Measuring the Power of Learning.®

Research Memorandum
ETS RM–18-02

Interview With an Avatar: A Real-Time Cloud-Based Virtual Dialog Agent for Educational and Job Training Applications

Vikram Ramanarayanan

David Pautler

Patrick Lange

David Suendermann-Oeft

March 2018

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Interview With an Avatar: A Real-Time Cloud-Based Virtual Dialog Agent for Educational and Job Training Applications

Vikram Ramanarayanan, David Pautler, Patrick Lange, and David Suendermann-Oeft
Educational Testing Service, Princeton, New Jersey

March 2018

Corresponding author: V. Ramanarayanan, *E-mail:* vramanarayanan@ets.org

Suggested citation: Ramanarayanan, V., Pautler, D., Lange, P., & Suendermann-Oeft, D. (2018). *Interview with an avatar: A real-time cloud-based virtual dialog agent for educational and job training applications* (Research Memorandum No. RM-18-02). Princeton, NJ: Educational Testing Service.

Action Editor: Anatassia Loukina

Reviewers: Malcolm Bauer and Mark Hakkinen

Abstract

Recent advances in immersive computing technology have the potential to accelerate development of engaging intelligent agents that can guide one or multiple phases of learner instruction as well as formative or summative assessment. In this spirit, we present a multimodal dialog system equipped with a virtual human avatar interlocutor. The agent, developed in Unity with WebGL support, leverages the Help Assistant–Language-Enabled and Free open-source cloud-based standard-compliant dialog framework. We designed and implemented a conversational job interview task based on the proposed framework. In this scenario, the avatar plays the role of an interviewer and responds to user input in real time to provide an immersive user experience. Initial crowdsourcing experiments demonstrate the promise of this approach.

Key words: immersive computing technology, multimodal dialog system, intelligent agents

Acknowledgments

We are grateful to Robert Mundkowsky and Dmytro Galochkin for help with the system engineering. We also thank Eugene Tsuprun, Nehal Sadek, Elizabeth Bredlau, Juliet Marlier, Lydia Rieck, Keelan Evanini, Hillary Molloy, and other members of the ETS Research team for contributions toward the interview item design and for suggestions for system development.

Dialog systems are nowadays becoming increasingly multimodal—interactions that were hitherto mostly based on voice and text (López-Cózar, Callejas, Griol, & Quesada, 2015) have increasingly started to encompass other input–output (I/O) modalities, such as video (Bohus & Horvitz, 2010), gesture (Bohus & Horvitz, 2010; Neßelrath & Alexandersson, 2009), electronic ink (Hastie, Johnston, & Ehlen, 2002; Johnston et al., 2002), avatars or virtual agents (DeMara et al., 2008; Swartout et al., 2013; Swartout et al., 2010), and even embodied agents like robots (Gorostiza et al., 2006; Yu, Bohus, & Horvitz, 2015), among others. The use of avatars, virtual agents, and robotic agents in particular has been popular, as it allows for a more immersive conversational experience.

The research and development community has done significant research in this field, which has led to the development of multiple software platforms and solutions for implementing embodied agents (Baldassarri, Cerezo, & Seron, 2008; Kawamoto et al., 2004; Kethuneni, August, & Vales, 2009; Rist et al., 2004; Thiebaux, Marsella, Marshall, & Kallmann, 2008). More recently, there has also been a push toward developing embodied virtual agents that are empathetic (Fung et al., 2016) and that are directed toward specific educational applications such as language learning (Lee, Noh, Lee, Lee, & Lee, 2010), including the possibility of targeted feedback to participants (Hoque, Courgeon, Martin, Mutlu, & Picard, 2013). In this research memorandum, we focus on applications similar to the latter case in the educational domain; in particular, we examine the challenge of developing a virtual dialog agent that can serve as a job interviewer for workforce training applications.

Although avatars and other interactive virtual agents remain exciting frontiers of research and development in the interaction technology domain, many challenges and design considerations have to be taken into account as well. For instance, the degree of realism and immersiveness of the interaction experience can elicit varying behaviors and responses from users, depending on the nature and design of the virtual interlocutor (Astrid, Krämer, Gratch, & Kang, 2010). Furthermore, inclusion of virtual avatars and agents into conversational interfaces introduces additional technical infrastucture requirements and constraints, including, but not limited to, additional memory resources, adequately high

network connection speeds (especially for cloud-based solutions), and support software. That being said, the advent of low-cost cloud-based distributed computing infrastructure and the growing number of open-source software solutions in this area are alleviating some of these challenges.

