A Modular, Multimodal Open-Source Virtual Interviewer Dialog Agent

Kirby Cofino
American University
Washington, DC
kirbygc00@gmail.com

Vikram Ramanarayanan
Educational Testing Service R&D
San Francisco, CA
vramanarayanan@ets.org

Patrick Lange
Educational Testing Service R&D
San Francisco, CA
plange@ets.org

David Pautler
Educational Testing Service R&D
San Francisco, CA
dpautler@ets.org

David Suendermann-Oeft
Educational Testing Service R&D
San Francisco, California
suendermann-oeft@ets.org

Keelan Evanini
Educational Testing Service R&D
Princeton, NJ
kevanini@ets.org

ABSTRACT

We present an open-source multimodal dialog system equipped with a virtual human avatar interlocutor. The agent, rigged in Blender and developed in Unity with WebGL support, interfaces with the HALEF open-source cloud-based standard-compliant dialog framework. To demonstrate the capabilities of the system, we designed and implemented a conversational job interview scenario where the avatar plays the role of an interviewer and responds to user input in real-time to provide an immersive user experience.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Discourse, dialogue and pragmatics**;

KEYWORDS

avatar, dialog system, open-source, multimodal, virtual agent

ACM Reference Format:
Kirby Cofino, Vikram Ramanarayanan, Patrick Lange, David Pautler, David Suendermann-Oeft, and Keelan Evanini. 2017. A Modular, Multimodal Open-Source Virtual Interviewer Dialog Agent. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3136755.3143034

1 INTRODUCTION

There has been significant work in the research and development community on the use of avatars, virtual agents and robotic agents to enable a more immersive conversational experience. This effort has led to the development of multiple software platforms and solutions for implementing embodied agents [1, 4, 5, 9, 10]. More recently, there has also been a push towards developing embodied virtual agents that are empathetic [2] and are directed toward specific educational applications such as language learning [6], including the possibility of targeted feedback to participants [3]. Here we demonstrate a fully open-source virtual dialog agent that can serve as a job interviewer for workforce training applications.

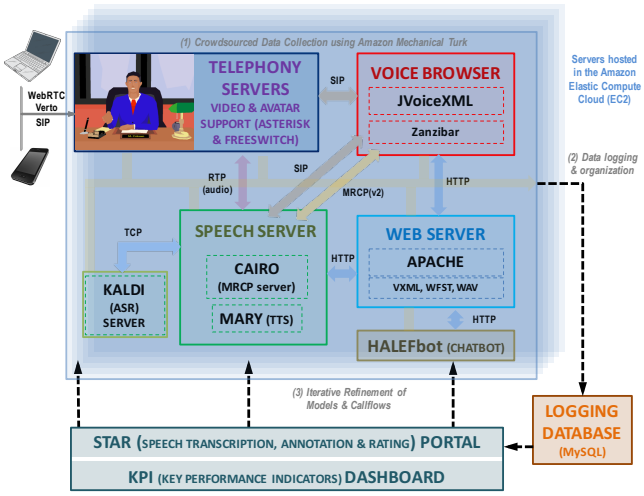

Figure 1: The HALEF multimodal dialog framework with virtual avatar support for educational learning and assessment applications (©Authors).

2 SYSTEM DESIGN AND IMPLEMENTATION

This section first describes our existing dialog framework. It then explores the creation of a prototypical avatar using Blender[1] and Unity3D[2], and proceeds to describe how such an avatar can be interfaced with the HALEF dialog system for an interview application.

2.1 The HALEF Dialog System

The multimodal HALEF[3] dialog system depicted in Figure 1 leverages different open-source components to form a spoken dialog

[1] https://www.blender.org/
[2] https://unity3d.com/
[3] http://halef.org

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3143034

Figure 2: Screenshot of the virtual agent described in this study (©Authors).

system (SDS) framework that is cloud-based, modular and standards-compliant. For more details on the architectural components, please refer to [8].

