Proceedings of the Workshop on Architectures and Standards for Intelligent Virtual Agents at IVA 2014

Boston, August 26, 2014

Angelo Cafaro, Arno Hartholt, and Herwin van Welbergen (eds.)

Proceedings of the Workshop on Architectures and Standards for Intelligent Virtual Agents at IVA 2014

Foreword

The scope of building a complete intelligent virtual agent (IVA) is too vast for a single research group. It requires interdisciplinary collaborations between research groups and reuse of existing components. Based off the SAIBA framework for multimodal behavior generation, an important current research direction for the IVA-community deals with facilitating the collaboration between groups and reuse of each other's work by using modular architectures and interface standards. In this workshop, in light of emergent technologies and the variety of IVA applications, we want to discuss the standardization level provided by SAIBA and understand whether it is still capable of supporting the next generation of IVAs. We aim at improving current points of standardization and identifying new architectural elements and functionalities that require standardization.

These proceedings contain position papers of six participating research groups who have each presented an overview of their state-of-the-art in architectures and standards for IVAs and their visions for future IVA architectures, standardization, and collaborations. We hope for a fruitful discussion on these topics during the workshop.

Angelo Cafaro, Arno Hartholt and Herwin van Welbergen

Organizing Committees

Workshop Co-Chair

Angelo Cafaro, (CNRS-LTCI, TELECOM ParisTech)
Arno Hartholt (Institute for Creative Technologies, University of Southern California)
Herwin van Welbergen (Social Cognitive Systems Group, CITEC, Bielefeld University)

Program Committee

Catherine Pelachaud, CNRS-LTCI, TELECOM ParisTech
David Schlangen, Bielefeld University
Dirk Heylen, University of Twente
Hannes Högni Vilhjálmsson, Reykjavík University
Justine Cassell, Carnegie Mellon University
Radosław Niewiadomski, University of Genoa
Stefan Kopp, Bielefeld University
Stefan Scherer, University of Southern California
Timothy Bickmore, Northeastern University

External Reviewers

Yoichi Matsuyama, Carnegie Mellon University
Alex Papangelis, Carnegie Mellon University

Acknowledgements

Herwin van Welbergen is funded by the German Federal Ministry of Education and Research (BMBF) within the Leading-Edge Cluster Competition, managed by the Project Management Agency Karlsruhe (PTKA).

Table of Contents

A Shared, Modular Architecture for Developing Virtual Humans

Arno Hartholt [1], David Traum[1], Stacy Marsella[2], Louis-Philippe Morency[1], Ari Shapiro[1], and Jonathan Gratch[1]

[1]USC Institute for Creative Technologies, Los Angeles, USA
[2]Northeastern University, Boston, USA

1 Main Research Themes

Realizing the full potential of intelligent virtual agents requires compelling characters that can engage users in meaningful and realistic social interactions, and an ability to develop these characters effectively and efficiently. Advances are needed in individual capabilities, but perhaps more importantly, fundamental questions remain as to how best to integrate these capabilities into a single framework that allows us to efficiently create characters that can engage users in meaningful and realistic social interactions. This integration requires in-depth, inter-disciplinary understanding few individuals, or even teams of individuals, possess.

Our research is focused on understanding the relationship between individual capabilities, how they strengthen each other within larger systems and which sets of minimum and desired permutations can be defined for different types of systems and domains. This research is often conducted within the context of a common, modular framework that contains a mix of research and commercial technologies to offer full coverage of subareas including speech recognition, audio-visual sensing, natural language processing, dialogue management, nonverbal behavior generation & realization, text-to-speech and rendering.

Many of our characters are so-called *question-answering agents*. They offer a user-driven, interview-type style of conversation where a user question is answered with a character answer. These systems explore how to best provide particular information within a given domain, often through a set of pre-defined responses from one or more agents. Examples are SGT Star [1], Boston Museum of Science Guides [20], Virtual Patients [10] and Gunslinger [7], covering military, education, medical and entertainment domains.

Another focus has been on modeling verbal and nonverbal back-channeling behavior through *virtual listeners*, in order to establish rapport and increase speaker fluency and engagement. The Rapport project is an example of this [5].

Virtual interviewers put the focus on gathering information from or assessing the user. Shifting some of the conversational burden to the agent requires increased dialogue management and natural language understanding capabilities. The SimCoach system [17], for example, is a web-based guide that helps navigate healthcare related resources, which uses a forward looking, reward seeking dialogue manager [15]. Combining advanced dialogue management with audio-visual sensing results in SimSensei, an agent who aids in recognition of psychological distress [4]. SimSensei enables an engaging face-to-face interaction where the character reacts to the perceived user state and intent through its own speech and gestures.

Many of these capabilities are available for the research community through the ICT Virtual Human Toolkit[1]. It provides a solid basis for the rapid development of new virtual humans, but also serves as an integrated research platform to enable context-sensitive research in any of the Virtual Human subfields, taking advantage of and examining the impact on other system modules.

2 Current Architectures and Standards

Many of the virtual humans developed at the University of Southern California Institute for Creative Technologies are instantiation of a more general Virtual Human Architecture, see Figure 1. It defines at an abstract level the capabilities of a virtual human and how these interact. Not every system will include all capabilities and some will implement them to a greater or lesser extent. The architecture allows for multiple implementations of a certain capability and simple substitution of one implementation for another during run-time, facilitating the exploration of alternative models for realizing individual capabilities. Thus the general architecture can be specialized in many different ways.

Capabilities are realized through specific modules, most of which communicate with each other through a custom messaging system called VHMsg, build on top

[1] https://vhtoolkit.ict.usc.edu

of ActiveMQ[2]. Messages are typically broadcasted, although there are intended consumers. Libraries have been built for several languages, including Java, C++, C#, Lisp and TCL, so that developers have a wide latitude in developing new modules that can communicate with the rest of the system. There is a standard set of message types used by existing modules [6], and it is very easy to create new message types. The basis for this architecture is SAIBA[3] which was extended to cover additional areas. It uses the Functional Markup Language (FML) [9], the Behavior Markup Language (BML) [11], and the Perception Markup Language (PML) [21].

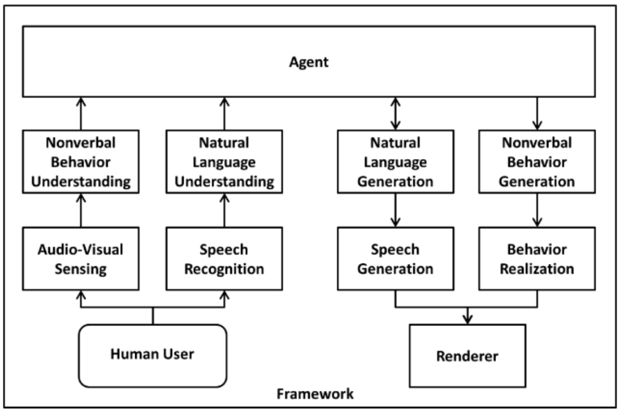

Agent
Nonverbal Behavior Understanding
Natural Language Understanding
Natural Language Generation
Nonverbal Behavior Generation
Audio-Visual Sensing
Speech Recognition
Speech Generation
Behavior Realization
Human User
Renderer
Framework

Fig. 1. The Virtual Human Architecture.

The relationship between capabilities and modules is not necessarily one to one. The Agent, for example, can range from rule-based systems [5], to a statistical text-classifier (the NPCEditor [13] which also serves as NLU and NLG), to cognitive architectures [23, 19], which can include sub-modules for task modeling, emotion modeling, and dialogue management. Other often used modules are MultiSense (audio-visual sensing) [21], FLoReS (dialogue manager) [15], Cerebella (nonverbal behavior generation) [14], and SmartBody (character animation) [22].

The strengths of the architecture lie in its modularity, extendibility and the wide range of integrated capabilities, offering a powerful basis for the rapid research and development of new implementations of both modules and systems. It has been validated by over two dozen systems, ranging from research prototypes to deployed applications. More information about the architecture, API, individual modules and capabilities, and created systems can be found in [6].

3 Future of Architectures and Standards for IVAs

We suggest that future efforts focus on achieving broader interoperability between systems created by separate research groups. For instance, while several BML realizers exist [3, 8, 16, 22, 24], they are not directly interchangeable due to an often less than strict implementation of the standard as well as many custom extensions. In addition, the transport layer is undefined, preventing messages between in-house and external modules to be sent and received easily.

Related, more thought should be given to sharing capabilities, data and assets throughout the community, to more easily leverage each other's work. This requires that standards and procedures move beyond merely an architecture and interface, and include explicit notions of the community, their systems and data, and associated methodologies. A layered approach of Community, Architecture & Interface and Implementation levels may aid in this goal.

Reference architectures like SAIBA should include a broader range of capabilities, including audio-visual sensing and natural language processing. They should also address challenges like continuous communication and processing between components rather than the current often sequential ones.

Finally, relationships between capabilities should be made more explicit and should be extendable and scalable. At run-time, the system and its components should be able to detect available capabilities as well as their level of sophistication and adapt accordingly, analogous to discoverable web services. For instance, a speech recognition capability could offer either discrete or continuous recognition, with optional prosody analysis; the remainder of the system should be able to work with the minimum provided capability as well as take advantage of richer input when available.

4 Suggestions for discussion

As mentioned in section three, our interest is in creating an environment in which collaboration can happen more effectively and efficiently, both on a technical and organizational level. Relevant topics:

- Integrating architectures and standards from related fields (e.g. ROS, OpenInterface, OpenCog, EmotionML, the Incremental Unit-architecture,);
- Missing standards (e.g. Context Markup Language (CML), messaging, etc.);
- Data, assets, knowledge & technology sharing;
- Design and development best practices.

