Embodied Conversational Agents and Interactive Virtual Humans for Training Simulators

Girija Chetty[1], *Matthew White*[2]

[1]University of Canberra
[2]Infinity Imaging Pty. Ltd.
Girija.chetty@canberra.edu.au, matthewwhite360@gmail.com

Abstract

Embodied Conversational Agents (ECA) and Interactive Virtual Humans (IVH) based on 3D Virtual and Augmented reality technologies can be immensely beneficial for human to human communications and interpersonal skills training for a range of skills including sales pitching, negotiation, leadership, interviewing, and communicating with empathy and cultural sensitivities. They provide an opportunity to practice the skills needed for communicating with humans in difficult environments, with virtual simulated scenarios. And these complex environments can include health care contexts, sales, retail and customer service contexts, or communication with people with different cultural and ethnic backgrounds. At HCT research Centre at University of Canberra, we have been investigating several new technology frameworks and algorithmic techniques for building ECAs and IVHs based on integrating Artificial Intelligence (AI), deep machine learning, cloud computing, 3D virtual and augmented reality technologies, as training simulators for different application contexts. In this paper, we present the details in terms of the research, technology and implementation for modelling different persona for ECAs and IVHs. A multilevel architecture with adaptable functional modules based on the customized persona for different virtual environments, their implementation using an integrated cloud based software platform and its evaluation is discussed.

Index Terms: multimodal integration, audio-visual, Embodied Conversational Agents, Virtual Humans

1. Introduction

In order to skill the professionals in various application contexts requiring effective interpersonal skills development involving human to human communication, there is a need to provide simulated scenarios for practice, so subtle nuances and complexities involved in communicating the message effectively is delivered efficiently. Some of these situations and contexts could involve health care contexts such as in chronic disease management, where patient needs constant support and motivation from a large group of care team to be able to monitor and track the disease progression, and help reverse the disease, or a totally different sales pitch and customer service context, where there is a need to deliver the message in a short span, capture the customer's interest and needs, and close the deal. Though these are completely different application scenarios, the common needs in these application contexts is complexities associated with these environments, in ability of humans to perform effectively and efficiently, and role of human centered information and communication technologies in these complex spaces.

In order to develop better interpersonal communication skills, futuristic training systems of the future will need to leverage multiple complementary technologies, including Artificial Intelligence (AI), deep machine learning, cloud computing, 3D virtual and augmented reality technologies, to simulate all aspects of a virtual world, from the physics of EVAs and IVHs to realistic human behavior. For the Multimodal systems project at the HCT research Centre in University of Canberra, one particular focus is on building high fidelity ECAs and IVHs, that are customized to different application environments. These agents would provide a social and human focus to training and simulate different persona, including guides, mentors, competitors, teammates, patients, and customers or other roles that require interactive face-to-face interaction and provide a powerful mechanism for training interpersonal skills and experiential learning. Existing virtual world environments, mostly based on military simulations and computer games, mostly focus on environments, and attempt to enhance the look and feel and photorealism of embodied environments, but not much on the human aspects, particularly in emulating the different personas and their behaviors. These personae need to be embedded appropriately in the design of embodied conversational agents and interactive virtual humans so that they can engage in a dialogue and conversation, and convey varying aspects of intelligence, and provide challenging simulation scenarios for skills training, team training or strategy and tactical training. There has been a growing need in recent years to train effective interpersonal, communication, leadership, negotiation, cultural awareness and interviewing skills for professionals working in different application contexts. The goal of the Multimodal Systems project is to fill this gap in these complex training environments, as the acquisition of domain specific interpersonal skills requires experiential learning with a vast knowledge of the various aspects of human behavior, and it is hard to acquire in new, inexperienced or inter-disciplinary recruits. Availability of novel, cutting edge human centered information and communication technologies, based on 3D virtual and augmented reality technologies for building embodied conversation agents and interactive virtual humans, can be promising in terms of their capability to interact with trainees at different interpersonal level with different personae, and emulate complex human behavior and challenging domain and application-specific situations. By incorporating this type of intelligent human behavior in the virtual characters and virtual simulated environments, it is possible to provide

experiential learning opportunities to new recruits, and cater to a wide range of training contexts, that currently require labor-intensive live exercises, role playing, or are taught in non-experiential environments, such as classroom contexts. However, this requires use of novel approaches and techniques in creating realistic and engaging characters that convey three main characteristics:

