

Final Report of NSF-funded Workshop:

Accelerating Progress in Perceptive Animated Interfaces and Virtual Humans

Ron Cole

University of Colorado Boulder

http://cslr.colorado.edu/beginweb/2004animated_wkshp/

Abstract

In 2003, a proposal was submitted to Dr. Mary Harper at the NSF by Ron Cole, Javier Movellan and Jonathan Gratch to organize a workshop on Perceptive Animated Interfaces incorporating Virtual Humans. The workshop was held April 9-10, 2004 in San Diego California. This report describes the importance of research on virtual humans, the objectives of the workshop, the main workshop activities and outcomes and recommends activities that can foster research leading to a new generation of more effective and engaging human computer interfaces that incorporate natural dialog interaction with virtual humans.

1. Why Virtual Humans?

Advances in computing, communication and human language technologies suggest the feasibility of inventing a new generation of human computer interfaces that engage individuals in natural face-to-face conversational interaction with intelligent animated characters. These perceptive animated interfaces will incorporate virtual humans that interact with people much like people interact with each other during face-to-face conversational interaction. Virtual human interfaces will use language processing and machine perception technologies to interpret the user's speech, facial expressions, gaze and hand and body gestures. Lifelike computer characters, with personality and attitude, will orient to the user's location in space and provide real time feedback while the user is speaking through head nods, facial expressions and other behaviors) and interpret the speaker's auditory and visual behaviors to infer the user's intentions and cognitive state. The animated agents will produce natural and expressive speech accompanied by contextually appropriate facial expressions and gestures consistent with the agent's unique personality. We propose work to establish a vital research community to stimulate and enable research and development of perceptive animated interfaces.

Perceptive animated interfaces incorporating virtual humans will be of great value to society, as they will revolutionize learning, training, therapy, interpersonal electronic communication, information access and retrieval and online transactions. The advent of intelligent animated agents will present unprecedented opportunities to engage and empower individuals to learn new skills, communicate more effectively, and increase

their participation in the emerging information society. They can help people learn to read or to speak, and can liberate teachers from some routine teaching tasks and help them tailor the learning process to the specific needs of each student.

The invention of virtual humans provides a new and exciting task domain for multidisciplinary research leading to development and integration of converging technologies to improve human performance. For example, inventing virtual humans and assessing their effectiveness in different tasks (e.g., Web guides, science tutors, job counselors or therapists) requires the integration of new ideas and technologies about the realization of personalities through communication behaviors, and new architectures that can handle real-time interaction between individuals and agents across modalities operating at different time scales. Virtual humans, once invented, will become an invaluable research tool for designing more engaging and effective interfaces and for understanding human behavior, as it will be possible to manipulate the verbal and nonverbal behaviors of virtual humans in various contexts (learning, training, therapy, persuasion, etc.) to measure and understand the affects of these behaviors on individuals' experiences and behaviors. In fact, invention of virtual humans will enable researchers in social and clinical psychology to conduct experiments that are difficult or impossible to do today with human experimenters, since virtual humans can be programmed to produce verbal and nonverbal behaviors, such as eye gaze and gestures, which are largely unconscious during human communication (and therefore difficult to control by human "stooges in experiments).

The Challenge of Virtual Humans

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, 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, while the listener provides simultaneous auditory and visual feedback to the speaker (e.g., "I agree," "I'm puzzled," "I want to speak."). 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 and backchannel cues, all at the same time.

Inventing systems that model the social dynamics of face-to-face communication in different social contexts is a daunting task. The system must simultaneously interpret and produce auditory and visual signals. 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, while a shift of gaze can indicate that a response is expected. Paralinguistic information is also critical, since

the prosodic contour 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.

Developing such systems requires advances in speech recognition, natural language generation and synthesis, facial animation, recognition of facial expressions and gestures, dialogue interaction and imparting personalities to computer agents. As well, realizing these scenarios requires a deeper understanding of the nature of human communication and human computer interaction. Most importantly, achieving these advances in knowledge and technology requires a community of researchers willing to work in an interdisciplinary manner and willing to go beyond the boundaries of well-established research communities. Speech researchers, for example, need to go beyond their traditional area of expertise and interact with computer vision researchers, psychologists, and computer animators. The rudiments of such a community are already established but are in dramatic need for consolidation.

