

# Technologies that power pedagogical agents and visions for the future

Sarel van Vuuren

Center for Spoken Language Research  
University of Colorado Boulder

## Abstract

This article presents a vision of intelligent learning systems that use sensitive and effective pedagogical agents to deliver personal and individualized tutoring or therapy. The systems and their underlying technologies need to be powerful, yet flexible enough to deliver a wide range of media-rich interventions to individuals with vastly different needs across environments as diverse as classrooms, clinics, community centers and homes. We describe challenges faced by researchers to develop such agents and systems, and describe the capabilities of current systems designed to teach children to read or provide speech therapy to individuals with Parkinson's disease or aphasia.

## Introduction

Imagine a world where intelligent learning systems use conversational pedagogical agents to provide independent, immersive and effective tutoring, assistance or speech therapy. Modeled after expert teachers and therapists, they will be able to keep costs low while tirelessly delivering structured, adaptive, and individualized instruction to large numbers of users, whether their goals are to learn a new skill (e.g. reading) or to recapture or improve an existing skill (e.g. the ability to converse in a dialog or speak clearly).

Research points to the potential benefit, power and utility of such systems. For example, one-on-one tutoring or small group instruction is particularly effective relative to classroom instruction (Bloom, 1984). Yet, many of today's classrooms have many students with poor reading and yet have high student-to-teacher ratios. The situation is similar for patients with speech disorders who could benefit from proven treatments, with many unable to afford or get access to individualized speech therapy services. To address this problem and need, we are developing systems that can augment classroom instruction or therapy in cost effective ways, by using a pedagogical agent that behaves as much as possible like a sensitive and effective teacher or therapist.

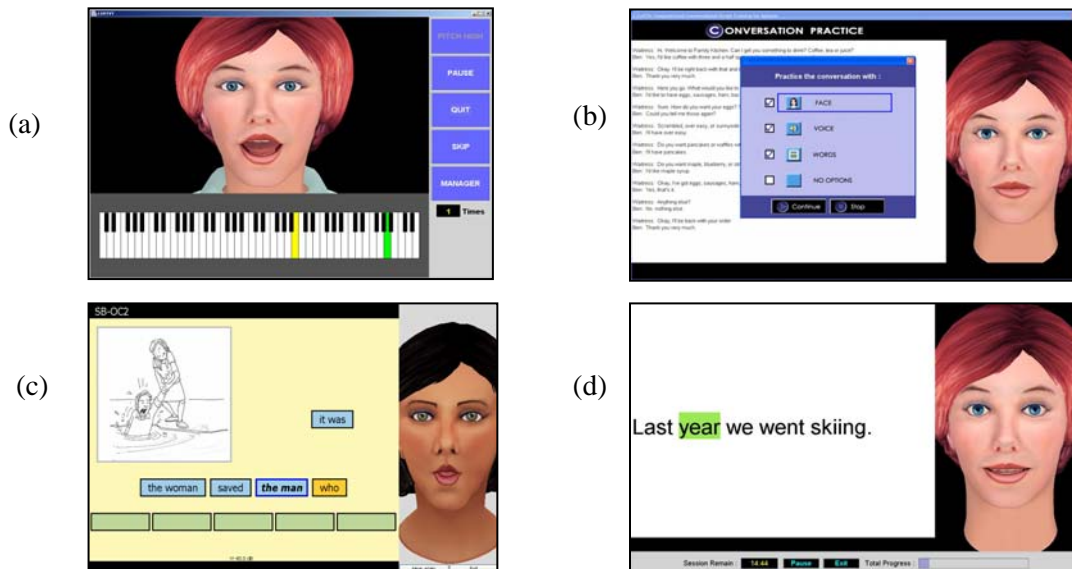
Inventing systems capable of supporting natural face-to-face communication with virtual humans over computer networks is akin to developing a car with 500 horsepower (representing the computational requirements of the component speech, language and animation technologies) that gets 500 miles per gallon (representing the ability to serve multiple users in real time via broadband networks). We describe some of the challenges faced by researchers to develop such systems, give examples from our own systems, and describe capabilities that are necessary to make such systems powerful and effective.