[2] http://activemq.apache.org/

[3] http://www.mindmakers.org/projects/saiba/wiki

References

1. Artstein R, Gandhe S, Leuski A, Traum DR. Field Testing of an interactive question-answering character. ELRA, LREC (2008)
2. Bickmore, T.W., Schulman D., Shaw, G., DTask & LiteBody: Open Source, Standards-based Tools for Building Web-deployed Embodied Conversational Agents, Intelligent Virtual Agents, PP. 425-431 (2009)
3. Julia Campbell, Matthew Hays, Mark Core, Mike Birch, Matthew Bosack, Richard E. Clark, Using Virtual Humans to Teach New Officers In Interservice/Industry Training, Simulation and Education Conference (I/ITSEC) 2011
4. David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Alesia Egan, Ed Fast, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Skip Rizzo, and Louis-Philippe Morency. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Paris, France, May 5-9, 2014.
5. Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, R. J. van der Werf, Louis-Philippe Morency, Virtual Rapport, IVA (2006)
6. Arno Hartholt, David Traum, Stacy C. Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, Jonathan Gratch, All Together Now: Introducing the Virtual Human Toolkit, at Intelligent Virtual Agents (IVA), 2013
7. Hartholt A., Gratch J., Weiss L., Leuski A., Morency L.P., Marsella S, At the Virtual Frontier: Introducing Gunslinger, a Multi-Character, Mixed-Reality, Story-Driven Experience IVA pp. 500-501 (2009)
8. Heloir, A. and Kipp, M.: A Realtime Engine for Interactive Embodied Agents, Intelligent Virtual Agents, pp. 393-404, 2009
9. Heylen, D.K.J., et al., The Next Step towards a Function Markup Language, at Intelligent Virtual Agents (IVA), 2008
10. Kenny, P.G., Parsons, T.D., Gratch, J. Rizzo, A., Evaluation of Justina: A Virtual Patient with PTSD, Lecture Notes in Computer Science, pp. 394-408 (2008)
11. Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R. Thórisson, Hannes Vilhjálmsson, Towards a Common Framework for Multimodal Generation: The Behavior Markup Language (LNAI), vol. 4133, pp. 205–217. (2006)
12. H. Chad Lane, Clara Cahill, Susan Foutz, Daniel Auerbach, Dan Noren, Catherine Lussenhop, William Swartout, The Effects of a Pedagogical Agent for Informal Science Education on Learner Behaviors and Self-efficacy, In Artificial Intelligence in Education, volume 7926, 2013
13. Leuski, A and Traum, D. NPCEditor: Creating virtual human dialogue using information retrieval techniques. AI Magazine, 32(2):42–56, (2011).
14. Stacy C. Marsella, Ari Shapiro, Andrew W. Feng, Yuyu Xu, Margaux Lhommet, Stefan Scherer, Towards Higher Quality Character Performance in Previz, Digital Production Symposium, Anaheim, CA, July, 2013
15. Morbini F., DeVault D., Sagae K., Gerten J., Nazarian A., Traum D., FLoReS: A Forward Looking, Reward Seeking, Dialogue Manager, Spoken Dialog Systems (2012)
16. I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio, B. De Carolis, GRETA. A Believable Embodied Conversational Agent, Multimodal Intelligent Information Presentation, (2005)
17. Albert Rizzo, Eric Forbell, Belinda Lange, John Galen Buckwalter, Josh Williams, Kenji Sagae, David Traum, SimCoach: An Online Intelligent Virtual Agent System for Breaking Down Barriers to Care for Service Members and Veterans, Chapter in Healing War Trauma (2012)
18. Albert Rizzo, Bruce Sheffield John, Brad Newman, Josh Williams, Arno Hartholt, Clarke Lethin, John Galen Buckwalter, Virtual Reality as a Tool for Delivering PTSD Exposure Therapy and Stress Resilience Training, In Military Behavioral Health, volume 1, 2012
19. Paul S. Rosenbloom, The Sigma Cognitive Architecture and System, Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB), 2013
20. Swartout W, Traum DR, Artstein R, Noren D, Debevec P, Bronnenkant K, et al. Ada and Grace: Toward Realistic and Engaging Virtual Museum Guides. IVA pp. 286–300 (2010)

21. Stefan Scherer, Stacy C. Marsella, Giota Stratou, Yuyu Xu, Fabrizio Morbini, Alesia Egan, Albert Rizzo, Louis-Philippe Morency, Perception Markup Language: Towards a Standardized Representation of Perceived Nonverbal Behaviors, IVA pp. 455–463 (2012)
22. Shapiro A., Building a Character Animation System, 4th Annual Conference on Motion in Games 2011, Edinburgh, UK, November 2011
23. David Traum, William Swartout, Jonathan Gratch and Stacy Marsella, Virtual Humans for non-team interaction training, In International Conference on Autonomous Agents and Multiagent Systems (AAMAS) Workshop on Creating Bonds with Humanoids, 2005.
24. Van Welbergen, H., Reidsma, D., Ruttkay, Z.M., Zwiers, J.: Elckerlyc: A BML realizer for continuous, multimodal interaction with a virtual human. Multimodal UI (2010)

Biographical Sketch

Arno Hartholt is a Computer Scientist at the University of Southern California (USC) Institute for Creative Technologies (ICT) where he leads the virtual human technology integration group as well as the Institute's central asset production & pipeline group. As such, he is responsible for much of the technology, art, processes and procedures related to virtual humans and associated systems. He has a leading role on a wide variety of research prototypes and applications in areas ranging from medical education and military training to intelligent tutoring and serious games. Hartholt studied computer science at the University of Twente in the Netherlands where he got his master's degree. He worked at several IT companies, from large multi-nationals to early start-ups, before accepting a position at USC ICT in 2005. As one of the main integration software engineers within the virtual humans project, Hartholt has developed a variety of technologies, with a focus on task modeling, natural language processing and knowledge representation. He can be reached at hartholt@ict.usc.edu.

David Traum is a principal scientist at ICT and a research faculty member at the Department of Computer Science at USC. At ICT, Traum leads the Natural Language Dialogue Group, which consists of seven Ph.D.s, four students, and four other researchers. The group engages in research in all aspects of natural language dialogue, including dialogue management, spoken and natural language understanding and generation and dialogue evaluation. The group collaborates with others at ICT and elsewhere on integrated virtual humans, and transitioning natural language dialogue capability for use in training and other interactive applications. Traum's research focuses on dialogue communication between human and artificial agents. He has engaged in theoretical, implementational and empirical approaches to the problem, studying human-human natural language and multi-modal dialogue, as well as building a number of dialogue systems to communicate with human users. He has pioneered several research thrusts in computational dialogue modeling, including computational models of grounding (how common ground is established through conversation), the information state approach to dialogue, multiparty dialogue, and non-cooperative dialogue. Traum is author of over 200 technical articles, is a founding editor of the Journal Dialogue and Discourse, has chaired and served on many conference program committees, and is currently the president emeritus of SIGDIAL, the international special interest group in discourse and dialogue. He earned his Ph.D. in computer science at University of Rochester in 1994.

Stacy Marsella has a Ph.D. from Rutgers University with a focus on AI and human problem solving. He is well known for his work in computational models of human cognition and emotion. He also has extensive experience in the design and construction of simulations of social interaction for a variety of research, education and analysis applications. This includes his work on virtual humans for immersive training environments such as ICT's MRE and SASO-ST systems and the DARPA-sponsored Tactical Language system. He leads several projects related to virtual humans, including SmartBody, a virtual human animation system, Cerebella, a nonverbal behavior generation system, and PsychSim, a model of social interaction based on theory-of-mind modeling as well as being co-developer of the EMA emotion model with Jon Gratch. He has also worked on psychotherapeutic applications of emotion models, including his work on Carmen's Bright Ideas, a system that teaches coping strategies to parents of cancer patients. Marsella plays a leadership role in organizing conferences on virtual humans, social intelligence and emotion modeling, has over 150 technical articles and is on the editorial boards of the Journal of Experimental And Theoretical Artificial Intelligence, IEEE Transactions on Affective Computing and Journal of Intercultural Communication. He is member of the Association for the Advancement of Artificial Intelligence (AAAI), a fellow in the Society of Experimental Social Psychologists, and a member in the International Society for Research on Emotions.

Louis-Philippe Morency is a research assistant professor in the Department of Computer Science at the University of Southern California (USC) Viterbi School of Engineering and research scientist at the USC Institute for Creative Technologies, where he leads the Multimodal Communication and Machine Learning Laboratory (MultiComp Lab). He received his doctoral and master's degrees from MIT's Computer Science and Artificial Intelligence Laboratory. His research interests are in computational study of nonverbal social communication, a multi-disciplinary research topic that overlays the fields of multimodal interaction, computer vision, machine learning, social psychology and artificial intelligence. Morency was selected in 2008 by IEEE Intelligent Systems as one of the *Ten to Watch* for the future of AI research. He received six best paper awards in multiple ACM- and IEEE-sponsored conferences for his work on context-based gesture recognition, multimodal probabilistic fusion and computational modeling of human communication dynamics. His work has been featured in *The Economist*, *New Scientist* and *Fast Company* magazines.

Ari Shapiro has nearly two decades of professional experience in the computer field as an engineer, consultant, manager and scientist. He currently works as a research scientist at the USC Institute for Creative Technologies, where his focus is on synthesizing realistic animation for virtual characters. At ICT, he heads the team for the SmartBody application, which serves as an animation system for synchronizing speech, facial animation, body motion and gesturing for many of ICT’s real time virtual human systems. For several years, he worked on character animation tools and algorithms in the research and development departments of visual effects and video games companies such as Industrial Light and Magic, LucasArts and Rhythm and Hues Studios. He has worked on many feature-length films, and holds film credits in The Incredible Hulk and Alvin and the Chipmunks 2. In addition, he holds video games credits in the Star Wars: The Force Unleashed series. Shapiro has published many academic articles in the field of computer graphics in animation for virtual characters, and is a five-time SIGGRAPH speaker. He completed his Ph.D. in computer science at UCLA in 2007 in the field of computer graphics with a dissertation on character animation using motion capture, physics and machine learning. He holds an M.S. in computer science from UCLA, and a B.A. in computer science from the University of California, Santa Cruz.

Jonathan Gratch’s research focuses on virtual humans and computational models of emotion. He studies the relationship between cognition and emotion, the cognitive processes underlying emotional responses, and the influence of emotion on decision-making and physical behavior. He is the director for virtual humans research at the USC Institute for Creative Technologies, a research professor of computer science and psychology, and co-director of USC’s Computational Emotion Group. He completed his Ph.D. in computer science at the University of Illinois in Urbana-Champaign in 1995. A recent emphasis of his work is on social emotions, emphasizing the role of contingent nonverbal behavior in the co-construction of emotional trajectories between interaction partners. His research has been supported by the National Science Foundation, DARPA, AFOSR and RDECOM. Along with ICT’s Stacy Marsella, Gratch received the Association for Computing Machinery’s Special Interest Group on Artificial Intelligence (ACM/SIGART) 2010 Autonomous Agents Research Award, an annual award for excellence for researchers influencing the field of autonomous agents. Gratch is the editor-in-chief of the journal IEEE Transactions on Affective Computing, a member of the editorial boards of the journals Emotion Review, and Journal of Autonomous Agents and Multiagent Systems. He is the former president of the HUMAINE Association for Research on Emotions and Human- Machine Interaction (now known as the Association for the Advancement of Affective Computing), and a frequent organizer of conferences and workshops on emotion and virtual humans. He belongs to the American Association for Artificial Intelligence (AAAI) and the International Society for Research on Emotion (ISRE). Gratch is the author of over 250 technical articles.

Acknowledgments

We’d like to acknowledge the members of the USC ICT Virtual Humans group for their dedication and contributions. The effort described here has been sponsored by the U.S. Army. Any opinions, content or information presented does not necessarily reflect the position or the policy of the United States Government; no official endorsement should be inferred.

Architectures and Standards for IVAs at the Social Cognitive Systems Group

Herwin van Welbergen, Kirsten Bergmann, Hendrik Buschmeier, Sebastian Kahl, Iwan de Kok, Amir Sadeghipour, Ramin Yaghoubzadeh, and Stefan Kopp

Social Cognitive Systems Group – CITEC and Faculty of Technology
Bielefeld University, Bielefeld, Germany

1 Main Research Themes

The 'Social Cognitive Systems' group explores how cognitive systems can be intelligent, socially adept interaction partners that allow a fluent and coordinated interaction with humans. To that end we develop methods to model the behavioral, perceptual-motor, and cognitive mechanisms of embodied human-like communication and cooperation. We apply and evaluate them in human–machine interaction scenarios with Intelligent Virtual Agents (IVAs). Scenarios range from embedding IVAs in traditional mouse-keyboard interfaces to virtual coaches to virtual assistants for elderly and cognitively impaired users, to cognitive models for investigating the semantic coordination of speech and gesture production and to computational models of dialog coordination based on linguistic feedback.

