Towards Interactive Training with an Avatar-based Human-Computer Interface

Ronald F. DeMara, Avelino J. Gonzalez,
Victor Hung, Carlos Leon-Barth, Raul A. Dookhoo
Intelligent Systems Laboratory
University of Central Florida
Orlando, FL
demara@mail.ucf.edu, gonzalez@mail.ucf.edu,
victor.hung@isl.ucf.edu, jleonbar@mail.ucf.edu,
raul1864@hotmail.com

Steve Jones, Andrew Johnson,
Jason Leigh, Luc Renambot,
Sangyoon Lee, Gordon Carlson
Electronic Visualization Laboratory
University of Illinois at Chicago
Chicago, IL
sjones@uic.edu, lenzman@mac.com
spiff@uic.edu, renambot@uic.edu
slee14@uic.edu, gordycarlson@gmail.com

ABSTRACT

The development of avatars has significant potential to enhance realism, automation capability, and effectiveness across a variety of training environments. Project Lifelike is a three-year National Science Foundation effort whose objective is to develop and evaluate realistic avatar interfaces as portals to intelligent programs capable of relaying knowledge and training skills. This interface aims towards support of spoken dialog within a limited domain and capabilities for learning to maintain its knowledge current and accurate. Research objectives focus on the integration of speaker-independent continuous speech recognition technology with a context-based dialog system and real-time graphics rendering capability derived from live subject motion capture traces. The motion capture traces are used by the avatar to provide spoken interaction with gestural expressions. This paper describes the first phase of the Lifelike project which developed an interactive avatar prototype of a National Science Foundation (NSF) program manager, Dr. Alex Schwarzkopf, for whom a contextual graph representation of domain knowledge was created. A Graphical Asset Production Pipeline was developed to allow digitization of the facial characteristics and physical movements of Dr. Schwarzkopf. Next, an example subset of his knowledge of NSF protocols was encoded in a grammar-based speech interpretation system and context-based reasoning system. These systems were integrated with the Lifelike Responsive Avatar Framework to enable the avatar to receive spoken input and generate appropriate verbal and non-verbal responses. The system demonstrates conveyance of knowledge within a limited domain such as NSF project reporting requirements. Work toward improving the realism of the avatar, long-term efforts toward creating a toolbox for generalization to other training applications, and results of evaluation of how users respond to different characteristics that contribute to realism in an avatar are discussed.

ABOUT THE AUTHORS

Ronald F. DeMara is a Professor in the School of Electrical Engineering and Computer Science at the University of Central Florida (UCF). His research interests are in Computer Architecture with emphasis on Real-time Intelligent Systems.

Avelino J. Gonzalez is a Professor of the School of Electrical Engineering and Computer Science at UCF. His research interests focus on the areas of artificial intelligence, context-based behavior and representation, temporal reasoning, intelligent diagnostics and expert systems.

Andrew Johnson is an Associate Professor in the Department of Computer Science and member of the Electronic Visualization Laboratory at the University of Illinois at Chicago (EVL/UIC). His research focuses on the

development and effective use of advanced visualization displays, including virtual reality displays, auto-stereo displays, high-resolution walls and tables, for scientific discovery and in formal and informal education.

Steve Jones is Associate Dean for Liberal Arts and Sciences, Professor of Communication and Research Associate in the EVL/UIC. He has authored and edited numerous books, including Society Online, CyberSociety, Virtual Culture, Doing Internet Research, and the Encyclopedia of New Media. His research interests include interaction in virtual environments, the social consequences of new media, internet studies, and the history of communication.

Jason Leigh is an Associate Professor of Computer Science and director of the EVL/UIC. Leigh is a co-founder of VRCO, the GeoWall Consortium and the Global Lambda Visualization Facility. Leigh currently leads the visualization and collaboration research on the National Science Foundation's OptIPuter project, and has led EVL's Tele-Immersion research agenda since 1995. His main area of interest is in developing collaboration technologies and techniques for supporting a wide range of applications ranging from the remote exploration of large-scale data, education and entertainment.

Luc Renambot received the Ph.D. degree at the University of Rennes-1 (France) in 2000, conducting research on parallel rendering algorithms for illumination simulation. Then holding a Postdoctoral position at the Free University of Amsterdam, till 2002, he worked on brining education and scientific visualization to virtual reality environments. Since 2003, he joined EVL/UIC first as a PostDoc and now as Research Assistant Professor, where his research topics include high-resolution displays, computer graphics, parallel computing, and high speed networking.