## Animated Pedagogical Agents that Teach and Conduct Therapy

Animated pedagogical agents and the learning systems built around them has been the subject of much research and development. Several applications have been described, e.g. Johnson et al. (2000) and Gratch et al. (2002) list many examples, and several studies have investigated the benefits of learning with a pedagogical agent, e.g. Moreno et al. (2001, 2004), Barker (2003), Baylor et al. (2003), Graesser et al. (2005), and Cole et al. (2006). Recent work has focused on modeling, understanding and improving social, pedagogical and conversational interaction. For example, Baylor & Kim (2005) analyze instructional roles, Graesser et al. (2005) and Cole et al. (2003) describe distributed systems using mixed-initiative dialogue, Bickmore & Cassell (2005) investigate social dialogue, while Gratch & Marsella (2005) study the presentation of emotion and affect.

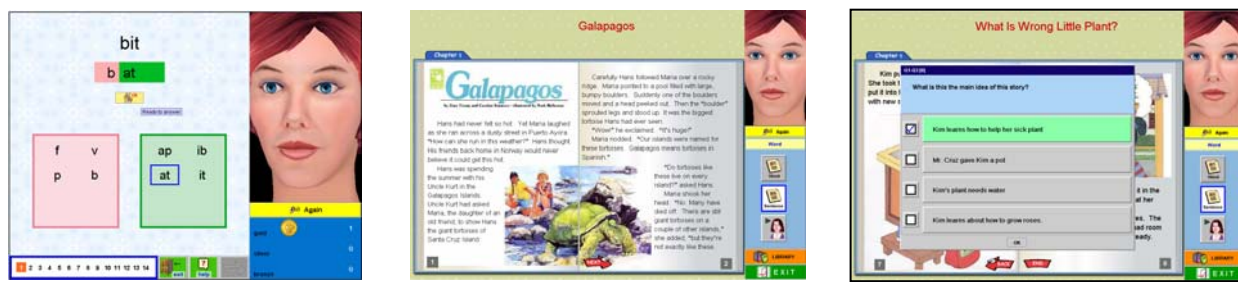
In our lab at the Center for Spoken Language Research (CSLR) we have studied and developed computer-based learning systems that use pedagogical agents to teach reading and to conduct speech therapy. These systems (a) faithfully implement an instructional program or treatment, and (b) use an animated pedagogical agent to emulate, to the extent possible, the behaviors of an expert teacher or therapist conducting the treatment. In each case, we worked with experts who developed the original program or treatment to design an initial prototype, and then refined the system through a series of “design-and-test” cycles with users partaking in this effort. This process, called participatory design, ensures treatment fidelity and good user experiences. Independent users of our systems have given them high ratings, saying the agent is believable and helps them learn (Cole, Wise & Van Vuuren, 2006).

We are developing virtual speech therapy systems for four independent treatments (Figure 1), including a system for individuals with Parkinson’s Disease (*LSVT<sup>TM</sup>-VT*: Halpern et al., 2006, Cole, et al., 2006), and separate treatments for individuals with aphasia (*Sentactics<sup>TM</sup>-VT*: Thompson, 2003; Van Vuuren et al., 2005; *ORLA<sup>TM</sup>-VT*: Cherney, 1995; and *C-Costa<sup>TM</sup>-VT*: Cherney et al., 2005).



**Figure 1.** Screen images of virtual therapists for (a) Parkinson disease and (b-d) aphasia interventions. (a) A pitch exercise in the LSVT-VT. (b) Cue selection during script practice in the C-Costa-VT. (c) Treatment for underlying forms exercise in the Sentactics-VT. (d) Sentence reading exercise in the ORLA-VT.