The rest of the research memorandum first presents our Help Assistant–Language-Enabled and Free (HALEF)[1] multimodal dialog framework and how we designed and interfaced an avatar with it. We then briefly describe the design of an initial user interaction study with Amazon Mechanical Turk participants and present the results of that study.

System Design and Implementation

This section first describes our existing dialog framework. It then explores the creation of a prototypical avatar in Unity3D[2] and proceeds to describe how such an avatar can be interfaced with HALEF dialogic applications.

The Help Assistant–Language-Enabled and Free Dialog System

The multimodal HALEF dialog system depicted in Figure 1 leverages different open-source components to form a spoken dialog system (SDS) framework that is cloud based, modular, and standards compliant. For more details on the architectural components, please refer to prior publications (Ramanarayanan, Suendermann-Oeft, Lange, Mundkowsky, et al., 2016; Yu et al., 2016).

Interfacing the Avatar With the Help Assistant–Language-Enabled and Free System

In this section, we describe the data flow to and from the avatar-based multimodal HALEF system. Callers use a Web browser–based interface to call into the system. This Web application is written in HTML, CSS, and Javascript. The Media Capture and Streams application programming interface[3] enables access to the computer's audio and video input devices via the Web browser. We use WebRTC[4] and Verto, FreeSWITCH's implementation

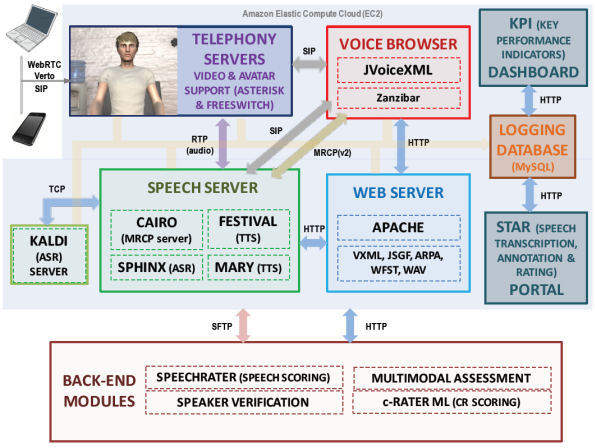

Figure 1. The Help Assistant–Language-Enabled and Free multimodal dialog framework with virtual avatar support for educational learning and assessment applications.

for signaling, to send video and audio to FreeSWITCH and receive audio back from FreeSWITCH. We deploy an Apache server to host all resources, including the Unity3D WebGL build of the avatar that the user loads into the browser. When the call comes in from the user, HALEF starts the dialog with an audio prompt that flows out of the HALEF system via Asterisk over Session Initiation Protocol (SIP)/Real-Time Transport Protocol (RTP) to FreeSWITCH. FreeSWITCH then sends the audio to the Web browser via WebRTC. The user then gives a response to the system that flows through WebRTC to FreeSWITCH and then through SIP/RTP to Asterisk. During the teleconference, the user's video and audio interactions are continuously streamed and recorded. We also use a message server implemented in Node.js[5] to receive commands from the Web server (specified in the VXML code) and forward them to the avatar run time setup in the users browser page. These commands allow us to puppeteer the avatar and trigger different behaviors at specific points in the callflow that blend smoothly with the avatar's default idling behavior—examples include raising a hand or nodding and shaking the head.

Once the Asterisk server receives the call, it establishes a SIP session with the voice browser, which in turn initiates a SIP session with the speech server. The servers negotiate the necessary streaming protocols via Session Description Protocol and set up the required

RTP connections for the media streams. Completion of this process successfully establishes a communication channel between the user and HALEF's components. Once the session is established, Asterisk streams audio via RTP to the speech server. The voice browser will then fetch the VXML code from the Web server based on the extension that has been called. It identifies the resources that the speech server will need to prepare for dialog turn and requests them via MRCPv2 from the speech server. A typical dialog turn follows the following pattern: The voice browser requests speech recognition from the speech server. When the caller starts speaking, the voice activity detector fires and identifies voiced portions of the speech; then, the speech is sent to the Kaldi ASR engine, which starts the decoding process. When the voice activity detector finds that the caller has finished speaking, the recognition result is sent back to the voice browser, which processes it and sends this answer to the spoken language understanding module. The output of the spoken language understanding module is subsequently sent to the dialog manager, which evaluates and generates VXML code with the final response to be played back (or spoken out by the speech synthesizer, if there is no prerecorded audio). The voice browser then interprets this VXML code and sends a synthesis/playback request to the speech server with the response. The speech server plays back/synthesizes the response and passes the result back via RTP to Asterisk, which forwards the audio signal to the user. Once the voice browser parses a VXML page that indicates the end of the application, the voice browser initiates a cleanup to close all open channels and resources. This ends the SIP session with Asterisk, which finally triggers Asterisk to send an end-of-call signal to the user.

