Time to go ONLINE! A Modular Framework for Building Internet-based Socially Interactive Agents

Mihai Polceanu*
polceanu@enib.fr
ENIB
Brest, France

Christine Lisetti
lisetti@cis.fiu.edu
Florida International University
Miami, USA

Figure 1: System diagram illustrating main components of the EEVA framework: the modular JavaScript mainframe (a) which controls the client side application by interconnecting the character input, output and non-verbal behavior model (NVBM) with the scenario logic (b) and dialogue content (c).

ABSTRACT

Although socially interactive agents have emerged as a new metaphor for human-computer interaction, they are, to date, absent from the Internet. We describe the design choices, implementation, and challenges in building EEVA, the first fully integrated platform-independent framework for deploying realistic 3D web-based social agents: with real-time multimodal perception of, and response to, the user's verbal and non-verbal social cues, EEVA agents are capable of communicating rich customizable content to users in real time, while building and maintaining users' profiles for long-term interactions. The modularity of the EEVA framework enables it to be used as a testbed for agents' social communication model development of increasing performance and sophistication (e.g. building rapport, expressing empathy). We discuss the framework's feasibility by analyzing the response time of the system over the Internet, in the context of a health intervention built using EEVA authoring functionalities.

*Current affiliation: University of Greenwich, UK.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

IVA '19, July 2–5, 2019, PARIS, France

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6672-4/19/07.

<https://doi.org/10.1145/3308532.3329452>

CCS CONCEPTS

• **Computing methodologies** → **Intelligent agents**; • **Computer systems organization** → **Real-time system architecture**; • **Applied computing** → **Health care information systems**.

KEYWORDS

web-based 3D character; multimodal interaction; real-time virtual counseling

ACM Reference Format:

Mihai Polceanu and Christine Lisetti. 2019. Time to go ONLINE! A Modular Framework for Building Internet-based Socially Interactive Agents. In *ACM International Conference on Intelligent Virtual Agents (IVA '19)*, July 2–5, 2019, PARIS, France. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3308532.3329452>

1 MOTIVATION

As human-computer interaction (HCI) has become increasingly present in daily life contexts involving socio-emotional content (e.g. medicine, education, entertainment), socially interactive virtual agents – also known as Embodied Conversational Agents (ECA) or as Intelligent Virtual Agents (IVA) – have emerged over the past decade as a new metaphor for HCI to address users' need for natural interfaces simulating human-human conversations.

Building an IVA, however, is no easy feat and presents many interdisciplinary challenges. Whereas having socially appropriate interactions can be challenging even for humans at times, generating artificial social behaviors requires a mix of technology, psychology and art. Indeed, social appropriateness during dialogues, requires (apart from choosing an appropriate topic) knowing how to use

different channels of communication to establish and maintain rapport via verbal and non-verbal cues [9, 22], such as respectful eye contact [10], motor mimicry and synchronous postures [3], expression of facial and other nonverbal social cues that are congruent with verbal utterances and emotional states, among others. In spite of such complexity, IVA researchers have leveraged latest progress in affective computing [2, 5, 6, 8] to build agents with social cues and responses [1, 10, 13, 16, 18, 18, 24] towards becoming able to establish some rapport [11] and express empathy [12, 14, 17].

However, IVA development did not scale with the now ubiquitous connected devices and latest progress on 3D graphics that can be rendered on internet browsers (Figure 2). Whereas a few attempts have been made to build web-based 3D ECAs [7, 15, 21], their implementation is still very rudimentary, and none provide an integrated framework for web-based IVA development, including social cues modeling and dialog generation.

Virtual agents have been found promising in eliciting user's motivation [4] and have been shown to be better accepted by users than text-only interventions [14] but, in order to be effective, such agents need to be easily accessible (via common communication devices, at any time), usable (have an easy to use interface), enjoyable (provide a positive user experience), responsive to user's emotional behaviors (establish and maintain rapport) and scalable (accommodate an increasing number of users without computational overhead).

2 THE EEVA FRAMEWORK

EEVA is a modular framework (shown in Fig. 1) that we developed with the purpose of being easy to configure and to extend to a wide range of scenarios. The framework consists of three main components, each of which is detailed in the following sub-sections. In brief: first, the *application layer* consists in a modular client-side JavaScript mainframe (Fig. 1.a) which has the role of controlling the multi-modal user interface on the client side, e.g. WebGL rendering of the ECA (WebGL has recently become the standard technology for 3D interactive web applications), audio/video input, graphical user interface (GUI) interaction and communication with services such as speech recognition and synthesis. Secondly, the JavaScript mainframe handles execution of a scenario (such as a health intervention) which is described in the *logic layer* (Fig. 1.b) – a collection of state-machines that are created by developers using a visual programming interface. Finally, the scenario states may be configured to pull various information (e.g. phrases for the ECA to speak, slides to show) from the *data layer* (Fig. 1.c) – a database of content meant to be communicated to the user, created using an additional authoring tool.