2.2 Building the Avatar

We modeled, textured, rigged, and animated the avatar in Blender. We modularized the animations into lip movements, head movements, arm and chest movements, and leg movements. In this way, we were able to mix and match the correct animations to play in order to give a naturalistic appearance to the avatar inside of Unity, which has a layering functionality to its animator. To sync the lips with the dialog we used Papagayo studio[4] to insert key frames into our animation actions (clips) in Blender. These key frames manipulate ten different shape keys we made on the model to correspond to ten different phoneme categories. In addition to this, we hand-animated the arms and hands to correspond with the dialog by emphasizing different ideas through motion.

We designed an interview task that instructs callers to act as a job candidate in an interview with a virtual interviewer agent. Please see [7] for a detailed callflow schematic corresponding to this task. As part of the task, the participant clicks a webpage button to start a call with the system and then proceeds to answer the sequence of questions posed by the virtual avatar interviewer.

2.3 Interfacing the Avatar with HALEF

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 API[5] enables access to the computer's audio and video input devices via the web browser. We use WebRTC[6] and Verto, FreeSWITCH's implementation 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 SIP/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 used a message server implemented in Python to receive commands from the webserver (specified in the VXML code) and forward them to the avatar runtime setup in the user's 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.

[4] http://lostmarble.com/papagayo/
[5] https://www.w3.org/TR/mediacapture-streams
[6] http://www.w3.org/TR/webrtc/

3 CONCLUSIONS

We have presented an open-source virtual agent that can seamlessly interface with an existing open-source modular cloud-based multimodal dialog system to create immersive interactive experiences.

4 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 as well as suggestions for system development.

REFERENCES

[1] Sandra Baldassarri, Eva Cerezo, and Francisco J Seron. 2008. Maxine: A platform for embodied animated agents. *Computers & Graphics* 32, 4 (2008), 430–437.
[2] Pascale Fung, Dario Bertero, Yan Wan, Anik Dey, Ricky Ho Yin Chan, Farhad Bin Siddique, Yang Yang, Chien-Sheng Wu, and Ruixi Lin. 2016. Towards Empathetic Human-Robot Interactions. *arXiv preprint arXiv:1605.04072* (2016).
[3] Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. 2013. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 697–706.
[4] SKSEA James and Ian Vales. 2009. Personal healthcare assistant/companion in virtual world. In *AAAI Fall Symposium Series*.
[5] Shin-ichi Kawamoto, Hiroshi Shimodaira, Tsuneo Nitta, Takuya Nishimoto, Satoshi Nakamura, Katsunobu Itou, Shigeo Morishima, Tatsuo Yotsukura, Atsuhiko Kai, Akinobu Lee, et al. 2004. Galatea: Open-source software for developing anthropomorphic spoken dialog agents. In *Life-Like Characters*. Springer, 187–211.
[6] Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, and Gary Geunbae Lee. 2010. Cognitive effects of robot-assisted language learning on oral skills. In *INTERSPEECH 2010 Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*.
[7] Vikram Ramanarayanan, David Suendermann-Oeft, Alexei Ivanov, and Keelan Evanini. 2015. A distributed cloud-based dialog system for conversational application development. In *16th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2015), Prague, Czech Republic*.
[8] Vikram Ramanarayanan, David Suendermann-Oeft, Patrick Lange, Robert Mundkowsky, Aliaksei Ivanou, Zhou Yu, Yao Qian, and Keelan Evanini. 2016. Assembling the jigsaw: How multiple W3C standards are synergistically combined in the HALEF multimodal dialog system. In *Multimodal Interaction with W3C Standards: Towards Natural User Interfaces to Everything*. Springer, to appear.
[9] Thomas Rist, Elisabeth André, Stephan Baldes, Patrick Gebhard, Martin Klesen, Michael Kipp, Peter Rist, and Markus Schmitt. 2004. A review of the development of embodied presentation agents and their application fields. In *Life-Like Characters*. Springer, 377–404.
[10] Marcus Thiebaux, Stacy Marsella, Andrew N Marshall, and Marcelo Kallmann. 2008. Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 151–158.