In addition to developing the technologies that enable virtual humans to converse with people, there are a host of research issues related to their representation and integration in complex systems. According to Swartout et al (2006): “Achieving human-level intelligence in cognitive systems requires a number of core capabilities, including planning, belief representation, communication ability, emotional reasoning, and most importantly, a way to integrate these capabilities. And yet, for many researchers, software integration is often regarded as a kind of necessary evil – something to make sure that all the research components of a large system fit together and interoperate properly – but not something that is likely to contribute new research insights or suggest new solutions. We have found, on the contrary, that ...the integration process has raised new research issues and at the same time has suggested new approaches to longstanding issues.”

State of the Field

There is a solid theoretical basis to motivated research and development of virtual humans. A significant (and growing) body of research indicates that computer programs provide more engaging, satisfying and effective experiences when their design is based on knowledge about human social conventions and the dynamics of human communication (Reeves and Nass, 1996; Nass & Brave, 2005). Such interfaces are intuitive, pleasing and effective because they foster *social agency*—they enable us to interact with computer programs like we interact with other people using highly overlearned (and largely unconscious) behaviors that we apply to daily social interactions. While theory and research suggest that social agency plays a fundamental role in interaction with all media (Reeves and Nass, 1996), social agency is realized most effectively and powerfully when made manifest through voices and faces (Reeves and

Nass, 1996; Nass & Brave, 2005; Mayer, 2001; Moreno, 2001, Atkinson, 2002; Baylor et al., 2003, 2005; Wang et al., 2005).

Virtual human systems are rapidly becoming a reality. Only a few years ago, programs that incorporated virtual humans were laboratory systems, and few had scaled to real world applications. Moreover, systems that were deployed were limited to interactions with avatars using mouse clicks. Today, several programs have been developed that incorporate spoken dialog interaction with virtual humans in real world applications, with impressive results. At the University of Colorado, Ron Cole and his colleagues have developed programs that use a virtual human to teach children to read, and to conduct speech therapy for individuals with Parkinson disease and aphasia (van Vuuren, in press; Cole et. al., in press-a, in press-b). These programs have undergone clinical trials and have demonstrated efficacy. The reading tutor is currently in use in over 30 kindergarten, first grade and second grade classrooms in 5 school districts in Colorado. Descriptions of these systems can be found at the workshop web site at http://cslr.colorado.edu/beginweb/2004animated_wkshp/readings.html. At USC, Lewis Johnson and his colleague have developed a program that enables users to control the nonverbal behaviors of their avatar, a US soldier, and to interact with other avatars, such as Iraqi civilians, through spoken dialog interaction. This program, which teaches both spoken language and social skills for communicating and building trust with Iraqis, is being used by the US military to train soldiers who will serve in Iraq. A description of this program can be found at http://www.isi.edu/isd/carte/proj_tactlang/index.html.

While the work at Colorado and USC have produced effective systems that use virtual humans in real world applications, these systems have emerged from multi-year, multimillion dollar efforts. Moreover, these systems have not resulted in community resources that can foster research by a community of researchers. The reality is that the expertise and infrastructure required to develop effective and scalable virtual human systems resides in just a few laboratories.

The development of interfaces that incorporate virtual humans requires collaboration among researchers in many areas—psychologists, linguists, speech scientists, engineers and computer scientists with multidisciplinary expertise in human communication, interface design, speech and language technologies, dialogue modeling and management, computer vision and computer animation. While individual researchers, research labs and existing research communities represent knowledge and skills in each of these areas, no research community exists today that strives to focus the necessary multidisciplinary resources on research and development of perceptive animated interfaces incorporating virtual humans.

In general, the situation today is that researchers in different fields are working more or less independently in separate areas within psychology, cognitive science, linguistics and computer science studying problems related to human communication, expression of emotions and gestures during communication, spoken dialogue systems, computer vision, computer animation and gesture recognition. Together, these fields provide a significant

base of knowledge, technologies and methods that are critical to development of virtual humans.

In addition, a number of research communities are emerging that focus on different aspects of the general problem of face-to-face communication with intelligent animated agents. For example, there are active (and largely independent) groups of researchers investigating audio-visual speech processing, face and gesture recognition, affective computing, perceptive interfaces and character animation. These groups tend to get together via conference workshops or annual meetings that run in collaboration with larger, more established conferences.