- **Believable ECAs/IVHs -** that they display enough realism and intelligence, to create an illusion of human-like behavior so that the humans will be drawn into the scenario and have transformative experiences.
- **Responsive ECAs/IVHs** – that they respond to the human users and to the associated events around them, which can essentially influence the user's actions, and incorporate rich communication dynamics, explaining the responses to the scenario.
- **Interpretable ECAs/IVHs** – that the users can interpret their responses to situations, mainly the dynamic cognitive and emotional states, using the explicit verbal and subtle nonverbal cues and gestures, that people use to communicate and understand each other.

Thus, the design of ECAs and IVHs need to address several complex requirements, with an aim to not just simply create an illusion of their inner behavior, but must respond to a dynamically unfolding scenario appropriately, and convey their inner and outward behavior accurately and clearly. The recent advances in Artificial Intelligence (AI), deep machine learning, cloud computing, data science, virtual and augmented realities, and 3D computer graphics and animation is promising in achieving the goals of building better ECAs and IVHs – that are more intelligent and can perceive and respond to events in the virtual world much better. The availability of novel algorithmic approaches and computational frameworks provide better opportunities for construction and plan coordination between humans and virtual agents, along with expression of more realistic gestures and emotions, for a meaningful spoken dialogue and conversation between humans and artificial agents, including the subtle nonverbal communication, such as eye contact, gaze aversion and facial expressions and gestures that accompany human speech. While there has been significant research literature on each of these individual components, there has been on no previous work on integrating all these capabilities into emulating different personae for ECAs and IVHs and the complex interplay between humans and these personae for different real-world application contexts.

Rest of this paper is organized as follows. Next section describes the background and related work, and details of proposed technology framework and functional architecture is presented in Section 3. The details of implementation and evaluation is presented in Section 4, and the paper concludes with summary of outcomes achieved and plans for further research in Section 5.

2. Background and Related Work

The integration of AI, machine learning and computer vision, speech and graphics for ECA and IVH is currently considered as the advanced research project of its kind in the world, and requires truly multidisciplinary efforts and collaborations, between disciplines of traditional artificial intelligence (Anderson and Lebiere, 1998; Laird, 2001), computer graphics (Lee and Waters, 1995; Perlin, 1995; Becheiraz and Thalmann, 1996; Rousseau and Hayes-Roth, 1996; Kalra and Magnenat-Thalmann, 1998; Brand, 1999), and behavior of social and behavioural sciences (Frijda, 1987; Wiggins, 1996). There are also some standalone efforts in computer science and information technology research communities to advance technological frameworks for building ECAs and IVHs, focused on different application contexts, including including health care, training, tutoring, retail, sales, marketing, and entertainment. Some of these works include research on embodied conversational agents (Gratch, 2002, Cassell, Bickmore et al., 2000). Cassell and her colleagues have built systems supporting face-to-face conversations between virtual humans (Cassell, Pelachaud et al., 1994) and between human and a virtual human. Some of application specific ECA's include their most recent agent, Rea (Cassell, Bickmore et al., 2000), simulating a real estate agent, and conversing with human users about available apartments and homes. Some other recent systems for team training purposes have also applied artificial intelligence (Bindiganavale, Schuler et al., 2000), including the PuppetMaster [Marsella et al, 1998], an automated assistant to a human instructor for large-scale simulation-based training, and the AETS [Zachary et al, 1998] which monitors a team of human students as they run through a mission simulation using the actual tactical workstations aboard a ship, with detailed cognitive models of each team member to track and remediate their performance. There are also been several research efforts on creating animated pedagogical agents (Lester, Stone et al., 1999), requiring the user to communicate with the agents through menus. The system developed by Rea (Cassell, Bickmore et al., 2000) supports spoken dialogue, but does not have sophisticated natural language understanding capabilities, and hence it is limited to understanding a small set of utterances. The system developed by (Bindiganavale, Schuler et al., 2000) include sophisticated natural language understanding, but they have no capabilities for dialogue with users; they only accept instructions. In addition, there were some efforts in including emotional state in the agents. (Berkowtiz, 2000) developed an approach to include the person's emotions in the characters, modelling the influence of emotional state in their decision making, and reflecting in terms of actions, memory, attention, and body language of characters. (Ball and Breese, 2000; Poggi and Pelachaud, 2000) experimented with emotions in animated agents, but these models of emotion do not quite fully use the advances in AI, machine learning and graphics. The use of models of emotion in ECAs were also attempted, such as those reported by Gratch and Marsella's EMotion and Adaptation (EMA) model (Marsella and Gratch, 2003; Gratch and Marsella, 2004).