2 Current Architectures and Standards

We aim to develop IVAs that can achieve the same high interactivity and real-time responsiveness as their human conversation partners. To this end, we have developed several IVA components that provide incremental and adaptive behavior generation:

- A multimodal memory component realized as a spreading activation model of semantic coordination for speech-gesture production (Bergmann et al., 2013).
- An incrementalized natural language generation system based on the SPUD framework (Buschmeier et al., 2012).
- A behavior planner for iconic gestures (Bergmann and Kopp, 2009).
- A BML 1.0 realizer capable of realizing behavior in an incremental and highly adaptable fashion (AsapRealizer; van Welbergen et al. (2014)).
- An information-state based incremental dialog manager (yet unpublished) capable of handling uncertain input.

The architecture in which we combine these components follows the SAIBA reference architecture. It makes use of BML 1.0 for behavior realization. We have not standardized (via FML) the communication between our Intent Planner and Behavior Planner yet.

We also use several external components in our IVA architectures, both commercial and developed by other research groups. For many of these we have multiple alternatives that offer different trade-offs between recognition/synthesis quality, reactivity, and control. For **automatic speech recognition** we use the SDKs of either Windows Speech Recognition (Microsoft) or Dragon NaturallySpeaking (Nuance), both in their incremental mode. For **speech synthesis** we use CereVoice (CereProc), MaryTTS (Schröder and Trouvain, 2003) or its incremental version Inprotk_iSS (Baumann and Schlangen, 2012). We can **track** the users' eyes and head with faceLAB 5 (SeeingMachines) as well as their face with SHORE (Fraunhofer IIS). **Audio processing** is either done with openSMILE (Eyben et al., 2010) or custom processing pipelines. For **dialog management**, we are also looking into OpenDial (Lison, 2014).

Our components are written in various programming languages, may run on different operating systems and on different computers. Furthermore, they allow the delivery of input processing results or construction and modification of behavior realization plans in an incremental manner. To manage both incrementally and connectivity our middleware framework IPAACA (`http://purl.org/net/ipaaca`) implements the Incremental Unit (IU) architecture (Schlangen and Skantze, 2011) and embeds it in a message oriented middleware (RSB; Wienke and Wrede (2011)).

3 Future Architectures and Standards

3.1 Short Term

The SAIBA architecture has helped us in providing a common terminology for behavior generation for IVAs and specifically in defining a standardized interface for behavior realization. We propose to enhance the standardization of terminology and interfaces provided by SAIBA to provide a full reference architecture for IVAs. To satisfy our requirements on fluent behavior re-

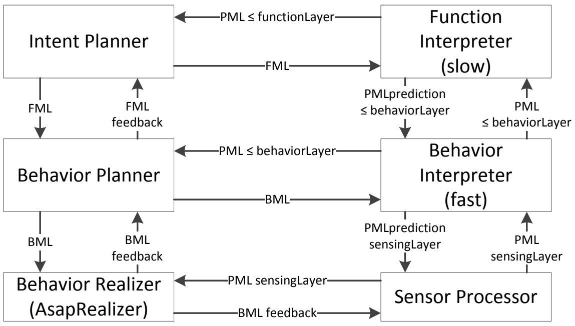


Figure 1: The Asap architecture (Kopp et al., 2014).

alization such an architecture should encompass at least behavior generation, input processing, a bi-directional coordination between input processing and output generation on multiple levels and provide support for incremental processing on all levels.

Our Articulated Social Agents Platform (Asap) satisfies these requirements. It embeds the SAIBA architecture (left side of Fig. 1) and enhances it with a matching sub-architecture for input processing and a close bi-directional coordination between input processing and output generation. Asap's input processing is inspired by the Perception Markup Language (PML) proposal (Scherer et al., 2012). We define explicit interpreters for each PML layer and use PML messages to communicate between the layers. Additionally, Asap enables a top-down information flow in input processing. For example, the Function Interpreter may communicate to the Behavior Interpreter that listening behaviors (e.g., nodding, saying 'uh-huh') may occur in the near future if the user is in a listening functional state. Input processing modules can also profit from generation modules. For example, the Intent Planner can communicate to the Function Interpreter that it just opened an adjacency pair, from which the Function Interpreter can assess that its complement may be uttered by the user in the near future. Links from input processing to behavior generation allow the generation of behavior based on different levels of understanding (e.g., reactive vs. intentional behavior). These links enable the feedback loops proposed in related work (e.g., Zwiers et al. (2011); Bevacqua et al. (2009)). Each Planner and Interpreter in Asap follows the IU-architecture (Schlangen and Skantze, 2011) for communication between its inner processes: processes fill IUs incrementally with their output and may read partial output of other processes via IUs. Our BML extensions allow incremental and adaptive behavior construction in AsapRealizer.

3.2 Long term

The long term goal we are working toward is to base the architecture for our IVAs on universal (i.e., less problem-specific) and cognitively motivated principles.

As an example, we are working on fully incremental production and recognition processes in order to allow for fast and flexible adaptivity e.g, in the face of dialog feedback (Buschmeier et al., 2012). We further work on representations and decision making mechanisms that consider uncertainty – which is inherent in the recognized and interpreted user input as well as in the intended effects of an agent's behaviors and actions – as valuable information instead of as a mere nuisance. We also investigate cognitively plausible approaches to behavioral interpretations based on predictive matching of sensomotorically grounded motor plans. As a first step in this direction, we have developed a computational cognitive model that allows an IVA to be engaged in gestural interaction with human interlocutors, while simulating mirroring mechanisms such as priming and imitation learning (Sadeghipour and Kopp, 2011).

4 Suggestions for Discussion

1. A new reference architecture for IVAs: During the workshop, we would like to gather the requirements and design a first version of a new reference architecture for IVAs. Our requirements include handling both input processing and output generation, the coordination these processes on multiple levels and incremental processing of input and output. In addition to drawing such an architecture and defining its terminology, we would like to set the agenda to further define shared interfaces between its modules (e.g., PML).

2. Organizing IVA challenges: The Gathering of Animated Lifelike Agents (GALA) festival provided awards for demos with IVAs and aimed to stimulate student work on IVAs, but did not foster the development and comparison of reusable IVA components. To this end we propose more focused challenges aimed at the development of specific components (within a reference architecture). Inspiration for such challenges can be found in related fields such as domestic robotics (RoboCup@Home; `http://www.ai.rug.nl/robocupathome/`), natural language generation (GIVE; Byron et al. (2007)), and speech synthesis (the Blizzard Challenge; Black and Tokuda (2005)).

3. How to share and combine smaller components: Many interesting components for IVAs that are smaller than, e.g., a full Behavior Planner have been developed over the years in isolated projects and experiments. We are interested in discussing how such smaller implementations can be embedded in the larger effort of designing full IVAs, especially in the design of a Behavior Planner. Challenges may help guide IVA component development in such a way that it fits a full IVA architecture. Inspiration for this might be found in the related field of robotics, where the Robot Operating System (ROS; Quigley et al. (2009)) has provided a rich infrastructure for sharing over 3000 robotics-components.

References

T. Baumann and D. Schlangen. 2012. Inpro_iSS: A component for just-in-time incremental speech synthesis. In *Proceedings of the ACL System Demonstrations*, pages 103–108, Jeju Island, South Korea.

K. Bergmann and S. Kopp. 2009. Gnetic – Using Bayesian Decision Networks for iconic gesture generation. In *Intelligent Virtual Agents*, volume 5773 of *LNCS*, pages 76–89. Springer, Berlin, Germany.

K. Bergmann, S. Kahl, and S. Kopp. 2013. Modeling the semantic coordination of speech and gesture under cognitive and linguistic constraints. In *Intelligent Virtual Agents*, volume 8108 of *LNCS*, pages 203–216. Springer, Berlin, Germany.

E. Bevacqua, E. Prepin, R. de Sevin, R. Niewiadomski, and C. Pelachaud. 2009. Reactive behaviors in SAIBA architecture. In *Proceedings of the AAMAS 2009 Workshop 'Towards a Standard Markup Language for Embodied Dialogue Acts'*, pages 9–12, Budapest, Hungary.

A. W. Black and K. Tokuda. 2005. The Blizzard challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proceedings of Interspeech 2005*, pages 77–80, Lisbon, Portugal.

H. Buschmeier, T. Baumann, B. Dosch, S. Kopp, and D. Schlangen. 2012. Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 295–303, Seoul, South Korea.

D. Byron, A. Koller, J. Oberlander, L. Stoia, and K. Striegnitz. 2007. Generating instructions in virtual environments (GIVE): A challenge and evaluation testbed for NLG. In *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, pages 3–4, Arlington, VA, USA.

F. Eyben, M. Wöllmer, and B. Schuller. 2010. openSMILE – The Munich versatile and fast open-source audio feature extractor. In *Proceedings of the International Conference on Multimedia*, pages 1459–1462, Florence, Italy.

S. Kopp, H. van Welbergen, R. Yaghoubzadeh, and H. Buschmeier. 2014. An architecture for fluid real-time conversational agents: Integrating incremental output generation and input processing. *Journal on Multimodal User Interfaces*, 8:97–108.

P. Lison. 2014. *Structured Probabilistic Modelling for Dialogue Management*. Ph.D. thesis, University of Oslo, Oslo, Norway.

M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng. 2009. ROS: An open-source Robot Operating System. In *Proceedings of the ICRA 2009 Workshop on Open Source Software*, Kobe, Japan.

A. Sadeghipour and S. Kopp. 2011. Embodied gesture processing: Motor-based perception-action integration in social artificial agents. *Cognitive Computation*, 3:419–435.

S. Scherer, S. Marsella, G. Stratou, Y. Xu, F. Morbini, A. Egan, A. S. Rizzo, and L.-P. Morency. 2012. Perception Markup Language: Towards a standardized representation of perceived nonverbal behaviors. In *Intelligent Virtual Agents*, volume 7502 of *LNCS*, pages 455–463. Springer, Berlin, Germany.

D. Schlangen and G. Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2:83–111.

M. Schröder and J. Trouvain. 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6:365–377.

H. van Welbergen, R. Yaghoubzadeh, and S. Kopp. 2014. AsapRealizer 2.0: The next steps in fluent behavior realization for ECAs. In *Intelligent Virtual Agents*, LNCS. Springer, Berlin, Germany. To appear.

J. Wienke and S. Wrede. 2011. A middleware for collaborative research in experimental robotics. In *2011 IEEE/SICE International Symposium on System Integration*, pages 1183–1190, Kyoto, Japan.

J. Zwiers, H. van Welbergen, and D. Reidsma. 2011. Continuous interaction within the SAIBA framework. In *Intelligent Virtual Agents*, volume 6895 of *LNCS*, pages 324–330. Springer, Berlin, Germany.

Biographical Sketches

Herwin van Welbergen is a postdoctoral researcher at CITEC, Bielefeld University. Herwin's research deals with the development user interfaces (ranging from social robots to virtual humans to classical mouse-keyboard UIs) that allow a very fluent interaction with a human user. Current he focuses both on the general architecture design of such user interfaces and specifically on multimodal behavior realization that allows fluent interaction.

Kirsten Bergmann is a postdoctoral researcher at CITEC, Bielefeld University. Her research interests include multimodal human communication, data-based and cognitive modeling of human communication skills, and the application of such models in virtual humans to support humans in learning etc.