We are also developing virtual tutors for reading instruction (*Foundations to Literacy*<sup>TM</sup>: Wise et al., in press; Cole, Wise & Van Vuuren, 2006), reading assessment (*ICARE*<sup>TM</sup>: Wise, 2005) and assistive services. *Foundations to Literacy* is a comprehensive, scientifically-based program that teaches reading and comprehension skills to students in Kindergarten through second grade. A virtual tutor, *Marni*, guides students through reading skill exercises and books (Figure 2). During the last three years, this program has been fielded in over 50 classrooms in Colorado schools with summative evaluation of the program using standardized tests showing significant gains in letter and word recognition in Kindergarten and first grade.



**Figure 2.** Screen images of the *Foundations to Literacy* program showing reading skills activity, interactive book, and multiple choice question activity.

## The Challenge of Creating Sensitive and Effective Pedagogical Agents

We believe that a pedagogical agent will produce satisfying and effective learning outcomes if it appears to possess and effectively employs the essential *perceptive* and *generative* behaviors that are used by an expert tutor or therapist in a particular learning domain. To create agents that behave in this way, we have to overcome two major challenges.

The first challenge is the limitations of current technologies. To create agents that behave like sensitive and effective teachers or therapists, it is necessary to invent machine *perception* and machine *generation* technologies that can support natural face-to-face communication. Machine perception technologies enable the computer system to recognize and understand speech, locate faces, recognize expressions, track eye gaze and recognize gestures. Machine generation technologies enable the system to produce the movements that a teacher would produce, including visual speech, head and eye movements, facial expressions, and body gestures. The agent must combine these technologies to *listen* to the user and understand her speech, *perceive* and *anticipate* the user's actions such as moving the mouse pointer, *see* the user to interpret her head and eye movements, facial expressions and other gestures, and *react* to the user in appropriate ways, with speech, and gestures. While spoken dialog, computer vision, language generation and character animation technologies are advancing rapidly, they are not yet capable of supporting the complex nuances between speaker and listener during tutoring or therapy.

The second challenge is the lack of knowledge about how to use these technologies to emulate the social dynamics of face-to-face communication during learning and therapy. Even if the underlying machine perception technologies were capable of recognizing words, facial expressions, eye gaze and gestures in real time with high accuracy, and machine generation technologies were capable of producing all auditory and visual behaviors of the pedagogical

To appear in Special Issue of Educational Technology, 2006.

agent in real time in response to user behaviors, the problem still would remain of knowing how to interpret these behaviors and knowing how to respond to them. The best we can do today is to observe and record experts interacting with students or patients, and attempt to model their behaviors with our systems.

## **Capabilities of Pedagogical Agents in Intelligent Learning Systems**

Users expect responsive and functional applications that are easy to access in a variety of environments such as at school, in a clinic or at home. We in turn must be able field these applications with minimum effort and ease of maintenance. Here we discuss the capabilities of our systems to accommodate these requirements.

