition of discrete utterances produced by specific individuals to commercial deployment of graceful, speaker-independent conversational interfaces. Progress in spoken language systems can be attributed to sustained funding from federal research agencies and industrial labs, and to the efforts of a dedicated community of researchers who have worked together to establish common research goals, to define task domains and performance objectives, to develop and share critical infrastructure, and to define and apply rigorous evaluation methodologies to a set of increasingly challenging task domains.

Inventing perceptive animated interfaces also requires the combined efforts of a community of dedicated researchers with shared research objectives, accessible research tools and corpora, common task domains, benchmark systems, and evaluation metrics for measuring progress. At present, no such community exists. Fortunately, there are many significant efforts at laboratories worldwide focused on understanding, classifying, and recognizing emotions when people listen and speak; on understanding hand and body gestures during speech communication; and on improving speech recognition and speech production by integrating auditory and visual modalities. Unfortunately, researchers in each of these areas tend to move in their own circles and often meet at small workshops run in collaboration with larger, more established conferences. By bringing theorists, researchers, and technologists from these communities together, new ideas, experiments, and theories will emerge from considering fundamental issues that cut across traditional disciplines and that are not currently addressed by computer science and engineering, linguists, cognitive science, or psychology.

## Appendix

The CSLR at CU is developing the interactive reading tutors described here; the speech and language technologies in the CU Communicator spoken dialogue system; and the CU Animate system, a research and authoring environment for real-time animation of full bodied characters that speak, gesture and emote. Computer vision technologies, including face tracking, gaze tracking and emotion classification are being developed by J. Movellan, M. S. Bartlett, and J. Hershey at the Machine Perception Laboratory at the University of California, San Diego. Research on improving animation and integrating animation and computer vision is conducted through close collaboration between the University of California, San Diego, and CU. The Institute for Cognitive Science at CU is developing language processing technologies (e.g., improvements to LSA) in the context of interactive learning system for comprehension training and essay generation. The Boulder Valley School District Department of Special Education, headed by Dr. J. Riordan, works closely with the project team to conduct participatory design activities, integrate reading tutors into classrooms, and enable evaluation of the tools. Prof. W. Kintsch, Director of the Institute for Cognitive Science, is Principal Investigator of the CLT.

## Acknowledgment

The authors would like to thank the following people for their invaluable contributions to the reading tutor project: D. Caccamise, L. Corson, T. Durham, W. Kintsch, E. Kintsch, N. Ngampatipatpong, L. Snyder, T. Streumph, J. Tuantranont, and B. Wise.

## REFERENCES


[1] D. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press, 1998.

[2] D. Massaro and M. Cohen, "Speech perception in hearing-impaired perceivers: Synergy of multiple modalities," *J. Speech, Lang., Hear. Sci.*, vol. 42, pp. 21–41, 1999.

[3] L. Barker, "Computer-assisted vocabulary acquisition: The CSLU vocabulary tutor in oral-deaf education," *J. Deaf Stud. Deaf Educ.*, vol. 8, no. 2, pp. 187–198, 2003.

[4] F. Thomas and O. Johnston, *Disney Animation: The Illusion of Life*. New York: Hyperion, 1995.

[5] B. Reeves and C. Nash, *The Media Equation: How People Treat Computers, Televisions and New Media Like Real People and Places*. New York: Cambridge Univ. Press, 1996.

[6] J. Allbeck and N. Badler, "Toward representing agent behaviors modified by personality and emotion," presented at the 1st Int. Joint Conf. Autonomous Agents and Multi-Agent Systems, Bologna, Italy, 2002.

[7] N. Badler, J. Allbeck, L. Zhao, and M. Byun, "Representing and parameterizing agent behaviors," in *Proc. Computer Animation*, 2002, pp. 133–143.

[8] J. Cassell, H. Vilhjalmsson, and T. Bickmore, "Beat: The behavior expression animation toolkit," in *Proc. ACM SIGGRAPH*, 2001, pp. 477–486.

[9] J. Gratch and S. Marsella, "Tears and fears: Modeling emotions and emotional behaviors in synthetic agents," in *Proc. Autonomous Agents*, 2001, pp. 278–285.

[10] A. B. Loyall, "Believable agents: Building interactive personalities," Ph.D. dissertation, Dept. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, 1997.

