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- An information-state based incremental dialog manager (yet unpublished) capable of handling uncertain input.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-

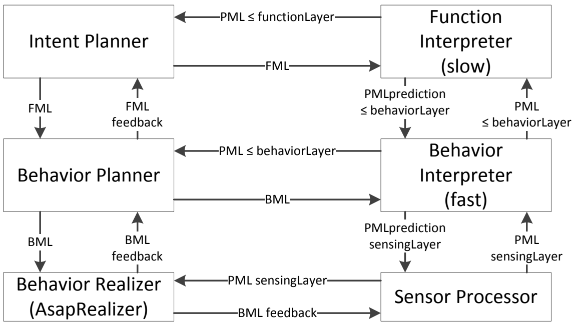

Figure 1: The Asap architecture (Kopp et al., 2014).

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).