Hendrik Buschmeier is a PhD-student at CITEC, Bielefeld University. He is interested in dialog phenomena and the mechanisms underlying dialog processing. Right now, Hendrik works on a computational model of dialog coordination based on linguistic feedback and adaptive language production.

Sebastian Kahl is a PhD-student in the Sociable Agents Group at CITEC, Bielefeld University. He is interested in the predictive and dynamic aspects of multimodal meaning representations. His current work entails the development of a spreading-activation based multimodal memory unit of a speech and gesture production system.

Iwan de Kok is a postdoctoral researcher at CITEC, Bielefeld University. He received his MSc and his PhD in Human Media Interaction from the University of Twente, The Netherlands. He is currently working on an incremental dialog system for a virtual coach. His further research interests lie in social signal processing, conversational behavior synthesis and virtual agents.

Amir Sadeghipour is a Ph.D. student at CITEC, Bielefeld University. He is interested in modeling the cognitive processes underlying humans communicative behavior, with a focus on hand-arm gestures. He has developed computational cognitive models which simulate and model the cognitive processes for gesture perception and production.

Ramin Yaghoubzadeh is a PhD student at CITEC, Bielefeld University. His research is on robust multimodal spoken dialog systems to assist older people and people with cognitive impairments. He also likes to tinker with the low-level technical details at times, and likes languages.

Stefan Kopp is head of the Sociable Agents Group at CITEC and professor of Computer Science at Bielefeld University. He explores the cognitive mechanisms of social interaction, both empirically and computationally, in order to build better artificial cooperation and communication partners.

Acknowledgments

This research is supported by the German Research Foundation (DFG) via the Center of Excellence EXC 277 ‘Cognitive Interaction Technology’ (CITEC), the Collaborative Research Center SFB 673 ‘Alignment in Communication,’ as well as by the German Federal Ministry of Education and Research (BMBF) within the ‘Leading-Edge Cluster Competition’ (managed by PTKA).

From Thalamus to Skene: High-level behaviour planning and managing for mixed-reality characters

Tiago Ribeiro, André Pereira, Eugenio di Tullio, Patrícia Alves-Oliveira, and Ana Paiva

INESC-ID & Instituto Superior Técnico, Universidade de Lisboa

1 Main Research Themes

This paper presents work developed by the authors during the last three years on interactive scenarios featuring expressive robots that embody intelligent virtual agents (IVAs). Our work has focused on understanding how a robotic character's behaviour can be modelled and expressed in a body-independent form while providing users with continuous interactions. It has contributed to both the LIREC[1] and current EMOTE[2] EU FP7 Projects.

1.1 LIREC - Thalamus and EMYS-Risk

During the LIREC project, we developed a socially interactive scenario in which the EMYS robot plays the Risk boardgame with three human players (Pereira et al., 2014). The Risk game was implemented as a multimedia application on a touch-table. EMYS takes into account the game state, history of interactions and perception of the environment in order to be perceived as a socially present artificial opponent. Its behaviour is managed by our own EmysToolkit software and an initial version of Thalamus, which handled the scheduling and synchronization of BML speech and face actions. The system was inspired by SAIBA (Kopp et al., 2006), using expressive utterances to model the intention of the robot. These were composed of dialogue acts annotated with animation instructions, and triggered from EMYS's artificial intelligence.

1.2 BML and Perception

While acting in a real-world setting, the behaviour of a character depends on external sources (perception) in order to pose correctly towards the users. We ran a case-study with two robots that interacted by using BML while perceiving the external environment (Ribeiro et al., 2012). The study showed how the imperfection of the robot's sensors makes it risky to schedule BML events and expect regular consistency.

1.3 EMOTE - Thalamus as a Censys system

The EMOTE project aims at developing empathic robotic tutors that can interact with school children through multimedia applications (scenarios). Our requirements for the IVA architecture are now for it to be independent of the type of robot used and to allow to reuse components in different scenarios. We also felt the need for a flexible way of structuring the mind and the body of the agent, especially because of the different types of physical devices that our agent is composed of (robot, external sensors, touch-device).

With this in mind, we created the Censys model of agents, which proposes that there is no need to explicitly define a Mind or a Body in an agent. These components can actually be built out of several interacting processes, which exchange information (Ribeiro et al., 2013). Thalamus was therefore restructured in order to become a modular component-integration framework aimed at IVAs (Ribeiro et al., 2014).

1.4 EMOTE - Skene

Skene is a semi-autonomous behaviour planner that is being developed in the EMOTE project. It is designed to be reused in different applications and embodiments, and in SAIBA fits as a behaviour planner. As we described previously, our interaction environment consists of both robots and other devices (e.g., multi-touch table (MTT)). Skene is the component in which all of them meet. Its input is a high-level behaviour description language (Skene Utterances) and perception information (such as target locations). Its output consists both of scheduled BML and non-BML actions (like sounds or application commands).

Skene Utterances essentially drive some state machines and behaviour generation mechanisms. Such functionality was inspired by the expressive utterances previously used on EMYS but now include instructions for Gazing, Pointing, Waving, Animating, Sound, Head-Nodding and Application instructions. The following is an example of a Skene Utterance:

```
<GAZE(student)> <WAVE(throughMap)>
```

[1] www.lirec.eu

[2] www.emote-project.eu

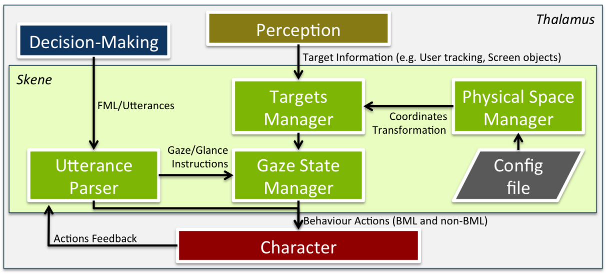


Figure 1: Skene's components.

```
Every map has a <POINT(scale)> scale
that <HEADNOD(1)> can help you. Some
maps represent <ANIMATE(metaphWorld)>
a country, while others represent
smaller areas like this one
<ANIMATE(metaphDichotomicRight)>.
```

The language was developed alongside with non-technical partners from psychology working in EMOTE who provided us with results of human-human interaction analysis to inform on the design of the agent's behaviour (Alves-Oliveira et al., 2014). It is being designed to be simple, human-readable and to provide a link between the psychologists' annotation process and the character's behaviour design.

Skene includes some semi-automated behaviour, such as a gaze-state machine that can keep the character following specific targets while providing it with additional behaviour (e.g gaze-aversion). This gaze-state machine is being developed based on annotations from human-human interactions combined with existing literature (Andrist et al., 2014).

By considering a representation of the Environment, Skene is not bounded to a specific physical set-up. The Physical Space Manager (PSM) keeps a model of the environment surrounding the character. Pointing, gazing or waving all require to know the targets' physical position in relation to the character. The PSM is configured with the resolution, dimensions and position of the MTT relative to the robot. The application informs it about the screen coordinates of relevant GUI and game elements; these coordinates can then be converted to angles to be used by expressive behaviours like Gazing or Pointing.

2 Current Architectures and Standards

The systems we described have followed on SAIBA, with some refinements as we illustrate in Figure 2. We have used both Wizard-of-Oz and autonomous modules to generate and feed Utterances to Skene. These Utterances drive the selection and automated generation of behaviour, both regarding the character's expressiveness and its actions in some virtual environment. It also considers the perception from both the virtual and the real environment in order to keep the behaviour consistent with the surroundings. Our main addition to the SAIBA model is the consideration of the interaction environment as part of the whole architecture. Our architecture is being used in two EMOTE scenarios, in a collaboration with EPFL[3], and on some MSc projects at IST.

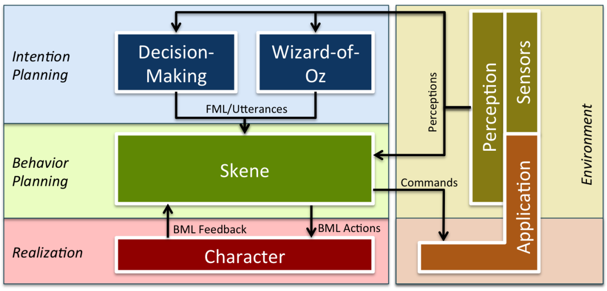


Figure 2: A general structure of our SAIBA-based IVAs.

3 Future Architectures and Standards for IVAs

Our vision is that a successful IVA architecture will allow us to take components from one IVA, and reuse them in another IVA and application. That requires a standardization of interfaces through which the components communicate and the definition of roles/functions of some components. We have been using Thalamus as the backbone for our robotic IVAs, which by following the Censys model, gives us flexibility to separate components and reuse them across applications. Although neither Censys nor Thalamus are architectures, inspiration from Censys might bring forward some of the flexibility we envision to other SAIBA-based systems and the future of IVAs. We also believe that the representation and connection with the surrounding environment must be considered by such an architecture, as the future will bring mixed-reality IVAs that interact with the users through both through some physical form and multimedia applications.

4 Suggestions for Discussion

The following list contains the topics we would like to discuss during this workshop:

- PML and FML syntax;
- How robots and environments fit into SAIBA;
- BML on non-anthropomorphic characters;
- Designing and developing scenarios with non-technical partners (e.g. from psychology and arts).

[3] École polytechnique fédérale de Lausanne

References

P. Alves-Oliveira, S. Janarthanam, A. M. Candeias, A. Deshmukh, T. Ribeiro, H. Hastie, A. Paiva, and R. Aylett. 2014. Towards Dialogue Dimensions for a Robotic Tutor in Collaborative Learning Scenarios. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*.

S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu. 2014. Conversational Gaze Aversion for Human-like Robots. *ACM/IEEE International Conference on Human-Robot Interaction*, pages 25–32.

S. Kopp, B. Krenn, and S. Marsella. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents*, pages 205–217.

A. Pereira, R. Prada, and A. Paiva. 2014. Improving social presence in human-agent interaction. In *Proceedings of the 32nd ACM Conference on Human Factors in Computing Systems*, pages 1449–1458. ACM.

T. Ribeiro, M. Vala, and A. Paiva. 2012. Thalamus: Closing the mind-body loop in interactive embodied characters. In *12nd International Conference on Intelligent Virtual Agents*, pages 189–195.

T. Ribeiro, M. Vala, and A. Paiva. 2013. Censys: A Model for Distributed Embodied Cognition. In *13th International Conference on Intelligent Virtual Agents*, pages 58–67.

T. Ribeiro, E. D. Tullio, L. J. Corrigan, A. Jones, F. Papadopoulos, R. Aylett, G. Castellano, and A. Paiva. 2014. Developing Interactive Embodied Characters using the Thalamus Framework: A Collaborative Approach. In *14th International Conference on Intelligent Virtual Agents*.

Biographical Sketch

Tiago Ribeiro is an eclectic researcher seeking harmony between arts and technology. He has been collaborating internationally on research projects like LIREC and EMOTE, and also with CMU, focusing especially on animation of expressive robots. He is currently in an early stage of obtaining his PhD, in which he pursues artist-oriented intelligent robot animation.

André Pereira is a multidisciplinary researcher focused in the design, implementation and evaluation of social robots. His research is mainly focused in studying how to create socially present board game opponents.