[11] S. Marcella and J. Gratch, "Modeling the interplay of emotions and plans in multi-agent simulations," presented at the 23rd Annu. Conf. Cognitive Science Society, Edinburgh, U.K., 2001.

[12] J. Gratch, J. Rickel, E. André, N. Badler, J. Cassell, and E. Petajan, "Creating interactive virtual humans: Some assembly required," *IEEE Intell. Syst.*, vol. 17, pp. 54–63, July/Aug. 2002.

[13] W. L. Johnson, J. W. Rickel, and J. C. Lester, "Animated pedagogical agents: Face-to-face interaction in interactive learning environments," *Int. J. Artif. Intell. Educ.*, vol. 11, pp. 47–78, 2000.

[14] J. Cassell, T. Stocky, T. Bickmore, Y. Gao, Y. Nakano, K. Ryokai, D. Tversky, C. Vaucelle, and H. Vilhjálmsson, "MACK: Media lab autonomous conversational kiosk," presented at the Imagina: Intelligent Autonomous Agents, Monte Carlo, Monaco, 2002.

[15] J. Gustafson, "Developing multimodal spoken dialogue systems—empirical studies of spoken human-computer interactions," Ph.D. dissertation, Dept. Speech, Music and Hearing, Royal Inst. Technol., , Stockholm, Sweden, 2002.

[16] M. Blomberg, R. Carlson, K. Elenius, B. Granström, J. Gustafson, S. Hunnicutt, R. Lindell, and L. Neovius, "An experimental dialogue system: WAXHOLM," in *Proc. Eurospeech*, vol. 3, 1993, pp. 1867–1870.

[17] J. Bertenstam, M. Blomberg, R. Carlson, K. Elenius, B. Granström, J. Gustafson, S. Hunnicutt, J. Högberg, R. Lindell, L. Neovius, A. de Serpa-Leita, L. Nord, and N. N. Ström, "The Waxholm system—A progress report," in *Proc. Spoken Dialogue Systems*, Vigsø, Denmark, 1995, pp. 81–84.

[18] J. Gustafson, M. Lundeberg, and J. Liljencrants, "Experiences from the development of August—a multimodal spoken dialogue system," presented at the IDS'99, Workshop Interactive Dialogue in Multimodal Systems, Kloster Irsee, Germany, 1999.

[19] M. Lundeberg and J. Beskow, "Developing a 3D-agent for the August dialogue system," presented at the Audio-Visual Speech Processing Conf., Santa Cruz, CA, 1999.

[20] J. Gustafson, L. Bell, J. Boye, J. Edlund, and M. Wiren, "Constraint manipulation and visualization in a multimodal dialogue system," presented at the ISCA Workshop Multi-Modal Dialogue Mobile Environments, Kloster Irsee, Germany, 2002.

[21] L. Barker and T. Weston. (2003) Designing, implementing, and assessing Web-based learning modules. [Online]. Available: http://www.colorado.edu/ATLAS/evaluation/papers_fld/module_pg.htm

[22] J. Cassell, "Toward a model of technology and literacy development: Story listening systems," *Appl. Development. Psych.*, to be published.

[23] R. Cole, T. Carmell, P. Connors, M. Macon, J. Wouters, J. de Villiers, A. Tarachow, D. Massaro, M. Cohen, J. Beskow, J. Yang, U. Meier, A. Waibel, P. Stone, G. Fortier, A. Davis, and C. Soland, "Intelligent animated agents for interactive language training," presented at the ESCA Workshop Speech Technology in Language Learning, Stockholm, Sweden, 1998.

[24] R. Cole, D. Massaro, J. de Villiers, B. Rundle, K. Shobaki, J. Wouters, M. Cohen, J. Beskow, P. Stone, P. Connors, A. Tarachow, and D. Solcher, "New tools for interactive speech and language training: Using animated conversational agents in the classrooms of profoundly deaf children," in *Proc. ESCA/SOCRATES Workshop Method and Tool Innovations Speech Science Education*, London, U.K., 1999, pp. 45–52.

[25] A. Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, and R. Kreuz, "AutoTutor: A simulation of a human tutor," *J. Cognitive Syst. Res.*, vol. 1, pp. 35–51, 1999, to be published.