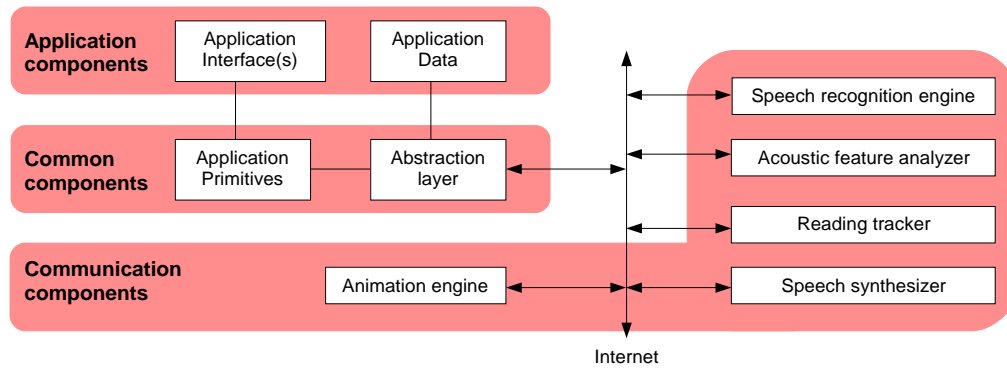
### **Agent Capabilities**

What distinguishes our learning systems is their emulation of human communication capabilities—our systems have computer interfaces that allow a learning experience with a conversational animated agent capable of ‘talking’ and ‘listening’. The system maintains a model of the user, including an individualized lesson plan and record of user progress. The program adapts to the skills of the user and monitors and records all interaction for review within a management environment. The user is presented with a task via media objects (e.g., text, illustrations, animations), with instructions, modeling and feedback provided by a pedagogical agent that produces accurate visual speech synchronized with a recorded voice. The user interacts with the system via mouse clicks, keystrokes or speech. The speech output can be processed to extract measures of the user’s speech such as pitch or loudness, and passed to an automatic speech recognizer to recognize words and extract meaning. The information extracted from the audio signal can be used to provide real time feedback to a patient on the quality of their utterances, or to provide real time feedback to a student who is reading out loud.

In our reading applications, the agent ‘talks’ to the student, introduces and explains each lesson, provides instructions and explanatory feedback within reading activities and books, pronounces words when the student clicks on them, narrates stories, and asks comprehension questions, while providing verbal and nonverbal cues (head movements) to encourage and reinforce learning. The system adapts to the user’s skill level and responds in predictive and anticipatory manner based on user behavior (reading position, location of mouse pointer, history and progress). The agent can track the mouse pointer by appearing to turn toward it or look at a region of the screen where it expects the user to look or click. The system ‘listens’ to and follows a user reading out loud. In our speech therapy applications, the agent provides encouraging and empathic real time feedback to the patient using verbal cues, head movements and expressions based on measurements of the loudness, duration and fundamental frequency of their utterances.

### **System capabilities**

Our systems operate in a client-server configuration, with the server maintaining a model of the user, including a lesson plan and detailed progress data. We explain the power, flexibility and capabilities of these systems in terms of their high-level functional components (Figure 3).

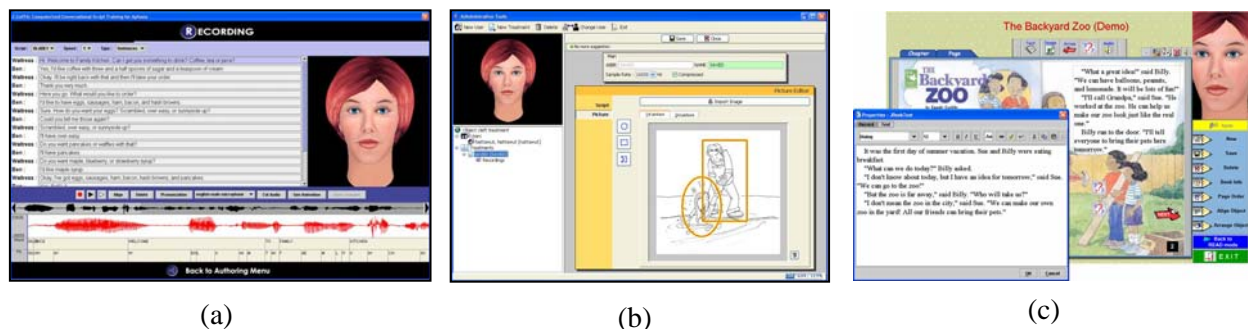


**Figure 3.** System architecture and functional components.

**Application components.** These components are programmatic and knowledge specific, and generally different for each application. Consisting of *application interfaces* and *application data*, they are designed in collaboration with knowledge experts.

The *application interface*, as the functional and graphical embodiment of the application, provides the look, feel and interaction of the application and includes a *player*, *authoring tools*, a *management environment* and *monitoring tools*. The player interfaces with an AI engine which determines the next system state, based on current state, user model, and user input. Authoring tools (Figure 4) allow developers and users to control the behavior of the agent and create, edit, and manipulate media objects in the application. The management environment displays and organizes user data for review and analysis. All interactions between the user and the system are recorded. The monitoring tools allow remote management and teacher or clinician oversight.