Gordon Carlson has a Masters in Communication and Adult Education from Oregon State University. He is currently working on his Ph.D. degree in new media communication at the UIC. His research focuses on how technologies mediate communication as well as how people talk about technology at the expert and user level.

Raul A. Dookhoo is a Research Assistant and Master of Science student in the Computer Science Department at the University of Central Florida. His interests are Voice Recognition and Automatic Grammar Generation.

Victor Hung is a Ph.D. student in Computer Engineering at UCF. His work entails developing a natural language dialog system based on the Context-Based Reasoning paradigm. He is currently stationed as a Research Assistant at the National Science Foundation headquarters in Arlington, Virginia.

Sangyoon Lee is Ph.D. student in the Department of Computer Science at the UIC. Lee received a Diploma, MS in Architecture from YONSEI University, Seoul, Korea in 1999 and an MFA from school of Art and Design at UIC in 2006. His research work focuses on Computer Graphics/Visualization and Human Computer Interaction (HCI).

Carlos Leon-Barth is a Research Assistant and Ph.D. student in the School of Electrical Engineering and Computer Science at UCF. He earned a BS (1993) in Electrical Engineering from University of Florida, and an MS (1996) in Computer Engineering from UCF. He is a former Engineer for IBM Global Services from 1996-2001. His research interests are Voice Recognition, Real-time Database Systems and Simulation and Modeling.

Towards Interactive Training with an Avatar-based Human-Computer Interface

Ronald F. DeMara, Avelino J. Gonzalez,
Victor Hung, Carlos Leon-Barth, Raul A. Dookhoo
Intelligent Systems Laboratory
University of Central Florida
Orlando, FL
demara@mail.ucf.edu, gonzalez@mail.ucf.edu,
victor.hung@isl.ucf.edu, jleonbar@mail.ucf.edu,
raul1864@hotmail.com

Steve Jones, Andrew Johnson,
Jason Leigh, Luc Renambot,
Sangyoon Lee, Gordon Carlson
Electronic Visualization Laboratory
University of Illinois at Chicago
Chicago, IL
sjones@uic.edu, lenzman@mac.com
spiff@uic.edu, renambot@uic.edu
slee14@uic.edu, gordycarlson@gmail.com

PROJECT OBJECTIVE

The Lifelike project is investigating intelligent avatar interfaces suitable for a range of question answering and training applications. The project objective is to enable domain-specific conversation with a realistic avatar supported by an intelligent engine capable of online learning. We address three specific characteristics of this interface that increase naturalness in interaction for various narrative and tutorial-style training applications:

1. a life-like embodiment of a particular person which a trainee can orally address,
2. the use of non-verbal cues including expressions by the avatar, as well as real-time trainee location tracking and a means for the user to designate a focal point within the virtual world, and
3. a knowledge-driven backend that can respond intelligently to questions and learn through its interactions.

Many existing Decision Support Systems (DSS) for tactical and training purposes rely heavily on traditional keyboard-style input devices and display their output in the form of written text or schematic/graphic representations such as maps and charts. On the other hand, an avatar-based interface in certain applications can reduce the trainee's machine interface workload and allow the trainee to more completely focus on the task being trained rather than on the user interface to the DSS.

Initially, research in this project developed a traditional keyboard/text interface for a DSS called *AlexDSS* by Sherwell, Gonzalez, & Nguyen (2005). AlexDSS was developed for NSF's Industry/University Cooperative Research Center (I/UCRC) program by providing a means to capture, preserve, and reuse the expertise of retiring NSF program director Dr. Alex Schwarzkopf in the form of a contextual graph representation. The goal of the AlexDSS project was to advise a querying user regarding programmatic and funding issues using the same knowledge and reasoning used by Dr. Schwarzkopf. That phase of the project has ended and the current users of the text-based AlexDSS system include Dr. Schwarzkopf's successor, as well as other NSF center directors who would otherwise regularly seek guidance from him in person.

With AlexDSS as a starting point, the objective of the first phase of the Lifelike project was to create an avatar-based interface for user interaction. We have developed an avatar for this purpose, one that supports a limited dialogue, and have evaluated several operational effectiveness aspects of the current implementation with numerous users. We refer to this avatar as *AlexAvatar*. The following sections describe the design of the current AlexAvatar, production generalization approaches using a graphical asset pipeline, observed system performance, and progress toward the eventual goal of a completely realistic avatar.

