A MODULAR OPEN-SOURCE STANDARD-COMPLIANT DIALOG SYSTEM FRAMEWORK WITH VIDEO SUPPORT

Vikram Ramanarayanan[†], *Zhou Yu*[‡], *Robert Mundkowsky*[†], *Patrick Lange*[†],
Alexei V. Ivanov[†], *Alan W. Black*[‡] *and David Suendermann-Oeft*[†]

[†] Educational Testing Service (ETS) R&D, San Francisco, CA
[‡]Carnegie Mellon University, Pittsburgh, PA
<vramanarayanan,rmundkowsky,plange,aivanou,suendermann-oeft>@ets.org, <zhouyu,awb>@cs.cmu.edu

We present HALEF (Help Assistant–Language-Enabled and Free), an open-source cloud-compatible multimodal dialog system that can be used with different plug-and-play back-end application modules and includes support for video interfacing via web browser. The system is compliant with multiple World Wide Web Consortium standards while maintaining an open codebase to encourage progressive development and a common standard testbed for multimodal dialog system development and benchmarking. The system can be deployed toward a versatile range of potential use cases, including intelligent tutoring, learning and assessment applications, and interactive voice response (IVR) systems, among others.

The HALEF framework is composed of the following distributed open-source modules that have been described in detail in previous publications [1, 2]:

- An Asterisk [3] telephony server that is compatible with SIP (Session Initiation Protocol), PSTN (Public Switched Telephone Network) standards, and acts as a public branching exchange (PBX).
- A Freeswitch telephony server [4] that is compatible with SIP and WebRTC (Web Real-Time Communications) standards, and allows support for voice and video communication via web browser.
- A voice browser – JVoiceXML [5] – that is compatible with VoiceXML 2.1 and can process SIP traffic, and incorporates support for multiple grammar standards such as JSGF (Java Speech Grammar Format), ARPA (Advanced Research Projects Agency) and WFST (Weighted Finited State Transducer).
- An MRCP (Media Resource Control Protocol) speech server [6] – Cairo – which allows the voice browser to initiate SIP or RTP (Real-time Transport Protocol) connections from/to the telephony server and incorporates multiple speech recognizers (Sphinx [7], Kaldi [8]) and synthesizers (Mary [9], Festival [10]).
- An Apache Tomcat-based web server[1] that can host dynamic VoiceXML pages, web services, and media libraries containing grammars and audio files.
- OpenVXML[2], a VoiceXML-based voice application authoring suite that generates dynamic web applications that can be housed on the web server.
- A MySQL[3] database server for storing call-log information.
- A Speech Transcription, Annotation and Rating (STAR) portal that allows one to listen to and transcribe full-call recordings, rate them on a variety of dimensions such as caller experience and latency, and perform various semantic annnotation tasks required to train automatic speech recognition and spoken language understanding modules.

We will illustrate the basic architecture and components of the HALEF spoken dialog system using example applications that are currently deployed in the educational domain. One such case is that of a job interview application where potential users call into the system via web interface, and have to respond to questions posed to them by an automated interviewer, based on which their communication skills can be assessed.

1. REFERENCES

[1] Vikram Ramanarayanan, David Suendermann-Oeft, Alexei Ivanov, and Keelan Evanini, "A distributed cloud-based dialog system for conversational application development," in *16th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2015), Prague, Czech Republic*. 2015.

[2] David Suendermann-Oeft, Vikram Ramanarayanan, Moritz Teckenbrock, Felix Neutatz, and Dennis Schmidt, "HALEF: an open-source standard-compliant telephony-based modular spoken dialog system–A review and an outlook," in *Proc. of the IWSDS Workshop 2015, Busan, South Korea*. 2015.

[3] J. van Meggelen, J. Smith, and L. Madsen, *Asterisk: The Future of Telephony*, O'Reilly, Sebastopol, USA, 2009.

[4] Anthony Minessale and Darren Schreiber, *FreeSWITCH Cookbook*, Packt Publishing Ltd, 2012.