[26] S. Marsella, "Pedagogical soap," in *Socially Intelligent Agents: The Human in the Loop, AAAI Fall 2000 Symp.*, K. Dautenhahn, Ed., 2000, pp. 107–112.

[27] K. Ryokai, C. Vaucelle, and J. Cassell, "Virtual peers as partners in storytelling and literacy learning," *J. Comput. Assisted Learn.*, vol. 19, no. 2, pp. 195–208, 2003.

[28] P. Stone, "Revolutionizing language instruction in oral deaf education," presented at the Int. Conf. Phonetic Sciences, San Francisco, CA, 1999.

[29] B. Bloom, "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring," *Educ. Res.*, vol. 13, pp. 4–16, 1984.

[30] P. Cohen, J. Kulik, and C. Kulik, "Educational outcomes of tutoring: A meta-analysis of findings," *Amer. Educ. Res. J.*, vol. 19, pp. 237–248, 1982.

[31] A. Graesser, N. Person, and J. Magliano, "Collaborative dialog patterns in naturalistic one-on-one tutoring," *Appl. Cognitive Psychol.*, vol. 9, pp. 359–387, 1995.

[32] D. Dehn and S. van Mulken, "The impact of animated interface agents: A review of empirical research," *Int. J. Human-Comput. Stud.*, vol. 52, pp. 1–22, 2000.

[33] A. Graesser, K. Moreno, and J. Marineau, "AutoTutor improves deep learning of computer literacy: Is it the dialog or the talking head?," presented at the 2003 Int. Conf. Artificial Intelligence Education, Sydney, Australia.

[34] A. Graesser, M. Ventura, G. Jackson, J. Mueller, X. Hu, and N. Person, "The impact of conversational navigational guides on the learning, use, and perceptions of users of a web site," presented at the AAAI Spring Symp. Agent-Mediated Knowledge Management, Menlo Park, CA, 2003.

[35] A. Graesser, K. VanLehn, C. Rose, P. Jordan, and D. Harter, "Intelligent tutoring systems with conversational dialogue," *AI Mag.*, vol. 22, pp. 39–51, 2001.

[36] N. Person, A. Graesser, L. Bautista, E. Mathews, and TRG, "Evaluating student learning gains in two versions of AutoTutor," in *Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future*, J. Moore, C. Redfield, and W. Johnson, Eds. Amsterdam, The Netherlands: OIS, 2001, pp. 286–293.

[37] T. Bickmore and J. Cassell, "Social dialogue with embodied conversational agents," in Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems, J. van Kuppevelt, L. Dybkjaer, and N. Bernsen, Eds. New York: Kluwer, to be published.

[38] R. Cole, "Tools for research and education in speech science," presented at the Int. Conf. Phonetic Sciences, San Francisco, CA, 1999.

[39] S. Sutton, D. Novick, R. Cole, and M. Fanty, "Building 10 000 spoken-dialogue systems," presented at the Int. Conf. Spoken Language Processing, Philadelphia, PA, 1996.

[40] R. Cole, S. Sutton, Y. Yan, P. Vermeulen, and M. Fanty, "Accessible technology for interactive systems: A new approach to spoken language research," presented at the Int. Conf. Acoustics, Speech and Signal Processing, Seattle, WA, 1998.

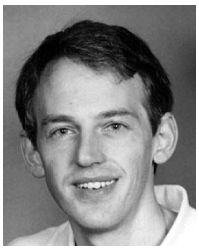
[41] S. Sutton and R. Cole, "Universal speech tools: The CSLU toolkit," in *Proc. Int. Conf. Spoken Language Processing*, 1998, pp. 3221–3224.

[42] P. Connors, A. Davis, G. Fortier, K. Gilley, B. Rundle, C. Soland, and A. Tarachow, "Participatory design: Classroom applications and experiences," presented at the Int. Conf. Phonetic Sciences, San Francisco, CA, 1999.

