INTERSPEECH 2016: Show & Tell Contribution
September 8–12, 2016, San Francisco, USA

Open Language Interface for Voice Exploitation (OLIVE)

Aaron Lawson, Mitchell McLaren, Harry Bratt, Martin Graciarena, Horacio Franco, Christopher George, Allen Stauffer, Christopher Bartels, Julien VanHout

SRI International, California, USA

aaron.lawson@sri.com

Abstract

We propose to demonstrate the Open Language Interface for Voice Exploitation (OLIVE) speech-processing system, which SRI International developed under the DARPA Robust Automatic Transcription of Speech (RATS) program. The technology underlying OLIVE was designed to achieve robustness to high levels of noise and distortion for speech activity detection (SAD), speaker identification (SID), language and dialect identification (LID), and keyword spotting (KWS). Our demonstration will show OLIVE performing those four tasks. We will also demonstrate SRI's speaker recognition capability live on a mobile phone for visitors to interact with.

Index Terms: speech activity detection, speaker and language identification, keyword spotting

1. Proposal

We propose to demonstrate the OLIVE speech-processing system, which SRI International developed under the DARPA Robust Automatic Transcription of Speech (RATS) program. The technology underlying OLIVE was developed to achieve robustness to high levels of noise and distortion. The specific tasks that we will demonstrate include: **speech activity detection (SAD)** [1] (detecting the presence of speech, not just an open channel); **speaker identification (SID)** [2] (finding and/or tracking speakers of interest); **language and dialect identification (LID)** [3] (detecting languages and dialects from a set of languages of interest); and **keyword spotting (KWS)** [4] (detecting specific keywords and phrases).

The OLIVE speech-processing system is based on client/server architecture. Clients connect to the server through a ZeroMQ message-passing application programming interface (API), which can connect to multiple user interfaces or external systems. The graphical user interface (GUI) clients are written in Java, and the core speech-processing engine is written in Python. Our demonstration will show two graphical user interfaces, reflecting (1) a forensic or close analysis use case and (2) a triage or "big data" use case.

The forensic use case (figure 1 below) enables users to select audio segments to perform speaker identification, speech activity detection, and language identification. Users can also add to or create new speaker ID or language ID models. To help with speaker enrollments, functionality is included that enables users to diarize a file given a small (5–10 second) snippet of the talker's speech. This feature is being extended to language ID, and other features are being adapted for speech activity detection, to identify keywords, and to provide a phonetic transcript in the file. The triage use case is when a significant amount of waveforms need to be processed. The system first discards the waveforms which contain no speech, and then sends the rest for processing by the speaker, language and keyword spotting plugins.

We will demonstrate two ways to use the OLIVE system, as shown in figure 2 below. The first way is with the two graphical user interfaces (forensic/close and triage/big-data GUIs). The second way is using the OLIVE API to integrate the server with an existing system.

This technology is under continuous development and refinement based on user feedback, and OLIVE is designed such that adding new capabilities is practically automatic, once the underlying algorithms are coded as plugins.

2. Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0037. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Distribution Statement A: Approved for Public Release, Distribution Unlimited.

3. References

[1] Graciarena, M, Alwan, A, Ellis, D, Franco, H, Ferrer, L, Hansen, JHL, Janin, A, Lee, BS, Lei, Y, Mitra, V, Morgan, N, Sadjadi, SO, Tsai, TJ, Scheffer, N, Tan, LN & Williams, B 2013, "All for one: Feature combination for highly channel-degraded speech activity detection," *Proc. of INTERSPEECH*, pp. 709–713, 14th Annual INTERSPEECH 2013, Lyon, France, 25–29 August.

[2] M. McLaren, N. Scheffer, M. Graciarena, L. Ferrer, and Y. Lei, "Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion," *Proc. ICASSP*, 2013, pp. 6773–6777.

[3] Y. Lei, L. Ferrer, A. Lawson, M. McLaren, and N. Scheffer, "Application of convolutional neural networks to language identification in noisy conditions," *Proc. Odyssey-14*, Joensuu, Finland, June 2014.

[4] J van Hout, V Mitra, Y Lei, D Vergyri, M Graciarena, A Mandal, H Franco, "Recent improvements in SRI's keyword detection system for noisy audio," *Proc. of Interspeech*, Baixas, France, 2014, pp. 1727–1731.