Building an Avatar in Unity3D

We used the Unity3D game-building developer environment to customize an avatar for our purposes. Because designing and building an avatar from scratch is both time and labor intensive, we leveraged existing assets from Morph3d[6] (which were not open source but can be obtained for a reasonable cost). In addition, we used Mixamo[7] to customize the imported assets with a few desired animations that might be required of an automated

interviewer virtual agent, such as nodding of the head and movement of the arms. Figure 1 shows a thumbnail of the final avatar with which callers interact.

Crowdsourcing Experiments

We used Amazon Mechanical Turk for our crowdsourcing experiments aimed at investigating user experience with the avatar interview item. Crowdsourcing, and particularly Amazon's Mechanical Turk, has been used in the past for assessing SDSs and for collecting interactions with SDSs (Jurcıcek et al., 2011; McGraw, Lee, Hetherington, Seneff, & Glass, 2010; Ramanarayanan, Suendermann-Oeft, Lange, Ivanov, et al., 2016; Rayner, Frank, Chua, Tsourakis, & Bouillon, 2011). In addition to reading instructions and calling in to the system, users were requested to fill out a 2- to 3-minute survey regarding the interaction. There were no particular restrictions on who could do the spoken dialog task, as we did not want to constrain the pool of people calling in to the system initially.

We designed an interview task that instructs callers to act as a job candidate in an interview with a virtual interviewer agent. Please see Ramanarayanan, Suendermann-Oeft, Ivanov, and Evanini (2015) for a detailed call flow schematic corresponding to this task. As part of the task, the participant clicks a Web page button to start a call with the system and then proceeds to answer the sequence of questions the virtual avatar interviewer poses.

Observations and Analysis

In an initial deployment of the avatar-based interview on Amazon Mechanical Turk, we collected approximately 56 calls, out of which approximately half were disrupted by technical issues (including long loading times for the Unity avatar application, network connection issues, etc.). To better understand the usability of the system, we asked all callers to rate various aspects of their interactions with the system on a scale from 1 to 5, 1 being least satisfactory and 5 being most satisfactory (see Figure 2).

We observed that out of 28 callers who were able to interact for more than 20 seconds, most were able to complete the call successfully. Callers also rated the audio intelligibility, engagement, and appropriateness of the system responses highly, with a

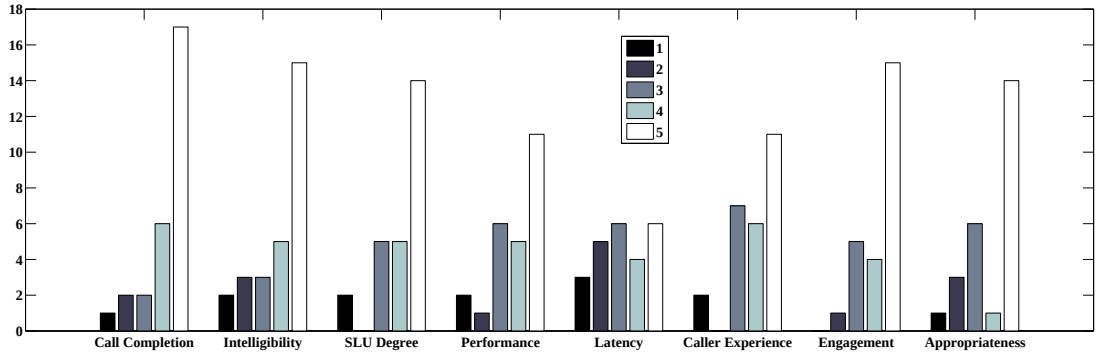

Figure 2. User experience ratings for all calls into the system that were more than 20 seconds long. Ratings are on a 5-point Likert scale ranging from 1 (*least satisfactory*) to 5 (*most satisfactory*). Different rating dimensions are presented along the x axis, whereas the y axis shows the number of calls. Please see the text for more details.

median rating of 5. System latency impacted user experience most negatively, with a median rating of 3. Latency issues, along with a suboptimal degree of spoken language understanding and overall system performance (a measure of whether the system performed per user expectations and if responses were appropriate), were likely instrumental in contributing to a median overall caller experience rating of 4, instead of the desired maximum rating. Nonetheless, these are encouraging numbers, given that this was only an initial study.

Discussion and Outlook

We have described enhancements to the HALEF dialog framework that allow one to puppeteer a virtual human avatar designed using the Unity3D integrated developer environment. Although we have successfully deployed a conversational interview item with such an avatar-rigged setup among crowdsourced participants, there is much room for improvement and further research.