Eugenio Di Tullio obtained his MSc in Computer Science at Universit degli Studi di Bari Aldo Moro, University of Bari (Italy) in 2012. Since then he worked at INESC-ID on the Machinima Project and joined the Emote project in January 2014.

Patrícia Alves-Oliveira is graduated in Clinical and Health Psychology by Instituto Superior de Psicologia Aplicada and Universidade de Aveiro and has done work in the area of human sexuality and emotion with an evolutionary point of view. She is currently a research assistant in INESC-ID at Instituto Superior Técnico, Technical University of Lisbon, working in human-robot interaction. Her interests regard the improvement of human-robot interaction, enabling a balanced future co-existence.

Ana Paiva is a research group leader of GAIPS at INESC-ID and an Associate Professor at Instituto Superior Técnico, Technical University of Lisbon. Her research is focused on the affective elements in the interactions between users and computers with concrete applications in robotics, education and entertainment.

Acknowledgments

This work was partially supported by the European Commission (EC) and was funded by the EU FP7 ICT-317923 project EMOTE and partially supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under the project PEst-OE/EEI/LA0021/2013. The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

Suggestions for Extending SAIBA with the VIB Platform

Florian Pecune[1], Angelo Cafaro[1], Mathieu Chollet[2], Pierre Philippe[2], and Catherine Pelachaud[1]

[1]CNRS-LTCI, Télécom ParisTech and [2]Institut Mines-Télécom, Télécom ParisTech, CNRS-LTCI

1 Main Research Themes

Virtual Interactive Behavior (VIB) is a SAIBA compliant platform which supports the creation of socio-emotional ECAs. It takes as input utterances augmented with communicative functions and emotional states specified in FML-APML (Carolis et al., 2004). The ECA spoken utterances are enriched by nonverbal behaviors NVB (gaze, facial expression, gesture). The choice of NVB can be modulated by the definition of the *dynamicline* that is associated to each ECA (Mancini and Pelachaud, 2008). The dynamicline specifies the preferred modalities and the expressive parameters of each modality (Huang and Pelachaud, 2012). These parameters act on the quality of execution of a behavior such as its speed and acceleration, its fluidity and amplitude.

VIB allows the agent to be an active interactant (speaker or listener). The ECA can decide which social attitudes to display towards its conversation partners (Pecune, 2013). These social attitudes can be shown by the choice of its intentions (Callejas et al., 2014), the type of the ECAs reactions (e.g. exhibit a polite or amused smile (Ochs and Pelachaud, 2013)) and its capacity to be temporally aligned (Prepin et al., 2013). For example, an interpersonal attitude can be chosen at intentional level and the supporting utterances (Chollet et al., 2014) and multimodal behaviors (Ravenet et al., 2013) are produced. While most of the behavior models integrated within VIB are procedural, Ding has used machine learning techniques to drive the multimodal behaviors of the agent when saying emotional speech (Ding et al., 2013) and laughing (Ding et al., 2014).

We are currently working on endowing the agent with the capacity to show its engagement during the interaction by choosing its conversation topics (Glas in (Ochs et al., 2013)) and by making use of hetero-repetition, on expanding the expressivity model by analysing a large database of motion capture data of expressive multimodal behaviors (Fourati and Pelachaud, 2014), and on modeling group behavior during conversations.

2 Current Architectures and Standards

VIB has a modular and extensible architecture. Each module represents an ECA's functionality, therefore one ECA is defined by a particular set of modules as depicted with three examples in Fig. 1. Different formats are adopted to describe the actions that an ECA can perform, these formats represent information ranging from the cognitive level (eg. communicative intentions) to the physical level (eg. skeleton animations). These information are exchanged in the form of events and represent the input and output of the modules. Each module in VIB automatically processes the received events.

We developed a graphical user interface (GUI) that allows a user to design, instantiate and connect the modules (e.g. FML/BML reader, behavior planner, 3D engine player, gesture editor, etc.).

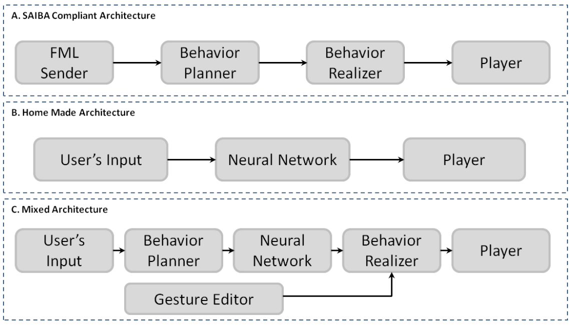


Figure 1: Three example architectures created in VIB

Core Modules These modules include a SAIBA compliant *Behavior Planner* and *Realizer* that work with our own FML-APML specification and extended BML file formats respectively. The Behavior Realizer outputs are keyframes performers used to compute the real-time animation of the agent's body and face. An external player (i.e. 3D engine) can be plugged into the system. This supports the rendering of the 3D environment with different engines (e.g. Ogre3D or Unity3D). Different agents or users can be represented in VIB and share the same player, therefore the same 3D environment.

Other Modules The libraries of movements used are not fixed and may be updated continuously. For this purpose, "editors" modules can be used to define the facial expressions by action units (AU) (Ekman and Friesen, 1978), the gestures (Kendon, 2004; McNeill, 1992), the mapping between AU and FAP (MPEG-4, 2014), the shape of hands, etc. Network communications modules have been developed to interface with external software or to exchange events between different instances of the platform over several machines. Currently, different APIs are used such as ActiveMQ and Thrift. A Neural Network module provides an opportunity for the user to create neurons and connect them via a specific GUI. It can be used to create real-time expressions by motor resonance.

3 Future Architectures and Standards for IVAs

User Behavior Perception Many ECA systems attempted to recognize user behavior during the interaction (e.g. the SEMAINE project (Bevacqua et al., 2012)). There exist commercial and academic software for behavior recognition (e.g. head tracking with OpenCV) and mental states (e.g. emotion recognition with SHORE). However, these are often heterogeneous and require *ad-hoc* development of interfaces for integration into ECA systems. We advocate for a standard representation of the ouptut provided by these applications. The Perception Markup Language (PML) (Scherer et al., 2012), represents a first attempt. However, in addition to the certainty value and the sensory layer, PML redefines behavioral and functional levels (i.e. *behavior* and *function* layers), which could simply reuse the BML and FML specifications.

Multi-party and Multi-floor Interaction In face-to-face communication a person might be engaged in more than one conversation at the same time with different roles. For example, someone could listen to a talk and speak to the person seated next by. This moves from dyadic settings towards more complex multi-party scenarios that should be represented both at functional (i.e. FML) and behavioral (i.e. BML) levels. An FML specification addressing this aspect has been proposed by (Cafaro et al., 2014) but it has not been adopted yet by any ECA system. We also think that representing more complex configurations at functional level and not only with BML may be important to affect the subsequent produced multimodal behavior. A few examples are *1-to-many* (e.g. describing a public speech) and *many-to-many* (two groups interacting as a whole with each other).

Transforming from FML to BML in SAIBA SAIBA currently does not specify how FML should be transformed into BML. We believe that this aspect cannot be left aside of the framework with individual researchers providing their *"home made"* solutions, as this may critically impact the flexibility of integration into other systems. Previous ECA systems have mainly adopted two strategies to solve this problem that are broadly categorized as **data-driven** or **procedural** approaches (cf. introduction of Chapter 6 in (Cafaro, 2014) for a review). In general, our vision is that SAIBA should not only provide standardized interface languages but also techniques and modules that enable to properly transfer between SAIBA components the information represented by these interface languages.

4 Suggestions for Discussion

Raw Perception vs. Attention We emphasized the importance of separating low level user's behavior recognition and its interpretation. An interesting aspect is also the distinction between raw environment perception and the actual information processed by an ECA depending on its attention level. In (Balint and Allbeck, 2013), agents' perceptions are limited by their senses (i.e. solely within their field of view). However, the agent's attention level might filter out some raw information. In a crowded scene, for example, an ECA in face-to-face interaction might exclude some auditory or visual raw perceptions (e.g. other agents walking by) since its attention is focused on the interactant, but another agent or an object (e.g. a car moving fast close by) might trigger an attention shift. We suggest a discussion on how to separate raw perception and attention with emphasis on how to model attention level.

Dealing with Reactive Behaviors Reactive behaviors are part of the interaction. For instance, an agent might bend over to avoid a ball coming to him, or scream after seeing a spider landing on its shoulder. According to (Scherer, 2001), a quicker response is needed for those behavior, may be bypassing the functional planning part of the SAIBA pipeline. The question is whether these low-level quick reactions would fit in SAIBA or should be modeled externally. (Bevacqua et al., 2008) attempted to model this aspect, but on behavioral level (e.g. BML). Our question is what happens at the functional level? If an immediate reaction is required, should the pending intention be canceled, or re-scheduled to be accomplished later?

From BML to Animation Similarly to transforming FML to BML, transforming planned BML into low level parameters ready to be executed as animations by BML Realizer might be problematic. Currently there is no guarantee that playing the same BML block on two different realizers will lead to the same result. Is there a way for designers to be assured that the behaviors they are creating will be played the same way, whatever the agent could be? One solution to address this problem might be to set an additional standardization layer between the Planner and the Realizer components in SAIBA which could provide lower level information (e.g. joint rotations or animation parameters).

References

T. Balint and J. Allbeck. 2013. Whats going on? multi-sense attention for virtual agents. In R. Aylett, B. Krenn, C. Pelachaud, and H. Shimodaira, editors, *Intelligent Virtual Agents*, volume 8108 of *Lecture Notes in Computer Science*, pages 349–357. Springer Berlin Heidelberg.

E. Bevacqua, K. Prepin, E. de Sevin, R. Niewiadomski, and C. Pelachaud. 2008. Reactive behaviors in saiba architecture. In *AAMAS 2009 Workshop Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 9–12.

E. Bevacqua, E. Sevin, S. Hyniewska, and C. Pelachaud. 2012. A listener model: introducing personality traits. *Journal on Multimodal User Interfaces*, 6(1-2):27–38.

A. Cafaro, H. Vilhjálmsson, T. Bickmore, D. Heylen, and C. Pelachaud. 2014. Representing communicative functions in saiba with a unified function markup language. In *Proceedings of the 14th International Conference on Intelligent Virtual Agents (to appear)*, IVA '14.

A. Cafaro. 2014. *First Impressions in Human-Agent Virtual Encounters*. Ph.D. thesis, Center for Analysis and Design of Intelligent Agents, Reykjavik University, Iceland.

Z. Callejas, B. Ravenet, M. Ochs, and C. Pelachaud. 2014. A computational model of social attitudes for a virtual recruiter. In *International Conference on Autonomous Agent and Multi-Agent Systems (AAMAS)*.

B. D. Carolis, C. Pelachaud, I. Poggi, and M. Steedman. 2004. Apml, a markup language for believable behavior generation. In H. Prendinger and M. Ishizuka, editors, *Life-like characters*, Cognitive Technologies, pages 65–86. Springer.

M. Chollet, M. Ochs, and C. Pelachaud. 2014. Mining a multimodal corpus for non-verbal signals sequences conveying attitudes. In *Language Resources and Evaluation Conference (LREC)*.

Y. Ding, C. Pelachaud, and T. Artires. 2013. Modeling multimodal behaviors from speech prosody. In *13th International Conference of Intelligent Virtual Agents - IVA*.