Unlike traditional systems, EEVA's design follows common modularity patterns found in robotics platforms (such as ROS [20]), enabling to create collections of modules to cover a variety of application use-cases, such as using a different browser, adapting to various internet bandwidth limitations, interaction capabilities, or interfacing with additional hardware such as robotic platforms [19] (Fig. 2), among others.

2.1 Application layer

The backbone of the client-side application is a JavaScript framework that manages the creation of a collection of modules and the

Figure 2: EEVA running on different platforms: (a) desktop, (b) mobile phone, (c) autonomous robot [19], (d) smartwatch concept.

communication routes between them. Each module then implements various functionalities, including:

- obtaining input from the user, such as asking for and acquiring permission to access microphone and camera, processing input information (e.g. extracting facial expressions, analyzing users' responses);
- deciding how to respond to the user (e.g. what verbal utterances should the agent express, and what non-verbal behaviors should it portray; and
- responding to the user through a multi-modal 3D embodied ECA, with speech synthesis, non-verbal behaviors and multimedia content (e.g. text, images and videos).

The result is essentially an interactive web application which can run on various platforms from desktop, to mobile phone, to autonomous robotic agent, to potential smart watch integrations (Fig. 2) via a compatible¹ internet browser, while allowing diverse modules like new characters, data-driven behavior models (using for example TensorFlow.js), dialogue systems or even hardware [19] to be plugged into the framework. Scalability is solved by distributing the framework core to end-user devices (3D character, perception, behavior, etc), leading to no significant overhead with adding more users. Compute-intensive functionality (speech/face recognition) can be done asynchronously via web services or with functionalities built into the browser.

2.2 Logic Component and Domain Scenario Authoring Components

With EEVA, building scenarios to be interpreted by the application component consists in defining the structure of a state machine (SM) (note that SM should not to be confused with finite state automata, we use Turing-complete JavaScript abstraction layer).

Once created with EEVA's graphical authoring tools (see Fig. 1, administration tools), the collection of SMs and content that make up the scenarios are saved in JSON format to the database (MongoDB), and can be modified or extended at a later time. In order for the ECA to play a scenario, a file is dynamically generated by the system which contains the translation of the JSON SMs into pure JavaScript code that can be executed on the client-side. We use the SMs to easily author an intervention, but this does not prevent

¹Compatibility consists in WebGL capability and WebRTC [23] compliance; most modern browsers now implement these standards.

the framework from integrating a different dialogue management system if so needed; the SM subsystem consists in one of several modules in the EEVA framework.

3 RUNTIME EVALUATION OVER THE WEB

We evaluated the realtime performance of our framework with a health agent which we build using EEVA framework and authoring tool, with the same motivational interviewing (MI) behavior change intervention described in [14]. The system was deployed on a custom built server running Ubuntu 14.04 LTS with a Quad-core 2GHz CPU and 4GB RAM and measured the response time of critical system components on connections between two continents (see Table 1). The server was configured with SSL/TSL encryption, in order to satisfy the WebRTC security standards for accessing the user's camera and microphone. We tested two main types of network connections available to the public: broadband and 4G mobile data. All experiments were performed under "first run" conditions - *i.e.* no cache mechanism was used, to simulate the first time a new user would connect to the system. The majority of the launch time consists in loading the 3D visual data (the 3D character and surrounding scene), taking on average 30 and 25 seconds on 4G and Broadband respectively, while the lightweight version amounts to 16 and 10 seconds respectively. Each component is asynchronously loaded, thereby limiting the additional overhead.

Table 1: Average response time and standard deviation analysis for EEVA (in milliseconds); 4G/Broadband connections over the Internet between North America and Europe; caching disabled (first run).

Functionality	4G mobile data	Broadband internet
Unity 3D character	30018 +/- 663	24626 +/- 1910
Light 3D character	16612 +/- 1471	10934 +/- 2008
TTS (sentence)	939 +/- 381	551 +/- 141
TTS (word)	72 +/- 40	44 +/- 23
Speech recognition	~30 (Offline processing)	
Entire HTTP request	1124 +/- 166	784 +/- 66
DOM loading	2313 +/- 80	1635 +/- 224

The experiments (Table 1) show that the main distributed functionalities of the EEVA framework do allow real-time interaction and acceptable loading times even for the first run. Scaling up to a large number of users does not imply a linear increase of server capacity, as after the download, most critical components run on the client-side, while only some lightweight communication with the server is performed.

4 FUTURE WORK ADDING EMPATHIC CUES

Future work will involve carrying out experiments and evaluations of nonverbal models of behavior by end-users of the health agent system, in terms of the realism of the IVA behaviors, as well as the end-users' perceived sense of rapport with the IVA delivering the health intervention. Backchannels can also be integrated directly on the client side to account for real-time reaction to the user.

We will also investigate data-driven approaches to modeling the character's behavior and deploying them in directly into the client-side environment, to be integrated with the EEVA mainframe.

ACKNOWLEDGMENTS

Funded by the National Science Foundation (NSF) award #1423260. Thanks to G. Ruiz, D. Rivero, S. Lunn, E. Henley, S. Bolivar for their contribution, and to anonymous reviewers for valuable feedback.