RELATED WORK

The notion of interactive agents has existed since the inception of the computing age. Idealistic visions of these agents are often rife with extraordinary capabilities, yet state-of-the-art technology yields only a sliver of these expectations. The 2000's presented an evolution of these embodied agents, beginning with the work of Cassell et al (2000), whose conversational playmate, Sam, gave insight into the effectiveness of a human-computer interaction in a physically immersive environment. Although Sam was not an autonomous entity, its creators exhibited the idea that even a child

could feel comfortable when interacting with a machine-based being. Bickmore and Picard (2004) presented their studies with Laura, a personal trainer agent. An early prototype of dialog-based agents, Laura's interactions with the user could be considered a one-sided question and answer session, with the agent controlling the 'conversation.' The primary result of Bickmore and Picard's work was the concept that a *caring* embodied agent proved more effective than one of indifference.

The latter half of the decade sought more ambitious goals in creating interactive agents. Lee et al (2005) experimented with using robots as conversational agents. An animatronics penguin, Mel, posed as an expert for a hypothetical product. In this work, Lee et al (2005) supported the notion that humans could indeed interact with a physically engaging and conversationally interactive machine. This idea was further demonstrated in Kenny et al (2007) with the Sergeant Blackwell conversational agent. With a more sophisticated dialog system than Mel, Sergeant Blackwell's capabilities for conversation provide the user with a more natural human-computer interaction. Kenny et al's agent, however, appears to lack a cognitive model within its dialog management mechanisms.

EXPERIMENTAL AVATAR PROTOTYPE

We have emphasized realistic appearance, behavior, and audio generation capabilities for the AlexAvatar in order to gain the credibility of the trainee with whom it communicates. The avatar has been developed to not only resemble a human being in general, but in particular a specific individual to whom the trainee has spoken previously in person. In our case, that person is Dr. Schwarzkopf, who directed NSF's I/UCRC program. Research effort has been devoted to an attempt to have the avatar appear not only to be realistic with respect to typical human behavior in general, but specifically with mannerisms compatible with those of Dr. Schwarzkopf.

The avatar-based interface executes on top of an engine that knows the operational details of the I/UCRC program as well as domain knowledge (i.e., AlexDSS). It can also remember each user and its interaction with the user as episodic knowledge. A future version will incorporate the ability to access user historical information to recall conversations of users involved in the I/UCRC program who have exchanged communications with the AlexAvatar. This can enable new levels of automated After-Action Review and trainee recall capabilities within the military training process.

As shown in Figure 1, the system has been developed using an LCD widescreen panel that can reproduce a human-like figure in its original 1:1 scaled size, a headset-based hypercardiod microphone for narrow field filtering of spoken user input, and audio speakers for generating the avatar's spoken output. This system allows trainee interaction with the digitized image (an actively rendered digital representation) of Dr. Schwarzkopf to access the AlexDSS knowledge engine.

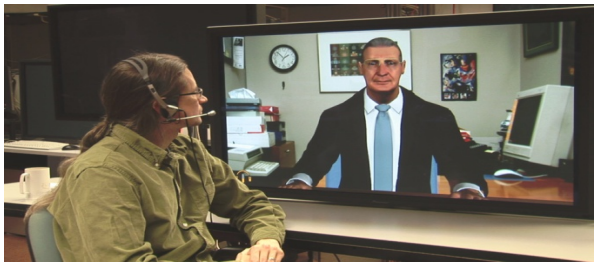

Figure 1. Lifelike System Prototype - AlexAvatar

Recognition of speech from untrained users was emphasized during development, so there is no induction dialog required to train the system. Thus, any person affiliated with the I/UCRC program can approach the system and initiate a conversation with the AlexAvatar that it will interpret and respond in a human-like fashion with information regarding NSF supplemental funding opportunities. The I/UCRC program was selected as the training domain for the demonstration prototype. For example, Figure 1 shows a user conversing with AlexAvatar sitting in its office.

This initial version of the avatar incorporates gestures and movements that Dr. Schwarzkopf uses in real life. These were obtained using the motion capture studio techniques developed below. Using active text-to-speech generation, the avatar's voice, while not identical to Dr. Schwarzkopf's, resembles the tone of a senior official. While it is possible to more effectively emulate Dr. Schwarzkopf's voice characteristics using advanced commercially-available technology, it is currently cost prohibitive for the initial phase of the prototype. In summary, our experimental scenario offers the capability for a user interested in exploring selected aspects of NSF I/UCRC project management to engage in an oral conversation with a virtual expert who imparts the sought domain knowledge.