But none of these systems use an integrated approach involving integration of AI, machine learning, 3D virtual and augmented reality technologies, addressing embodiment of different personas in their characters. Further, they fall short of domain specific persona adaptation in terms of real spoken dialogue between virtual humans and actual human, gestures and emotions associated with each persona and context. In this work, we aim to address this gap, and propose a new approach to address some of these shortcomings, by using cutting edge technology framework and architecture, described next.

3. Multilevel ECA/IVH Architecture

Imagine a simulated sales training exercise where the characters you interact display almost human persona– they could be fashion conscious, or price aware, or more towards a professional or casual look and feel in shopping for a retail item. In this scenario, the avatars or persona should depict the actual human like behavior, including their speech, visual gestures, body language, and other subtle non-verbal signals. Moreover, based on their persona, they need to engage in a dialogue with human and converse in languages and dialect, with each avatar or agent responding differently in the dialogue session with the human. They must understand their location in the world they are in, and can reason about what to do, and they need to exhibit emotions based on the gender, age and preferences. This is easier said than done, and very complex. However, such a simulator could open whole new horizons for training and simulation. By integrating a wide set of technology frameworks including Artificial Intelligence (AI), deep machine learning, cloud computing, 3D virtual and augmented reality technologies, to build ECAs, IVHs and environment around them, the training simulators can mimic a broad range of human behaviors and conversational characters for different application scenarios. To the best of our knowledge none of the prior works have addressed this research question in an integrated cohesive manner. The goal of the proposed work being pursued in our lab is two-fold, to perform advanced technology research in areas that lead to a fully realistic embodied conversational agents, interactive virtual humans and artificial synthetic worlds and environments, and to adapt them for implementation for different personae for different application scenarios including educating and training environments, retail, marketing and customer service environments, as well as health, disease monitoring, and well-being environments. These two goals complement each other; as the advances in technological frameworks and tools gets mature, they can drive innovation in applications development.

This iterative research and development aim to address some of the important questions in this area, including:

- How realistic do the ECA/IVH and their persona need to be, to be more effective,
- How believable do they need to be?
- How many layers of persona specific behaviors can be modelled and displayed in them?
- How can they be adapted for different application environments, and which capability in their persona need to be dominating in each application context?

Without an integrated technology architecture, some of these conflicting requirements are difficult questions to answer, but with a minimal functional design and an evolving multilevel architecture we propose here (Figure 1), it is possible to address each question step by step, and develop an effective design, development, integration, testing and deployment workflow, and then progressively add the layers or levels of functionality to address the other complex questions.

The minimalist design for such a multilevel architecture follows the same paradigm as Belief-Desire-Intention (BDI) style agents with a sense-think-act cycle involving three levels is as shown in Figure 1, and outlined below:

- **Cognitive Thinking Level**: This the innermost level, containing the cognitive component. For each ECA/IVH, there is one unique cognitive level based on the persona they need to emulate. This is the brain or mind of agent, which does the decision making based on the inputs, goals and desired behavior. For each persona the combination of different components that make up this unique persona, including cognitive aspects, dialogue and question-answer responses, finite state transition machines, the verbal audio-visual scripts, and the synchronization with gestural and non-verbal cues differ.