[43] J. Payson. (2001) Look who's talking. Prime Time Thursday, ABC Television Network. [Online]. Available: http://oak.colorado.edu/~spradhan/download/Ron-Videos/ABC-Primetime
[44] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, "Galaxy-II: A reference architecture for conversational system development," in *Proc. Int. Conf. Spoken Language Processing*, vol. 3, 1997, pp. 931–934.
[45] S. Bayer, C. Doran, and B. George, "Exploring speech-enabling dialogue with the Galaxy communicator infrastructure," in *Proc. Human Language Technology Conf.*, 2001, pp. 114–116.
[46] W. Ward and B. Pellom, "The CU Communicator system," presented at the IEEE Workshop Automatic Speech Recognition and Understanding, Keystone, CO, 1999.
[47] B. Pellom, W. Ward, and S. Pradhan, "The CU Communicator: An architecture for dialogue systems," presented at the Int. Conf. Spoken Language Processing, Beijing, China, 2000.
[48] B. Pellom, W. Ward, J. Hansen, K. Hacioglu, J. Zhang, X. Yu, and S. Pradhan, "University of Colorado dialog systems for travel and navigation," presented at the Human Language Technology Conf., San Diego, CA, 2001.
[49] (2002) Dialog System for Travel and Navigation. CU Communicator System. [Online]. Available: http://communicator.colorado.edu
[50] J. Zhang, W. Ward, B. Pellom, X. Yu, and K. Hacioglu, "Improvements in audio processing and language modeling in the CU Communicator," in Proc. of Eurospeech, Aalborg, Denmark, 2001.
[51] J. Zhang, W. Ward, and B. Pellom, "Phone based voice activity detection using online Bayesian adaptation with conjugate normal distributions," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, 2002, pp. I-321–I-324.
[52] B. Pellom, "Sonic: The University of Colorado continuous speech recognizer," Center for Spoken Language Research, Univ. Colorado, Boulder, Tech. Rep. No. TR-CSLR-2001–01, 2001.
[53] B. Pellom and K. Hacioglu, "Recent improvements in the Sonic ASR system for noisy speech: The SPINE task," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2003, pp. I-4–I-7.
[54] O. Salor, B. Pellom, T. Ciloglu, K. Hacioglu, and M. Demirekler, "On developing new text and audio corpora and speech recognition tools for the Turkish language," presented at the Int. Conf. Spoken Language Processing, Denver, CO, 2002.
[55] W. Ward, "Extracting information from spontaneous speech," presented at the Int. Conf. Spoken Language Processing, Tokyo, Japan, 1994.
[56] P. Taylor, A. Black, and R. Caley, "The architecture of the festival speech synthesis system," in *Proc. 3rd ESCA Workshop Speech Synthesis*, 1998, pp. 147–151.
[57] J. Ma, J. Yan, and R. Cole, "CU animate: Tools for enabling conversations with animated characters," presented at the Int. Conf. Spoken Language Processing, Denver, CO, 2002.
[58] J. Movellan, J. Susskind, and J. Hershey, "Global search, real time tracking of non-rigid objects using integral banks of integral images," Machine Perception Lab., Univ. California, San Diego, Tech. Rep. No. TR 2002.04, 2002.
[59] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 23–38, Jan. 1998.
[60] J. F. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade, "Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding," *Psychophysiology*, vol. 36, pp. 35–43, 1999.
[61] Emotion tracker (2002). [Online]. Available: http://markov.ucsd.edu/~movellan/mplab
[62] Galaxy Communicator download (2002). [Online]. Available: http://communicator.sourceforge.net
[63] CU Communicator download (2002). [Online]. Available: http://communicator.colorado.edu
[64] ITR reading tutor project (2002). [Online]. Available: http://cslr.colorado.edu/beginweb/reading/reading.html
[65] D. Steinhart, "Summary street: An intelligent tutoring system for improving student writing through the use of latent semantic analysis," Ph.D. dissertation, Dept. Psychol., Univ. Colorado, Boulder, 2001.
[66] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inform. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.
[67] T. Landauer and S. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge," *Psychol. Rev.*, vol. 104, pp. 211–240, 1997.
[68] E. Kintsch, D. Steinhart, G. Stahl, C. Matthews, R. Lamb, and LRG, "Developing summarization skills through the use of LSA-based feedback," *Interactive Learn. Environ.*, vol. 8, pp. 87–109, 2000.
[69] Colorado literacy tutor (2002). [Online]. Available: http://www.colit.org

**Ronald Cole** received the B.A. degree in psychology from the University of Rochester, Rochester, NY, in 1967 and the Ph.D. in psychology from the University of California, Riverside, in 1971.