[1] http://tomcat.apache.org/

[2] https://github.com/OpenMethods/OpenVXML

[3] https://www.mysql.com/

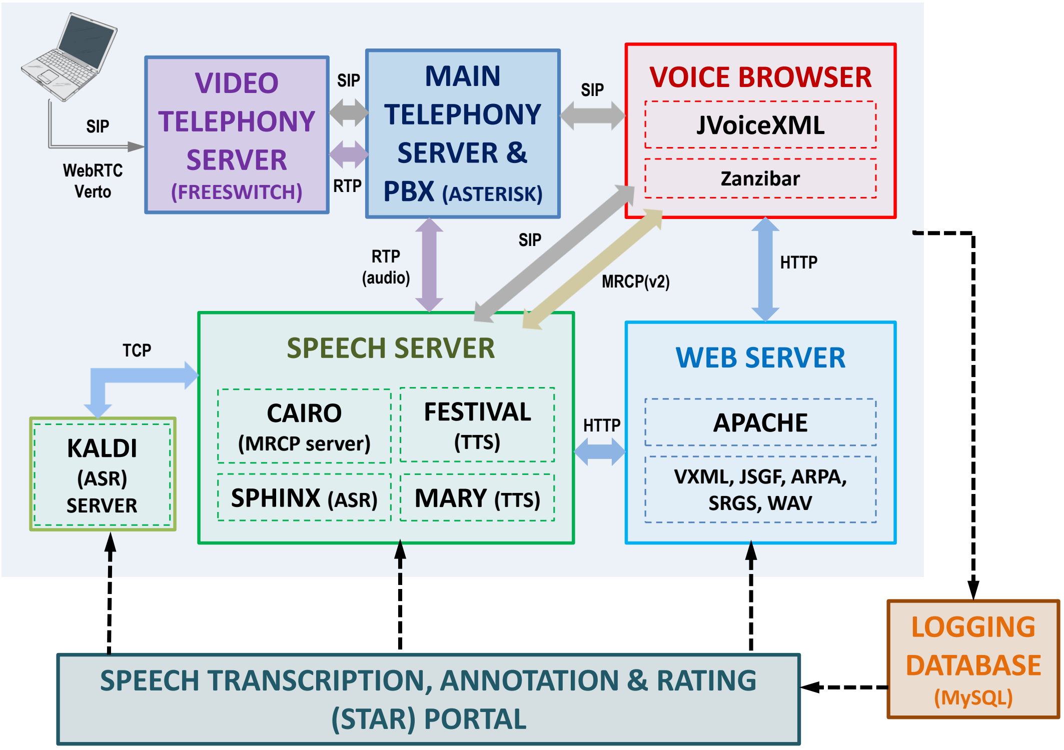

Fig. 1. System architecture of the HALEF multimodal dialog system depicting the various modular open-source components.

[5] D. Schnelle-Walka, S. Radomski, and M. Mühlhäuser, “JVoiceXML as a Modality Component in the W3C Multimodal Architecture,” *Journal on Multimodal User Interfaces*, vol. 7, pp. 183–194, 2013.

[6] D. Prylipko, D. Schnelle-Walka, S. Lord, and A. Wendemuth, “Zanzibar OpenIVR: An Open-Source Framework for Development of Spoken Dialog Systems,” in *Proc. of the TSD Workshop*, Pilsen, Czech Republic, 2011.

[7] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf, “The CMU SPHINX-4 Speech Recognition System,” in *Proc. of the ICASSP’03*, Hong Kong, China, 2003.

[8] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, “The Kaldi speech recognition toolkit,” in *Proc. of the ASRU Workshop*, Hawaii, USA, 2011.

[9] Marc Schröder and Jürgen Trouvain, “The german text-to-speech synthesis system mary: A tool for research, development and teaching,” *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[10] P. Taylor, A. Black, and R. Caley, “The Architecture of the Festival Speech Synthesis System,” in *Proc. of the ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.