- **Sensor Fusion Level**: This is the second level where different types of sensors try to integrate with different components, with the cognitive level components on one end, and higher-level components to make the ECA/IVH to interact effectively. Based on the embodied persona, this could involve a combination of different input and output processing components. Input could include vision, speech, and different environmental IoT (Internet of Things) sensors, and output could include verbal speech, facial emotions, facial and, body gestures, and actions the character would performs, for example the style of speech, accent, facial and body gestures, dressing and walking styles. This is the most important layer as the ECA/IVH persona is perceived by external world with an efficient design and implementation of this layer.

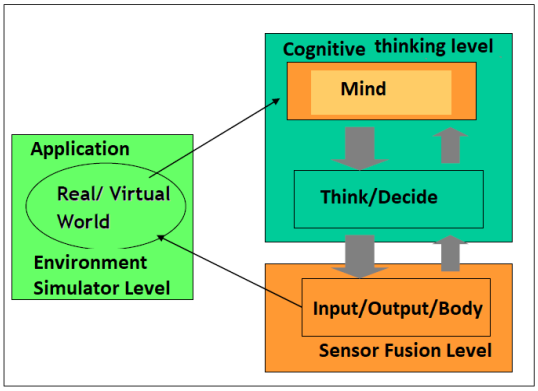


Figure 1: *Multilevel ECA/IVH Architecture*

- **Environment Simulator Level:** This level is external environment level where ECAs/IVHs would be interacting with different simulated application contexts with their different personas. This includes the 3D Virtual and Augmented reality environments that create the artificial synthetic worlds corresponding to each application context, and the characters or ECA/IVHs interact and communicate in this world. This level also includes the application domain specific background characters, any scenario management interfaces, and any form of after-action review. Any input from the real world, including locations, positioning and co-ordinates of other agents, and devices such as cameras, microphones and navigational aids would be part of this environment simulator level.

Each of the three levels in multilevel architecture consists of different functional modules, based on application context and the virtual/augmented reality environment associated with application context, the input and output sensors and actuators (both virtual and actual/IoT sensor and actuators), and the cloud based AI/deep learning and decision making cognitive model. For example, a virtual training simulator architecture would consist of functional modules shown in Figure 2 below.

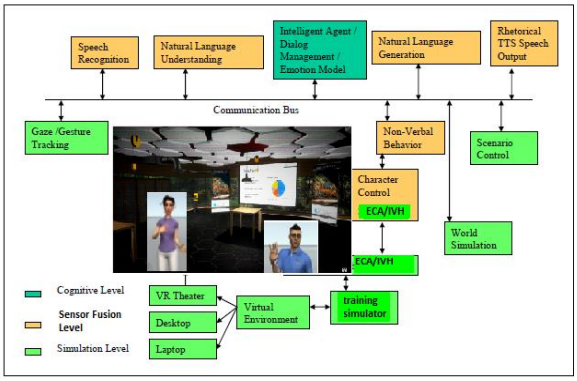
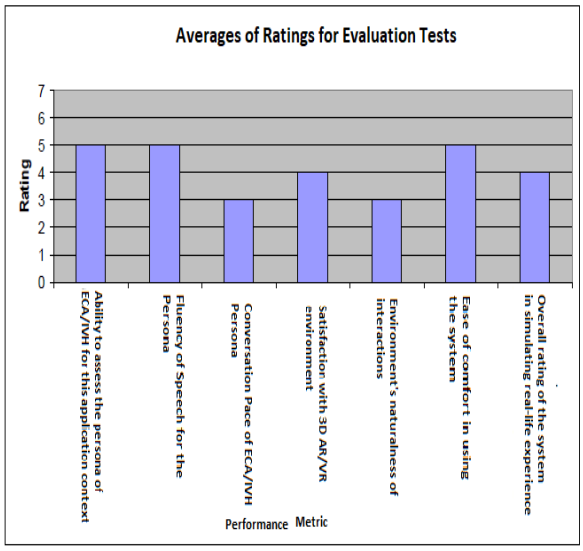


Figure 2: *Different Function Modules for a training simulator use case for multilevel architecture*