Y. Ding, K. Prepin, J. HUANG, C. Pelachaud, and T. Artires. 2014. Laughter animation synthesis. In *International Conference on Autonomous Agent and Multi-Agent Systems (AAMAS)*.

P. Ekman and W. Friesen. 1978. *The Facial Action Coding System: A Technique For The Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, USA.

N. Fourati and C. Pelachaud. 2014. Emilya: Emotional body expression in daily actions database. In *Language Resources and Evaluation Conference (LREC)*.

J. Huang and C. Pelachaud. 2012. Expressive body animation pipeline for virtual agent. In *proceedings of 12th International Conference of Intelligent Virtual Agents - IVA*, pages 355–362.

A. Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

M. Mancini and C. Pelachaud. 2008. Distinctiveness in multimodal behaviors. In *7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, pages 159–166, Estoril, Portugal.

D. McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

MPEG-4. 2014. `http://mpeg.chiariglione.org/standards/mpeg-4`.

M. Ochs and C. Pelachaud. 2013. Socially aware virtual characters: The social signal of smiles. *IEEE Signal Process. Mag.*, 30(2):128–132.

M. Ochs, Y. Ding, N. Fourati, M. Chollet, B. Ravenet, F. Pecune, N. Glas, K. Prpin, C. Clavel, and C. Pelachaud. 2013. Vers des agents conversationnels anims socio-affectifs. In *Interaction Humain-Machine (IHM'13)*.

F. Pecune. 2013. Toward a computational model of social relations for artificial companions. In *Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII*, pages 677–682.

K. Prepin, M. Ochs, and C. Pelachaud. 2013. Beyond backchannels: co-construction of dyadic stancce by reciprocal reinforcement of smiles between virtual agents. In *International Conference CogSci (Annual Conference of the Cognitive Science Society)*.

B. Ravenet, M. Ochs, and C. Pelachaud. 2013. From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. In *13th International Conference of Intelligent Virtual Agents - IVA*, pages 263–274.

S. Scherer, S. Marsella, G. Stratou, Y. Xu, F. Morbini, A. Egan, A. Rizzo, and L.-P. Morency. 2012. Perception markup language: Towards a standardized representation of perceived nonverbal behaviors. In Y. Nakano, M. Neff, A. Paiva, and M. Walker, editors, *Intelligent Virtual Agents*, volume 7502 of *Lecture Notes in Computer Science*, pages 455–463. Springer Berlin Heidelberg.

K. R. Scherer. 2001. Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, 92:120.

Biographical Sketches

Florian Pecune is a PhD candidate at the LTCI laboratory of Telecom Paristech. His research activities are focused on designing Embodied Conversational Agents able to adapt their decision making in particular social contexts. As part of the ANR MoCA project, which aims at creating a connected world of artificial companions, he proposed a model to compute the social attitude of an agent according to its goals and beliefs.

Angelo Cafaro is a postdoctoral researcher at CNRS-LTCI, Telecom ParisTech. He is doing research in the area of embodied conversational agents and serious game environments with emphasis on social interaction, group behavior and expression of social attitudes. Angelo is part of the EU FP7 Verve project, which aims at developing serious games to support the treatment of elderly people who are at risk of social exclusion. He obtained his Ph.D. from Reykjavik University in 2014. His dissertation dealt with analyzing and modeling human nonverbal communicative behavior exhibited by a virtual agent in a first greeting encounter with the user. In his dissertation he also proposed a SAIBA compliant computational model featuring a unified specification for the Function Markup Language (FML). More information is available on his personal webpage: `www.angelocafaro.info`.

Mathieu Chollet is a PhD candidate at the LTCI laboratory of Telecom Paristech. His research activities are focused on Embodied Conversational Agents and their applications for social skills training. As part of the EU FP7 TARDIS project, which aims at improving youngsters' job interview skills, he proposed behavior models of attitude expression for virtual recruiters. Additionally, he was involved as a Visiting Researcher at the Institute for Creative Technologies where he designed an interactive virtual audience architecture for public speaking training. Personal webpage: `http://perso.telecom-paristech.fr/~mchollet/`

Pierre Philippe is a research engineer at the LTCI laboratory of Telecom ParisTech. He received a Master degree in Computer Science and a M.S. degree in Artificial Intelligence from Brussels Polytechnic School. He worked as a consultant in Computer Associates, then as a research engineer in IRIDIA lab at Brussels University (Belgium) and in COIN lab at Skövde University (Sweden). He is currently modelling and programming modules for the Greta platform. His research interest includes Embodied Conversational Agents, emotions modelling and cognitive architectures.

Catherine Pelachaud is a Director of Research at CNRS in the laboratory LTCI of Telecom ParisTech. Her research interest includes embodied conversational agent, nonverbal communication (face, gaze, and gesture), expressive behaviors and socio-emotional agents.

E-Mail Contacts (`@telecom-paristech.fr`)

Florian Pecune: `florian.pecune`
Angelo Cafaro: `angelo.cafaro`
Mathieu Chollet: `mathieu.chollet`
Pierre Philippe: `pierre.philippe`
Catherine Pelachaud: `catherine.pelachaud`

UTEP's AGENT Architecture

Iván Gris, David Novick, Diego A. Rivera, Mario Gutiérrez

The University of Texas at El Paso

1 Main Research Themes

Two related goals of IVA research involve increasing the believability and perceived trustworthiness of agents and increasing the user's sense of engagement. In our research, we use human-human interaction as a model for agent behaviors that build rapport between humans and agents. Our paralinguistic model of rapport (Novick & Gris, 2014) comprises a sense of emotional connection, a sense of mutual understanding, and a sense of physical connection. Our research, which focuses primarily on the sense of physical connection, seeks to increase the naturalness of non-verbal interaction to correspondingly increase human-IVA rapport (Tickle-Degnan & Rosenthal, 1987; Huang et al., 2011). For these studies, we have developed two agents, with variable non-verbal behaviors, in immersive games designed to maintain users' engagement.

1.1 Familiarity Agent

This agent is part of a longer-term project to provide IVAs with behaviors that enable them to build and maintain rapport with their human partners. We focus on paralinguistic behaviors, especially nonverbal behaviors, and their role in communicating rapport. The Our IVA guides its players through a speech-controlled game called "Escape from the Castle of the Vampire King (see Figure 1), through which we measure the familiarity between humans and agents across two interaction sessions. We studied whether increasing amplitude of nonverbal paralinguistic behaviors leads to an increased perception of physical connectedness between humans and ECAs (Novick & Gris, 2013; Gris, Novick, Gutierrez & Rivera, in press).

1.2 Two-Way Virtual Rapport Agent

This agent is a work in progress in which we make use of full-body gesture recognition to create a sense of physical connection between users and agents. To achieve this, we guide users through a jungle-island survival scenario (see Figure 2). We focus on collaborative physical activities that take place in the virtual environment but require the user's physical action, such as lighting a fire or spearing a fish.

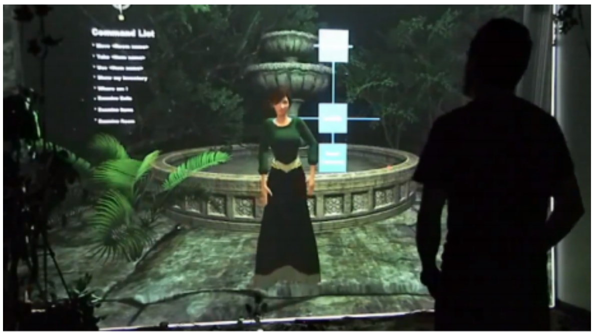


Figure 1. Interaction with the "Escape from the Castle of the Vampire King" agent.

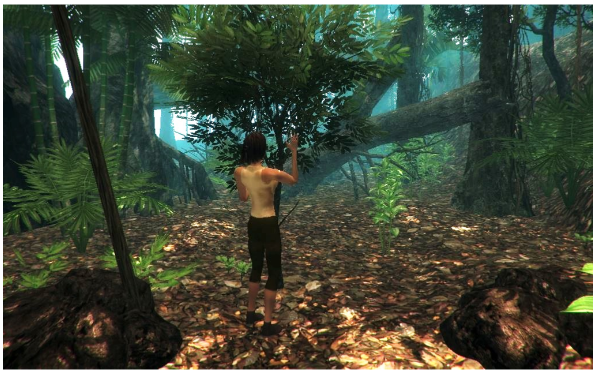


Figure 2. Jungle survival agent.

2 Current Architectures and Standards

We developed our agents independently of SAIBA, BML, and FML, in part because we wanted to understand agent development from the ground up and in part because our research requires some less-common features, such as recognition of full body gestures, blending animations, and agent portability and reuse. The project would likely have benefited from using PBL as a foundation for gesture recognition.

Our implementation uses Unity 4, a Microsoft Kinect, and the Windows Speech SDK, interfaced and net-worked with each other and synchronized to handle the agent's complex behavior. Figure 3 presents an implementation-level outline of the system's components and their relationships. We use the Unity 4 game engine to display our agents and Unity's Mecanim system to create an extensive array of animations.

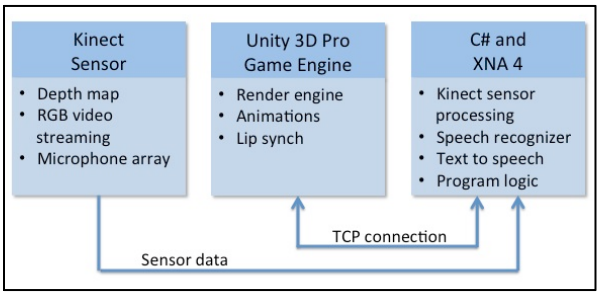


Figure 3. Implementation-level architecture of the software-hardware interfaces and feature handling.

Animations are played by a state graph that follows user-specified parameters of when an animation should start, end, or blend with another animation. Multiple animations can be blended to obtain a completely different animation in real time and give the user the impression that the agent never moves in exactly the same way twice. Animations are divided layers that can control different parts of the body, so multiple animations can be played at the same time and affect different limbs of the agent. The animations are played when the system decodes a message sent by the dialog tree that has the information about the specific animation to be played, the length of the dialog that the agent will say, and the position where the agent and the player should be.

To describe interactive scenes, we developed a dialog interpreter that parses an XML document and links dialog states through conditionals. The file includes the responses anticipated by the systems at relevant parts of the scene. After the interpreter compiles the file, it builds a dialog tree that contains the relationships of the dialogs segments through the storyline.

3 Future of Architectures and Standards for IVAs

Current generations of users are accustomed to hyper-realistic videogame characters and movie assets. A key aspect of these advanced-gen characters is the fluid and realistic way in which they move, which is usually handmade through motion capture. IVAs are a step behind, as their movement is generated on the fly. We would like to see a standard that enables research groups to exchange animation sets or subsets (animation of specific joints) that can be blended at a later time to create unlimited specialized movements.'

Another key feature of IVAs is the real-time perception of a user's full-body gestures. For most IVAs, this ability is limited, partly because the necessary setup is more elaborate, requiring more than a computer and a screen. Now that it is possible to track skeleton joint movement, gesture recognition can be more than a set of image-analysis algorithms with limited tracking capabilities, complicated setups, and long render times. We hope that a standard can be set to capture gestures or poses based on skeleton data, so that studied gestures can easily be ported into the recognizers and applied across agents.