SYSTEM DESIGN

As shown in Figure 2, the avatar is supported by integrated components consisting of a Speech Recognizer (SR), Dialog Manager, and Speech Generation module. These roughly correspond to the avatar's tasks of hearing, understanding, and responding, respectively. Each module executes as a separate thread and communicates with the others using a handshaking protocol created to enforce synchronization constraints between threads. Voice input from the microphone headset is provided to the Lifelike Recognizer module that performs speaker-independent continuous speech recognition on the input waveform to produce two forms of recognized speech data. The first result is a list of the most likely domain-specific concepts detected using a domain-specific grammar-driven recognizer. The second form of recognized speech data produced is an ASCII text of the English words in the phrase as recognized using a generic non-customized grammar-free lexicon.

The Dialog Manager uses both sets of data to more fully disambiguate the speech input. The Dialog Manager also maintains the current context which indicates the focus of the conversation using domain-specific knowledge. It provides the current context back to the Recognizer to help focus the recognition task. The Dialog Manager also selects or generates a text response string to be spoken by the avatar which is passed to the Speech Output system to perform text-to-speech and gesture generation. Each system is described in detail below.

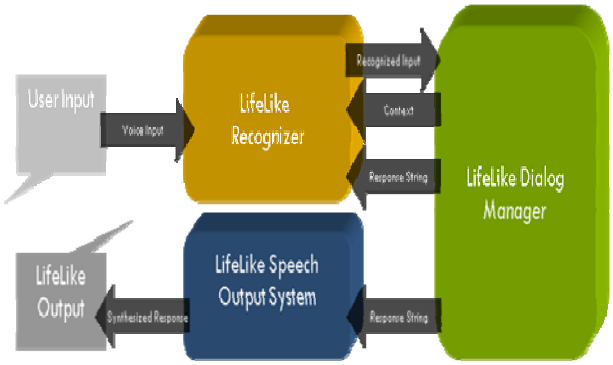

Figure 2. Lifelike System Diagram

Lifelike Recognizer

The Lifelike Recognizer module was designed using a layered model for modularity that can support a range of Commercial-Off-The-Shelf (COTS) speech recognition engines. Figure 3 depicts the Speech Recognizer Control (SRC), Speech Recognizer Engine (SRE), and Smart layers in the SR. The SRC resides at the lowest layer and allows use of compatible COTS recognition engines. The prototype is currently implemented using Microsoft Speech API (SAPI) version 5.1 from Microsoft Development Network (2008). The SRC translates the audio waveform into both textual formats with the help of Chant SpeechKit middleware from Chant Software (2008). Chant middleware components simplify the process using the Software Development Kits (SDKs) from multiple speech technology vendors to increase portability and reduce development time. The SR is activated by the Dialog Manager to initiate processing of microphone input only at appropriate times in the dialog to prevent miss-recognition due to spurious noise. Activation invokes the primary recognition strategy which is grammar-based. The grammars are used by the SR to provide a stream of domain-specific concepts with a certain degree of confidence as its primary output similar to research by Wendt, Fink, & Kummert (2002).

Simultaneously with the custom grammar-based resolution thread, a second instantiation of a SAPI process performs recognition in standard dictation mode using a non-domain-specific SAPI lexicon to attempt an audio-to-text conversion on the same waveform. Therefore, two independent recognition pathways are using multiple sets of COTS programming APIs, all of which are controlled by the SR module implemented in the C# language.

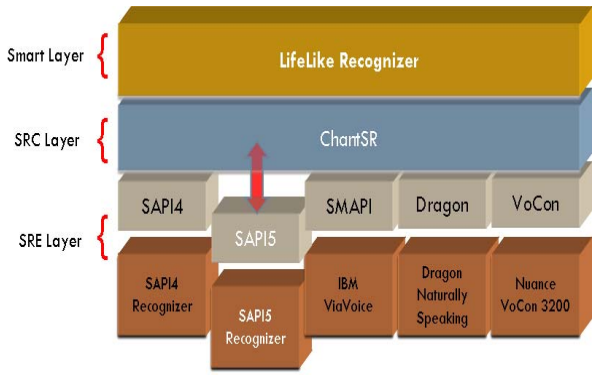

Figure 3. Layered Speech Recognition Architecture

Although the current version of the SR uses the Microsoft SAPI 5.1, the layered design using Chant allows compatibility with other recognizer engines such as Nuance's Dragon Naturally Speaking and IBM's ViaVoice. SAPI 5.1-compatible grammars provide a recognition framework for domain-specific entries in the lexicon. These are partitioned and

organized by context to reduce the size of the search space and improve recognition accuracy.