The improvement of user experience, specifically with respect to eliminating delays and improving synchronization between speech and avatar movements, will be a key focus of future work. Two technical challenges are worth noting. First, because the avatar is a

WebGL component in the browser page, commands sent to it are necessarily in a different stream than the audio to be played in sync with it; unless these streams happen to be packetized together (at a level in the network stack below the control of the processes involved), they will be vulnerable to network delays en route that can disrupt their synchronization. Second, because Unity WebGL components typically require minutes to load, hosting them in the browser page imposes a long initial wait. A potential solution to both challenges is to render the WebGL component on the server side before taking any calls and then streaming the avatar video and the speech audio as a unified stream. If the interaction style were to remain strictly turn based, in which the server does not interpret or act on user input until the user indicates that he or she has finished his or her turn (as virtually all current SDS and chatbot systems do), then the WebGL component could be replaced with video recordings of the avatar or of a human actor. However, our nascent work on tracking engagement from video of the user naturally leads away from strict turn taking, because it allows the avatar to perform back-channel addressee actions, such as facial expressions, *during* the user's turn without making an interruption. To maintain this option, we intend to continue work on an avatar that can be puppeteered through both its turn and the user's. In addition, we plan to collect more data from participants to improve the models and the accuracy of the system. Another interesting enhancement already in the works is to incorporate real-time engagement tracking into the system based on caller audio and video, which aims to deliver a more immersive and context-sensitive interaction experience to users.

References

Astrid, M., Krämer, N. C., Gratch, J., & Kang, S. H. (2010). It doesn't matter what you are! Explaining social effects of agents and avatars. *Computers in Human Behavior, 26*, 1641–1650.

Baldassarri, S., Cerezo, E., & Seron, F. J. (2008). Maxine: A platform for embodied animated agents. *Computers & Graphics, 32*, 430–437.

Bohus, D., & Horvitz, E. (2010). Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction* (pp. 5.1–5.8). New York, NY: ACM. https://doi.org/10.1145/1891903.1891910

DeMara, R. F., Gonzalez, A. J., Jones, S., Johnson, A., Hung, V., Leon-Barth, C., . . . Renambot, L. (2008, December). *Towards interactive training with an avatar-based human–computer interface*. Paper presented at the Interservice Industry Training, Simulation & Education Conference, Orlando, FL.

Fung, P., Bertero, D., Wan, Y., Dey, A., Chan, R. H. Y., Siddique, F. B., . . . Lin, R. (2016). *Towards empathetic human–robot interactions*. arXiv preprint arXiv:1605.04072.

Gorostiza, J. F., Barber, R., Khamis, A. M., Pacheco, M., Rivas, R., Corrales, A., . . . Salichs, M. A. (2006). Multimodal human–robot interaction framework for a personal robot. In *The 15th IEEE International Symposium on Robot and Human Interactive Communication, 2006* (pp. 39–44). New York, NY: IEEE.

Hastie, H. W., Johnston, M., & Ehlen, P. (2002). Context-sensitive help for multimodal dialogue. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces* (p. 93). New York, NY: IEEE.

Hoque, M. E., Courgeon, M., Martin, J.-C., Mutlu, B., & Picard, R. W. (2013). Mach: My automated conversation coach. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 697–706). New York, NY: ACM.

Johnston, M., Bangalore, S., Vasireddy, G., Stent, A., Ehlen, P., Walker, M., . . . Maloor,

P. (2002). Match: An architecture for multimodal dialogue systems. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 376–383). Stroudsburg, PA: Association for Computational Linguistics.

Jurcıcek, F., Keizer, S., Gašic, M., Mairesse, F., Thomson, B., Yu, K., & Young, S. (2011). Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk. In P. Cosi, R. De Mori, G. Di Fabbrizio, & R. Pieraccini (Eds.), *Proceedings of INTERSPEECH 2011, 12th annual confrence of the International Speech Communication Association* (pp. 3061–3064). Grenoble, France: International Speech Communication Association.

Kawamoto, S.-I., Shimodaira, H., Nitta, T., Nishimoto, T., Nakamura, S., Itou, K., . . . Lee, A. (2004). Galatea: Open-source software for developing anthropomorphic spoken dialog agents. In H. Prendinger & M. Ishizuka (Eds.), *Life-like characters* (pp. 187–211). New York, NY: Springer.

