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 – JV oiceXML [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 Finite 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 server1 that can host dynamic VoiceXML pages, web services, and media libraries containing grammars and audio files.
- OpenVXML2, a VoiceXML-based voice application authoring suite that generates dynamic web applications that can be housed on the web server.
- A MySQL3 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 annotation 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


1http://tomcat.apache.org/
2https://github.com/OpenMethods/OpenVXML
3https://www.mysql.com/
Fig. 1. System architecture of the HALEF multimodal dialog system depicting the various modular open-source components.