From 1970 to 1975, he was an Assistant Professor at the University of Waterloo, Waterloo, ON, Canada. In 1973, he was Visiting Professor at the University of Tel Aviv, Tel Aviv, Israel. From 1975 to 1988, he was Associate Professor and Senior Project Scientist at Carnegie Mellon University, Pittsburgh, PA. From 1988 to 1998, he was a Professor at the Oregon Graduate Institute, Beaverton, where he founded the Center for Spoken Language Understanding. Since 1998, he has been with the University of Colorado, Boulder, where he cofounded the Center for Spoken Language Research. He was Editor-in-Chief of *Survey of the State of the Art of Human Technology* (Cambridge, MA: Cambridge University Press, 1997," and Coauthor of "The challenge of spoken language systems: Research directions for the nineties" (*IEEE Trans. Speech Audio Processing*, vol. 1, pp. 1-21, Jan. 1995).

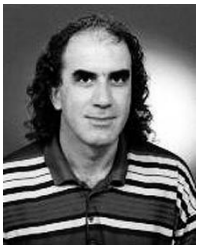


**Sarel van Vuuren** received the Ph.D. degree in electrical and computer engineering from the Oregon Graduate Institute of Science and Technology, Beaverton, in 1999, where he was a Fulbright Scholar (1994–1998).

In 1996, he was with Digital Equipment Corporation. From 1999 to 2001, he was with SpeechWorks International. He is currently Head of Research and Development for the Interactive Book and Literacy Tutor project at the Center for Spoken Language Research, University of Colorado, Boulder. He regularly acts as consultant to academia and industry. His current research interest is multimodal human-computer interfaces, involving studies in machine learning, signal processing, speaker and pronunciation verification, semantic analysis, computing architectures, and real-time animation.

**Bryan Pellom** (Member, IEEE) received the B.Sc. degree in computer and electrical engineering from Purdue University, West Lafayette, IN, in 1994 and the M.Sc and Ph.D. degrees in electrical engineering from Duke University, Durham, NC, in 1996 and 1998, respectively.

From 1999 to 2002, he was a Research Associate with the Center for Spoken Language Research (CSLR), University of Colorado, Boulder. His research activities were focused on automatic speech recognition, concatenative speech synthesis, and spoken dialog systems. Since 2002, he is a Research Assistant Professor in the Department of Computer Science and with the CSLR. His current research is focused in the area of large vocabulary speech recognition. In 2002, he was Technical Chair for the International Conference on Spoken Language Processing (ICSLP).

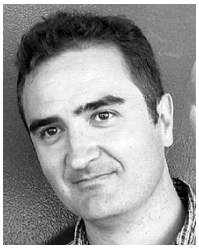


**Kadri Hacioglu** was born in Nicosia, Cyprus. He received the B.Sc., M.Sc., and Ph.D. degrees in electrical and electronic engineering from the Middle East Technical University, Ankara, Turkey, in 1980, 1984, and 1990, respectively.

After his two-year military service, in 1992, he joined the faculty of Eastern Mediterranean University, Magosa, North Cyprus, as an Assistant Professor, and became an Associate Professor in 1997. While there, he taught several classes on electronics, digital communications, speech processing and neural networks. During this time, he conducted research on applying fuzzy logic, neural networks, and genetic algorithms to signal processing and communications problems. From 1998 to 2000, he was a Visiting Professor in the Department of Computer Science, University of Colorado, Boulder. Here, he taught classes on neural networks and continued his research. Since 2000, he has been a Research Associate at the Center for Spoken Language Research, University of Colorado. He has authored or coauthored numerous papers and supervised a dozen M.Sc./Ph.D. theses. His current research interests are concept-based language modeling, speech understanding, natural language generation, and search methods in speech recognition/understanding. He also does research on multiuser detection and equalization in CDMA systems.

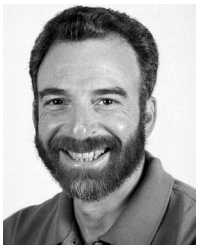
**Jiyong Ma** (Member, IEEE) received the B.S. degree in computational mathematics and the M.S. degree in thermal physics from Heilongjiang University, Harbin, China, in 1984 and 1988, respectively, and the Ph.D. degree from the Department of Computer Science, Harbin Institute of Technology, Harbin, China, in 1999.