An additional innovation in SR processing is that grammars are generated semi-automatically from a relational database that facilitates dialog development, maintenance, and portability. The SQL database contains the structural knowledge of concept instances, and the relationships between them, which are then extracted for dialog management. New speech information regarding the project can be added automatically the next time the SR module starts by invoking transfer of database content into grammars automatically by using a set of rewriting rules. A small embedded script reads the database and overwrites the grammars with a new set. The long-term objective, which has been partially met in the current prototype, is to generate new grammars for speech recognition as necessary without needing to compose all grammars and their variations manually. An extension to this technique could be used to refine grammars each time the system is used.

Dialog Manager

The role of the Dialog Manager (DM) is to interpret the text streams provided by the Speech Recognizer to determine the context, decide how to react to contextual shifts, and coordinate the communication between the subsystems accordingly. The DM system also attempts to disambiguate likely candidate phrases using the information already in its database concerning the users and the I/UCRC program.

Conversational goal management is achieved using a context-based approach. A *context* refers to a particular situation that is dictated by the configuration of internal and external circumstances such as the internal state of the conversation agent and the perceived state of the human trainee. For every context, there is an associated goal condition and a group of relevant actions that can be executed to achieve this condition. A *goal condition* is defined as an end state that an agent desires to reach to impart specific knowledge to the trainee.

It is imperative that a dialog system be able to properly manage conversational goals, as the user can have multiple goals and may introduce new goals at any time. Henceforth, the system must be able to service many goals at one time, as well as be prepared to take on more goals, unannounced. This need to be able to jump between different goals in real-time lends itself to the Context-Based Reasoning (CxBR) technique used by Stensrud, Barrett, Trinh and Gonzalez (2004). CxBR agents provide responses that are directly related to its active context. The fact that contexts correspond to accomplishing particular goals combined with the idea that conversational goals take on a very fluid nature yields the assertion that goal management can be facilitated using CxBR methods.

Figure 4 shows the architecture of the DM, which consists of three components: Speech Disambiguator, Knowledge Manager, and CxBR-based Dialog Manager. The Semantic Disambiguator serves as a listening comprehension filter, where heard noises (the SR output) are converted to conversationally-relevant content to be processed by the person, known as the Disambiguated Input. The Knowledge Manager acts as a person's rote memory. The Speech Disambiguator and the Dialog Manager send keyword-based requests to it as inputs, and the Knowledge Manager outputs relevant information in the form of a contextualized data base. The Dialog Manager serves to provide the proper responses output to the user. It takes in input from the Speech Disambiguator, as well as its own internal context-based mechanisms to determine this output response.

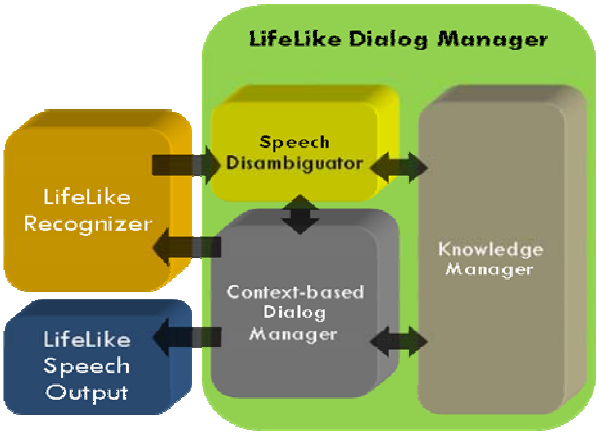

Figure 4. Dialog Manager Architecture

Goal management in the Lifelike Dialog Manager involves three parts: 1) goal recognition, 2) goal bookkeeping and 3) context topology. Goal recognition refers to the process of analyzing user input utterances to determine the proper conversational goal that is to be addressed. This is analogous to the context activation process in CxBR methods. Goal bookkeeping deals with keeping track of the identified goals in an ordered manner. Bookkeeping simply services the recognized goals in the order they are received, using a stack. Context topology refers to the entire set of speech acts of the conversation agent. This structure also includes the transitional actions when moving between contexts

when a goal shift is detected. The context topology carries out the responses needed to clear out the goal bookkeeping stack. Goal recognition is accomplished using linguistic analysis of each user utterance. This is similar to the inference engine found in CxBR systems, where conditioned predicate logic rules determine the active context according to the environmental state. The difference with the goal recognizer, however, is that the context is resolved using keywords and phrases that are extracted from a parts-of-speech parsing of input responses. With the aid of a contextually-organized knowledge base, the user utterance is interpreted, and the context associated with this understanding is activated.