REFERENCES

- [1] R. Amini, C. Lisetti, and G. Ruiz. 2015. HapFACS 3.0: FACS-based facial expression generator for 3D speaking virtual characters. *IEEE Transactions on Affective Computing* 6, 4 (2015). <https://doi.org/10.1109/TAFFC.2015.2432794>
- [2] Maryam Ashoori, Chunyan Miao, Majid Nili, and Mehdi Amoui. 2008. Economically inspired self-healing model for Multi-Agent Systems. In *Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2007*, Vol. 2. 261–264. <https://doi.org/10.1109/IAT.2007.80>
- [3] Janet B. Bavelas, Alex Black, Charles R. Lemery, and Jennifer Mullett. 1986. "I show how you feel": Motor mimicry as a communicative act. *Journal of Personality and Social Psychology* 50, 2 (1986), 322–329.
- [4] Amy L. Baylor. 2011. The design of motivational agents and avatars. *Educational Technology Research and Development* 59, 2 (2011), 291–300.
- [5] Rafael A. Calvo and Sidney D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing* 1, 1 (2010), 18–37. <https://doi.org/10.1109/T-AFFC.2010.1>
- [6] M. Courgeon and C. Clavel. 2013. MARC: A framework that features emotion models for facial animation during human-computer interaction. *Journal on Multimodal User Interfaces* 7, 4 (2013), 311–319.
- [7] Schroeder et al. 2018. Pocket Skills: A Conversational Mobile Web App To Support Dialectical Behavioral Therapy. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2018)* (2018), 1–15.
- [8] Wagner et al. 2013. The Social Signal Interpretation (SSI) Framework Multimodal Signal Processing and Recognition in Real-Time. In *Proceedings of the ACM Multimedia Conference*. 1–4.
- [9] JE Grahe. 1999. The importance of nonverbal cues in judging rapport. *Journal of Nonverbal behavior* 23, 4 (1999), 253–269.
- [10] Ouriel Grynszpan, Jean Claude Martin, and Philippe Fossati. 2017. Gaze leading is associated with liking. *Acta Psychologica* 173 (2017), 66–72. <https://doi.org/10.1016/j.actpsy.2016.12.006>
- [11] Lixing Huang, Louis-philippe Morency, and Jonathan Gratch. 2011. Virtual Rapport 2.0. In *International Conference on Intelligent Virtual Agents, Intelligence, Lecture Notes in Artificial Intelligence*. Springer-Verlag Berlin Heidelberg, 68–79.
- [12] Joris H. Janssen. 2012. A three-component framework for empathic technologies to augment human interaction. *Journal on Multimodal User Interfaces* 6, 3-4 (2012), 143–161. <https://doi.org/10.1007/s12193-012-0097-5>
- [13] Stefan Kopp, Brigitte Krenn, Stacy Marsella, A. Marshall, C. Pelachaud, H. Pirker, K. Thórisson, and H. Vilhjálmsdóttir. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents (Lecture Notes in Computer Science)*, Jonathan Gratch, Michael Young, Ruth Aylett, Daniel Ballin, and Patrick Olivier (Eds.), Vol. 4133. Springer, 205–217.
- [14] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishé. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM TMIS* 4, 4 (2013), 19.
- [15] Gerard Llorach and Josep Blat. 2017. Say Hi to Eliza. In *International Conference on Intelligent Virtual Agents*. 255–258. https://doi.org/10.1007/978-3-319-67401-8_34
- [16] Magalie Ochs, Catherine Pelachaud, and Gary McKeown. 2017. A User-Perception Based Approach to Create Smiling Embodied Conversational Agents. *ACM Transactions on Interactive Intelligent Systems* 7, 1 (2017), 1–33.
- [17] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in Virtual Agents and Robots. *ACM Transactions on Interactive Intelligent Systems* 7, 3 (2017). <https://doi.org/10.1145/2912150>
- [18] Catherine Pelachaud. 2009. Modelling multimodal expression of emotion in a virtual agent. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 364, 1535 (dec 2009), 3539–48.
- [19] Pedro Pena, Christine Lisetti, Mihai Polceanu, and Ubbo Visser. 2018. eEVA: Real-time Web-based Affective Agents for Human-Robot Interface. In *RoboCup 2018: Robot World Cup XXII*. Montreal, Canada, Springer International Publishing.
- [20] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: an open-source Robot Operating System. In *ICRA workshop on open source software*, Vol. 3. Kobe, Japan, 5.
- [21] Vikram Ramanarayanan, David Pautler, Patrick Lange, and David Suendermann-Oett. 2018. Interview with an avatar: A real-time cloud-based virtual dialog agent for educational and job training applications. Technical Report Research Memorandum No. RM-18-02. Princeton, NJ: Educational Testing Service. 1–8 pages.
- [22] L. Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological Inquiry* 1, 4 (1990), 285–293.
- [23] Website. 2019. WebRTC project. <https://webrtc.org>. Accessed Mar 2019.
- [24] U. Yasavur, C. Lisetti, and N. Rishé. 2014. Let's talk! speaking virtual counselor offers you a brief intervention. *Journal on Multimodal User Interfaces* 8, 4 (2014).