The *application data* consist of *program* data and rules, which codify expert knowledge and the learning process in a study plan or treatment protocol and drives the AI engine; *content* data, in the form of media objects; and *user* data, in the form of login, performance and personalization information. Application data are stored locally or remotely in relational databases.



**Figure 4.** Screen images of authoring tools to (a) create and record prompts spoken by the agent, (b) mark up interactive images, and (c) write interactive books.

To appear in Special Issue of Educational Technology, 2006.

**Communication components.** These components provide the perceptive and generative capabilities of the system with components for character animation, automatic speech recognition, reading tracking, speech feature analysis, speech synthesis, dialog management and text parsing. A detailed discussion can be found in (Cole, Van Vuuren, et al., 2003); here we briefly describe the first two components.

The *animation* component provides 3-D character animation with accurate facial emotions and anatomically correct movements of the lips, tongue and jaw during speech production. To create accurate visible speech (Ma et al., 2002, 2004), researchers in our lab recorded motion capture data of markers attached to a reading expert's lips and face while saying the common syllables and permissible diphone sequences present in English. The marker points in the sequences were mapped to the vertices and polygons of a 3-D model, and the data compressed and saved. In operation, audio to be spoken by the agent is automatically segmented into diphone units using a speech recognizer. The animation engine then selects, concatenates and interpolates suitable sequences from the library of saved sequences in an approach called *multi-unit concatenative synthesis*, finally rendering the concatenated sequence in synchrony with the audio.

The animation engine is also capable of producing facial expressions and emotions (e.g. sadness, fear, anger, disgust, joy and surprise) through morph targets designed to allow control over individual facial components. To control expressions and head movements, a markup language was developed (Ma et al., 2002) allowing expressions to be specified by type, time of occurrence and duration.

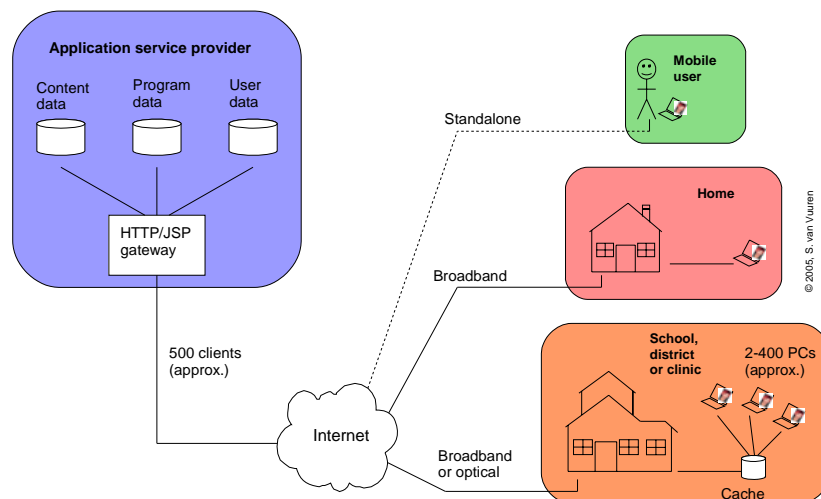
The *automatic speech recognition* component consists of a speech recognition engine (Pellom, 2001) and its acoustic and language model data. The recognizer receives spoken audio which it transcribes. English, Spanish, and altogether 16 languages are currently supported. Hagen and Pellom et al. (2004) have trained the recognizer on children's read speech achieving state-of-the-art performance with less than 7% word-error.

**Common components.** These components keep the application and communication components working together and include web-based installation, updating and software maintenance processes. They manage the system configuration and state, control communication, and provide functionalities for data management, logging and reporting. Written in JAVA code and designed to run natively on the user's PC—so as to avoid the need to run in a browser, these components ensure portability and stability. System communication and data management are handled transparently with communication conducted via HTTP or proxy server, avoiding problems with firewalls. Data can be accessed locally on the user's machine or remotely on a server and stored in a relational database, CD-ROM, or file system.