Responsive Avatar Framework

Creating a realistic active digital representation of a particular human being is a challenging and multifaceted task. Initially, investigations were conducted to identify and evaluate the interoperability of COTS packages for facial modeling, rendering of real-time graphics, motion-capture, and text-to-speech synthesis. The result was a customized Graphical Asset Production Pipeline which encapsulates the tasks needed to create a visual representation of a human character. Furthermore, the options and best practices for recording vocal mannerisms and non-verbal mannerisms were evaluated and identified.

FaceGen, a tool used by researchers Heinrichs, Muller & Tewes (2006) for face recognition, was used. FaceGen generates three-dimensional (3D) head and face models using front and side photographic images. The technique was used to develop the highly acclaimed video game, Oblivion (2008). Figure 5 shows the resulting 3D head of Alex Schwarzkopf. FaceGen provides a neutral face model that can be parametrically controlled to emulate almost any facial expression. In addition, FaceGen enables the user to control the gender, age, and race of the model. While this is a sufficient initial capability, much can still be done to improve the visual realism by applying more advanced techniques such as modeling the sub-surface light scattering properties of skin tissue as in research done by Donner & Jensen (2005).

Motion capture was performed at EVL with a new motion capture system equipped with eight high-resolution (Vicon MX-F40, 4 mega pixels) infrared tracking cameras as depicted in Figure 6. Motion capture is the most widely used approach for acquiring realistic human figure animation for the film and video game industries. The accomplishment thus far has been to capture a series of simple motions and enable our avatar to "re-enact" them. We anticipate altering the motion-captured data in real-time while the avatar is moving so that dynamic, naturalistic behaviors can be synthesized.

Figure 5. Model (Left: avatar, right: Alex)

Real-time graphics required selecting a capable engine with the right blend of features and within budget. After evaluation of several possible open source graphics engines, we decided to use Object-oriented Graphics Rendering Engine (OGRE) as our underlying graphics library for the avatar framework. OGRE provides both a high-level interface for interacting with graphical objects as well as a low-level shader control to create specialized visual effects that will be used in the future to create more realistic avatars. Its plug-in-based architecture also provides greater ability to interoperate with other software tools.

Figure 6. Motion Capture of Dr. Schwarzkopf

For the text to speech synthesis, Microsoft SAPI 5.1 was chosen as the Application Programming Interface (API). MS-SAPI provides an event generation mechanism to report the status of phoneme or word changes during the synthesis of voice in real time. These events can be used to create realistic phoneme-based lip animations. Furthermore, a number of commercial speech systems provide an interface to SAPI so that an application can transparently leverage a multitude of speech systems.

Figure 7 depicts the Lifelike Responsive Avatar Framework (LRAF) that controls the avatar and provides connectivity to the SR and DM. The LRAF drives the avatar's operation to create a realistic representation capable of speech input, emotive response, and vocal response. The LRAF has two separate input sources. One is the Lifelike Dialog Manager which provides sentences that are intended to be spoken by the avatar. The other is the user's behavioral information such as eye-gaze. Currently, as an approximation for eye-gaze tracking, we are using an infrared camera to track retro-reflective markers on a head-band. A current objective is to also track the user location when speaking with the avatar.

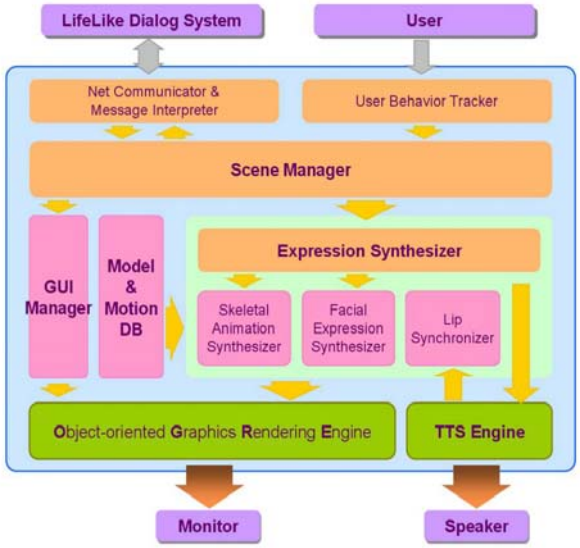

Figure 5. Lifelike Responsive Avatar Framework

The most significant component of the LRAF is the Expression Synthesizer which is responsible for taking the 3D models and applying the motion-capture data to produce a sequence of facial and body animations that fit the context of what has been spoken. Three major components of the Expression Synthesizer are: 1. the Skeletal Animation Synthesizer, 2. the Facial Expression Synthesizer, and 3. the Lip Synchronizer. The initial version of LRAF uses only static animation playback.