To accelerate research and development activities leading to new knowledge and technologies for inventing virtual humans, it is crucial to establish a community of researchers from all relevant disciplines that will work together to initiate and undertake the many tasks required to make these interfaces a reality. These activities include defining research goals and challenges, sharing knowledge about prevailing theories and methodologies in each discipline, proposing and designing system architectures for inventing and evaluating virtual humans, defining realistic and challenging task domains, building prototype systems as test beds for research, establishing evaluation criteria for measuring progress, and identifying and developing critical infrastructure that enables this work, and is accessible and available to all.

We thus organized a workshop, supported by a grant from the NSF, to bring together researchers from different disciplines to share their knowledge, expertise, tools and technologies, to identify the main barriers to inventing virtual humans, and to help plan a research agenda to accelerate progress. We also hoped that the workshop would foster new multidisciplinary collaborations which are necessary to establish a research community with a shared vision for virtual human research.

2. Workshop Objectives

The main goals of the workshop are (a) to understand prior work and research challenges required to develop perceptive animated interfaces and virtual humans, (b) determine practical steps and activities required; and (c) initiate a set of activities that will help establish a vital community of researchers who will work together to accelerate progress.

The workshop was designed to address any of the following questions:

- What is the state of scientific knowledge about perception, production and interpretation of auditory and visual behaviors during face-to-face communication? How are these behaviors influenced by task domain, social influences, and other variables? What knowledge can be applied immediately to the design of perceptive animated interfaces? What scientific knowledge is missing, and what research is required to gain this knowledge?

- What are the capabilities and limitation of technologies and methodologies currently used in research, development and evaluation of advanced dialogue systems? What sorts of perceptive animated interfaces can these technologies support today? What key research breakthroughs are needed in speech and language technologies, what is required to achieve these breakthroughs, and how will these breakthroughs translate into more effective perceptive animated interfaces?
 - What is the state of the art of computer vision technology relative to monitoring and interpreting visual behaviors to enable face-to-face communication with an animated character? What is the missing science? What key research breakthroughs are needed to enable perceptive interfaces? What research tools, corpora, and systems are currently available to enable research and development efforts? What new infrastructure is needed to conduct research? What effort and cost is required to develop this infrastructure?
 - What is the state of the art of animation technology? What research and development activities are needed to produce natural and contextually appropriate facial expressions, eye movements, and hand and body movements in different tasks? What infrastructure is required to achieve key research breakthroughs?
 - What architectures have been proposed or implemented to support real time dialogue interaction between users and virtual humans? Does the proposed system architecture and task domain enable researchers to achieve and measure key research objectives? How do we measure and evaluate the performance of these systems and system modules? How do we compare different systems? How do we measure progress over time?
- What systems could and should be developed to serve as test beds for research and development of perceptive animated interfaces? What task domain(s) should be selected?
- What resources—annotated corpora, research tools, etc.—are needed to study relevant communication behaviors between people or between people and machines, and to enable researchers to train and evaluate machine perception and generation algorithms? (In the appendix below, we explain the critical importance of developing corpora to enable research in perceptive animated interfaces, and provide examples of how development of corpora has accelerated progress in science.)
- What standards are required to assure interoperability of system components and real time interaction over communication channels?

- What metrics and methodologies are required to evaluate and compare systems and system components, and to measure progress within and between research and development sites over time?
- What concrete steps should the research community take to stimulate and sustain research, and to create a strong and enduring community that will realize the vision of perceptive animated interfaces?

3. Workshop Planning

Shortly after a grant was awarded to the University of Colorado to organize the workshop, PI Ron Cole and co-PIs Jonathan Gratch and Javier Movellan formed the workshop organizing committee, consisting of Amy Baylor, Justine Cassell, Ron Cole, Susan Duncan, Jonathan Gratch, Eric Hamilton and Javier Movellan. Links to web sites of each member of the organizing committee can be found at http://cslr.colorado.edu/beginweb/2004animated_wkshp/orgcommit.html.

Over a period of several months, the committee worked together to accomplish two main goals: select and invite participants to the workshop, and plan the workshop agenda.