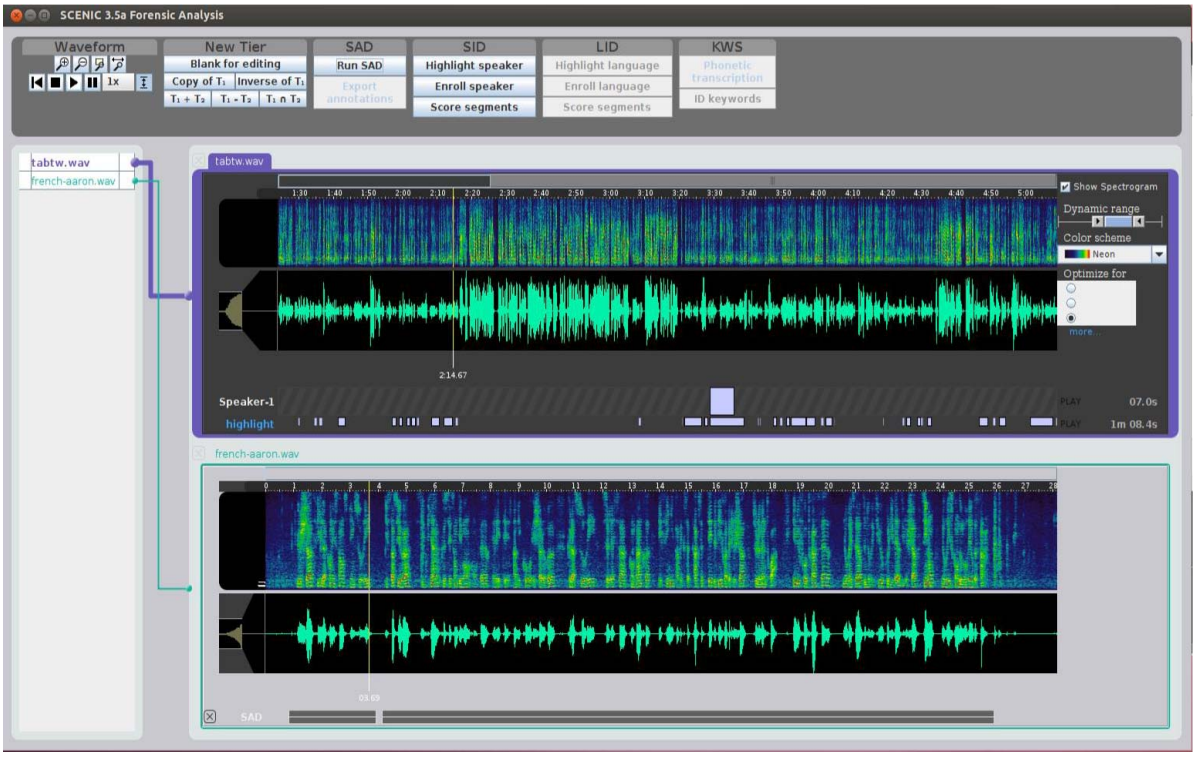

Figure 1: The OLIVE Forensic Analysis Interface enables close editing of audio files, enrollment of new speakers, scoring of segments, speech activity segmentation, and semi-supervised speaker diarization.

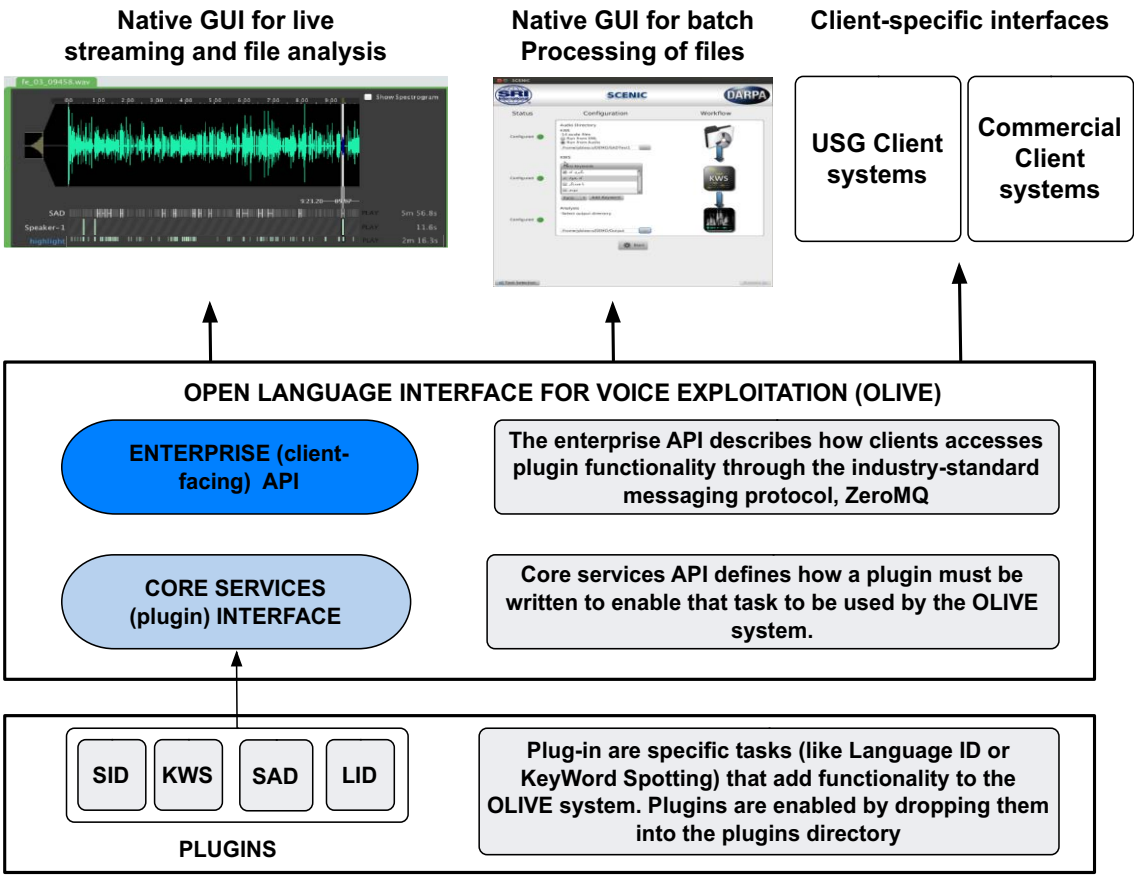

Figure 2: The OLIVE system, showing the three main components, from top to bottom: the graphical user interfaces, the OLIVE system, and the task-specific plugins.