However, research is progressing toward more complicated algorithmic control of the animations. For example, the motion-captured skeletal animations could be varied (through exaggeration or attenuation) to correspond to the avatar's subtle emotional changes. This will enable a more complex and believable personality model of avatar behavior, as found by Kshirsagar and Magnenat-Thalmann (2002).

DEMO AND AVATAR ACCEPTANCE RESULTS

The initial demonstration system allowed any person affiliated with the I/UCRC program to initiate a conversation with the avatar without prior training. The prototype language lexicon includes proper names for about 200 members of the NSF I/UCRC program who belong to various universities and/or hospitals. Each university may have multiple centers that sponsor different engineering research programs that NSF supports with funding. Accordingly, approximately 600 items such as university names and NSF centers are accounted for in the lexicon.

To study the user acceptance of an avatar in this environment, we conducted a study to determine which elements are most important in creating realistic and useful avatars as interfaces for question and answer systems. The test population consisted of thirty (n=30) students from the University of Illinois at Chicago and these subjects represented numerous ages and academic backgrounds. Nine subjects were female. The approach employed the classic experimental design model and was based on the work of Koon (2006) and Garau et al (2002) whereby within each specific experimental trial (called a pairing) an independent variable is manipulated in hopes of causing a direct result in a dependant variable. The study displayed ten short videos of our AlexAvatar, discussing various programmatic aspects of NSF. Segments were about 30 seconds long and were paired (5 pairs) to test for the following variables: gaze (eye contact), head motion, body motion, voice, and a baseline comparison of the current version of the avatar with a version from a year earlier. Subjects answered a set of seven Likert scale questions, a useful tool for studying satisfaction of user interface systems by Epps, Close et al (2007). The changes in answers from one version of the avatar to the next in each pair were tracked to determine the significance of each variable. Open ended questions were also asked to determine the justification for the subjects' responses.

The two most significant variables were body motion and the baseline test. Subjects clearly preferred our motion-captured body motion compared to a still avatar. All seven questions showed a preference for the motion; paired t-tests showed significant shifts in the mean responses on the Likert questions ($\bar{x}$ ranges from 0.633 to 1.533) with all seven having two tail statistical significance scores better than .011.

Even stronger results were found within the baseline comparison pairing; all seven means shifted strongly in favor of the newer avatar ($\bar{x}$ ranges from 1.53 to 2.33)

with two tail statistical significance scores of 0.01 or lower. The baseline test compares two versions of the avatar. The first was created roughly a year ago and the second is a recent revision. The early one lacks structured or purposeful motions while the second incorporates motion capture data from Alex, the basis for our character. The Text-To-Speech (TTS) system of the first avatar is several years old while the newer avatar incorporates the latest TTS voice technology and runs on the most recent version of Microsoft SAPI, a voice synthesis engine. Finally, the textures, model, and background of the newer avatar are more detailed and precise creating a more compelling and realistic looking character. These changes represent significant improvements in image, movement, and sound. Subjects prefer body motion to a still avatar and there is strong evidence that our work over the last year has yielded a substantially improved avatar.

The most intriguing result was the statistical tie in preference between a TTS and pre-recorded content. Results indicated that 13 preferred TTS and 15 preferred a pre-recorded voice (1 tie); 14 thought TTS was more realistic and 13 thought pre-recorded was more realistic (2 ties). Reasons for each preference were varied and complex indicating a strong need for further research into this area. Figures 8 and 9 represent the results regarding voice preference, recorded or TTS.

English is their first language * Which Voice Video Did You Prefer?

Crosstabulation

Count

		Which Voice Video Did You Prefer?		
		TTS	Recorded	Total
English is their first language	English	8	8	16
	ESL	5	7	12
	Total	13	15	28

Figure 6. Avatar Voice Preference

Sex of the subject * Which Voice Video is More Realistic?

Crosstabulation

Count

		Which Voice Video is More Realistic?		
		TTS	Recorded	Total
Sex of the subject	Male	10	8	18
	Female	4	5	9
	Total	14	13	27

Figure 7. Avatar Voice Realism

FUTURE WORK

More testing of the system is being undertaken to improve its performance. In our particular test scenario, AlexAvatar used direct query to identify a user whose identity could be found in a database containing 200 names and last names of people related to the I/UCRC program. The test AlexAvatar system was 100% effective after it asked the user for his identity not more than three times. By controlling the context of the conversation in real-time, the appropriate group of grammars to be used is selected for that context. This minimizes the speech recognition search space, improves performance and minimizes duplication of grammar items.