The participants were selected from the U.S., Canada and Europe to include individuals conducting research into virtual humans (e.g., Norman Badler, Amy Baylor, Justine Cassell, Jonathan Gratch, Lewis Johnson) computer scientists and engineers researching and developing the technologies that power virtual humans (e.g., Wayne Ward, Jiyong Ma, Eric Petajan, Javier Movellan) and psychologists and psycholinguists conducting research in areas of interface design, face to face human communication and speech and gesture (e.g., Janet Bavelas, Susan Duncan, David McNeill, Clifford Nass). A full list of workshop participants can be found at:

http://cslr.colorado.edu/beginweb/2004animated_wkshp/attendees.php.

4. Workshop Activities and Outcomes

The two-day agenda was carefully designed by the organizing committee to provide a common grounding for all workshop participants through (a) a series of short presentations by leaders in the field, (b) a “poster” session in which some of the participants presented demonstrations of systems that incorporated virtual humans, (c) two breakout sessions, with groups in each session focusing on specific topics. The workshop agenda can be found at:

http://cslr.colorado.edu/beginweb/2004animated_wkshp/agenda.html.

During the workshop, the agenda was modified dynamically by the organizing committee in response to discussions in plenary sessions. For example, due to time constraints, technology demonstrations, which were originally scheduled as individual presentations, were moved to a single open session to save time and enable more interaction among participants. Similarly, the topics of the breakout groups were decided in plenary meetings. We believe that these mid-course modifications helped facilitated communication on topics of most relevance to the workshop attendees.

Three of the four breakout groups that met on the second day of the workshop submitted written reports in the months following the workshop. These reports are presented verbatim in the appendix.

5. Recommendations

Insights about the steps required to establish a vital research community to accelerate research and development of virtual humans can be guided by prior efforts that produced successful outcomes. Lessons can be learned, for example, from DARPA speech and natural language processing programs, which span over three decades. Because development of speech and natural language processing technologies were judged to be in the national interest, DARPA designed and funded programs that brought together researchers who worked together to define challenging tasks and develop systems within 5 years to achieve targeted levels of performance. The research community then worked together to identify the infrastructure needed to develop the proposed systems, and developed rigorous evaluation methodologies and metrics to measure and compare performance of different systems and to measure progress of all systems over time.

One of the key lessons learned from these programs is the critical importance of infrastructure, and the remarkable amount of work required to produce it. In the area of speech recognition (which is just one component of a virtual human) infrastructure includes annotated speech corpora, pronunciation dictionaries, lexicons, and tools for training and evaluating speech recognition systems. Development of infrastructure in each of these areas required many thousands of hours of work.

The participants in the breakout groups at the Virtual Humans Workshop posed a wide variety of exciting and important questions that can be answered only through research. Many of these questions require the integration of virtual humans into specific applications in which characteristics or behaviors of virtual humans are manipulated. Thus, research tools must be developed that combine state of the art machine perception and generation technologies with authoring tools for designing a wide variety of applications with precise control of virtual humans.

While a few laboratories have acquired the expertise and resources to develop virtual human systems, the infrastructure and expertise required to invent and evaluate virtual humans is vested in relatively few laboratories. It is in everyone's best interest to accelerate invention of virtual humans by gaining fresh perspectives from researchers across many different disciplines, and this can be accomplished by providing them with the resources to stimulate and sustain new multidisciplinary research efforts.

We thus encourage interested researchers to work together to submit proposals to NSF CISE CRI and MRI (and other relevant) programs to develop community resources in support of virtual human research. Interested researchers should also investigate programs at NIH, DARPA and Education that focus, respectively, on areas of health, training and education. There are many active programs in these agencies that are

designed to support both basic research and system development that could incorporate virtual humans and could provide support for development of community resources to foster virtual human research. It is our experience that program managers at NIH and Department of Education are accessible and eager to discuss and encourage new research ideas, including research programs that incorporate virtual humans.

We also encourage the NSF to explore partnerships with other agencies to develop joint programs that support research and development of virtual humans. As virtual human systems have the potential to provide immersive and effective outcomes in math and science education, in clinical treatments, and for training personnel in industry, homeland security and the military, there may be great interest in other agencies in developing new programs that bring scientists from different disciplines together to conduct basic research leading to more powerful and beneficial tools and applications in education, health and training. As basic research is required to invent and integrate the core technologies that underlie virtual humans, and to discover and model the behaviors that make virtual humans engaging and effective in different application domains, agencies that deal with education, health and national defense may be willing to explore new initiatives with the NSF to acquire the knowledge and technologies needed to realize the potential benefits of these systems.