Kethuneni, S., August, S. E., & Vales, J. I. (2009). Personal healthcare assistant/companion in virtual world. In *AAAI Fall Symposium Series* (pp. 41–42). Retrieved from `https://www.aaai.org/ocs/index.php/FSS/FSS09/paper/download/972/1181`

Lee, S., Noh, H., Lee, J., Lee, K., & Lee, G. G. (2010, September). *Cognitive effects of robot-assisted language learning on oral skills.* Paper presented at the INTERSPEECH 2010 Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology, Tokyo, Japan.

López-Cózar, R., Callejas, Z., Griol, D., & Quesada, J. F. (2015). Review of spoken dialogue systems. *Loquens, 1*(2), e012.

McGraw, I., Lee, C.-Y., Hetherington, I. L., Seneff, S., & Glass, J. (2010). Collecting voices from the cloud. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *LREC 2010: Seventh international conference on language resources and evaluation* (pp. 1576-1583). Paris, France: European Language Resources Association.

Neßelrath, R., & Alexandersson, J. (2009). A 3D gesture recognition system for

multimodal dialog systems. In A. Jönsson,J. Alexandersson, D. Traum, & I. Zukerman (Eds.), *Proceedings of the twenty-first international joint conference on artificial intelligence (IJCAI09)* (pp. 46–51). Palo Alto, CA: AAAI Press.

Ramanarayanan, V., Suendermann-Oeft, D., Ivanov, A., & Evanini, K. (2015). *A distributed cloudbased dialog system for conversational application development.* Paper presented at the 16th annual SIGdial Meeting on Discourse and Dialogue, Prague, Czech Republic.

Ramanarayanan, V., Suendermann-Oeft, D., Lange, P., Ivanov, A. V., Evanini, K., Yu, Z., . . . Qian, Y. (2016). *Bootstrapping development of a cloud-based spoken dialog system in the educational domain from scratch using crowdsourced data* (Research Report No. RR-16-16). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12105

Ramanarayanan, V., Suendermann-Oeft, D., Lange, P., Mundkowsky, R., Ivanou, A., Yu, Z., . . . Evanini, K. (2016). Assembling the jigsaw: How multiple W3C standards are synergistically combined in the HALEF multimodal dialog system. In D. D. Dahl (Ed.), *Multimodal interaction with W3C standards: Towards natural user interfaces to everything* (pp. 295–310). New York, NY: Springer.

Rayner, E., Frank, I., Chua, C., Tsourakis, N., & Bouillon, P. (2011). *For a fistful of dollars: Using crowd-sourcing to evaluate a spoken language call application.* In *Proceedings of the ISCA International Workshop on Speech and Language Technology in Education (SLaTE 2011)* (pp. 117–120). Grenoble, France: International Speech Communication Association.

Rist, T., André, E., Baldes, S., Gebhard, P., Klesen, M., Kipp, M., . . . Schmitt, M. (2004). A review of the development of embodied presentation agents and their application fields. In H. Prendinger & M. Ishizuka (Eds.), *Life-like characters: Tools, affective functions, and applications* (pp. 377–404). New York, NY: Springer.

Swartout, W., Artstein, R., Forbell, E., Foutz, S., Lane, H. C., Lange, B., . . . Traum, D. (2013). Virtual humans for learning. *AI Magazine, 34*(4), 13–30.

Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., . . .

White, K. (2010). Ada and Grace: Toward realistic and engaging virtual museum guides. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent virtual agents: Proceedings of the 10th international conference* (pp. 286–300). New York, NY: Springer.

Thiebaux, M., Marsella, S., Marshall, A. N., & Kallmann, M. (2008). Smartbody: Behavior realization for embodied conversational agents. In L. Padgham (Ed.), *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems* (Vol. 1, pp. 151–158). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Yu, Z., Bohus, D., & Horvitz, E. (2015). Incremental coordination: Attention-centric speech production in a physically situated conversational agent. In *16th annual meeting of the Special Interest Group on Discourse and Dialogue* (p. 402–406). Stroudsburg, PA: Association for Computational Linguistics.

Yu, Z., Ramanarayanan, V., Mundkowsky, R., Lange, P., Ivanov, A., Black, A. W., & Suendermann-Oeft, D. (2016). *Multimodal HALEF: An open-source modular Web-based multimodal dialog framework*. In K. Jokinen & G. Wilcock (Eds.), *Lecture notes in electrical engineering: Vol. 427. Dialogues with social robots* (pp. 233–244). Singapore: Springer. https://doi.org/10.1007/978-981-10-2585-3_18

Notes

[1] http://halef.org/

[2] https://unity3d.com/

[3] https://www.w3.org/TR/mediacapture-streams

[4] http://www.w3.org/TR/webrtc/

[5] https://nodejs.org/en/

[6] https://www.morph3d.com/

[7] https://www.mixamo.com/