## Deployment scenarios

The current systems run on off-the-shelf PCs and laptops. The architecture has been designed to enable real time interaction with multiple users, with optional communication between clinicians and multiple clients. Various deployment scenarios are supported (Figure 5) ranging from standalone configurations, where a mobile user might be using the system on a laptop, to home users connecting to an application service provider over a broad-band connection, to schools, clinics, and other organizational users using the system over a local area network. While we have

deployed our systems in each of these configurations, we have not yet scaled, tested and optimized web-based deployments with several hundred simultaneous users per server.



**Figure 5.** Deployment scenarios supported by the system architecture.

## Summary

We provided a brief overview of efforts at CSLR to build and deploy intelligent learning systems with affective interfaces that use conversational pedagogical agents to provide personal and individualized tutoring or therapy. While available technologies can support fairly sophisticated applications, produce satisfactory user experiences and effective treatment outcomes, advances in human language technologies and our state of knowledge about how to use these technologies must be achieved in order to model the social dynamics of effective and sensitive teachers. Nevertheless, our work demonstrate that by designing powerful and extensible server-client architectures, by harnessing the power of current off the shelf computers, and by managing communication of media and data between servers and clients efficiently it is possible to realize the potential of pedagogical agents. As these systems are being deployed, we expect them to dramatically benefit large numbers of users and to provide cost effective solutions to individualized one-on-one tutoring and therapy in ways that augment regular instruction.

**Acknowledgments.** This work was supported in part by grants from the Coleman Foundation, National Science Foundation, Department of Education and National Institutes of Health: NSF/ITR IIS-0086107, NSF/IERI EIA-012\1201, NICHD/IERI 1R01HD-44276.01, IES R305G040097, NIH 1R21DC007377-01, NIH 5R21DC006078-02, NIDRR H133B031127, NIDRR H133G040269, NIDRR H133E040019. We'd like to thank Ron Cole, Barbara Wise and our other colleagues, cited herein, for their invaluable contributions and feedback.

To appear in Special Issue of Educational Technology, 2006.

## References

- Barker, L. (2003). Computer-Assisted vocabulary acquisition: The CSLU vocabulary tutor in oral-deaf education, *Journal of Deaf Studies and Deaf Education*, vol. 8, no. 2, pp. 187-198.
- Baylor, A. L. & Ryu, J. (2003). Does the presence of image and animation enhance pedagogical agent persona? *Journal of Educational Computing Research*, 28(4), 373-395.
- Baylor, A. L. & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education*, 15(1).
- Bickmore, T., Cassell, J. (2005) "Social Dialogue with Embodied Conversational Agents" In J. van Kuppevelt, L. Dybkjaer, & N. Bernsen (eds.), *Advances in Natural, Multimodal Dialogue Systems*. New York: Kluwer Academic.
- Bloom, B.S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring, *Educational Researcher* 13, pp. 4-16.
- Cherney, L. (1995). Efficacy of oral reading in the treatment of two patients with chronic Broca's aphasia. *Topics in Stroke Rehabilitation*, 2(1), 57-67.
- Cherney, L., Halper, A., Babbit, E., Holland, A., Cole, R., Van Vuuren, S., Ngampatipatpong, N. (2005). "Learning to Converse: Script Training, Virtual Tutors, and Aphasia Treatment", ASHA, San Diego.
- Cole, R., Wise, B., Van Vuuren., S. (2006). How Marni teaches children to read. *Educational Technology* [THIS JOURNAL].
- Cole, R., Halpern, A., Lorraine, R., Van Vuuren, S., Ngampatipatpong, N., Yan, J. (2006). A Virtual Speech Therapist for Individuals with Parkinson's Disease. *Educational Technology* [THIS JOURNAL].
- Cole, R., Van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W., & Yan, J. (2003). Perceptive Animated Interfaces: First Steps Toward a New Paradigm for Human-Computer Interaction, *Proceedings of the IEEE: Special Issue on Human-Computer Multimodal Interface*, 91 (9), pp. 1391-1405, Sept., 2003.
- Graesser, A.C., Chipman, P., Haynes, B.C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education*, 48, 612-618.
- Gratch, J., Rickel, J., André, E., Badler, N., Cassell, J., and Petajan, E. (2002). Creating interactive virtual humans: Some assembly required, *IEEE Intelligent Systems*, vol. 17, no. 4, pp. 54-63, July/August 2002. Reeves, B. and Nass, C. (1996). *The Media Equation: How people treat computers, television, and new media like real people and places*, NY: Cambridge University Press.
- Gratch, J. and Marsella, S. (2005). "Some Lessons for Emotion Psychology for the Design of Lifelike Characters," *Journal of Applied Artificial Intelligence (special issue on Educational Agents - Beyond Virtual Tutors)*, vol. 19(3-4), pp. 215-233.
- Hagen, A., Pellom, B., Van Vuuren, S., and Cole, R. (2004). Advances in Children's Speech Recognition within an Interactive Literacy Tutor, in *Proceedings of the Human Language*