References

- Atkinson, R. K. (2002) Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology*, 94 (2), 416-427.
- 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).
- Cole, R., Halpern, A., Ramig, L., Van Vuuren, S., Ngampatipatpong, N., and Yan, J. (In press). A Virtual Speech Therapist for Individuals with Parkinson Disease. Special Issue of *Educational Technology Magazine*.
- Cole, R., Van Vuuren, S., Pellom, B., Hacıoglu, 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.
- Cole, R., van Vuuren, S., and Wise, B. (In press) How Marni teaches children to read. Special Issue of *Educational Technology Magazine*.
- Johnson, W.L., Vilhjalmsson, H. and Marsella, M. (2005) Serious Games for Language Learning: How Much Game, How Much AI? 12th International Conference on Artificial Intelligence in Education , July 18-22, Amsterdam, The Netherlands.

Mayer, R. (2001) *Multimedia Learning*. Cambridge, UK: Cambridge University Press.

Moreno, R., Mayer, R.E., Spires, H.A., Lester, J.C., (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), 177–213.

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.

Swartout, W. Gratch, J., Hill, R., Hovy, E., Marsella, S., Rickel, J. and Traum, D. “Toward Virtual Humans,” in *AI Magazine*, v.27(1), 2006
www.ict.usc.edu/~gratch/AI-mag06.pdf

Van Vuuren, S. (In press). Animated Pedagogical Agents that Teach and Conduct Therapy: Vision, Challenges and Capabilities. Special Issue of *Educational Technology Magazine*.

Appendix

Breakout Group Reports

Should virtual humans be realistic? (Amy Baylor, scribe)

The initial focus of this discussion was regarding the importance of distinguishing virtual human (VH) *realism* versus *naturalness*. The group discussed that VH realism refers to its superficial features – i.e., how it appears to the user. In contrast, VH naturalness refers to whether it functions (or behaves) naturally. Cassell stressed (and others agreed) that naturalness is more important. The group discussed the importance of considering these as multiple dimensions of realism, particularly as perceived by users within the particular context and intended purpose of the VH.

The need for designers and developers to carefully consider the desired outcome of the VH before moving considering realism issues was a key theme of the discussion. For example:

- Is the VH for a short intervention or a long-term relationship?
- What is the intrinsic value for employing a virtual human?
- What role of VH is most appropriate (e.g., co-learner, expert, motivator, peer, non-human)?
- Is realism even an important issue to consider given the context?

There was concern that some are just developing the “best” (i.e., most realistic and intelligent) VHS without considering first the overall macro-level purpose of the VH within its intended context.

Consequently, while evaluation of VHS is needed from both a macro- and micro-level, the group concurred that there may be a greater need at this point in time for more breadth (macro-level evaluation) in VH research than depth (micro-level evaluation). It was discussed that a good way to frame macro-level research questions would be to start with predictions based on human literature and social psychology. The importance of considering the length of time that the user works with the VH was also discussed as an important consideration in assessing first impression vs. long-term goals. A central theme of the discussion focused upon the need for VH evaluation to be guided by a clear sense of what would make the VH effective for its intended purpose(s) (e.g., so that user “hangs out longer” in environment, or learns more, or is more engaged).

With respect to empirical results, Baylor reported her research group’s findings that employing realistic VH images (as contrasted to more cartoon-like images) may result in significantly greater learning outcomes for college students. This is in contrast with Norman’s speculations in the past that VHS that are too realistic may be distracting to users and elicit exaggerated expectations. When given a choice, Baylor also discussed her findings that African-American college students are significantly more motivated to work with Black VHS (much more so than White students working with White VHS). Nass stated that this “intra-ethnic honesty” is consistent with other work done with children with humans in the area of medical evaluation. Baylor also discussed a somewhat

disturbing trend where both male and female students tend to report a more positive learning experience (and sometimes even learn more) when working with male VHs.

Databases (Javier Movellan, scribe)

The break-out group on databases was composed of Susan Duncan, Javier Movellan, Eric Petajan, and Jianxia Xue. Most of our meeting was devoted to the issue of coding schemes for annotation of multi-modal databases of mouth, face, hand, and body motion and position data, such as would be relevant for tracking and modeling the sorts of behaviors that occur in human interaction.