A follow-up demonstration system is currently being developed to expand the feature set of the first AlexAvatar prototype. In essence, it expands upon the original design by adding two key features: temporal awareness and multi-user capability.

Temporal awareness refers to the idea that the avatar can retain and recall chronologically relevant user-specific data. For example, the avatar is being enhanced to reference episodic memories concerning previous conversations in the current conversation. The multi-user feature currently under development will allow the avatar to conduct a group meeting, where two or more human users wearing headsets can interact with the avatar, which will turn to face each person being addressed.

ACKNOWLEDGEMENTS

This research is supported by NSF Collaborative Research award 0703927.

REFERENCES

Bickmore, T. W. and Picard, R. W. (2004). Towards caring machines. *Computer Human Interaction*. Retrieved June 1, 2008 from http://affect.media.mit.edu/pdfs/04.bickmore-picard-chi.pdf

Cassell, J., Ananny, M., Basu, A., Bickmore, T., Chong, P., Mellis, D., Ryokai, K., Smith, J., Vilhjálmsson, H. & Yan, H. (2000). Shared reality: Physical collaboration with a virtual peer. *Proceedings of CHI 2000*.

Chant Software. (2008). Integrate Speech Technology for Hands-free Operation. Retrieved June 6, 2007, from

http://www.chant.net/Products/SpeechKit/Default.aspx.

Donner, C. and Jensen, H. W. (2005). Light diffusion in multi-layered translucent materials. *ACM SIGGRAPH 2005 Papers*. Los Angeles. California. ACM.

Epps and Close. (2007*).* A study of co-worker awareness in remote collaboration over a shared application. *CHI '07 extended abstracts on Human factors in computing systems*. San Jose, CA. New York: ACM.

Garau, M., Slater, M., Bee, S. & Sasse, M. A. (2001). the impact of eye gaze on communication using humanoid avatars. Proceedings of the *SIGCHI conference on Human factors in computing systems*. Seattle, Washington. New York: ACM.

Heinrichs, A., Muller, M. and Tewes, A. (2006). Emergent Graphs with PCA features for Improved Face Recognition. *Institutfür Neuroinformatik*, Ruhr-Universität, D–44780 Bochum, Germany

Kenny, P., Hartholt, A., Gratch, J., Swartout, W., Traum, D., Marsela, S. and Piepol, D. (2007). Building Interactive Virtual Humans for Training Environments. *I/ITSEC 2007*.

Koon, K. (2006). A case study of icon-scenario based animated menu's concept development. Proceedings of the 8th conference on *Human-computer interaction with mobile devices and services.* Pp. 177-180.

Kshirsagar, S. and Magnenat-Thalmann, N. (2002). A Multilayer Personality Model. *Proceedings of the 2nd international symposium on Smart graphics*. Hawthorne, New York, ACM.

Lee, C., Sidner, C. and Kidd, C. (2003). Engagement During Dialogues with Robots. *AAAI Spring Symposia.*

Microsoft Developer Network.(2008). Speech API Overview (SAPI 5.3). Retrieved, June 6, 2007, from http://msdn.microsoft.com/en-us/library/ms720151(VS.85).aspx

Ney H. & Ortmanns, S., (1999). Dynamic programming search for continuous speech recognition, *IEEE Signal Processing Magazine*, 16, 64–83.

Oblivion: Oblivion (2008) Retrieved June 1, 2008, from the Wiki:
http://www.uesp.net/wiki/Oblivion:Oblivion

OGRE 3D: Open Source Graphics Engine. Available from http://www.ogre3d.org, visited on January 2006.

Sherwell, B. W., Gonzalez, A. J. & Nguyen, J. (2005). Contextual Implementation of Human Problem - solving Knowledge in a Real-World Decision Support System. Proceedings of the Conference on *Behavior Representation in Modeling and Simulation*, Los Angeles, CA, May.

Stensrud, B. S., Barrett, G. C., Trinh, V. C. & Gonzalez, A. J. (2004). Context-Based Reasoning: A Revised Specification. *FLAIRS Conference* 2004.

Wendt, S., Fink, G. & Kummert, F. (2002). Dynamic Search-Space Pruning for Time-Constrained Speech Recognition (2002). Proceedings from ICSLP-2002*: The 7th International Conference on Spoken Language Processing*, 377-380.