To appear in Special Issue of Educational Technology, 2006.

Technology Conference / National Chapter of the Association of Computational Linguistics, Boston, May 2004.

Halpern, A.E., Cole, R., Ramig, L.O., Yan, J., Petska, J., Vuuren, S., Spielman, J., (2006).

“Virtual Speech Therapists - Expanding the Horizons of Speech Treatment for Parkinson’s disease,” accepted for presentation at the Conference on Motor Speech, March 23-26, 2006, Austin, Texas.

Johnson, W.L., Rickel, J.W., and Lester, J.C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments, *International Journal of Artificial Intelligence in Education*, vol. 11, pp. 47-78.

Ma, J., Yan, J., and Cole, R. (2002). CU Animate: Tools for Enabling Conversations with Animated Characters, in International Conference on Spoken Language Processing (ICSLP), Denver, Sept 2002., Sep, 2002.

Ma, J., Cole, R., Pellom, B., Ward, W., and Wise, B. (2004). Accurate Automatic Visible Speech Synthesis of Arbitrary 3D Models Based on Concatenation of Di-Viseme Motion Capture Data. *Journal of Computer Animation and Virtual Worlds*, Vol. 15(5), pp. 485-500.

Moreno, R., Mayer, R., Spires, H., and Lester, J. (2001). The Case for Social Agency in Computer-Based Teaching: Do Students Learn More Deeply When They Interact with Animated Pedagogical Agents? *Cognition and Instruction*, 19(2), pp. 177-213.

Moreno, R. (2004). Animated pedagogical agents in educational technology. *Educational Technology*, 44 (6), 23-30.

Pellom, B. (2001) "SONIC: The University of Colorado Continuous Speech Recognizer", Technical Report TR-CSLR-2001-01, CSLR, University of Colorado, March.

Thompson, C., Shapiro, L., Kiran, S., & Sobeks, J. (2003). The role of syntactic complexity in treatment of sentence deficits in agrammatic aphasia: The complexity account of treatment efficacy (CATE), *Journal of Speech, Language, and Hearing Research*, 46, 591-605.

Van Vuuren, S., Ngampatipatpong, N. (2005). Aphasia Virtual Clinician User Guide V1.0. Available by request: sareel@colorado.edu.

Wise, B. (2005). Developing an Independent and Adaptive Reading Evaluation, Invited talk at the 56th Annual Meeting of the International Dyslexia Association, Technology strand, Nov. 13, Denver, Colorado.

Wise, B., Cole, R., Van Vuuren, S., Schwartz, S., Snyder, L., Ngampatipatpong, N., Tuantranont, J., & Pellom, B. (In press). Learning to Read with a Virtual Tutor: Foundational exercises and interactive books. In Kinzer, C. & Verhoeven, L. (Eds). *Interactive Literacy Education*. Mahwah, NJ: Lawrence Erlbaum.

Wise, B., Cole, R.A., Van Vuuren, S. (2005). *Foundations to Literacy: Teaching children to Read Through Conversational Interaction with a Virtual Teacher*, Invited talk at 56th Annual Meeting of the International Dyslexia Association, Technology strand, Nov. 13, Denver, Colorado.