More broadly, we see a need to leverage standards and shared tools through more effective organization of the research community. Our experience suggests that newcomers to the field would be well served by having a single place from which to obtain standards and tools rather than having to visit our research groups' individual sites.

4 Suggestions for Discussion

Our approach for IVA design is modular. Our agents recognize full-body gesture recognition and produce nonverbal behaviors using multiple technologies that we later synchronize through a network. We are interested in how to develop these modular technologies, how can we share them across research groups, and how we can address difficult interaction problems with them. Specific questions include:

- How can shared resources (e.g., BML, PML, SABIA) support extension to novel modes of interaction?
- Given the popularity and accessibility of the latest gaming technologies, how can we make use of them to develop quick, inexpensive and portable IVA prototypes?
- With the increased IVA functionality and interactive capabilities, what requirements and standards should we take into consideration when developing agents that interact not only with humans but with other agents at the same time?
- Can we define a clear separation of IVAs' architectural components between their domain and the agent's features (e.g., gesture recognition, speech recognition, virtual environments, gesture and pose handling, AI elements)? How can we use this to create IVAs, and specific features that can be shared across research groups and disciplines?

References

Gris, I. Adaptive virtual rapport for embodied conversational agents. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013.

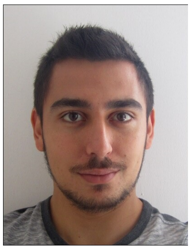
Huang, L., Morency, L.-P., and Gratch, J. Virtual rapport 2.0. In Intelligent Virtual Agents, *Lecture Notes in Computer Science*, Vol. 6895, chapter 8, pages 68-79. Springer, Berlin, Heidelberg, 2011.

Novick, D., and Gris, I. Building rapport between human and ECA: a pilot study. In Proceedings of HCI International 2014, Heraklion, Greece, July, 2014*, Lecture Notes in Computer Science,* vol. 8511, 472-480.

Gris I., Novick, D., Rivera D.A., Gutierrez, M. Recorded speech, virtual environments, and the effectiveness of embodied conversational agents. In *Proceedings of Intelligent Virtual Agents, 2014*, in press.

Tickle-Degnen, L., and Rosenthal, R. Group rapport and nonverbal behavior. In *Group processes and intergroup relations*, pages 113-136. Sage, Newbury Park, CA, 1987.

Biographical Sketches



Ivan Gris is a Ph.D. student at the University of Texas at El Paso working under the supervision of Dr. David Novick. He works in the Interactive Systems Group as part of the Advanced aGent ENgagement Team as the project manager for developing full body embodied conversational agents and immersive, interactive environments. When he is not working on his dissertation, he is working on the research and development for two start-up companies he created. One of these companies is a tech start-up that uses embodied conversational agents, animatronics, scene development, and visual and special effects to develop a unique role-playing themed experience.

Diego A. Rivera is an undergraduate student majoring in Computer Science. He works with the Advanced aGent ENgagement Team as a lead animator. He works integrating and developing the systems that handle the animations for the agents using Unity 3D. In addition he is the main programmer of the dialog interpreter and several other tools and pipelines that control the agent's behavior. Recently Diego obtained an internship on the Institute for Creative Technologies at the University of South-California where he is working under the supervision of Dr. Chad Lane as an integration programmer for Virtual Human and Intelligent Tutoring Research.

Mario Gutierrez is a Master's student at the University of Texas at El Paso. Mario is the lead programmer for the Advanced aGent Engagement Team. He is currently working on his Master's project, which consists of implementing a memory knowledge base for an ECA. He aims to develop agents with memory of previous interactions and dialogs with the users, and then recall them at the appropriate time and in a correct maner.

David Novick is the Mike Loya Distinguished Chair and Professor of Computer Science at the University of Texas at El Paso. He earned his J.D. at Harvard University in 1977 and his Ph.D. in Computer and Information Science at the University of Oregon in 1988. His research has included work on turn-taking and gaze in artificial agents. He built UTEP's Advanced aGent ENgagement Team, whose research focuses on interactive systems and, especially, building rapport in multimodal conversation.

We Never Stop Behaving: The Challenge of Specifying and Integrating Continuous Behavior

Hannes Högni Vilhjálmsson[1], Elías Ingi Björgvinsson[1], Hafdís Erla Helgadóttir[1], and Stefán Ólafsson[1]

[1]CADIA, School of Computer Science, Reykjavik University

1 Main Research Themes

The Socially Expressive Computing group aims to build virtual environments where social interaction is both effective and visually convincing. What particularly sets our work apart is the belief that to reach this goal all social bodies, regardless of whether they are avatars representing human users or characters controlled by autonomous agents, need to automate social awarenes and reactivity to the social environment (Vilhjálmsson, 2014).

This paper is an exposition of the idea that it is not trivial to bring typical Embodied Conversational Agent (ECA) architectures from a situation where they support one-on-one conversations with human users into virtual environments populated with other agents and avatars. This has implications for SAIBA because it grew out of the former situation. The following sections briefly review three phases of prior and current research that demonstrate this.

1.1 Phase 1: Discrete Control of Behavior

In the Embodied Conversational Agent REA (Cassell et al., 1999), the architecture was structured around the agent's response to discrete events generated by the actions of a human interlocutor (Figure 1). These multimodal events were fused into so-called frames which contained both interactional and propositional interpretations of what just happened. Those in turn created obligations, addressed by similar outgoing frames which generated schedules of supporting verbal and non-verbal actions.

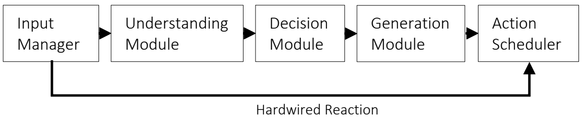


Figure 1: The REA architecture was a pipeline for generating a proper response to a user's input. A short-cut supported a quick hardwired reaction, but in the absence of user input events, nothing would happen

The multi-modal generation process, from text to be spoken to annotated intent to supporting behavior, got consolidated into a flexible tool called BEAT (Cassell et al., 2001). This was an early instance of FML and BML annotation and scheduling. Spark, which animated avatars based on the online chat messages exchanged by their users, used this approach (Vilhjálmsson, 2004). Avatars would idle until someone "spoke" and a multimodal performance ensued. The bursts of lively conversation were framed by awkward inactivity.

1.2 Phase 2: Coordinated Social Environment

The Tactical Language and Culture Training System (Johnson et al., 2004) was a virtual environment populated with interactive characters playing roles in scenarios for training human users in the use of foreign languages and cultural skills (Figure 2).

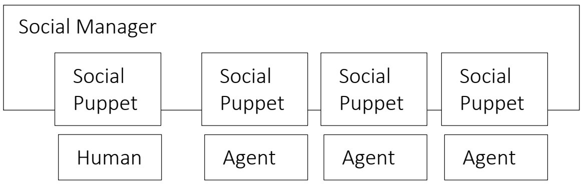


Figure 2: The social puppets of TLCTS could react to any other social puppet in a coordinated fashion thanks to a central manager. Emphasis was on initiating and breaking contact.

Bringing ECAs into an interactive game environment aggrevated the awkwardness inactive moments. People standing idly, waiting to be spoken to, were neither engaging nor natural. A way was needed to naturally manage the co-presence of multiple social bodies that would for example react to someone approaching or even just passing by. This was done with a centralized social manager that kept track of all *social puppets*, which essentially were the embodiments of the agents and users (Vilhjalmsson et al., 2007). The manager fed puppets with perceived FML events, mostly of the interactional kind, such as *A recognizes you* or *A requests turn in your group*. Each puppet would then provide the manager with an FML response, such as *Invite A to join* or *Give A the turn*, as well as animating the supporting nonverbal behavior.

While the social puppets could bounce around FML messages, potentially creating intricate patterns of interactional behavior, each behavior onset or change in idle state (each character could display a variety of idle motions based on context) was a discrete response to a discrete event. Sometimes these meaningful events would occur seconds or even minutes apart. Was no behavior necessary in between?

1.3 Phase 3: Continuous Motion Control and Emergence

People do not simply stop behaving. Our behavior is continuous and is modulated by a constantly changing environment. To capture this notion, it seemed natural to think of the body as being pushed or pulled by a an ever present force, not physical but an abstract one, kind of a *social* force.

Our first instantiation of this, CADIA Populus (Pedica and Vilhjalmsson, 2010) got completely rid of the conversation focused ECAs and neither considered FML or BML. Instead it focused on dynamic positioning and orientation of bodies in large social environments based on a steering behavior framework (Reynolds, 1999). We implemented such things as Kendon's F-formation system (Kendon, 1990) and reactions to the invasion of private space).

At any given moment one or more steering forces would motivate motion in any of the degrees of freedom (forces could be prioritized or combined using weights). These forces would belong to certain contexts, essentially implementing the social norms associated with particular situations. Only those in conversation would for example need to worry about maintaining their F-formation (Figure 3.

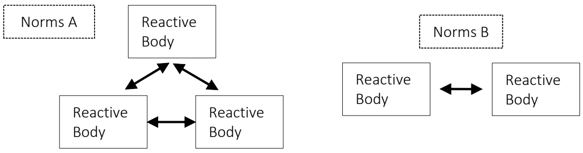


Figure 3: Social bodies are continuously influenced by social forces, tied to certain contexts such as group membership, and update their position and orientation accordingly

This approach resulted in a far more life-like environment (Pedica et al., 2010) where overall social order seemed to emerge from relatively simple rules, independently executed by every character.

2 Current Architectures and Standards

We have now been seeking ways to integrate our continuous steering approach with mechanisms that also support action planning. We have both integrated BDI systems (Thrainsson et al., 2011) and behavior trees (Pedica and Vilhjálmsson, 2012). We are currently combining a REA-like architecture with the behavior tree-based social steering mechanism we call Impulsion. This is taking place in a Unity 3D environment called Virtual Reykjavik (Figure 4). The work is not completed, e.g. our proposed FML (Cafaro et al., to appear) has not been fully integrated with Impulsion.

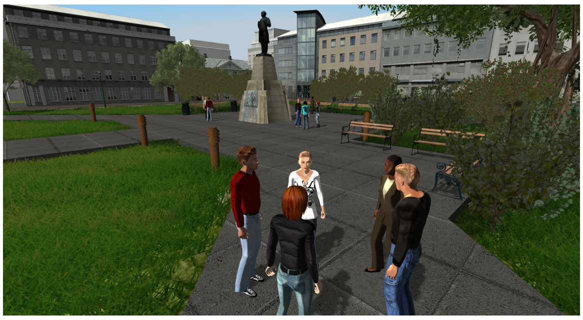


Figure 4: Groups gather in the central square of Virtual Reykjavik

3 Future Architectures and Standards for IVAs

The challenge that SAIBA faces is that for life-like characters we need a multitude of motion engines (e.g. for locomotion, for gatherings, for ”idling” and gazing), that continuously shape nonverbal behavior - *not on a timed schedule*, but as a reaction to a dynamic environment. To some extent we have addressed this by including BML commands that ”set” the state of such engines, but we have not specified what then occurs in any detail.

Furthermore, complex motion engines can be valuable assets and it seems that the SAIBA community could work on interfaces that support their migration between behavior realizers. Being able to shop for components such as gaze engines and gathering engines could save a lot of work.