Coding schemes can be categorized in a hierarchy where at the lower level would be actual audio and video sequences and the higher level would be annotations that describe semantic descriptions of large video sequences. MPEG-4 Face and Body Animation (FBA) and FACS are examples of low-level coding schemes. MPEG-4 FBA Face Animation Parameters (FAPs) code non-rigid facial movements in terms of displacements of a predefined set of points (e.g., points on the contours of the lips). Rigid head and body movements are coded as Euler angles of the different body parts. MPEG-4 FBA is designed for both coding and decoding (animation) i.e., it provides a standard for coding facial movement and a standard for animation of 3D models. In addition, the Viseme and Expression FAPs allow high level two visemes and two expressions to be specified with weighting factors for each video frame. FACS on the other hand is designed to code facial expressions, not to decode them. FACS provides a basis of 44 expression morphs that involve spatially localized regions of the face.

Harmonizing FACS and MPEG-4 FAPs would provide the FACS community with the representation that they are familiar with plus the synchronized low level motion capture data provided by FAPs. One potential difficulty in this harmonization is the fact that in some cases FACS codes rely heavily on textural aspects (e.g., wrinkles and shadows) not easily captured by the current MPEG4 FAP's standard which is based on movement of a collection of points on the face. The converse is also true. Speech related lip movements, for example, are not coded in FACS at the resolution level needed for realistic animation.

The group also discussed the challenges involved in generating the MPEG-4 parameter representations of hand and body motions and positions (gestures) that could comprise the intermediate level of representation between low-level video and high-level semantic labeling data on video sequences. Both FACS and MPEG4 may provide very valuable intermediate representations to enable analysis of human behavior but analysts should be aware that both methods may lose information they care about. Further, it is not yet clear what would be the best method for generating the sort of MPEG-4 parameter-based specifications of gestural behavior that would be most useful for modeling work on virtual humans.

Important issues surfaced later in our meeting leading to an interesting and contentious exchange of ideas. The human behavioral scientists present raised the concern that any form of coding brings with it an underlying philosophy about which dimensions of

behavior are and are not of interest. Javier pointed out that inclusion of low-level information (e.g., actual audio and video) would address this concern, as other levels of representation could then be reworked or abandoned as dictated by changing scientific knowledge and theoretical perspectives. The group also discussed the importance of including as much rich context in the low-level audio video data as practicable (i.e., multiple camcorder views capturing all interactants, as well as high-quality audio.), the point being that the low-level data itself could mislead if that context is not clearly captured.

The group discussed the benefits of shared databases that include multiple ways of coding. While the social sciences community may argue about the need for such datasets, the need is unquestionable for making progress in computer vision and computer animation. This may be in part due to the fact that both computer vision and computer animation address low-level issues about human behavior.

Susan and Javier expressed strong concerns about schemes that categorize body movement, which are intrinsically continuous, into discrete categories. By doing so, they may lose the information we should really care about. For example the mode in which a gesture is done may be of more interest than the gesture category itself. The way a sentence is spoken may be of more importance to social interaction than what is being said (a textual transcription). This is also an issue with motion capture approaches. For example, it is well known that motion capture, loses a great deal of the expressiveness of movement in the human body (e.g., the springiness of body fat).

Virtual Human Architectures (Jonathan Gratch, scribe)

The break-out group on architectures consisted of, Norm Badler, Jonathan Gratch, Mary Harper, Jintao Jiang, Lewis Johnson, Jiyong Ma, Matthew Stone, Paul Tepper, Matthew Turk and Wayne Ward. Our group considered the question of how to move toward shared architectures and tools for Virtual Humans, specifically focusing on the form such an architecture should take. Although there was consensus that such sharing would advance the field, there was also considerable agreement that significant technical and social hurdles must be overcome to achieve this goal. The discussion focused on each of these hurdles in turn.

Technical challenges: A number of technical challenges and uncertainties must be addressed before research groups could effectively share their tools. The general vehicle of such sharing seems clear: there should be some form of common architecture with well defined application programmer interfaces (APIs) that would allow groups to replace individual modules. The challenge in specifying such an architecture, however, arises from the considerable heterogeneity of research issues addressed by members of the virtual human community. The current state of the art was characterized as “4 + 1 of N integrations” of different functionalities with limited overlap in the functions integrated across research groups and often irreconcilable design decisions implicit in the underlying architectures. Two central issues dominated the discussion: locus-of-control and integration incrementality.