Table 1: *Performance Evaluation of ECA/IVH Architecture*



4. Evaluation and Performance

To assess the value of the proposed multilevel architecture for ECAs/IVHs based on AI, deep machine learning, augmented and virtual reality tools, an extensive set of simulator environments were created similar to the one shown in Figure 2, for different personas and use cases corresponding to training simulator context for health (chronic disease and mental health support), and retail sales training (customer persona simulator). The simulated environments and ECA/IVH were created with AWS cloud computing tool kits, including AWS gamelift, lumberyard, Sumerian hosts, Sagemaker, Lambda, DynamoDB, Poly speech, lex dialog bots and Cloud formation tools. The qualitative and quantitative evaluation of different functional modules for different use cases were done with feedback from customers or trainees. The feedback was collected both in terms of ratings on a 5-point Likert scale for different aspects of persona including speech, visual, gesture and dialogue components, as well open text comments. Table 1 gives a summary of average rating per metric.

5. Conclusions and Further Work

In this paper, we describe a novel technology framework for ECAs/IVHs based on integration of cloud-based AI, deep machine learning and 3D virtual and augmented reality environments. As presented in this paper, creation of personas for different use cases and application contexts involving embodied conversational agents, and interactive virtual humans is a complex undertaking, and involves multidisciplinary skills and collaboration. Designing a multi-level architecture and its implementation in terms of different functional modules that can be adapted for different use cases can lead to better user experience with simulated personas, based on performance evaluation done for some example use cases. Further work will involve enhancing the framework for creating ECAs/IVHs for more complex application environments, and enhancing the visualization and intelligence of the personas, so that they are believable in their appearance, language and behavior, are more responsive and interpretable to the users and simulator environments, ultimately leading to creation of compelling virtual environments.

6. References

[1] Anderson, J. R. and C. Lebiere (1998). The Atomic Components of Thought. Mahwah, NJ, Lawrence Erlbaum Associates. Ball, G. and J. Breese (2000). Emotion and Personality in a Conversational Character. Embodied Conversational Agents.

[2] J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Cambridge, MA, MIT Press: 189-219.

[3] Becheiraz, P. and D. Thalmann (1996). A Model of Nonverbal Communication and Interpersonal Relationships Between Virtual Actors. Computer Animation, IEEE Press.

[4] Berkowitz, L. (2000). Causes and Consequences of Feelings, Cambridge University Press.

[5] Bindiganavale, R., W. Schuler, et al. (2000). Dynamically Altering Agent Behaviors Using Natural Language Instructions. Fourth International Conference on Autonomous Agents, Barcelona, Spain, ACM Press.

[6] Brand, M. (1999). Voice puppetry. ACM SIGGRAPH, ACM Press/Addison-Wesley Publishing Co.

[7] Cassell, J., T. Bickmore, et al. (2000). Human conversation as a system framework: Designing embodied conversational agents. Embodied Conversational Agents.

[8] J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Boston, MIT Press: 29-63.

[9] Cassell, J., C. Pelachaud, et al. (1994). Animated Conversation: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. ACM SIGGRAPH,

[10] Reading, MA, Addison-Wesley. Cassell, J., Vilhj_almsson, H.H., Bickmore, T.W.: Beat: the behavior expression animation toolkit. In: Proceedings of SIGGRAPH. (2001) 477-486

[11] Frijda, N. (1987). "Emotion, cognitive structure, and action tendency." Cognition and Emotion 1: 115- 143.

[12] Gratch, J., Rickel, J., André, E., Badler, N., Cassell, J., Petajan, E.: Creating Interactive Virtual Humans: Some Assembly Required, IEEE Intelligent Systems, July/August, 54-63, (2002)

[13] Gratch, J. and S. Marsella (2004). "A domain independent framework for modeling emotion." Journal of Cognitive Systems Research 5(4): 269- 306.

[14] Gratch, J. and S. Marsella (2004). Evaluating the modeling and use of emotion in virtual humans. 3rd International Joint Conference on Autonomous Agents and Multiagent Systems, New York.

[15] Kalra, P. and N. Magnenat-Thalmann (1998). "Realtime Animation of Realistic Virtual Humans." IEEE Computer Graphics and Applications 18(5): 42-55.

[16] Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2007

[17] Kendon, A. Some functions of gaze direction in social interaction. Acta Psychologica, 26: 22-63, 1967.