From 1999 to 2001, he was a Postdoctoral Researcher in the Institute of Computing Technology, Chinese Academy of Sciences. He is currently with the Center for Spoken Language Research, University of Colorado, Boulder. He has published over 60 papers in scientific journals and conference proceedings. His current research is focused on multimodal intelligent computer interfaces, including computer vision and pattern recognition, computer animation, sign language recognition, and speaker and speech recognition.



**Javier Movellan** was born in Palencia, Spain. He received the B.S. degree from the Universidad Autonoma de Madrid, Madrid, Spain and the Ph.D. degree from the University of California, Berkeley, in 1989, where he was a Fulbright Scholar.

From 1989 to 1993, he was a Research Associate at Carnegie Mellon University, Pittsburgh, PA. From 1993 to 2001, he was an Assistant Professor at the University of California, San Diego (UCSD). He is currently an Associate Scientist at the Institute for Neural Computation and Director of the Machine Perception Laboratory, UCSD. He has been studying learning and perception by human and machine for more than 15 years. His work span studies in probability theory, machine learning, machine perception, experimental psychology, robotics, and developmental psychology. He founded UCSD's Machine Perception Laboratory in 1997 with the goal of understanding human and machine perception by developing machine perception systems that combine multiple sources of information (e.g., audio and video) and interact naturally with people, reading their lips, recognizing their facial expressions, and making inferences about cognitive and affective states. He is Founder of the Kolmogorov project, which provides a collection of open source tutorials and software on topics related to machine learning, machine perception, and statistics.



**Scott Schwartz** received the B.A. degree in speech and hearing sciences from the State University of New York, New Paltz, in 1980 and the M.A. degree in 1982 in speech and hearing from Ohio University and his Ph.D. in communication disorders at the University of Wisconsin-Madison.

He has worked as a Speech Language Pathologist in the public schools in Boston, MA, and. Madison, WI. Research and publications from his work in Madison, WI, are in the area of language development in children with Down's syndrome. He is currently a Speech Language Pathologist in the public schools in Boulder, CO, and the Liaison between the Boulder School District and the Center for Spoken Language Research at the University of Colorado, Boulder. In the Boulder, CO, school district Scott has also served as a special education and autism consultant. He also teaches at the University of Colorado in the area of phonological disorders in children and autism spectrum disorders.

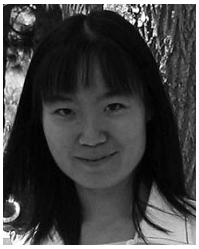
Dr. Schwartz is a member of the American Speech and Language Association and the Autism Society of America.

**David Wade-Stein** received the B.S. and M.S. degrees in computer science from the University of California, Santa Barbara, in 1986 and 1989, respectively, where his emphasis was fault-tolerant software systems, and the Ph.D. degree in psychology from the University of Colorado, Boulder, in 2001, where his emphasis was building and testing intelligent tutoring systems using latent semantic analysis.

He was with Culler Scientific Systems, Communication Machinery Corporation, and the Computer Science Department, University of California, Santa Barbara. He was also a Consultant for Digital Sound Corporation and the Speech Technology Laboratory, Santa Barbara, CA. He taught computer programming and system administration classes for Navy personnel at Pt. Mugu, CA. He is currently a Research Associate at the Center for Spoken Language Research, University of Colorado, Boulder, where he builds literacy tutors and interactive books and performs research on latent semantic analysis and semantic parsing.

**Wayne Ward** was born in Pensacola, FL, in 1951. He received the B.A. degree with a double major in mathematical science and psychology from Rice University, Houston, TX, in 1973 and the M.A. and Ph.D. degrees in psychology from the University of Colorado, Boulder, CO, in 1981 and 1984, respectively.

From 1986 to 1998, he was a Research Faculty Member in the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. He is currently a Research Professor at the Center for Spoken Language Research, University of Colorado, Boulder.



**Jie Yan** (Member, IEEE) received the Ph.D. degree in computer science at Harbin Institute of Technology, Harbin, China, in 1999.

From 1999 to 2001, she was an Associate Researcher in Microsoft Research, China. She is currently a Research Associate in the Center for Spoken Language Research, University of Colorado, Boulder. She has published more than 20 papers and two patents in the related area. Her research interests include multiview human face detection, tracking, and recognition; realistic human face and body animation; and human facial expression, lip motion, body language synthesis, and virtual reality techniques.