4 Suggestions for Discussion

In light of the previous discussion, the following topics are suggested for the workshop:

- How does SAIBA currently deal with continuous behavior?
- What is our experience with integrating SAIBA with continuous motion engines (e.g. for locomotion)? Is there inherent incompatability?
- Do we see value in addressing continuous motion within the SAIBA framework? Perhaps with a motion engine interface specification?

References

A. Cafaro, H. H. Vilhjalmsson, T. Bickmore, D. Heylen, and C. Pelachaud. to appear. Representing communicative functions in saiba with a unified function markup language. In *Proceedings of the 13th International Conference on Intelligent Virtual Agents*. Springer-Verlag.

J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. H. Vilhjalmsson, and H. Yan. 1999. Embodiment in conversational interfaces: Rea. In *Proceedings of CHI*, pages 520–527. ACM Press.

J. Cassell, H. H. Vilhjálmsson, and T. Bickmore. 2001. Beat: The behavior expression animation toolkit. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, pages 477–486, New York, NY, USA. ACM.

W. L. Johnson, S. Marsella, and H. H. Vilhjalmsson. 2004. The darwars tactical language training system. In *Proceedings of I/ITSEC*. SSA.

A. Kendon. 1990. *Conducting Interaction: Patterns of behavior in focused encounters*. Cambridge University Press, Cambridge, England.

C. Pedica and H. H. Vilhjalmsson. 2010. Spontaneous avatar behavior for human territoriality. *Appl. Artif. Intell.*, 24(6):575–593, July.

C. Pedica and H. H. Vilhjálmsson. 2012. Lifelike interactive characters with behavior trees for social territorial intelligence. In *ACM SIGGRAPH 2012 Posters*, SIGGRAPH '12, pages 32:1–32:1, New York, NY, USA. ACM.

C. Pedica, H. H. Vilhjalmsson, and M. Larusdottir. 2010. Avatars in conversation: The importance of simulating territorial behavior. In A. et al., editor, *in Proceedings of the 10th International Conference on Intelligent Virtual Agents*, volume 6356 of *Lecture Notes in Computer Science*, pages 336–342. Springer Berlin Heidelberg.

C. W. Reynolds. 1999. Steering behaviors for autonomous characters. In *in the proceedings of Game Developers Conference*, pages 763–782. Miller Freeman Game Group.

P. R. Thrainsson, A. L. Petursson, and H. H. Vilhjálmsson. 2011. Dynamic planning for agents in games using social norms and emotions. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents*, IVA'11, pages 473–474, Berlin, Heidelberg. Springer-Verlag.

H. H. Vilhjalmsson, C. Merchant, and P. Samtani. 2007. Social puppets: Towards modular social animation for agents and avatars. In *Proceedings of the 2Nd International Conference on Online Communities and Social Computing*, OCSC'07, pages 192–201, Berlin, Heidelberg. Springer-Verlag.

H. H. Vilhjálmsson. 2004. Animating conversation in online games. In M. Rauterberg, editor, *Entertainment Computing ICEC 2004*, volume 3166 of *Lecture Notes in Computer Science*, pages 139–150. Springer Berlin Heidelberg.

H. H. Vilhjálmsson. 2014. Automation of avatar behavior. In J. Tanenbaum, M. S. el Nasr, and N. M., editors, *Nonverbal Communication in Virtual Worlds: Understanding and Designing Expressive Characters*, pages 255–266. ETC Press.

Biographical Sketch

Hannes Högni Vilhjálmsson is an Associate Professor of Computer Science at Reykjavik University where he directs the Center for Analysis and Design of Intelligent Agents (CADIA) and leads the Socially Expressive Computing (SECOM) group (`http:\\secom.ru.is`). Prior to RU, Vilhjálmsson was the technical director on the Tactical Language and Culture Training project at USC/ISI and one of the main architects of the early REA ECA and the BEAT nonverbal behavior toolkit at the MIT Media Lab. He can be reached at: `hannes@ru.is`.

Elías Ingi Björgvinsson is enrolled in the Language Technology Masters Program, an interdisciplinary program across University of Iceland and Reykjavík University. He has a BA in Theoretical Linguistics (Icelandic and German). He is a member of the Socialy Expressive Computing (SECOM) group where he works on spoken interaction in Icelandic. He can be reached at: `eliasb13@ru.is`

Hafdís Erla Helgadóttir is currently pursuing her BSc in Computer Science at Reykjavík University, where she just finished her first year. As a member of SECOM, she has been working on a research project over the summer concerning intelligent agents displaying dominant and submissive nonverbal behaviour and the effects of dominance on turntaking. AI is of particular interest to her, inspired by her fascination with games and their making. She can be contacted at: hafdis13@ru.is

Stefán Ólafsson holds BAs in English and Chinese from the University of Iceland. He is currently a masters student in the Language Technologies program at the University of Iceland and Reykjavík University. He is a member of the Socially Expressive Computing group at CADIA where he focuses on dialog systems and cognitive architecture. His email is stefanola13@ru.is

Discussion Summary

H. Buschmeier, A. Cafaro, M. Chollet, P. De Loor, I. Gris, A. Hartholt, M. Jégou, S. Moon, E. Nouri, D. Novick, F. Pecune, F. Popineau, B. Ravenet, A. Robb, H. H. Vilhjálmsson, H. van Welbergen, A. A. Zadeh, and R. Zhao

The discussion groups focused on the following three topics:

1. Planning and realization of multimodal behavior;
2. Sharing assets, resources and knowledge;
3. Describing contextual information.

1 Planning and realization of multimodal behavior

In the first discussion group we focused on the increasing demand in novel IVA applications for continuous rather than discrete multimodal behavior planning and realization. In the SAIBA framework the Behavior Markup Language (BML) supports the specification of discrete behavior (e.g. gestures, speech, posture shifts) and their synchronization. Currently, however, it is impossible to specify behavior in BML that in nature is exhibited in a continuous fashion (e.g. maintaining interpersonal distance, walking in a crowded environment). Related to this, incremental approaches have emerged that compose reactive behavior from small chunks of BML – which can be regarded as partially realizing such continuous behavior. A definition for incremental and continuous behavior is needed and the distinction between the two should be clarified. Then, at the design level, it remains a question if and how BML should support the specification of continuous behavior.

During the discussion some ideas came up for such a specification. Firstly, two layers seem to be needed. A BML representation layer that enables the description of discrete behavior (extending the current BML standard), and an additional layer dealing with continuous behavior representation. This approach may lead to conflicts between the two layers that need to be handled. For example, in a group interaction there could be a BML chunk continuously updating interpersonal space, i.e. locomotion behavior, and another chunk that requires a different movement, i.e. movement towards a given destination. Conceptually, realization capabilities can be seen as shared resources that need to be allocated among competing BML chunks that require them, and some mechanism needs to be designed to determine which chunks are allocated which resources.

Secondly, an overall control logic is required to enable, disable and modify behaviors in the continuous behavior layer itself and to select behaviors, for example, according to priorities. To this end, a scripting language embedding BML descriptors could be introduced which supports these enabling/disabling/modification operations and can additionally chain together behavior using loops in a similar fashion as in imperative programming languages (e.g. for repeating gaze behavior).

Inspiration for the specification of such a two-level specification language and the corresponding realizer implementation may be drawn from design patterns in software engineering or in other fields (e.g. crowd simulation). Since continuous realization requires adaptation of ongoing motion and speech, we established that real-time animation and incremental text-to-speech experts are key figures that need to be involved in the design and implementation process.

We scheduled an agenda for the short and long term. In the short term we plan to organize a collaborative meeting where participants will model a given scenario featuring an interaction that involves both continuous and discrete behavior (e.g. similarly to the Cheeseburger scenario discussed in the 2010 FML workshop at ICT in USA). The goal will be to describe the interaction using the current BML specification, and propose new descriptors when needed, taking into account incremental and continuous behavior. Starting from this exercise, the gaps discussed above, will be formally defined, and a sketch design for the control logic (i.e. logic dealing with the control of discrete and continuous behavior) will be defined. This will be the basis for a preliminary new version of BML and a shared library of control logic patterns (i.e. scripts).

In the long term, all the gaps in the BML standard should be implemented in a new BML 2.0 version. Challenges for this new BML version consist of specifying continuous behavior and providing the specification means required for resolving conflicts between behaviors and allocating their required resources.

2 Sharing assets, resources and knowledge

In the second discussion group we focused on how to best facilitate the sharing of a variety of resources, including capabilities, assets, data and best practices.

First, this requires a common understanding of terms, and to a certain extent technologies and methodologies, in order to make sharing possible. This then allows for the decision of the scale of what's being shared, from individual assets to tools to modules. In the long term, the definition of an overall common theoretical and technical framework based on common standards would act as an organizing principle for available elements. It was felt that the most usable elements to share would be animations, software modules (e.g. a dialogue manager), and experimental data sets and results.

Second, there needs to be a common portal for these shared elements. In its most basic form, this could be an overview of available resources, both academic and commercial. This would provide the additional benefit of acting as an archive that can preserve current views and capabilities for future reference. In a more elaborate form, actual elements created by the research community can be uploaded and shared. Each element would require a minimum set of meta data, including developer info, ease of use, technical readiness level, category, used standards, relevant papers, etc. Software will need additional information, including a defined API and documentation and tutorials. Social elements should be available, providing the ability to comment on and rank elements as well as answer questions and facilitate discussion.

For the short term, this requires a more formal set of requirements, prioritized with a focus on a baseline functionality that can grow over time. Organizers and project leads will have to be identified, as well as possible sources of funding. Explicit attention needs to be paid on defining how to be successful; it is difficult and time consuming to initiate, grow and sustain a community. A central location and forum will have to be set up to facilitate initial discussions. These points should be formalized within a white paper and circulated and updated amongst a small group of interested parties acquired through the IVA emailing list. The finalized white paper can be the starting point for engaging funding sources.

In the long term, an ontology and related formal descriptions can be drafted related to categories of shared resources, target audiences, the overall landscape and associated topology. Meta data will have to be specified and templates set up for lessons learned, tutorials, documentation, etc. If the community grows to a critical mass, the hope would be to be able to connect researchers with other audiences, including application developers and the general public, resulting in potential study participants, end-user feedback, industry collaborations and funding.

3 Describing contextual information

In the third discussion group, we discussed about memory systems and representation of contextual information. In particular, in long term user-agent interactions, it becomes important to keep information about the context of the interactions and build agents with memory capabilities (e.g. relational agents). Indeed, communicating with agents that do not remember who is the user or what he/she said earlier might break the illusion, and lead to an unpleasant interaction. However, there is yet no standardized system allowing virtual agents to keep records of the interactions and manage them. Many knowledge base systems already exist, but each of them use their own process to store information, and to retrieve it.

Then, for the short term, we agreed to design standards regarding these two questions: (1) How do we store the information? We should first try to focus on how to abstract information according to a particular context. Each interaction should be summarized in different chunks of information and store in a knowledge base. Each chunk of information should also be linked to a set of contextual tags, so it can be retrieved later. (2) How do we infer which information should be used during the interaction? The process of retrieving and using chunks of information according to the interaction should be standardized as well.

For the long-term, we should try to focus on a third question, about forgetting the information: How long should we keep the information in the system? Since agents that remember everything might seem creepy, we considered two situations: If the human has to be immersed in a virtual world, the agents he encountered should be granted an accurate but short-termed memory. If the agents are introduced in the real world, their memory should last longer, but should be less accurate.