[18] Kenny, P., Parsons, T., Gratch J., Leuski, A., Rizzo A.: Virtual Patients for Clinical Therapist Skills Training. 7th International Conference on Intelligent Virtual Agents, pp 197-210, Paris France. (2007).

[19] Knublauch, H., Fergerson, R. W., Noy, N. F., Musen, M. A. The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. Third International Semantic Web Conference, Hiroshima, Japan, 2004.

[20] Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. Computer Animation and Virtual Worlds 15(1) (2004) 39-52.

[21] Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thorisson, K., Vilhjalmsson, H: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language,. 6th International Conference on Intelligent Virtual Agents (Marina del Rey, CA, August 21-23 2006).

[22] Laird, J. E. (2001). It Knows What You're Going To Do: Adding Anticipation to a Quakebot. Proceedings of the Fifth International Conference on Autonomous Agents, Montreal, Canada, ACM Press.

[23] Lee, Y. and K. Waters (1995). "Realistic Modeling for Facial Animation." SIGGRAPH. Lee, J., Marsella, S: Nonverbal Behavior Generator for Embodied Conversational Agents. 6th International Conference on Intelligent Virtual Agents, pp 243-255, Marina del Rey, CA. (2006).

[24] Lester, J. C., B. A. Stone, et al. (1999). "Lifelike Pedagogical Agents for Mixed-Initiative Problem Solving in Constructivist Learning Environments." User Modeling and User-Adapted Instruction 9(1-2): 1-44.

[25] Leuski, A., Patel, R., Traum, D., Kennedy B.: Building effective question answering characters. (2006) In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, Sydney, Australia.

[26] Marsella, S. and J. Gratch (2003). Modeling coping behaviors in virtual humans: Don't worry, be happy. Second International Joint Conference on Autonomous Agents and Multi-agent Systems, Melbourne, Australia.

[27] Marsella, S. & Johnson, W.L. (1998) Intelligent Tutoring Systems, Springer-Verlag. Lecture Notes in Computer Science

[28] Pellom, B.: Sonic: The University of Colorado continuous speech recognizer. Technical Report TRCSLR- 2001-01, University of Colorado, Boulder, CO (2001)

[29] Perlin, K. (1995). "Real Time Responsive Animation with Personality." IEEE Trans. on Visualization and Computer Graphics 1(1): 5-15.

[30] Poggi, I. and C. Pelachaud (2000). Emotional Meaning and Expression in Performative Faces. Affective Interactions: Towards a New Generation of Computer Interfaces. A. Paiva. Berlin, Springer-Verlag: 182-195.

[31] Rickel, J., Gratch, J., Hill, R., Marsella, S., Swartout, W.: Steve Goes to Bosnia: Towards a New Generation of Virtual Humans for Interactive Experiences. In AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment, Stanford University, CA, March (2001)

[32] Rousseau, D. and B. Hayes-Roth (1996). Personality in Synthetic Agents. Stanford, CA, Knowledge Systems Laboratory, Stanford University.

[33] Sethy, A., Georgiou, P., Narayanan, S.: Building topic specific language models from webdata using competitive models. In: Proceedings of EUROSPEECH, Lisbon, Portugal (2005)

[34] Swartout, W., Gratch, J., Hill, R., Hovy, E., Marsella, S., Rickel J., Traum, D.:Toward Virtual Humans, AI Magazine, v.27(1). (2006)

[35] Thiebaux, M., Marshall, A., Marsella, S., Fast, E., Hill, A., Kallmann, M., Kenny, P., Lee, J., SmartBody Behavior Realization for Embodied Conversational Agents, Submitted to IVA07, (2007), Paris, France.

[36] Wiggins, J. S. (1996). The Five-Factor Model of Personality: Theoretical Perspectives. New York, The Guilford Press.

[37] Zachary, W., Cannon-Bowers, J., Bilazarian, P., Krecker, D.,Lardieri, P., & Burns, J. (1999). The Advanced Embedded Training System (AETS): An intelligent embedded tutorin system for tactical team training. International Journal of Artificial Intelligence in Education, 10, 257-277.