In terms of locus-of-control, the group argued as to what should be the “heartbeat” of the virtual human. Most existing architectures have some central and inflexible component that is essence the “sync signal” for the virtual human and that, in large measure, all other components must synchronize their behavior with. For example, in many conversation-focused systems, the speech synthesis system is the sole source of timing constraints to which behavior must be synchronized. This stands in contrast to human behavior where speech production may be paused or slowed in response to physical events, conversational back channeling, or mechanical constraints on the body’s ability to gesture. Other more environmentally situated systems can synchronize behavior to physical events and user feedback, but then face a challenge in reconciling these timing constraints with speech synthesis. One possibility is to allow multiple heartbeats (allow each system to have its own clock) but this may make it hard to impose consistency on the overall behavior of the agent.

The problem of reconciling constraints will be a general problem in any architecture, but it is made worse by the fact that many critical modules are limited in their ability to accept or produce such time information in their APIs: one couldn’t impose a single heartbeat, even if one wanted to. For example, most concatenative synthesizers have only crude ways of altering the timing of synthesized speech (particularly if one wishes to retain the conveyed meaning), and speech recognition systems rarely provide the incremental feedback that could drive more reactive back channeling behaviors. Thus, one possible solution would be to provide more explicit ability to import and export timing information into the various virtual human components so they could be synced to a single signal.

A second technical challenge concerns the question of integration incrementality. Many the modules underlying virtual human behavior were developed in isolation to perform specific tasks and don’t accept incremental input or generate incremental results. This exacerbates the scheduling problems discussed above and contributes to an overall lack of responsiveness in the virtual human. For example, one could have more flexibility in reconciling timing constraints the speech synthesizer operated on smaller units and one could provide more responsive feedback if the speech recognition system provided partial results. In general, the group felt this issue could be resolved with existing technology. For example, one could simply change the way people typically use modules (current synthesizers could generate speech at the phrase level, though there may be minor timing and coarticulation anomalies). One could also change the APIs of existing modules to provide more incremental input and output.

Social Challenges: The virtual human community is also faced with social challenges that in many ways have contributed to the current heterogeneity of research goal and tools. Virtual human research is an interdisciplinary effort spanning several well-established research communities, each with its own theoretical concerns, conferences, journals and funding sources. In contrast to these individual disciplines, the virtual human “community” is maintained through a series of small workshops that bring together “4 + 1 of N” research areas on some specific subtopic of interest. There is no virtual human

journal, no virtual human conference, and few significant research grants with the explicit goal of advancing virtual humans (rather, the goal is to advance intelligent tutoring, behavioral animation, natural language processing, etc.). There have been several attempts to agree on a single workshop to attend, but each of these is perceived as not sharing the values of a significant portion of the community: the conference on autonomous agents and multi-agent systems is perceived as too “agent” centric, Computer Animation and Social Agents is viewed as too animation centric, etc. (Mary Harper suggested the International Conference on Multimodal Interfaces might be neutral territory.) All this suggest a need to promote, articulate, and fund a set of shared and agreed upon research goals

There was considerable disagreement on how to overcome these social obstacles, but we agreed that both top-down and bottom-up efforts are necessary. From the top-down, an NSF or DARPA sponsored effort could facilitate interaction through a suitably crafted all encompassing research program. The group discussed what such a program might look like; what tasks might embody the wide range of research goals largely shared by the virtual human community. Possibly a range of tasks could be contrasted: multi-way communication vs. 1-on-1 ongoing conversation vs. a real-life Microsoft paperclip (advisor on the side). Norm Badler suggested having a mechanism where virtual humans from different research groups would have to have a conversation with each other. But progress must also come from the bottom up. For example, Lewis Johnson could get together with Wayne Ward and sort out why current speech technology doesn't meet his needs. Various mechanisms could help foster such interactions. Research groups could exchange graduate students across disciplines to promote shared understanding; a common mailing list could improve community awareness; a series of tutorials at key conferences could build understanding and bridges between related venues.

Although these social and technical obstacles are significant, we are confident that they will, and in many ways are already being overcome. Just in the last two years since USC hosted a virtual human workshop on shared architectures, there has been considerable growth in the virtual human community. We have grown in size, in expertise, but also in a shared understanding of the limitation of our individual perspectives and technology. Increasingly, the consensus view is we can not go it alone. With a little effort on each of our parts, we won't have to.